frontiers | Research Topics

Network bioscience
volume II

Edited by
Marco Pellegrini, Marco Antoniotti and Bud Mishra

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-3203-4
DOI 10.3389/978-2-8325-3203-4

Network bioscience volume II

Topic editors

Marco Pellegrini — Institute of Informatics and Telematics, Department of Engineering, ICT and Technology for Energy and Transport, National Research Council (CNR), Italy

Marco Antoniotti — University of Milano-Bicocca, Italy

Bud Mishra — New York University, United States

Citation

Pellegrini, M., Antoniotti, M., Mishra, B., eds. (2023). *Network bioscience volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3203-4

Table of contents

Frontiers in Genetics

TYPE Editorial
PUBLISHED 26 July 2023
DOI 10.3389/fgene.2023.1256025

OPEN ACCESS

EDITED AND REVIEWED BY
Richard D. Emes,
Nottingham Trent University,
United Kingdom

*CORRESPONDENCE
Marco Pellegrini,
marco.pellegrini@iit.cnr.it

RECEIVED 10 July 2023
ACCEPTED 17 July 2023
PUBLISHED 26 July 2023

CITATION
Antoniotti M, Mishra B and Pellegrini M (2023), Editorial: Network bioscience Volume II.
Front. Genet. 14:1256025.
doi: 10.3389/fgene.2023.1256025

Editorial: Network bioscience Volume II

Marco Antoniotti[1,2], Bud Mishra[3] and Marco Pellegrini[4]*

[1]Data and Computational Biology Laboratory, Dipartimento di Informatica, Sistemistica e Comunicazione, Università Degli Studi di Milano-Bicocca, Milan, Italy, [2]B4-Bicocca Bioinformatics Biostatistics and Bioimaging Centre, Università Degli Studi di Milano-Bicocca, Milan, Italy, [3]Courant Institute of Mathematical Sciences, New York University, New York, NY, United States, [4]Consiglio Nazionale Delle Ricerche, Pisa, Italy

KEYWORDS

network bioscience, gene regulatory networks, causal networks, knowledge graphs, metabolic graph models

Editorial on the Research Topic
Network bioscience Volume II

Network biology is based on the intuition that the quantitative modeling and algorithmic tools of network theory offer new possibilities to understand, model, and simulate the cell's internal organization and evolution, fundamentally altering our view of cell biology. As network biology has been gaining ground and recognition in the last 20 years, the scope of its application, while still well grounded in molecular biology and genetics, has moved steadily from tackling fundamental biological questions towards translational medicine, including modeling of diseases and applications in drug design and drug action prediction.

This Research Topic *Network Bioscience Vol II* follows in the track of the first one *Network Bioscience* completed in 2019 (Antoniotti et al., 2019), and it aims at collecting cutting-edge research on the many guises of network bioscience.

The papers contained in the present Research Topic are examples of how network and graph analysis can be used to elucidate various aspects of biological systems from inferring missing annotations, handling heterogeneous data types, including the vast literature available online, understanding metabolic dynamics, phenotype-genotype linking, to relationships assessment among diverse omics data for drug design and drug repositioning, to a deeper understanding of modularity in gene networks.

Among the recent trends with a potential of high impact, a most notable one is the incorporation of causality considerations and concepts within the classical network models so to make better use of perturbation data that are currently not exploited to their full potential. In particular such hybrid causal network models help bridging the gap between *descriptive* and *actionable network models*, the former successfully describe biological systems as they are, the latter allows us to formulate questions and find answers within the vast scope of *what-if*, counterfactual, worlds.

Papers presentation

The papers collected in this Research Topic are roughly grouped as follows:

- "Foundational" papers,
- Analysis of particular biomedical problems,
- Algorithms and Tools.

Five foundational papers in this collection tackle in innovative ways basic issues in the mathematical modeling of bio-networks, covering an overview of relevant modularity concepts, to incorporating causality and biologically plausible sparsity assumptions in gene regulatory networks (GRNs), as well as empirically-found motif sub-network distributions.

Alcalá-Corona et al. bridge a notable gap between the perspective on community detection and network modularity derived from statistical physics and network science on the one hand, and its adaptation and application in biological research, on the other hand.

Maheshwari et al. present a general biological network inference method that combines the discovery of a parsimonious network structure and the identification of Boolean functions that determine the dynamics of the system. The method uses a causal logic framework to assimilate indirect information obtained from perturbation experiments and infer relationships that have not yet been documented experimentally.

Seçilmiş et al. note that many gene regulatory networks (GRNs) use sparsity definitions that are independent of other relevant biological properties expected from GRNs. They thus provide a general approach for identifying GRN that are both biologically accurate and structurally sparse GRN, within the entire space of possible GRNs, by selection criteria based on Akaike and Bayesian Information Criterion (AIC and BIC) adapted to the task of GRN inference.

Zhivkoplias et al. developed a novel motif-based preferential attachment algorithm, *FFLatt*, that aims at constructing a gene-proteins gene-regulatory network (GRN) rich in feed-forward loop (FFL) which are network motifs known to be significantly enriched in experimentally validated GRN.

In cancer driver gene identification, it is often assumed that a driver mutation is less likely to occur in case of an earlier mutation that has common functionality in the same molecular pathway (mutual exclusivity—ME). Ahmed et al. note that the current mutual exclusivity tests lack a network-centric view and thus fail to model key aspects of the problem. Thus they propose a network-centric framework to evaluate the pairwise significance values found by statistical ME tests and correct potential biases.

Two articles apply network techniques in the area of optimization of antibody design for drug design and to challenging modeling of the immune system's role for a specific relevant complex condition (human infertility).

In the context of SARS-CoV-2 studies, Gross and Sharan tackle a fundamental problem of growing importance for antibody design that is the identification of mutations of concern of the Spike protein with high escape probability.

Taraschi et al. use a network-based approach to explore the etiopathogenic mechanisms involved in human hypofertility and infertility, aiming at understanding the involvement of the immune system.

Several of the articles describe advances in designing and providing the scientific community with increasingly powerful tools and algorithms capable of making the best use of the vast amount of heterogeneous and often noisy and incomplete biological data available from online repositories.

Castresana-Aguirre et al. note that when analyzing the association between a gene set and a pathway an issue that is generally ignored is that gene sets often represent multiple pathways. They experimentally found that pre-clustering of genes can be beneficial in this association studies by increasing the sensitivity of pathway analysis methods and by providing deeper insights into biological mechanisms related to the phenotype under study.

Di Maria et al. introduce *BioTAGME*, a system for the inference of novel knowledge and new hypotheses from the current biomedical literature analysis by constructing an extensive Knowledge Graph modeling relations among biological terms and phrases extracted from titles and abstracts of papers available in PubMed.

Semantic knowledge graphs (KGs) are increasingly used to combine unstructured human-curated full-text literature data and structured gene expression data from biomedical databases. In this area, Gurbuz et al. here demonstrate how KGs can be used to find new indications for existing drug targets in order to accelerate the process of launching a new drug for a disease on the market.

Often newly sequenced prokaryotic genomes have poor initial gene functional annotation and missing metabolism pathway gene assignments. To counter these shortcomings, Lu et al. developed *PPA-GCN*, a prokaryotic pathways assignment framework based on graph convolutional network, to assist functional pathway assignments.

Galvão Ferrarini et al. describe a novel tool, *Totoro*, that aims at predicting the metabolic reactions that are most likely active during the transient states of a metabolic network as a result of network perturbation simulations.

Zhao et al. propose a computational method called *LncPNet* to predict potential lncRNA–protein interactions based on spatial embedding a lncRNA–protein heterogeneous network into a collection of low-dimensional latent representations.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was partially supported by the CRUK/AIRC/FC-AECC Accelerator Award #22790, *Single Cell Cancer Evolution in the Clinic*, by the FAQC 2021 *Competitive Funding Program* of the Università degli Studi di Milano-Bicocca, Milan, Italy, and by the European Commission PPPA-AIPC-2021 project "Towards an UNIque approach for artificial intelligence data-driven solutions to fight Childhood cAncer FOR EUrope," LC-01815952/101052609, UNICA4EU. This work has been partially supported with co-funding of the European Union—Next-Generation EU, in the context of the National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), CUP: B83C22003920001. This work has been partially supported by the National Cancer Institute Physical Sciences-Oncology Center Grant U54 CA193313-01.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Reference

Antoniotti, M., Mishra, B., and Pellegrini, M. (2019). Editorial: Network bioscience. *Front. Genet.* 10, 1160. doi:10.3389/fgene.2019.01160

frontiers
in Genetics

REVIEW
published: 14 September 2021
doi: 10.3389/fgene.2021.701331

Modularity in Biological Networks

Sergio Antonio Alcalá-Corona [1,2], *Santiago Sandoval-Motta* [1,2,3], *Jesús Espinal-Enríquez* [1,2] *and Enrique Hernández-Lemus* [1,2]*

[1] *Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico,* [2] *Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico,* [3] *National Council on Science and Technology, Mexico City, Mexico*

Network modeling, from the ecological to the molecular scale has become an essential tool for studying the structure, dynamics and complex behavior of living systems. Graph representations of the relationships between biological components open up a wide variety of methods for discovering the mechanistic and functional properties of biological systems. Many biological networks are organized into a modular structure, so methods to discover such modules are essential if we are to understand the biological system as a whole. However, most of the methods used in biology to this end, have a limited applicability, as they are very specific to the system they were developed for. Conversely, from the statistical physics and network science perspective, graph modularity has been theoretically studied and several methods of a very general nature have been developed. It is our perspective that in particular for the modularity detection problem, biology and theoretical physics/network science are less connected than they should. The central goal of this review is to provide the necessary background and present the most applicable and pertinent methods for community detection in a way that motivates their further usage in biological research.

Keywords: modularity, community structure, motifs, biological networks, systems biology

OPEN ACCESS

Edited by:
Marco Pellegrini,
Institute of Computer Science and Telematics, Italian National Research Council, Italy

Reviewed by:
Dario Ghersi,
University of Nebraska Omaha, United States
Vinícius Vieira,
Universidade Federal de São João del-Rei, Brazil

****Correspondence:***
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *27 April 2021*
Accepted: *23 August 2021*
Published: *14 September 2021*

Citation:
Alcalá-Corona SA, Sandoval-Motta S, Espinal-Enríquez J and Hernández-Lemus E (2021) Modularity in Biological Networks. Front. Genet. 12:701331. doi: 10.3389/fgene.2021.701331

1. INTRODUCTION

The field of Systems Biology has many branches that focus on studying networks. It is common to encounter in the literature terms such as metabolic networks, transcriptional networks, protein-protein interaction networks, etc. These networks are graph-theoretical constructs composed of nodes and edges that aim to describe the integrated state of a biological system. Nodes represent the elements of the system, while edges represent the relation between any two of these elements. Depending on the scale of the biological entities at hand, a network can describe systems such as: ecological systems where each node is a biological entity itself; an organism with nodes being organs or groups of organs; tissues or individual cells with genes, proteins, organelles, and metabolites interacting with each other; down even to the level of amino acids interacting to build a protein. Networks facilitate the identification of relevant entities and interactions through the use of theoretical and computational analysis over experimental data. These analyses aim to make predictions, or at least detailed and accurate descriptions of the underlying biological systems. Since one of the most common applications of complex systems in biology is the representation of biological interactions as edges or links of a network, the connectivity or interaction structure of such a network is of utmost importance. This structure is known as the topology of the network and in biological systems it is usually not random. This means that who is connected to whom is relevant, and the distribution of links is arguably related to the particular functionality of such systems.

In biological systems, modularity has been associated with properties such as robustness (Aldana and Cluzel, 2003), mainly derived from the Boolean network approach (Kauffman, 1969). The concept of robustness is related to the ability of a system to withstand perturbations and retain its functionality, whichever it may be (Aldana et al., 2007). Examples of robustness in a biological system can be observed in biochemical networks (Barkai and Leibler, 1997; Morohashi et al., 2002), signaling networks (Igoshin et al., 2007; Espinal et al., 2011; Espinal-Enríquez et al., 2017) and other complex biosystems. For instance, in prokaryotic organisms, sigma factors, despite their structural similarity, regulate different sets of genes, but the regulatory function of a dysfunctional sigma factor can be reassigned to other sigma factors making the organism functional (Torres-Sosa et al., 2012). Another example of modularity arises when a set of genes is regulated by the same transcriptional factor (set known as a *regulon*). It has been proposed that these sets of genes can give rise to functional modules in *Pseudomonas aeruginosa* (Schulz et al., 2015) and that such modules are essential for the adaptation and survival under challenging environments.

One goal of studying biological systems as networks is to understand how the interconnectedness and function of each element derives in a system-level behavior. In order to uncover these features one can look into the *design principles* of the network. This means, to try to uncover the particular patterns present in the network's topology, such as the ways the nodes are connected to each other; the functional groups they belong to; or if nodes with a particular function agglomerate in subgroups. Topological features, of course, are only partially responsible for the actual design principles of biological systems. Connectivity features common of biological networks, such as the approximate scale-free nature of their connectivity distributions, hierarchical and modular organization, set the stage for functional features to emerge. Such functional features are a consequence of the underlying organizational structure of the systems, their physiological setting and environmental constraints. Regarding network connectivity, it is known that the organization patterns of large complex networks are often composed of structural sub-units often called modules or communities (Girvan and Newman, 2002). Communities and modules in the present context are interchangeable terms, however in this manuscript we will use the latter term as we believe it has a similar meaning over a large number of disciplines, with the possible exception of the Social Sciences and Mathematics.

2. MODULARITY IN BIOLOGICAL SYSTEMS

So what is a module? Despite there is still no consensus on what defines a module, a generally accepted notion is that it corresponds to a tightly interconnected set of edges in a network. Intuitively, the density of connections inside any so-called module (*within-connections*) must be significantly higher than the density of connections with other modules (*between-connections*) (Thieffry and Romero, 1999; Girvan and Newman, 2002; Clauset et al., 2004; Palla et al., 2005). Modularity has been helpful in many biological fields and can even be useful in exploratory research (Serban, 2020). In the following sections, we will present and discuss the latest developments of modularity research in biological systems as well as the necessary concepts and formal definitions to understand and promote the usage of several modularity detection algorithms in the biological sciences (Didier et al., 2018; Li et al., 2019).

2.1. Emergence of Modularity

In order to perform their vital functions and at the same time comply with changing environmental conditions, living systems must possess a high degree of internal organization. A likely scheme to attain such a sophisticated degree of organization is through the coupling of diverse biological processes, which creates the needed correlations among their internal and external constraints to perform a certain task. This theory is known as the *networks of processes* (Clarke and Mittenthal, 1992) and suggests that modules can be thought as clusters of coupled elements that work under certain constraints. It also states that organisms can be studied as super-modules (e.g., networks) made up of several interplaying modules that adapt as a whole to changes in their environment. Under this scheme, modularity can be thought of as a very effective way to prioritize and optimize the correct functioning of living systems, which are undoubtedly subject to changing environmental conditions or even to entropic decay.

The question of how modularity emerges in biological networks has no definitive answer yet, either. It has been shown that dynamical networks, which include temporal processes occurring in the whole spatial structure of the network, can give rise to modular behavior when driven by growth, duplication and diversification. These duplication-centered dynamic models emerge from the fact that if some parts of a system undergo duplication, the new system will be more modular than the original (Lorenz et al., 2011). How modularity emerges is closely related to the question of how and why it is preserved across so many biological systems (Kashtan and Alon, 2005; Gibson, 2016). This question has been addressed in evolutionary/developmental biology (evo/devo) and in molecular systems biology as a kind of intersection point between both disciplines. It has been argued that there is indeed a relationship between modularity and controllability (Constantino and Daoutidis, 2019).

Despite underlying mutational mechanisms have been proposed to explain the emergence of modularity, selection and other evolutionary forces have also been part of this discussion (Wagner et al., 2001, 2007; Espinosa-Soto and Wagner, 2010; Clune et al., 2013; Friedlander et al., 2013; Banerjee et al., 2017; Verd et al., 2019; Jaeger and Monk, 2021), as are ecological factors such as spatial distribution and population dynamics (Gilarranz, 2020). Biological modularity arise in the contexts of dynamical process that may even challenge compartmentalization and cause the breakdown of modularity or its rearrangement (Valverde, 2017; Wang et al., 2021).

In the next section, we will discuss the different notions of modularity –particularly those more closely related to the modular organization at the molecular, functional and cellular levels– and their application to a wide diversity of biological phenomena.

2.2. Applications of Network Modularity

One clear example of application of network theory in biology is the study of Gene Regulatory Networks (GRNs) (Davidson and Levin, 2005). These networks can be conceptualized as control systems that drive whole-genome expression patterns (Hernández-Lemus et al., 2019). This coordinated expression is attained through the orchestrated expression of transcription factors and other regulatory molecules like siRNAs, histones, etc. The wider availability of high throughput technologies has sprouted a new wave of modularity research in GRNs. After the completion of the human genome project (HGP), and following the pioneering work of Kauffman (1969) and Britten and Davidson (1969) in the late 1960s, transcriptional regulation module discovery has become an extremely fruitful research field. For instance, it has been demonstrated that modularity can emerge as a consequence of gene co-expression in GRNs; by associating the functions of these genes and their regulators, it has been argued that gene co-expression may confer functional advantages to the organisms, as genes with related functions are likely regulated in a similar manner (Solé et al., 2002; Narula et al., 2010). Gene functionality of several genes with no prior functional description has already been predicted (Segal et al., 2003; Lee et al., 2004; Tanay et al., 2004). Also, by integrating gene expression levels with the modular structure, it was possible to build a comprehensive map of gene regulation for a whole organism (Zhu et al., 2008).

Community structure and modularity in metabolic networks is another important research field. Many biochemical interventions and biotechnological applications depend on modularity, and with the advent of synthetic biology, the use of modules will probably escalate in the near future, driven by the possibility to evolve engineered biological systems (Parter et al., 2007). Modularity in metabolic networks has been extensively explored since the pioneering work by Ravasz et al. (2002) where through the reconstruction of 43 metabolic networks from different organisms, they found that scale free topologies were ubiquitous. Briefly, in these networks the probability distribution of connections on the network (degree-distribution) follows a power law, so that most nodes will end up with few connections and only a few nodes will end up with many. In this case, the studied networks had values of the scaling exponent around 2, and an average clustering coefficients (see section 3) about an order of magnitude larger than expected for scale free networks. This scaling exponent around 2, suggests that these networks are probably under a dynamical regime between that of an ordered system and the one of a chaotic one. This regime is known as *critical* and it has been observed in many different complex systems (Shmulevich et al., 2005). Another important theoretical contribution of this work is the introduction of the *topological overlap matrix* (Ravasz et al., 2002; Cheng et al., 2019).

The **interactome** (Sanchez et al., 1999) is a useful concept related to Protein-protein (physical) interaction (PPI) networks, which are also organized into functional subnetworks or modules. An interactome is defined as a biological network, which encompasses the complete set of molecular interactions in a particular cell. These interactions range from physical (as in PPI networks) to indirect, as is the case of epistatic or gene-gene interactions, and may even include edges defined by regulatory interactions like those of a GRN (Gómez-Romero et al., 2020). Even if interactomes seem to be less clearly defined than other biological networks, they may be used to represent processes that, although not completely understood, may be associated with some specific phenotypes. The *human disease network* (HUDiNE) (Goh et al., 2007) was actually created by using interactomes. HuDiNe, according to its creators is *a network of disorders and disease genes linked by known disorder–gene associations*. The observation that genes linked to similar diseases present a higher likelihood of sharing physical interactions between their products (e.g., PPI) and a higher correlation in their expression profiles, lead to the conclusion that such a network will likely display characteristic disorder-specific functional modules. This fact was corroborated by analyzing the topological structure of the HuDiNe (Goh et al., 2007). Since the release of HUDiNE, interactomes related to disease have been carefully curated and archived in structured databases, thus making possible the discovery of new *co-morbidities* from a molecular rather than epidemiological perspective (Menche et al., 2015).

In the case of human diseases, modular network decomposition has been applied to further our understanding of the interactions driving the emergence of several complex diseases (Sardiu et al., 2017; Tripathi et al., 2019; Lucchetta and Pellegrini, 2020). One good example is the work of De Matos Simoes and collaborators with cancer cells. By using a network modularity analysis, they showed that transmembrane proteins along with ion channel complexes and receptors play a significant role in the pathogenesis of B-cell lymphoma. The authors based their argument on the observation that central and peripheral layers in the modular decomposition of the networks may play different physiological roles. Hierarchical modular separation may then provide clues as to cross-regulatory phenomena in complex phenotypes. Specifically, they noted that these molecules act via the communication disruption between the intracellular regions and the peripheral regions of B cells (de Matos Simoes et al., 2012). In pancreatic cancer, the disruption of intracellular adhesion and cell-division cycles in the tumors were found to be driven by clearly defined transcriptional modules (Long et al., 2016). Also, network communities related to survival have been found in regulatory networks from hepatocellular carcinoma (Xu et al., 2016). Expression activity of the genes in such modules may contribute to timely stratification and tumor staging of liver cancer patients.

Other complex phenotypes have been dissected by analyzing the community structure of their underlying networks. During brain development, for example, it has been shown that the perinatal transition leads to modular reorganization of the brain, which is in turn associated with the development of new functions. This modularization is also correlated with specific gene sets whose expression are synchronously changing, as they share transcriptional regulators (Monzón-Sandoval et al., 2016). Similar methods have allowed the identification of distinctive molecular pathways that differentiate early and late-onset temporal lobe epilepsy in children (Moreira-Filho et al., 2015). These studies have pointed out that differentially expressed modules in early onset epilepsy are related to neural excitability and febrile seizures, whereas no neural excitability gene modules were found for late onset. These findings support the hypothesis that early onset epilepsies, even if accompanied by severe hippocampal damage, may present compensatory effects. This difference may set the basis for differentiated drug treatments.

Community structure in regulatory networks may also be useful to discover potential molecular targets to treat complex diseases (Muraro and Simmons, 2016). In coronary artery disease, for instance, modules associated with the hypertrophic cardiomyopathy pathway and membrane-related functions were detected (Liu et al., 2016). These pathways, the authors suggest, can provide a means to define a set of druggable process-specific targets (Ashrafian et al., 2011). Transcriptional modules associated with the response to allergens leading to seasonal allergic rhinitis have been also identified by Shi and collaborators (Shi et al., 2010). These modules revealed that the MAP kinase, B-cell receptor and toll-like receptor signaling pathways are crucial for the critical stages of allergic rhinitis. Regarding the role of gene regulation on viral pathogenicity and how it has been shaped by modular adaptation, it has been discussed how enhanced redundancy leads to robustness of the infectious phenotypes (Oliveira et al., 2013).

So far we have discussed several examples where finding modules in biological networks lead to a better understanding of the molecular and regulatory processes involved in certain phenotypes and behaviors. A relevant fraction of the modularity finding approaches used in network biology were developed with a particular biological question in mind. The methods thus developed were, in general, efficient to answer that kind of questions but resulted somehow lacking generalizability. We call these methods *ad hoc*, since they have been developed for a special purpose. Most of these methods are indeed quite useful on a case-by-case basis. However, since modularity analysis is a relevant problem in contemporary theoretical biology, it is desirable to have general methods, or at least methods with broad applicability, to help lay the conceptual foundations of biological modularity. We believe that a first step toward this aim consists in applying the general methods developed in graph theory and network science to biological questions and fine-tune them to account for known biological phenomena. In the next section, we will review several necessary concepts and useful methods for modularity detection that come from a more theoretical perspective. As such, these methods were developed to be useful under any, or at least several, quite general circumstances. We have also included a benchmark section, where we discuss how these algorithms stand against each other in the discovery of modules using both real and synthetic datasets. Although the field of modularity detection in biological systems is somewhat young, it has a long history in physics, and thus, many algorithms are already out there making impossible to review all of them. A later section will discuss the most relevant methods separated by the algorithm they are based on in the hopes that the reader will find some of them useful for their research.

3. NETWORK THEORY

In order to better understand the modularity detection methods that will follow, we will briefly define/recall a few important network properties. For a deeper coverage of these and several other properties we suggest the reader to look, for instance, at the review by Newman (2010). For an introductory lecture on the importance of networks in biology and their main applications besides modularity detection we suggest the review by Green et al. (2018).

3.1. Complex Networks: Concepts and Definitions

For the sake of clarity, we will briefly introduce some well-known definitions of network theoretical concepts.

DEFINITION 1. *A **network** is formally defined as a graph $G(V, E)$ over two sets: a set of nodes or vertices, $v_i \in V$, (e.g., bio-reactants), and a set of edges or links connecting such vertices ($e_i \in E$) (e.g., chemical reactions). The connectivity of the network is often represented by the **adjacency matrix** $\mathbb{A} = A_{i,j}$, where $A_{i,j} \neq 0$ implies an existing interaction between nodes v_i and v_j.*

DEFINITION 2. *The degree-distribution of a network refers to the distribution of the number of connections per node, and is defined as the number of connections a given node has to other nodes (called the* degree *of the node). Thus, **the degree distribution** is defined as the probability distribution of the degrees of all the nodes of the network. This measure is often used as an indicator of the relative importance of a particular node (Barabasi and Oltvai, 2004).*

*Mathematically: Let v_i^m be the set of vertices connected to a given vertex (a.k.a. node) m (i.e., $A_{i,m} \neq 0; \quad \forall v_i \in v_i^m$). We call v_i^m the **neighborhood** of vertex m. The size, or cardinality, of this set $C(e_i^m) = k_m$ is called the **degree** or **connectivity** of vertex m, also written as $deg(v_m)$.*

DEFINITION 3. *A **Network motif** is defined by a group of connected nodes (a sub-graph) that is prevalent in a network or in several networks. Each motif is thus associated with a particular pattern of interconectedness between vertices, and may reflect a framework in which particular functions are achieved efficiently. These patterns describe arrangements of interconnection that are present with a significantly higher frequency than in networks where nodes are randomly connected (Milo et al., 2002).*

DEFINITION 4. *Intuitively, **network modularity** consists in associating network nodes to different categories or subsets of the network. Assignment is based on connectivity patterns within the graph, rather than on some inherent node features. The formal definition of network modularity is still controversial, but we believe that by giving some enclosing definitions from graph theory, we can gain a deeper understanding of this concept and methods described below.*

DEFINITION 5. ***Full/Overlapping partition.*** *We may consider a set Z of disjoint subsets of a network $Z(V, E)$ so that $Z = Z_1 \dot{\bigcup} Z_2 \dot{\bigcup} \dots \dot{\bigcup} Z_k$. This is called a* full partition *of the network.*

If, on the other hand, we allow a non-empty intersection between the subsets $Z_i \bigcap Z_j \neq \emptyset$, we have $Z = \hat{Z}_1 \bigcup \hat{Z}_2 \bigcup \ldots \bigcup \hat{Z}_k$ which is called an overlapping partition *of the network.*

DEFINITION 6. ***Incomplete/Modular Partition.*** *We can also consider an incomplete partition of Z, i.e., one in which not every vertex in V is assigned to a subset. In this case we call $M \subset Z$ a* modular partition *of the network, $M = M_1 \bigcup M_2 \bigcup \cdots \bigcup M_k \subset Z$. The subsets M_i (which may or may not be overlapping) are called the* modules *of Z. There are several ways in which a network can be partitioned. Here lies the difficulty in defining modularity in complex networks: different definitions of modularity may induce different modular partitions of the network, which leads to different modularity measures.*

DEFINITION 7. *The* ***clustering coefficient*** *$CC(i)$ for a particular vertex i in a network is given by:*

$$CC(i) = \frac{\textit{number of triangles connected to i}}{\textit{number of possible triangles connected to i}} \tag{1}$$

Here, a triangle is a set of three fully interconnected nodes. Since $0 \leq CC(i) \leq 1$. Equation (1) can be rewritten as:

$$CC(i) = \frac{2\,E_i}{k_i\,(k_i - 1)} \tag{2}$$

Where E_i is the number of triangles centered in vertex i and k_i is the degree of that vertex.

Once we have an operative definition of clustering coefficient, its mean value is the average over all nodes i.

$$\langle CC \rangle = \frac{1}{N} \sum_{i}^{N} CC(i) \tag{3}$$

$\langle CC \rangle$ is a probabilistic measure of the abundance of triangles (not necessarily triads, but also higher order motifs) in the network.

Global measures such as the $\langle CC \rangle$ are computationally cheap (Fortunato, 2010). However, their utility is mostly restricted to the case of hierarchic modularity scenarios (modules within modules). Hierarchic modularity was originally defined as the property of self similarity in the module distribution in a large scale network, evidenced by a power-law behavior of the clustering coefficient $C(k) \sim k^{-1}$. This relation in turns involves the coexistence of a hierarchy of nodes with different degrees of *node-modularity* –as measured by the node-specific clustering coefficient–. In brief, under such assumptions, the higher a node connectivity k is, the smaller its clustering coefficient, which in the asymptotic regime gives rise to the inverse law, $1/k$.

3.2. Network Models: Types and Approaches

3.2.1. Weighted Networks

A weighted network is defined by the assignment of a weight for each of the edges of the network. These weights are established based on the type and strength of the interaction at hand. Interestingly, weighted networks have proven to further increase the reliability of the modules proposed. For instance, the weighted overlap measure (WOM) is a similarity measure that calculates the overlap between two sets weighted by their relative contribution to the overall (joint set) (Smith, 1985). The WOM has been used to define gene modules that are more cohesive than those obtained through unweighted networks though this is not always the case. Here a more *cohesive* module means that the average value of the inter-module clustering coefficient is higher than the average value of the network's clustering coefficient. Since its proposal, the WOM has been used to recover experimentally validated functional gene modules in cancer cells and in yeast (Zhang and Horvath, 2005). More importantly, it has been shown that modularity affects biological functions as the dynamics of the whole network is determined by the organizational patterns generated by the modules themselves. For example, bi-stable switches, where weighted edges are essential for bi-stability, are known to enhance regulatory feedback and feed-forward loops, which in turn are related to the ability of an organism to adapt to changing environments (Kashtan et al., 2009; Gyorgy and Del Vecchio, 2014).

The functional role of regulatory modules has proved to go beyond that of loops and motifs. By studying a transcriptional network of myeloid cells, Alcalá-Corona and coworkers showed that modules are consistently associated at the pathway level to sets of biological functions (Alcalá-Corona et al., 2016). Community structure has also proven to affect the dynamical behavior of the network (Qi and Ge, 2006). By analyzing simple models of gene regulation, Xu and Wang were able to fully decompose a complex network in terms of independent functional modules (Xu and Wang, 2010). Although clear cut decompositions are not likely to occur in a real biological networks due to pleiotropy, decompositions make possible to observe modular effects in an idealized way. For instance, they have been used to study the effects of the free scale topology and of hierarchical modularity on the large scale structure of GRNs (Zhan, 2007). When network structural properties are supplemented with appropriate dynamic behavior, robustness is enhanced (Aldana et al., 2007). This increase in robustness has been shown to be due to the presence of large attractor basins that lead to stable gene expression patterns (Sevim and Rikvold, 2008).

3.2.2. Multi-Level Networks

The advancement of graph theory along with interactomes gave rise to the concept of multi-layered networks. Multi-layered networks encompass several types of interactions and node types. However, in this *multiplex framework* interactions are integrated into different network layers and therefore more information about the real underlying phenomena can be retained (Didier et al., 2015). Adding extra dimensions to a graph can make the associated mathematical analyses more intricate and hinder the application of common topological approaches to study modularity. Nevertheless, it has been shown that real modules encountered in curated networks are better recovered with modular algorithms applied to multilayered networks, compared with the same algorithms applied to single-layer networks. A detailed mathematical framework for multilayer networks—introductory, though not elementary—is found in the comprehensive paper by De Domenico et al. (2013).

In addition to the multiple molecular levels of description of a phenomenon, multi-layered networks can be adapted to include multiple species which can be useful in disciplines such as in comparative genomics. This extended approach also has more robust scalability features than mono-layered networks (Ritchie et al., 2016). Multi-layered networks have enforced the development of new theoretical approaches need for discovering modularity such as the *Multiplex PageRank algorithm* (Iacovacci and Bianconi, 2016).

Another important feature of multi-layered networks is that they allow a direct analysis of the functional features of their subjacent modules (e.g., pathway-based strategies). This approach is useful for studying phenotypes that are naturally multi-layered, like those associated with genetic regulation where multiple different sources (e.g., transcription factors, chromatin, methylation, etc.) are responsible for the phenotype. For instance, through the use of a multi-layered network of transcription factors and microRNA co-targeting, along with protein-protein interaction and gene co-expression (Cantini et al., 2015) were able to find a set of cancer driver genes associated with the community structure of the network.

A related issue to that of multilayer networks is *multiscale modularity*. Despite highly connected nodes, or hubs, are often labeled as the most important nodes of a network, recent studies in the modular structure of the regulatory networks of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Staphylococcus aureus* revealed an unexpected relevance for low degree metabolites. By using flux balance analysis and graph theoretical methods, Samal et al. (2006) were able to discover connected clusters of low-degree metabolites. These large clusters of low degree nodes turned out to be over-represented in these metabolic networks so that a majority of the essential metabolic reactions could be characterized by just a few low degree metabolites. In this study, reactions whose fluxes were strongly correlated formed well-defined communities in metabolic networks of the organism. The large scale community structure, that is, the network modules conforming relatively large subnetworks, and the small scale modularity (partitions of small motifs), represent a complex interplay that has been shown to play an important role in metabolism under the assumption of hierarchical network organization (Gao et al., 2016). By introducing the concept of multiscale modularity, they propose that network community structure may be defined in several organizational levels, taking into account high and low degree nodes.

4. MODULARITY DETECTION ALGORITHMS

From the perspective of the statistical physics, computer science, computational sociology, network science and complex systems communities, there has been a significant amount of work devoted to solve the modular partition or community detection problem. Unlike what happened with biological networks, these methods aimed at reaching formal and theoretically-founded results with wide applicability. It is important to note that there is the possible drawback of losing some interpretability of the results in the quest for generality. However, it is our belief that these methods will prove useful for the biological community, as these approaches remain largely unknown and offer complementary views of the same problem. With this in mind, the following sections will be focused on introducing this second perspective to the community detection problem.

Classification of community detection algorithms depends on their approach to the graph partition problem. Although there is a wide variety of methods and algorithms to approach the problem of graph partitioning and network modularity detection, they often fall in one of five (quite general and sometimes overlapping) possible categories:

1. Methods based on data clustering
2. Methods based on optimization of the modular partition
3. Methods based on the spectral properties of the adjacency matrix
4. Random walk based and other dynamical algorithm methods
5. Stochastic block models

As we will see, there are advantages, disadvantages and limitations in all types of models. For this reason, it is wise to consider the features, applicability and benchmark performance before opting-in for a certain model.

4.1. Data Clustering-Based Methods

There are several methods based on measuring some significant statistical similarity or distance over the biological data. Some techniques have been developed to ascertain whether a set of proposed modules adequately represents the whole set of molecular determinants of a single disease, or closely related diseases.

For instance, in Menche et al. (2015), a topological method was devised in order to locate disease-related communities within the interactome (whole set of interactions in a particular cell). This method uses the overlap among communities of different pathologies to predict disease-disease associations. Although simple, this method has proved very useful and further improvements have been made to the initial algorithm, in particular on relation to the establishment of endo-phenotype models as discussed in Ghiassian et al. (2015) and Ghiassian et al. (2016).

One important limitation of clustering based methods rely on the challenge to determine the optimal number of clusters. The problem of an optimal number of clusters/modules is actually an open challenge in theoretical computer science and graph theory. Even approximate solutions often depend on the specifics of the algorithm used. Some methods as the ones based on spectral bisection have conditions to define an a priori number of clusters, while other methods like those based on structural properties, on dynamical process over the networks and those which have a stochastic component; may determine a number of clusters, based on their large and local structure of the network, an approach some consider to be more *natural*.

One relevant method for disease module detection is DIAMOND (Ghiassian et al., 2015). The theoretical ground for DIAMOND is that in incomplete interactomes *"diseases cannot be associated with topologically dense network communities"*, rather, the statistical significance of an interaction, meaning the weight of the link, is the relevant quantity used to characterize such modules. This highlights the impact of the node/link ratio in the establishment of interacting structure and then in biological function. By extending the ideas of the DIAMOND/HuDiNe approaches it is possible to analyze the relationship between drug targets and disease-proteins through a topological *proximity measure*. This measure quantifies the interactions between drugs and disease-proteins in the human disease interactome (Guney et al., 2016) and can be used as a proxy for therapeutic effect. This can be useful for establishing a basis for drug screening and repositioning and evaluation strategies. Another approach to detect modularity in the interactome was based on identifying joint patterns of gene expression and drug response (Chen and Zhang, 2016). This was done to gain further insight into the biochemical mechanisms of drug action that may drive the development of new therapeutic targets in cancer. Interactome modularity has allowed *de novo* design of therapeutic strategies in cancer and also allowed the creation of methods for drug repositioning analysis (Chen et al., 2016). Such methods are aimed at detecting multi-targeted drug candidates that may disable malignant cellular functions.

Several methods have been proposed to analyze community structure in PPI networks. Feature selection by clustering has been applied to real and synthetic interaction data revealing modules with increased biological significance for *E. coli* and yeast networks (Henriques and Madeira, 2016). A similar approach was used in the `NCMine` method (Tadaka and Kinoshita, 2016) which is implemented as a plug-in for the popular network visualization and analysis suite `Cytoscape` (Adamcsek et al., 2006; Su et al., 2010; van Dongen and Abreu-Goodger, 2012) and is based on a technique called near-clique mining that distinguishes nodes in a network as either "core" or "peripheral" to a given subnetwork. Topological Data analysis (TDA) has also been used to detect topological network modules in protein interaction networks. TDA encompasses several statistical methods like clustering and perturbation analysis to find structure in data. By deleting protein complexes of the *S. cerevisiae* INO80 protein interaction network and performing TDA, isolated modules that contain proteins with shared biological functions were discovered to belong to the same module, even if they mapped to distinct locations of the network (Sardiu et al., 2017).

Clustering using genetic algorithms has been also applied with certain success (Ramadan et al., 2016). In brief, an objective function is built for exclusive clustering (nodes belonging to a unique module) and overlapping clustering (a particular node or set of nodes can be as indicated by spectral clustering methods, see section 4.2). This function is then optimized by a replication/mutation/recombination genetic algorithm in order to detect modular components of the network identified as protein complexes. One approach to detect such modularity in GRNs is through phylogenetic profiling. This approach is based on the idea that the joint presence or joint absence of two traits across various species is used to infer a meaningful biological connection, such as involvement of two different proteins in the same biological pathway.

As it was mentioned, sometimes approaches made use of hybrid methods, such is the case, for instance, of the work by Servis and Clark (2021) that perform a cluster identification strategy by using modularity optimization to analyze chemical heterogeneity in complex solutions. We will abound on modularity optimization in the next subsection.

4.2. Methods Based on *Modularity* Optimization

Unlike the methods based on similarity of data, most of the methods take into account the large-scale structure of the network itself, defined by the edges between nodes, regardless of the source of the data (Newman, 2012). Such as the case of the methods based on and supported by some class of Modularity optimization (see Definition 8).

In order to categorize different modularity measures, we must distinguish between *local* and *global* methods that quantify and assess network modularity. Measures of local modularity emphasize scoring specific clusters or partitions of the network. This score considers the number of modules that are dense or sparsely connected in a given assignment (Reichardt and Bornholdt, 2006). The more dense connections are within a module and the more sparse the connections are from within a module to outside vertices, the higher the modularity score will be. The local modularity of a network is usually given as the score of the highest-scoring partition. Finding the best partition and evaluating its score solves the modularity problem completely, but it relies on comprehensive enumeration of partitions, a problem that often carries computationally prohibitive combinatorial burdens (Fortunato, 2010).

The case of *global* modularity of a network is different in the sense that global measures usually are computed without *a priori* computing the network partitions. Instead, this measure relies on other network properties such as the *average clustering coefficient* $\langle CC \rangle$. The rationale is that vertices that form a module should have adjacent neighbors, as they increase the modular density and induce the formation of *triangles* in the graph.

An important family of local modularity measures is based on the concept of *edge-betweenness*, a concept introduced to generalize the node-associated betweenness centrality measure. Edge betweenness is then defined as the number of shortest paths between pairs of nodes that run along a given edge. The more paths traverse pairs of nodes traversed by an edge, the more *central* the edge is for the global connectivity structure of the network (Freeman, 1977). The first algorithm that used this concept was proposed by Girvan and Newman (Newman and Girvan, 2004) and is a paradigmatic example of the application of local modularity measures. The method consists in disconnecting sets of vertices by removing edges with larger betweenness. This algorithm was applied to several simulated networks as well as a number of real networks with an *a priori* known modular structure with good overall performance. More importantly, Newman and Girvan also provided a formal measure of network modularity.

DEFINITION 8. *Given a network modular partition we have the following:*

$$Q = \sum_i (e_{ii} - a_i^2) = Tr(\mathbb{E}) - ||\mathbb{E}^2|| \tag{4}$$

Here, e_{ij} is the matrix element –from the modularity matrix $\mathbb{E}$– whose entries are defined as the fraction of all the edges in the network that connects nodes in the i module to the nodes in the module j, $a_i = \sum_j e_{ij}$. Notice that, for an arbitrary matrix $\mathbb{X}$, a norm is defined as $||\mathbb{X}|| = \sum_i \sum_j x_{ij}$.

Q is called the *Girvan-Newman modularity* of a network partition, or sometimes just the *Modularity*. Q measures the fraction of edges in the network connecting vertices within the same module or *community* (or *intra-community edge* ratio) and then subtracts form this fraction its expected value in a network with the same partition scheme over randomly connected nodes. $Q = 0$ implies that the partition's modularity is not better than random, whereas $Q = 1$ is indicative of a strong modular structure.

Modularity can also be rewritten (Clauset et al., 2004) as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j) \tag{5}$$

Where m is the total number of edges in the network. k_i is the degree for node i. A_{ij} is the adjacency matrix. C is an indicator function such that $C_i = C_j$ implies that nodes i and j belong to the same community, δ is Kronecker's delta function. This way, if two nodes i and$j$ belong to the same community $\delta(C_i, C_j) = 1$, otherwise $\delta(C_i, C_j) = 0$.

There is yet another (equivalent) way to represent the modularity Q that may result even more useful in practice (Fortunato and Barthelemy, 2007; Porter et al., 2009):

$$Q = \sum_{s=1}^{M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \tag{6}$$

The sum, over all M modules of the partition, l_s is the number of edges inside community s. L is the number of edges in the network and d_s is the total degree of nodes in module s.

These important ideas lead to the establishment of *Community Detection* as one of the foundational problems of Network Science (Newman and Girvan, 2003; Newman, 2004a; Kovács and Barabási, 2015). Maximization of modularity Q has been proposed as a central idea in several optimal network partition algorithms (Clauset et al., 2004; Newman, 2004b, 2006b). However, modularity optimization, also known as Q_{max} algorithms, are constrained by a resolution limit that depends on the overall size of the network and on the interconnection density of the modules, which may lead to failure of Q_{max} methods due to sub-optimal optimization caused by the presence of a multitude of local minima on the modularity function (Fortunato and Barthelemy, 2007).

A related issue with respect to large networks is that calculating the modularity score Q (see Equation 6) belongs to the family of `NP-Hard` or non-deterministic polynomial-time problems. The main characteristic of these problems is that they cannot be solved in polynomial-time, so they are computationally and time consuming, precluding its direct use on extremely large networks. Several heuristic approaches have been proposed to deal with this problem (Danon et al., 2005; Duch and Arenas, 2005; Guimera and Amaral, 2005; Newman, 2006b; Von Luxburg, 2007; Brandes et al., 2008). One particularly useful technique is known as the Louvain method (Blondel et al., 2008). This approach is based on a two-step heuristic: (1) a maximal modularity full partition is obtained by merging nodes in order to maximize modularity through a greedy method, (2) then a network is formed in which nodes are the modules from the first step. This stage is continued recursively until no further improvement in modularity can be obtained.

A whole new family of methods was developed after the introduction of the modularity measure Q. Most of these methods aimed to maximize either Q itself or some proper function of Q under the rationale that if one is able to find a partition that maximizes Q, the induced community structure would be optimal. In this family we can find the original works by Newman (2004b) as well as later refinements of his method, either by himself (Clauset et al., 2004; Newman, 2006b) or by others (Guimera et al., 2004; Duch and Arenas, 2005; Blondel et al., 2008; De Leo et al., 2013). However, since maximization of the Q-measure has a resolution limit that depends on the size of the network and the degree of interconnection between the modules, the method is not fail-safe (Fortunato and Barthelemy, 2007; Lancichinetti and Fortunato, 2011). Some recent implementations, however, have been developed to improve the results obtained under Q-optimization as is the case of the works by Medus and Dorso (2009), Khadivi et al. (2011), Gong et al. (2011), and (Bettinelli et al., 2012).

4.3. Spectral Graph Theory

Another family of algorithms is based on *Spectral graph theory*, which uses the analysis of the eigenvalues of the *adjacency matrix* or the *Laplacian matrix* of a graph. It consists in a transformation of the set of nodes into a set of points in a space whose coordinates are elements of eigenvectors, then the set of points can be clustered via standard techniques (Fortunato, 2010). The change of representation induced by the eigenvectors makes the cluster properties much more evident (Donath and Hoffman, 1972; Fiedler, 1973).

The analysis of the spectrum of the **Laplacian matrix** $\mathbb{L}$, is the most used approach in spectral clustering. This matrix can be derived from the adjacency matrix $\mathbb{A}$ of a network and it is constructed by reversing the signs of the non-diagonal entries and replacing the diagonal entries with the degree of the corresponding node (See **Figure 1**).

The Laplacian matrix can be written in block-diagonal form, that is, the nodes can be ordered in such a way that the Laplacian displays k square blocks along the diagonal, with some entries different from zero, and all other elements vanish. Each block is the Laplacian of the corresponding subgraph, so it has the trivial eigenvector $\vec{1}$ with components $(1, 1, 1, ..., 1, 1)$. Therefore, there are k degenerate eigenvectors with equal non-vanishing components in correspondence with the nodes of a block, whereas all other components are zero. In this way, from the components of the eigenvectors, it is possible to identify the connected components of the graph, and then based on this property, it is possible to find highly connected groups of nodes and the expected number of modules in which the network may be partitioned.

Since the values of the eigenvector components are close for nodes in the same community, it is possible to use them as coordinates, such that vertices turn into points in a metric space. So, for M eigenvectors, the nodes can embed in an M-dimensional space. Thus, modules appear as groups of points well-separated from each other (Donetti and Muñoz, 2004). Also, it is possible to use the Laplacian matrix property, in which, if the graph has g connected components, the largest g eigenvalues are equal to 1, with eigenvectors characterized by having equal-valued components for nodes belonging to the same component. Thus, the modules can be found by inspecting the components of the eigenvectors with eigenvalue 1 (Capocci et al., 2005).

Furthermore, in the context of *Spectral clustering*, there is a remarkable relationship introduced by Newman (Newman, 2006b), between *Modularity optimization* and the spectral properties of the *adjacency matrix* known as *Spectral optimization*. We can rewrite the Q optimization in terms of finding the spectrum of a particular matrix as we will see below.

Starting from Equation (5), it is possible to define the *modularity matrix* B_{ij} as:

$$\mathbb{B} = B_{ij} = \left(A_{ij} - \frac{k_i k_j}{2m} \right)$$

Now, let us suppose a particular *a partition* of a network into just **two** modules. Thus we can assign to each node, a quantity s_i, such as:

$$s_i = \begin{cases} +1, & \text{if a node } i \text{ belongs to group 1} \\ -1, & \text{if vertex } i \text{ belongs to group 2} \end{cases}$$

Thus, Q can conveniently be written in matrix form:

$$Q = \frac{1}{4m} \sum_{ij} B_{ij} s_i s_j = \frac{1}{4m} \vec{s}^T \mathbb{B} \vec{s} \tag{7}$$

where $\vec{s}$ is a column vector whose elements are s_i.

Then, in order to optimize this form of Q it is possible to perform the so-called *relaxation method* (that is, allowing its entries to take continuous values and retaining the norm of the vector), which is one of the standard methods for the approximate solution of vector optimization problems such as this one. Thus, by differentiating and imposing the constraint $|s| = \sqrt{n}$ or equivalently:

$$\sum_i s_i^2 = n$$

The modularity maximization problem is now straightforward. We now have a maximization problem with this norm as a constraint, or equivalently, $(n - \sum_i s_i^2) = 0$. This is done by introducing a *Lagrange multiplier* λ, and taking the partial derivative with respect to the components of the vector (one at a time) of the following expression:

$$\frac{\partial}{\partial s_{i=k,j=k}} \left[\sum_i \sum_j B_{ij} s_i s_j + \lambda \left(n - \sum_i s_i^2 \right) \right] = 0 \tag{8}$$

to obtain:

$$\left[\sum_i B_{ik} s_i + \sum_j B_{kj} s_j - 2\lambda s_k \right] = 0 \tag{9}$$

which leads to:

$$\sum_j B_{kj} s_j - \lambda s_k = 0$$

$$\sum_j B_{kj} s_j = \lambda s_k$$

for all k.

Which is in a matrix form an eigenvalue problem for the *modularity matrix*:

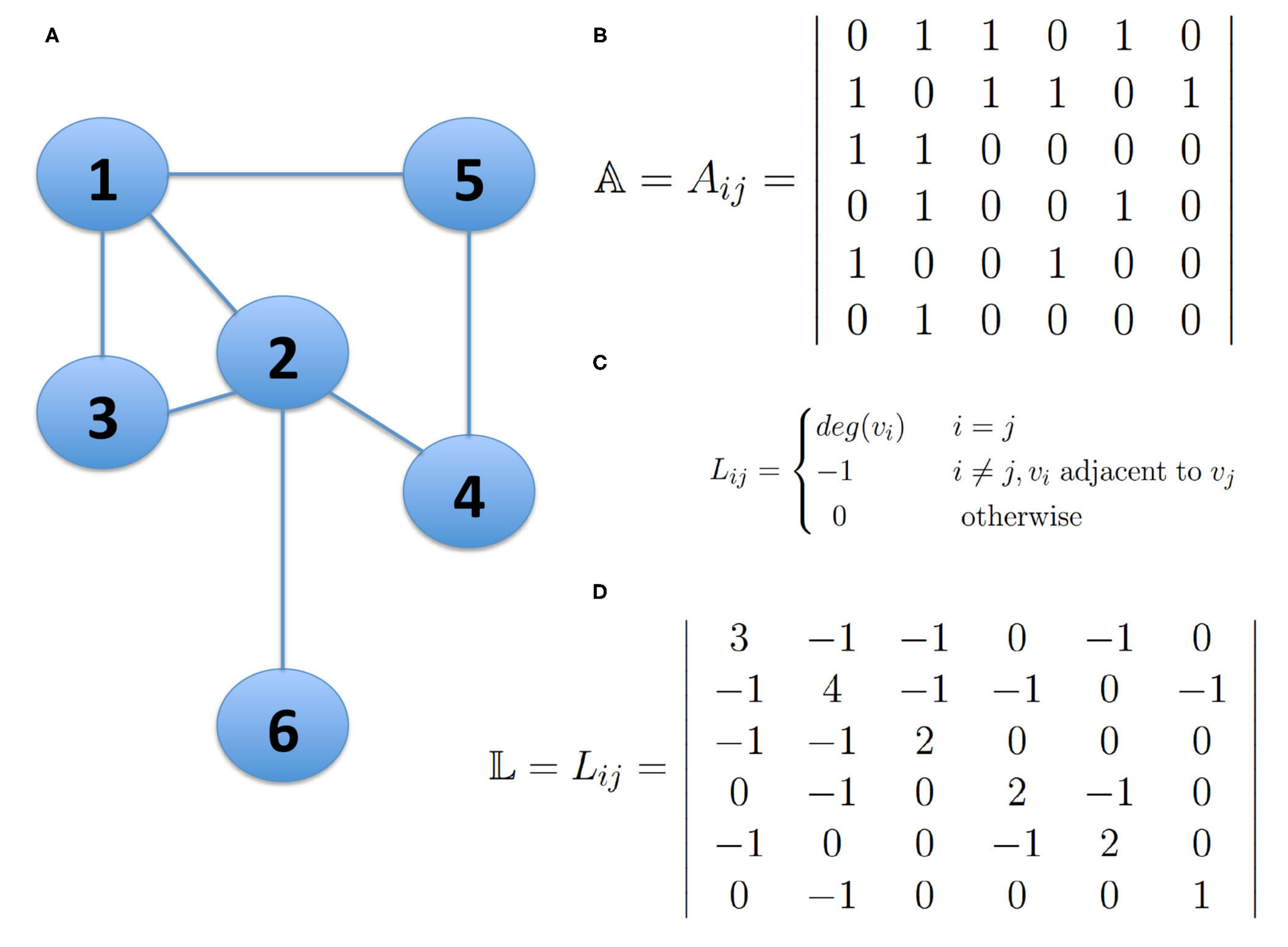
$$\mathbb{B} \vec{s} = \lambda \vec{s} \tag{10}$$

The value of λ that maximizes Q is the largest possible one, that is the dominant eigenvalue of the matrix $\mathbb{B}$.

It is worth mentioning, that similarly to this approach, the **spectral bisection method** (Barnes, 1982), uses the spectrum of the Laplacian matrix, to find partitions of a graph by dividing it recursively into two groups. Every partition of a graph with n nodes in two groups can be represented by an index vector $\vec{s}$, whose component s_i is $+1$ if a node i is in one group and $a1$ if it is in the other group. Then the cut size R of the partition of the graph in the two groups can be written as:

$$R = \frac{1}{4} \vec{s}^T \mathbb{L} \vec{s} \tag{11}$$

Finally, the *Modularity optimization* approach can be extended to a more than two modules, by writing an additional contribution



B

$$\mathbb{A} = A_{ij} = \begin{vmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{vmatrix}$$

C

$$L_{ij} = \begin{cases} deg(v_i) & i = j \\ -1 & i \neq j, v_i \text{ adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

D

$$\mathbb{L} = L_{ij} = \begin{vmatrix} 3 & -1 & -1 & 0 & -1 & 0 \\ -1 & 4 & -1 & -1 & 0 & -1 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 2 & -1 & 0 \\ -1 & 0 & 0 & -1 & 2 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{vmatrix}$$

FIGURE 1 | The Laplacian Matrix of a network. Panel **(A)** presents a small undirected network; Panel **(B)** shows the Adjacency Matrix $\mathbb{A}$ describing the network connectivity of the network in **(A)**; Panel **(C)** shows the definition of the Laplacian Matrix of a Network and panel **(D)** shows the Laplacian Matrix $\mathbb{L}$ of the network in **(A)**. The bold numbers represent the degree of node i, whenever i=j. This figure is intended for illustrative purposes, no *actual results* are presented.

ΔQ to the modularity upon further dividing a group g of size n_g in two as:

$$\Delta Q = \frac{1}{4m} \sum_{i,j \in g} \left[B_{ij} - \delta_{ij} \sum_{k \in g} B_{ik} \right] s_i s_j \tag{12}$$

$$\Delta Q = \frac{1}{4m} \vec{s}^{T} \mathbb{B}^{(g)} \vec{s} \tag{13}$$

where δ_{ij} is Kronecker's δ, and $\mathbb{B}^{(g)}$ is the $n_g \times n_g$ matrix with elements indexed by the labels i, j of nodes within group g. Because Equation (13) has the same form as Equation (7) it is possible to apply the spectral approach to this generalized *modularity matrix*, just as before, to maximize ΔQ.

In addition, the *modularity matrix* $\mathbb{B}$ also has always the trivial eigenvector $\vec{1}$ with eigenvalue zero (like the *Laplacian matrix*), because the sum of the elements of each row/column of the matrix vanishes. Thus, it is also possible to optimize modularity on bipartitions via *spectral bisection*, by replacing the Laplacian matrix with the modularity matrix (Newman, 2006a,b).

4.4. Random Walk Based Models

The use of random walks to find modules on a network is based on the somehow intuitive premise that a random walker moving on the network will spent more time inside modules—due to the high density of edges, thus many possible trajectories—than hoping from one module to another. A first approach to this problem was addressed by Zhou (2003) who used random walks to define a *distance* between pairs of nodes, assuming that there is a high likelihood that *closer* nodes—under this measure of distance—belong to the same module. Such distance was used to define global and local *attractor nodes* used to detect modules, i.e., minimal distance subnetworks. A different but related approach was taken by Pons and Latapy (2006) on a method called *Walktrap*. Here, distance is calculated via the probability that a random walk moves from one module to another on a fixed number of steps, then grouping nodes via hierarchical clustering.

A method based on the application of the Markov property of node-to-node walks called Markov Cluster algorithm (MCL) was developed by Van Dongen (2001). MCL simulates a diffusive process in the network. A *stochastic matrix* is obtained by dividing every entry of the adjacency matrix A_{ij} by the corresponding degree of node *i*. This stochastic matrix is used to calculate transition probabilities on a Markov random field. This method is quite elegant and comparatively easy to implement, however, its large computational complexity makes it difficult to apply in practice for real (large) networks (even in sparse cases).

As already mentioned, for large sparse networks also the standard versions of spectral based algorithms are suboptimal, in the sense that in some cases these fail to detect communities even when other algorithms such as belief propagation can do so. Efforts to improve these spectral theory methods have been made by resorting again to random walk dynamics, mainly through implementing non-backtracking random walks (the random walker cannot move backwards) over the network (Krzakala et al., 2013; Newman, 2013; Zhang and Newman, 2015). Other methods in the literature are built on ideas borrowed from non-linear dynamic processes, such as spin-coupling models with nearest neighbor interaction (Reichardt and Bornholdt, 2004), synchronized oscillators (Arenas et al., 2006; Arenas and Diaz-Guilera, 2007), as well as generalized random walks (Van Dongen, 2001; Zhou and Lipowsky, 2004; Pons and Latapy, 2005). Among this plethora of models, INFOMAP has been shown to be quite reliable and computationally efficient (Rosvall and Bergstrom, 2007, 2008).

The INFOMAP algorithm is founded on a clever combination of random walk dynamics and information theory. The main idea is to reach optimal compression of the information needed to describe the diffusion process of a set of random walkers. This is achieved by using the random walk *itself* as a proxy for the diffusion process via a sequential enumeration algorithm and the use of tools of information theory and computational linguistics.

In a nutshell, the approach is quite similar to the way we imprint location information on geographic maps of cities: you can map a large number of close-to-each-other streets into a neighborhood (“a module,” with its own description) and a series of close-by neighborhoods into a town. The larger the scale of these *urban modules*, the smaller the total amount of information needed for their description. In a similar way, the INFOMAP algorithm looks up for the *minimal description length* for the modular partition of a network. The best partition is the one that can be described with the minimal information.

In brief, the *description length* is a measure of the complexity of a given process. By using the description length is possible to characterize the trajectory of a random walk (or the trajectories for an ensemble of random walkers), in the form of the *map equation*:

$$L(M) = q_{\curvearrowleft} H(\mathcal{Q}) + \sum_{i=1}^{m} q_{\curvearrowleft} H(\mathcal{P}_i) \tag{14}$$

Here, $L(M)$ is the description length of an ensemble of random walkers moving through a given modular partition M. The first term $q_{\curvearrowleft} H(\mathcal{Q})$ represents the average number of bits needed to describe the movements from nodes in one module of the partition to nodes in another module, whereas the second term represents the information for the intramodule walks. Since by the coding theorem (Knuth, 1985), the information needed to characterize inside module walks is smaller, a minimal description length implies that most of the time walkers move inside modules of a given partition, thus optimizing modularity, allowing however for the presence of a number of intermodule hops. This method uses a *greedy* algorithm, so it can be applied quite efficiently even to large networks, directed or undirected. There are also INFOMAP implementations to find hierarchic modular structure (Rosvall and Bergstrom, 2011) and overlapped modules (Esquivel and Rosvall, 2011).

4.5. Stochastic Block Models

Statistical inference provides a powerful set of methodological tools useful in modularity detection. The usual way to proceed is by adjusting a *generative network model* to the experimental data. A stochastic block model (SBM) is by far, the most used model to generate networks with a modular structure. The essentials of the SBM are as follows:

The stochastic block model generates a number n of vertices of the network; the algorithm makes a partition of the vertex set $\{1, \ldots, n\}\{1, \ldots, n\}$ into q disjoint subsets $C_1, \ldots, C_q$ i.e., the modules. By starting with a symmetric $q \times q$ matrix P containing edge probabilities for all the possible connections. These probabilities must be known a priori. Then the SBM is generated by randomly sampling this edge set as follows: any two vertices $u \in C_i$ and $v \in C_j$ are connected by an edge with probability P_{ij}.

Modularity detection works out by optimizing the unnormalized log-likelihood that a given partition g of a graph G in q modules will be reproduced by the SBM (Karrer and Newman, 2011).

$$\mathcal{L}(G|g) = \sum_{i,j=1}^{q} e_{ij} \log\left(\frac{e_{ij}}{n_i n_j}\right) \tag{15}$$

Here $\mathcal{L}(G|g)$ is the log-likelihood for a partition g of a given network G to be produced by the standard SBM. e_{ij} is the number of edges connecting module i with module j of the partition, and n_i, n_j are the number of nodes in modules i and j respectively. The sum includes the case $i = j$. The strongest drawback of the method is that it requires *a priori* knowledge of the number q of modules in which the network has to be partitioned, although this limitation has been recently overcome by using a Bayesian formulation (Peixoto, 2018).

General SBM models (i.e., non-Bayesian) have been demonstrated to be formally equivalent to modularity optimization approaches that do not usually require a fixed number of modules for the partition (Newman, 2013). Despite this and the fact that maximum likelihood exact estimation is an *NP* problem—so all solutions are approximate—SBM models are still popular in statistics and machine learning algorithms.

As we have discussed in this section, topology based methods for modularity detection are robust, general and intelligible. They can also be benchmarked with experimentally available modular partitions. Such validation uses robust statistics, such as the ones given by normalized mutual information measures. The strength of these methods is that they do not rely *a priori* on any non-topological information, as they are based on the (weighted or un-weighted, directed or un-directed) connectivity as given by adjacency matrices. This is the basis of their generality and broad applicability, in particular to complex biological problems.

The fact that these methods do not need any prior knowledge—aside from the connectivity structure—does not preclude us to incorporate such information when available, to enhance our intuition and empower our predictions when applied to real large scale biological networks. For this reason we strongly believe that the popularization of these approaches within the computational and systems biology research settings will prove to be highly beneficial for both, the construction of more general approaches to study modularity in biology and for the further development of analytic methodologies in the theory of complex networks.

5. BENCHMARKING AND PERFORMANCE TESTS

Whenever several methods perform a similar task, benchmarking becomes necessary. However, as described in Tripathi et al. (2016), a large heterogeneity among different community structure discovery methods is often found. As many of the available methods for module discovery have been developed as *ad-hoc* solutions, they often lack reliability when applied to other biological systems. Also, the intrinsic complexity of biological modularity makes it hard for a single method to describe all types of modules correctly. Nevertheless, in the following section we will show how by resorting to theoretically sound and rigorous methods of comparison that do not rely on the specifics of a given biological system, one can attain precise measurements of performance for any module detection method.

5.1. Testing Performance and Scoring Measurements

Benchmarking community detection algorithms using real biological networks is not optimal, as it is not clear what the ideal partition is. However, real networks such as the social network of bottle-nose dolphins from Doubtful Sound (New Zealand) built and studied by Lusseau (2007), as well as the network of college football teams obtained by Girvan and Newman (Girvan and Newman, 2002) have been used for this purpose. Real biological network communities (also called *ground-truth communities*) are often inferred from non-topological studies carried out by network curators, which based on experimental observations (e.g., protein-protein interactions) define the network itself. As these methods rely only on observed data, it is possible that the resulting network is either incomplete or has spurious interactions. So how can one find these modules and relate them to particular functionalities, especially when such functionalities are unknown? One general approach is to use random network methods to test if the community or modular structure in our networks is valid and significant (Sah et al., 2014). One common approach consists in generating network models that satisfy the constraints imposed by the real networks (such as the connectivity, the number of nodes, etc.) and keep a graph structure that is as random as possible. These network realizations allow the use of a large set of tools already available to analyze the topology of random networks. In particular, they are useful for creating *null-models* that serve as a baseline to which we can compare the significance of our partition model. As such null models have been established, they can be used to test biological functional hypotheses. This generation of null models serves directly to generate scoring metrics that allow the comparison and selection of the best network partitions. These null-model networks may be generated synthetically, and this way we could test to what extent the algorithm is able to found the a-priori known communities.

There are two classic and widely used performance tests for community detection algorithms: the GN and the LFR (Fortunato, 2010), both of which belong to a class of methods generated under the *planted l-partition model* (Condon and Karp, 2001).

DEFINITION 9. *In the* ***planted l-partition model*** *a network with* $n = g \cdot l$ *nodes, is partitioned into* l *groups of* g *nodes each. Nodes in the same group are linked with a fixed probability* p_{in}*, whereas nodes in different groups are linked with probability* p_{out}*. Each module is then a random Erdös-Rényi network with* $p = p_{in}$ *and if every module were a node, the whole network would also be an Erdös-Rényi graph with* $p = p_{out}$*.*

For a subgraph representing a module or community C, the average connectivity degree will be given as $\langle k \rangle_{in} = p_{in}(g-1)$ *and the average external degree would be* $\langle k \rangle_{out} = g \cdot p_{out}(l-1)$ *(recall that for an Erdös-Rényi graph connected with probability p, the average degree is given as* $\langle k \rangle = p(n-1)$*). If these conditions hold, the average degree for the whole network is*

$$\langle k \rangle = p_{in}(g-1) + g \cdot p_{out}(l-1) \tag{16}$$

This way, if $\langle k \rangle_{in} > \langle k \rangle_{out}$ *(i.e., if the intra-module average degree is greater than the inter-module average degree), then the network will have well-defined community structure. This is equivalent to the intuitive definition of modularity, namely* $p_{in} > p_{out}$*.*

The GN test was designed by Girvan and Newman (Girvan and Newman, 2002) to test their community detection algorithm. It is a particular case of the *planted l-partition model* where the authors fixed $l = 4$ and $g = 32$ to get a network composed of 128 nodes forming 4 modules with 32 nodes each and an average degree of $\langle k \rangle = 16$. Within this framework link-density is adjusted by scanning the values of the average in-degree $\langle k \rangle_{in}$ and out-degree $\langle k \rangle_{out}$ to choose specific values to change the community structure for each network provided that $\langle k \rangle = \langle k \rangle_{in} + \langle k \rangle_{out} = 16$.

Under this model it is possible to have explicit expressions for the average in- and out- degrees, namely: $\langle k \rangle_{in} = p_{in}(g-1) = 31\,p_{in}$ and $\langle k \rangle_{out} = g \cdot p_{out}(l-1) = 96\,p_{out}$. By varying the values of p_{in} and p_{out} it is then possible to simulate networks with a stronger or weaker modularity. For instance, a clearly defined community structure is induced if $p_{in} \simeq 0.5$ or larger, whereas a value of $p_{in} \simeq 0.25$ or lesser precludes the existence of well-defined modules.

For this benchmark communities are well-defined for $\langle k \rangle_{in} > 8$. One of the advantages of the GN test is that by varying a single parameter in a pretty simple network it is possible to contrast different network partition methods. In order to test a particular method via the GN test one has to calculate a *similarity measure* between the partition of the GN network as given by this method against the natural partition of the network in four modules of the same size. A highly used similarity measure—proposed by Newman and Girvan (Girvan and Newman, 2002)—is the fraction of edges correctly classified, though a more objective measure can be the normalized mutual information between partitions (see Equation 17) (Arenas et al., 2008).

In spite of its simplicity and mathematical rigor, the GN test presents a couple of important shortcomings derived from unrealistic assumptions. First, all the nodes are expected to have the same degree. Second, all the communities must be of the same size. Clearly real complex networks, such as those encountered in biology, are characterized by long-tailed degree distributions or power law-like ones, and also by heterogeneous community sizes. Some improved versions of the GN method have been developed such as the one presented in Fan et al. (2007) where different weights are assigned to *inner* and *outer* edges, regarding their position in the communities.

The fact that the planted l-partition model generates mutually-interconnected Erdös-Renyi random graphs implies that all the nodes will have almost the same degree and all the communities will have exactly the same size. Of course, these two features do not match with what is observed in real networks. To tackle this problem, Lancichinetti et al. proposed the *LFR Benchmark test* (Lancichinetti et al., 2008). The LFR test assumes that the node degree distribution and the module size distribution follow a—more realistic—power law behavior. Each node shares a fraction $1 - \mu$ of its edges with nodes within its community and a fraction μ with nodes in other communities. Hence $0 \leq \mu \leq 1$ the mixing parameter is equivalent to a normalized version of the $\langle k \rangle_{out}$ used in the GN test. The LFR test was devised for undirected, unweighted networks, but there are implementations for directed, weighted graphs including the possibility to have overlapping communities (Lancichinetti and Fortunato, 2009a). Aside from purely computational costs, the main performance test for network community detection algorithms must establish a clear criterion to compare the degree of *similarity* between the modules discovered (i.e., the specific partition) by an algorithm and the real (in the test, a priori known) partition. There are several proposals in the complex network literature as how to measure similarity between different partitions (Meilă, 2007), some of them based on pair recounting and group coincidence counts (Fortunato, 2010).

Additionally, two widely used measures are the fraction of correctly classified edges and the normalized mutual information between partitions. The former was proposed by Girvan and Newman to test their algorithm, but can be generalized to other benchmark tests. The criteria for the correct classification is as follows: Each of the modules A_i of the partition found by the given algorithm is compared to all of the *actual* modules B_i, known a priori from the real network partition. When more than half of the nodes in one of these A_i correspond to those of a community B_i then A_i is considered to be correctly classified and no more comparisons between A_i and the rest of the B_is are carried out. In the contrary case (less than half corresponding nodes) or when the community A_i is smaller than half the size of the given B_i, then the module is compared to the rest of the B_i's until exhaustion. This criterion is quite stringent since there are cases in which one may consider that some of the nodes have been correctly classified by the algorithm but the measure (total node count divided by the size of the network to give a number between 0 and 1) rules them out.

DEFINITION 10. *The* ***normalized mutual information between partitions*** *(NMIBP) was proposed by Danon et al. as a similarity measure (Danon et al., 2005) built on ideas proposed by Ana and Jain (2003), Kuncheva and Hadjitodorov (2004).*

The rationale is that if two partitions are similar, very little information is needed to infer one partition given the other. One is able to calculate the mutual information between two partitions A and B by building a confusion matrix $\mathbb{N}$ *where rows correspond to the actual modules and columns correspond to the modules found by the given algorithm. The* N_{ij}*-th element of* $\mathbb{N}$ *is the number of nodes in a real (known a priori) community i that are also present in the community j detected by the algorithm. Since the partitions under comparison may have a different number of groups (the modules or communities),* $\mathbb{N}$ *is not necessarily a square matrix. This way the similarity between two partitions A and B is given by the normalized mutual information measure (NMI) as follows:*

$$NMI(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\, log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_A} N_{i.}\, log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{C_B} N_{.j}\, log\left(\frac{N_{.j}}{N}\right)} \quad (17)$$

Here, the number of actual modules (partition A) is denoted by C_A*, the number of modules found by the algorithm (partition B) is* C_B*, the sum over the row i of the matrix* $\mathbb{N} = N_{ij}$ *is* $N_{i.}$ *and the sum over column j is* $N_{.j}$ *and N is the total number of nodes. If the partitions A and B are identical, then* $NMI(A,B) = 1$*, whereas completely dissimilar partitions give* $NMI(A,B) = 0$*.*

This measure is highly used in the performance tests for community detection algorithms since it is highly sensitive as it quantifies explicitly the amount of information recovered by the algorithm from the original topological structure of the network (Lancichinetti and Fortunato, 2009b; Lancichinetti et al., 2011; Tripathi et al., 2016). The NMIBP measure can be used in the GN and LFR performance tests, both in standard and overlapping partitions (Lancichinetti and Fortunato, 2009a).

More recently there have been some other approaches that propose new benchmarks that provide actual techniques to determine which is the most suited algorithm in most circumstances based on observable properties of the network under consideration. Also considering the use of the mixing parameter μ and the Normalized Mutual Information measure (NMI) (Yang et al., 2016). There are also benchmarks based on novel methods that generate networks with topological properties found in empirical biological networks (Sah et al., 2014; Gilbert, 2015).

Despite the high performance of algorithms and methods shown on the artificial networks generated by benchmarks and its test with the μ (mixing factor), for example on the LFR test, an open question is, whether the methods with good results on benchmarks necessarily find meaningful modules in actual networks (Jebabli et al., 2018; Cherifi et al., 2019).

It may happen that the community structure found by some methods with high performance in benchmarks, does not necessarily correspond to correct ground-truth community structure—that is, the one based on real known node groups, or derived from some metadata or even identified by the node attributes—and vice versa. There could be a substantial difference between structural communities and metadata groups (Orman et al., 2012; Hric et al., 2014; Jebabli et al., 2018).

So, for a fair assessment of the performance of some methods, it is necessary to have a good match between the detected partition and the attribute-based partitioning for considering that a method is reliable. Both tests are complementary, and we recommend applying both of them to perform a complete and accurate assessment of an actual community structure.

Nonetheless, to overcome these limitations, exploiting the topological features of the so-called "*community graphs*" (where the nodes are the communities and the links represent their interactions) has been proposed to evaluate the algorithms; in contrast with metrics defined at node level that are fairly insensitive to the variation of the overall community structure. Thus, if the ground-truth community structure is available, it is possible to compare it vs. the one discovered by these algorithms by using these clustering-based metrics as has been proposed by some authors (Orman et al., 2012; Hric et al., 2014; Jebabli et al., 2018; Cherifi et al., 2019), where more emphasis has been put on the topology of the community structure.

In this direction, some modifications to the LFR benchmarks have been proposed to make generated networks more realistic (Orman et al., 2012). In this work, authors studied generated networks in terms of community-centered topological properties to evaluate some methods, they used such properties to compare community structures to rank the tested community detection algorithms. As well, recently da Fonseca Vieira et al. (2020) tested some representative state-of-the-art methods for overlapping community detection (Cherifi et al., 2019) with synthetic and real-world benchmark *Ground-Truth networks* showing that, although the methods can identify modular communities, they often miss many structural properties of the communities.

5.2. Good Performance Methods Commonly Applied to Biological Networks

Beyond presenting the benchmarking for the performance of the different algorithms, it is important to point out which methods we think are good for finding modules, given the biological question under consideration. The question of which algorithm is the best for biological networks is not easy to answer, it will depend on the context of the research question and the data on which the network is built.

However, two of these graph-theoretically-grounded, general purpose algorithms have been widely applied in biological networks with good and significant results, such methods are the **Louvain** (Blondel et al., 2008) and **Infomap** (Rosvall and Bergstrom, 2008). Both methods have good performance and accuracy scores, as we can see from the several artificial network bencharmking analyses (Lancichinetti et al., 2008, 2009; Lancichinetti and Fortunato, 2009a; Sah et al., 2014; Gilbert, 2015; Yang et al., 2016), as well as in *Ground-Truth networks* and also in terms of *community-centered topological properties* (Orman et al., 2012; Hric et al., 2014; Jebabli et al., 2018). In addition, both methods show good results and performance in biological networks, even in comparison with more recent methods (Mall et al., 2017b; Debnath et al., 2021). Furthermore, they also have been proved as standard methods to identify biologically meaningful modules in biological networks (Zheng et al., 2021) and even for evaluating significant topological differences between networks (Mall et al., 2017a). In addition, they have been incorporated on different Bioinformatic analysis suites and tools, as well as implemented in different programming languages widely used today, such as R, Python, MatLab, and C++ and incorporated into standard widely network analysis libraries such as `igraph`.

The *Louvain method* (Blondel et al., 2008) is by far the most widely used method in biological networks, showing significant results and meaningful modules (Praneenararat et al., 2011) even compared with newer methods in recent studies (Şen et al., 2014; Bennett et al., 2015; Rahiminejad et al., 2019; Calderer and Kuijjer, 2021). The method is indeed still widely used nowadays, for example, in the context of SARS-COV-2 analyses (Zheng et al., 2020). The efficiency and high performance of this method lie on its taking into account the whole structure of the network and searching for the best partition in an algorithmic greedy fashion. In addition, this method has been extended and applied to bipartite biological networks (Pesantez-Cabrera and Kalyanaraman, 2016; Calderer and Kuijjer, 2021) as well as to multilayer and multiplex biological networks (Mucha et al., 2010; Didier et al., 2015; Mittal and Bhatia, 2018).

On the other hand, *Infomap* is accepted as a very well-known method in module detection (Acharya et al., 2012) and even as a method for comparing the performance and accuracy of novel methods in biological networks (Lecca and Re, 2015), and has been incorporated in some bioinformatic layouts as a standard community detection framework (Aldecoa and Marín, 2014; Zhou and Xia, 2018; Farage et al., 2021). Moreover, has been widely adapted and extended by its authors in several ways to different kinds of networks and problems in community detection, for example, hierarchical module detection (Rosvall and Bergstrom, 2011), bipartite networks (Kheirkhahzadeh et al., 2016) and multilayer networks (De Domenico et al., 2015). In addition, these extensions have proved to give meaningful results in the context of biological networks as ecological networks (Pilosof et al., 2020; Farage et al., 2021), multiplex genetic datasets

(Mittal and Bhatia, 2018) and breast cancer networks (Alcalá-Corona et al., 2018a). The efficiency and high performance of Infomap lie in how information flow in a network can reveal the structure of it (Esquivel and Rosvall, 2011; Aslak et al., 2018; Eriksson et al., 2021), combined with a strategy of optimizing partitions such as the *Louvain method*, which make it one of the most robust and applicable methods for all kinds of networks and giving meaningful results (Kawamoto and Rosvall, 2015; Emmons and Mucha, 2019).

Finally, it is worth mentioning that other three methods have been demonstrated to be efficient and reliable in the context of biological networks in comparison with Infomap and Louvain: the **Spinglass Method** (Reichardt and Bornholdt, 2004, 2006), **OSLOM** (Lancichinetti et al., 2011), and **Label Propagation approach** (Garza and Schaeffer, 2019).

Thus, we can suggest **as a general strategy for community detection in biological networks to apply both Louvain and Infomap, in addition to one of these three latter methods and then consensing the partition by the Consensus Clustering approach** (Lancichinetti and Fortunato, 2012) to compute a unique community structure.

6. APPLICATION EXAMPLE: COMMUNITY DETECTION METHODS FOR CANCER NETWORKS

Network approaches have been extensively used for instance, to observe structural differences between cancer and non-cancer related networks (Reyna et al., 2020; Wang et al., 2020). These differences, often carry functional features that may help to understand such complex phenotypes (Miecznikowski et al., 2016; Drago-García et al., 2017; de Anda-Jáuregui et al., 2019; Dorantes-Gilardi et al., 2020).

Finding functional modules in cancer has been a matter of intense research. A common method to infer such modules resorts to the so-called *Weighted gene co-expression network analysis (WGCNA)* (Zhang and Horvath, 2005; Langfelder et al., 2008). In this method, Pearson correlation is used to evaluate pairwise gene co-expression. Such co-expression network can be decomposed into modules by using different methods.

For instance, in Ai et al. (2020), the authors used the dynamic tree cut method (Langfelder and Horvath, 2008) to infer modules in a microarray-based colorectal cancer (CRC) gene co-expression network. This method improves the classic hierarchical clustering that sets a fixed cutoff value. A dynamic branch cutting depending on the dendrogram shape is implemented. With this approach, Ai and cols., found that GUCA2A, GUCA2B, and CDH3 genes were highly correlated with the occurrence of CRC.

Along similar lines, WGCNA was used to analyze 182 CRC and 54 normal samples (Qiu et al., 2020). There, a k-means clustering was used to find modules, and the hub genes from those modules were separated into samples with high and low expression. The authors identified that overexpression of MYL9, MYLK, and CNN1 genes was associated with poorer outcome in CRC patients.

In breast cancer, efforts have been made to observe modules that may be underlying functional processes (Wilkinson and Huberman, 2004; Zhu et al., 2008; Cantini et al., 2015). It is widely known that breast cancer is a highly heterogeneous disease. This heterogeneity can be traced down to the genetic level (Alcalá-Corona et al., 2017).

Molecular subtyping provides a helpful tool to classify tumors by identifying common patterns in their genetic expression. One of the most used classification methods is PAM50 (Sørlie et al., 2001). Samples are grouped based on the molecular signature. With this method, breast cancer can be divided into four main differentiated subtypes: Luminal A, Luminal B, HER2+, and Basal-like. Each subtype has a different clinical and histopathological manifestation.

Network approaches to identify modules in breast cancer molecular subtypes has been a matter of intense research. For instance, the `infomap` algorithm has been used to reveal functional modules in HER2+ breast cancer transcriptional network (Alcalá-Corona et al., 2018b). Additionally, it has been observed that in the HER2+ tumors related network, a hierarchical modular structure appears (Alcalá-Corona et al., 2018a).

In basal-like breast cancer, network modularity has been used to observe functional modules and discern whether or not those modules are shared between the cancer and the non-cancer network (de Anda-Jáuregui et al., 2019). It has been observed that the basal breast cancer has a different distribution of module size between cancer and non-cancer networks (de Anda-Jáuregui et al., 2019). Additionally, those modules are composed of different genes.

In all those cases, cancer networks are formed by small connected same-chromosome gene components. Often, said components coincide with modules independent of the community detection method. However, this is not always the case. For example, in García-Cortés et al. (2021), for Luminal A breast cancer, an RNA-Seq-derived gene co-expression network was decomposed into communities by using four different methods: Fast greedy (Clauset et al., 2004), Infomap (Rosvall and Bergstrom, 2008), Leading eigenvector (Newman, 2006b) and Louvain (Blondel et al., 2008).

The aforementioned methods have different postulates and different approaches to detect communities. In that work (García-Cortés et al., 2021) it was demonstrated that, independent of the algorithm used to detect communities, the results were very similar in terms of the number of detected communities and the nature of the genes observed in each community.

Despite modules being quite similar, independently of the method to detect them (Jaccard indexes between modules obtained by the different methods, are larger than 0.95), the algorithm with optimal modularity was the Louvain method. Interestingly, Modularity is larger in the case of Luminal A network than the healthy network, for all methods.

An additional effect observed when comparing cancer and non-cancer derived networks, is a high proportion of same-chromosome gene-gene interactions in cancer phenotypes. On the other hand, healthy tissue-derived networks are composed of interactions between genes from any chromosome in a homogeneous fashion. This phenomenon has been called *loss of long-distance co-expression in cancer* (Espinal-Enríquez et al., 2017). This abrupt change has been reported for different tissues such as breast cancer (Espinal-Enríquez et al., 2017; de Anda-Jáuregui et al., 2019), each breast cancer molecular subtype (García-Cortés et al., 2020), clear cell renal carcinoma (Zamora-Fuentes et al., 2020), lung adenocarcinoma and lung sqamous cell carcinoma (Andonegui-Elguera et al., 2021). It is worth noticing that modularity has been used as an indirect measure of coordinated gene function (Solé et al., 2002; Segal et al., 2003; Lee et al., 2004; Tanay et al., 2004; Zhu et al., 2008). In this case, modules do not always represent gene function, but often act as a proxy for *spatial clustering* between genes from the same chromosome.

The studies just mentioned are just a handful instances, illustrating how network modularity determination is a becoming an essential approach to biological discovery.

7. CONCLUDING REMARKS

As we have already discussed, complexity in biological systems can be understood partially by using network approaches. Modularity is often an inherent component of complex biological networks. However relevant, network modularity discovery (or community detection, as is also called) is a daunting task. Its importance in theoretical biology, to describe the emergence of functional behaviors in biological systems, as well as its use in understanding the underlying principles behind such functionality make it a worthy tool in biology.

In the past years, a number of relevant approaches to this problem have been developed in the computational and systems biology settings. Most of these approaches, although extremely informative are built upon *Ad Hoc* assumptions and are thus not easy to generalize. Hence, they provide useful information, but are too specific. On then other hand, the network science and statistical physics research communities have been developing a series of quite general modularity detection algorithms. Here we present some of them, organized as *families* of methods, depending on their methodological foundations: (i) clustering algorithms, (ii) modularity optimization methods, (iii) methods based on the spectral properties of adjacency matrices, (iv) methods based on random walks and (v) methods based on stochastic block models. These broad families of methods along with the benchmarks that have been developed to evaluate their performance may constitute a relevant toolbox for the analysis of biological systems from a more general perspective. We argue that by resorting to these methods (freed from the design constraints typical of *Ad Hoc* methods) will allow to focus on the actual biology rather than on the method's specificities.

The problem of modularity and the discovery of functional communities in biological networks is an important emerging field of research. Omic high throughput technologies and the rise of computing power as well as the development of novel analytical algorithms have allowed the generation of bio-molecular network models at an unprecedented pace. This has led us with the need to develop theoretical and computational tools to extract biologically useful (e.g., functional or mechanistic) information from such large scale models. A wide variety of biological questions that can be answered—at least partially—by knowing the modular structure of the underlying networks, are being added to the current research scenario in the systems biology and genomics communities. A number of powerful mathematical and computational schemes to deal with modularity are also currently under development.

In the preceding review, we have discussed both, the biological problems and the computational approaches to the problem of modularity in complex bio-molecular networks. It is our sincere desire that works like this will stimulate the discussion between researchers in all the involved fields. A discussion that may in turn strengthen the ties of collaboration and ultimately leads to fruitful cross-fertilized scientific discoveries.

AUTHOR CONTRIBUTIONS

SA-C and EH-L: conceived the idea, contributed to the writing of the manuscript, and revised the manuscript. SS-M and JE-E: contributed to the writing of the manuscript and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by CONACYT (grant no. 179431/2012), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). Additional support has been granted by the National Laboratory of Complexity Sciences (grant no. 232647/2014 CONACYT). SS-M appreciates the support from CONACyT through the Catedras-CONACyT program. EH-L also acknowledges additional support from the 2016 Marcos Moshinsky Research Chair in the Physical Sciences. The funders have no role in the design or development of this work.

REFERENCES

Acharya, L., Judeh, T., and Zhu, D. (2012). "A survey of computational approaches to reconstruct and partition biological networks," in *Statistical and Machine Learning Approaches for Network Analysis*, eds M. Dehmer and S. C. Basak (New Jersey: Wiley), 1. doi: 10.1002/9781118346990.ch1

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023. doi: 10.1093/bioinformatics/btl039

Ai, D., Wang, Y., Li, X., and Pan, H. (2020). Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder. *Biomolecules* 10:1207. doi: 10.3390/biom10091207

Alcalá-Corona, S. A., de Anda-Jáuregui, G., Espinal-Enriquez, J., and Hernández-Lemus, E. (2017). Network modularity in breast cancer molecular subtypes. *Front. Physiol*. 8:915. doi: 10.3389/fphys.2017.00915

Alcalá-Corona, S. A., de Anda-Jáuregui, G., Espinal-Enriquez, J., Tovar, H., and Hernández-Lemus, E. (2018a). "Network modularity and hierarchical structure in breast cancer molecular subtypes," in *International Conference on Complex Systems* (Cham: Springer), 352–358. doi: 10.1007/978-3-319-96661-8_36

Alcalá-Corona, S. A., Espinal-Enriquez, J., De Anda Jáuregui, G., and Hernandez-Lemus, E. (2018b). The hierarchical modular structure of HER2+ breast cancer network. *Front. Physiol*. 9:1423. doi: 10.3389/fphys.2018.01423

Alcalá-Corona, S. A., Velázquez-Caldelas, T. E., Espinal-Enriquez, J., and Hernández-Lemus, E. (2016). Community structure reveals biologically functional modules in MEF2C transcriptional regulatory network. *Front. Physiol*. 7:184. doi: 10.3389/fphys.2016.00184

Aldana, M., Balleza, E., Kauffman, S., and Resendiz, O. (2007). Robustness and evolvability in genetic regulatory networks. *J. Theoret. Biol*. 245, 433–448. doi: 10.1016/j.jtbi.2006.10.027

Aldana, M., and Cluzel, P. (2003). A natural class of robust networks. *Proc. Natl. Acad. Sci. U.S.A*. 100, 8710–8714. doi: 10.1073/pnas.1536783100

Aldecoa, R., and Marin, I. (2014). Surpriseme: an integrated tool for network community structure characterization using surprise maximization. *Bioinformatics* 30, 1041–1042. doi: 10.1093/bioinformatics/btt741

Ana, L., and Jain, A. K. (2003). "Robust data clustering," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Madison, WI: IEEE), 2–128. doi: 10.1109/CVPR.2003.1211462

Andonegui-Elguera, S. D., Zamora-Fuentes, J. M., Espinal-Enriquez, J., and Hernández-Lemus, E. (2021). Loss of long distance co-expression in lung cancer. *Front. Genet*. 12:625741. doi: 10.3389/fgene.2021.625741

Arenas, A., and Diaz-Guilera, A. (2007). Synchronization and modularity in complex networks. *Eur. Phys. J. Spcl. Top*. 143, 19–25. doi: 10.1140/epjst/e2007-00066-2

Arenas, A., Díaz-Guilera, A., and Pérez-Vicente, C. J. (2006). Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett*. 96:114102. doi: 10.1103/PhysRevLett.96.114102

Arenas, A., Fernandez, A., Fortunato, S., and Gomez, S. (2008). Motif-based communities in complex networks. *J. Phys. A Math. Theoret*. 41:224001. doi: 10.1088/1751-8113/41/22/224001

Ashrafian, H., McKenna, W. J., and Watkins, H. (2011). Disease pathways and novel therapeutic targets in hypertrophic cardiomyopathy. *Circ. Res*. 109, 86–96. doi: 10.1161/CIRCRESAHA.111.242974

Aslak, U., Rosvall, M., and Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. *Phys. Rev. E* 97:062312. doi: 10.1103/PhysRevE.97.062312

Banerjee, K., Kolomeisky, A. B., and Igoshin, O. A. (2017). Accuracy of substrate selection by enzymes is controlled by kinetic discrimination. *J. Phys. Chem. Lett*. 8, 1552–1556. doi: 10.1021/acs.jpclett.7b00441

Barabasi, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*. 5, 101–113. doi: 10.1038/nrg1272

Barkai, N., and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature* 387, 913–917. doi: 10.1038/43199

Barnes, E. R. (1982). An algorithm for partitioning the nodes of a graph. *SIAM J. Alg. Disc. Meth*. 3, 541–550.

Bennett, L., Kittas, A., Muirhead, G., Papageorgiou, L. G., and Tsoka, S. (2015). Detection of composite communities in multiplex biological networks. *Sci. Rep*. 5, 1–12. doi: 10.1038/srep10345

Bettinelli, A., Hansen, P., and Liberti, L. (2012). Algorithm for parametric community detection in networks. *Phys. Rev. E* 86:016107. doi: 10.1103/PhysRevE.86.016107

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp*. 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., et al. (2008). On modularity clustering. *IEEE Trans. Knowl. Data Eng*. 20, 172–188. doi: 10.1109/TKDE.2007.190689

Britten, R. J., and Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science* 165, 349–357. doi: 10.1126/science.165.3891.349

Calderer, G., and Kuijjer, M. L. (2021). Community detection in large-scale bipartite biological networks. *Front. Genet*. 12:520. doi: 10.3389/fgene.2021.649440

Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep*. 5:17386. doi: 10.1038/srep17386

Capocci, A., Servedio, V. D., Caldarelli, G., and Colaiori, F. (2005). Detecting communities in large networks. *Phys. A Stat. Mech. Appl*. 352, 669–676. doi: 10.1016/j.physa.2004.12.050

Chen, H.-R., Sherr, D. H., Hu, Z., and DeLisi, C. (2016). A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer. *BMC Med. Genomics* 9:1. doi: 10.1186/s12920-016-0212-7

Chen, J., and Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 32, 1724–1732. doi: 10.1093/bioinformatics/btw059

Cheng, L., Liu, P., Wang, D., and Leung, K.-S. (2019). Exploiting locational and topological overlap model to identify modules in protein interaction networks. *BMC Bioinformatics* 20:23. doi: 10.1186/s12859-019-2598-7

Cherifi, H., Palla, G., Szymanski, B. K., and Lu, X. (2019). On community structure in complex networks: challenges and opportunities. *Appl. Network Sci*. 4, 1–35. doi: 10.1007/s41109-019-0238-9

Clarke, B. S., and Mittenthal, J. E. (1992). Modularity and reliability in the organization of organisms. *Bull. Math. Biol*. 54, 1–20. doi: 10.1016/S0092-8240(05)80173-9

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* 70:066111. doi: 10.1103/PhysRevE.70.066111

Clune, J., Mouret, J.-B., and Lipson, H. (2013). The evolutionary origins of modularity. *Proc. R. Soc. B Biol. Sci*. 280:20122863. doi: 10.1098/rspb.2012.2863

Condon, A., and Karp, R. M. (2001). Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms* 18, 116–140. doi: 10.1002/1098-2418(200103)18:2<116::AID-RSA1001>3.0.CO;2-2

Constantino, P. H., and Daoutidis, P. (2019). A control perspective on the evolution of biological modularity. *IFAC Pap. Online* 52, 172–177. doi: 10.1016/j.ifacol.2019.09.136

da Fonseca Vieira, V., Xavier, C. R., and Evsukoff, A. G. (2020). A comparative study of overlapping community detection methods from the perspective of the structural properties. *Appl. Network Sci*. 5, 1–42. doi: 10.1007/s41109-020-00289-9

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *J. Stat. Mech. Theory Exp*. 2005:P09008. doi: 10.1088/1742-5468/2005/09/P09008

Davidson, E., and Levin, M. (2005). Gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A*. 102, 4935–4935. doi: 10.1073/pnas.0502024102

de Anda-Jáuregui, G., Alcalá-Corona, S. A., Espinal-Enriquez, J., and Hernández-Lemus, E. (2019). Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Appl. Network Sci*. 4, 1–13. doi: 10.1007/s41109-019-0129-0

De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X* 5:011027. doi: 10.1103/PhysRevX.5.011027

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., et al. (2013). Mathematical formulation of multilayer networks. *Phys. Rev. X* 3:041022. doi: 10.1103/PhysRevX.3.041022

De Leo, V., Santoboni, G., Cerina, F., Mureddu, M., Secchi, L., and Chessa, A. (2013). Community core detection in transportation networks. *Phys. Rev. E* 88:042810. doi: 10.1103/PhysRevE.88.042810

de Matos Simoes, R., Tripathi, S., and Emmert-Streib, F. (2012). Organizational structure and the periphery of the gene regulatory network in B-cell lymphoma. *BMC Syst. Biol*. 6:1. doi: 10.1186/1752-0509-6-38

Debnath, S., Rakshit, S., Sengupta, K., and Plewczynski, D. (2021). "Biomolecular clusters identification in linear time complexity for biological networks," in *Proceedings of International Conference on Frontiers in Computing and Systems* (Singapore: Springer), 611–622. doi: 10.1007/978-981-15-7834-2_57

Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. *PeerJ* 3:1042. doi: 10.7717/peerj.1525

Didier, G., Valdeolivas, A., and Baudot, A. (2018). Identifying communities from multiplex biological networks by randomized optimization of modularity. *F1000Research* 7:1042. doi: 10.12688/f1000research.15486.1

Donath, W. E., and Hoffman, A. J. (1972). Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices. *IBM Tech. Disclosure Bull*. 15, 938–944.

Donetti, L., and Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *J. Stat. Mech. Theory Exp*. 2004:P10012. doi: 10.1088/1742-5468/2004/10/P10012

Dorantes-Gilardi, R., Garcia-Cortés, D., Hernández-Lemus, E., and Espinal-Enriquez, J. (2020). Multilayer approach reveals organizational principles disrupted in breast cancer co-expression networks. *Appl. Network Sci*. 5, 1–23. doi: 10.1007/s41109-020-00291-1

Drago-García, D., Espinal-Enriquez, J., and Hernández-Lemus, E. (2017). Network analysis of emt and met micro-rna regulation in breast cancer. *Sci. Rep*. 7:13534. doi: 10.1038/s41598-017-13903-1

Duch, J., and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Phys. Rev. E* 72:027104. doi: 10.1103/PhysRevE.72.027104

Emmons, S., and Mucha, P. J. (2019). Map equation with metadata: varying the role of attributes in community detection. *Phys. Rev. E* 100:022301. doi: 10.1103/PhysRevE.100.022301

Eriksson, A., Carletti, T., Lambiotte, R., Rojas, A., and Rosvall, M. (2021). Flow-based community detection in hypergraphs. *arXiv preprint arXiv:2105.04389.*

Espinal, J., Aldana, M., Guerrero, A., Wood, C., Darszon, A., and Martinez-Mekler, G. (2011). Discrete dynamics model for the speract-activated ca 2+ signaling network relevant to sperm motility. *PLoS ONE* 6:e22619. doi: 10.1371/journal.pone.0022619

Espinal-Enríquez, J., Fresno, C., Anda-Jáuregui, G., and Hernández-Lemus, E. (2017). RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep*. 7:1760. doi: 10.1038/s41598-017-01314-1

Espinal-Enriquez, J., Priego-Espinosa, D. A., Darszon, A., Beltrán, C., and Martinez-Mekler, G. (2017). Network model predicts that catsper is the main ca 2+ channel in the regulation of sea urchin sperm motility. *Sci. Rep*. 7, 1–14. doi: 10.1038/s41598-017-03857-9

Espinosa-Soto, C., and Wagner, A. (2010). Specialization can drive the evolution of modularity. *PLoS Comput. Biol*. 6:e1000719. doi: 10.1371/journal.pcbi.1000719

Esquivel, A. V., and Rosvall, M. (2011). Compression of flow can reveal overlapping-module organization in networks. *Phys. Rev. X* 1:021025. doi: 10.1103/PhysRevX.1.021025

Fan, Y., Li, M., Zhang, P., Wu, J., and Di, Z. (2007). Accuracy and precision of methods for community identification in weighted networks. *Phys. A Stat. Mech. Appl*. 377, 363–372. doi: 10.1016/j.physa.2006.11.036

Farage, C., Edler, D., Eklöf, A., Rosvall, M., and Pilosof, S. (2021). Identifying flow modules in ecological networks using infomap. *Methods Ecol. Evol*. 12, 778–786. doi: 10.1111/2041-210X.13569

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math. J*. 23, 298–305. doi: 10.21136/CMJ.1973.101168

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep*. 486, 75–174. doi: 10.1016/j.physrep.2009.11.002

Fortunato, S., and Barthelemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A*. 104, 36–41. doi: 10.1073/pnas.0605965104

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* 40, 35–41. doi: 10.2307/3033543

Friedlander, T., Mayo, A. E., Tlusty, T., and Alon, U. (2013). Mutation rules and the evolution of sparseness and modularity in biological systems. *PLoS ONE* 8:e70444. doi: 10.1371/journal.pone.0070444

Gao, S., Chen, A., Rahmani, A., Zeng, J., Tan, M., Alhajj, R., et al. (2016). Multi-scale modularity and motif distributional effect in metabolic networks. *Curr. Protein Peptide Sci*. 17, 82–92. doi: 10.2174/1389203716666150923104603

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernandez-Lemus, E., and Espinal-Enriquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol*. 10:1232. doi: 10.3389/fonc.2020.01232

García-Cortés, D. E., Hernandez-Lemus, E., and Espinal-Enriquez, J. (2021). Luminal a breast cancer co-expression network: structural and functional alterations. *Front. Genet*. 12:514. doi: 10.3389/fgene.2021.629475

Garza, S. E., and Schaeffer, S. E. (2019). Community detection with the label propagation algorithm: a survey. *Phys. A Stat. Mech. Appl*. 534:122058. doi: 10.1016/j.physa.2019.122058

Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol*. 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Ghiassian, S. D., Menche, J., Chasman, D. I., Giulianini, F., Wang, R., Ricchiuto, P., et al. (2016). Endophenotype network models: common core of complex diseases. *Sci. Rep*. 6:27414. doi: 10.1038/srep27414

Gibson, G. (2016). On the evaluation of module preservation. *Cell Syst*. 3, 17–19. doi: 10.1016/j.cels.2016.07.009

Gilarranz, L. J. (2020). Generic emergence of modularity in spatial networks. *Sci. Rep*. 10, 1–8. doi: 10.1038/s41598-020-65669-8

Gilbert, J. P. (2015). *A probabilistic model for the evaluation of module extraction algorithms in complex biological networks* (Ph.D. thesis). University of Nottingham, Nottingham, United Kingdom.

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A*. 99, 7821–7826. doi: 10.1073/pnas.122653799

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A*. 104, 8685–8690. doi: 10.1073/pnas.0701361104

Gómez-Romero, L., López-Reyes, K., and Hernández-Lemus, E. (2020). The large scale structure of human metabolism reveals resilience via extensive signaling crosstalk. *Front. Physiol*. 11:1667. doi: 10.3389/fphys.2020.588012

Gong, M., Fu, B., Jiao, L., and Du, H. (2011). Memetic algorithm for community detection in networks. *Phys. Rev. E* 84:056101. doi: 10.1103/PhysRevE.84.056101

Green, S., Şerban, M., Scholl, R., Jones, N., Brigandt, I., and Bechtel, W. (2018). Network analyses in systems biology: new strategies for dealing with biological complexity. *Synthese* 195, 1751–1777. doi: 10.1007/s11229-016-1307-6

Guimera, R., and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature* 433, 895–900. doi: 10.1038/nature03288

Guimera, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E* 70:025101. doi: 10.1103/PhysRevE.70.025101

Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based *in silico* drug efficacy screening. *Nat. Commun*. 7:10331. doi: 10.1038/ncomms10331

Gyorgy, A., and Del Vecchio, D. (2014). Modular composition of gene transcription networks. *PLoS Comput. Biol*. 10:e1003486. doi: 10.1371/journal.pcbi.1003486

Henriques, R., and Madeira, S. C. (2016). Bicnet: Flexible module discovery in large-scale biological networks using biclustering. *Algorithms Mol. Biol*. 11:1. doi: 10.1186/s13015-016-0074-8

Hernández-Lemus, E., Reyes-Gopar, H., Espinal-Enriquez, J., and Ochoa, S. (2019). The many faces of gene regulation in cancer: a computational oncogenomics outlook. *Genes* 10:865. doi: 10.3390/genes10110865

Hric, D., Darst, R. K., and Fortunato, S. (2014). Community detection in networks: structural communities versus ground truth. *Phys. Rev. E* 90:062805. doi: 10.1103/PhysRevE.90.062805

Iacovacci, J., and Bianconi, G. (2016). Extracting information from multiplex networks. *Chaos* 26:065306. doi: 10.1063/1.4953161

Igoshin, O. A., Brody, M. S., Price, C. W., and Savageau, M. A. (2007). Distinctive topologies of partner-switching signaling networks correlate with their physiological roles. *J. Mol. Biol*. 369, 1333–1352. doi: 10.1016/j.jmb.2007.04.021

Jaeger, J., and Monk, N. (2021). Dynamical modules in metabolism, cell and developmental biology. *Interface Focus* 11:20210011. doi: 10.1098/rsfs.2021.0011

Jebabli, M., Cherifi, H., Cherifi, C., and Hamouda, A. (2018). Community detection algorithm evaluation with ground-truth data. *Phys. A Stat. Mech. Appl*. 492, 651–706. doi: 10.1016/j.physa.2017.10.018

Karrer, B., and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* 83:016107. doi: 10.1103/PhysRevE.83.016107

Kashtan, N., and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci. U.S.A*. 102, 13773–13778. doi: 10.1073/pnas.0503610102

Kashtan, N., Parter, M., Dekel, E., Mayo, A. E., and Alon, U. (2009). Extinctions in heterogeneous environments and the evolution of modularity. *Evol. Int. J. Organ. Evol*. 63, 1964–1975. doi: 10.1111/j.1558-5646.2009.00684.x

Kauffman, S. (1969). Homeostasis and differentiation in random genetic control networks. *Nature* 224, 177–178. doi: 10.1038/224177a0

Kawamoto, T., and Rosvall, M. (2015). Estimating the resolution limit of the map equation in community detection. *Phys. Rev. E* 91:012809. doi: 10.1103/PhysRevE.91.012809

Khadivi, A., Rad, A. A., and Hasler, M. (2011). Network community-detection enhancement by proper weighting. *Phys. Rev. E* 83:046104. doi: 10.1103/PhysRevE.83.046104

Kheirkhahzadeh, M., Lancichinetti, A., and Rosvall, M. (2016). Efficient community detection of network flows for varying Markov times and bipartite networks. *Phys. Rev. E* 93:032309. doi: 10.1103/PhysRevE.93.032309

Knuth, D. E. (1985). Dynamic huffman coding. *J. Algorithms* 6, 163–180. doi: 10.1016/0196-6774(85)90036-7

Kovács, I. A., and Barabási, A.-L. (2015). Network science: destruction perfected. *Nature* 524, 38–39. doi: 10.1038/524038a

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., et al. (2013). Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. U.S.A*. 110, 20935–20940. doi: 10.1073/pnas.1312486110

Kuncheva, L. I., and Hadjitodorov, S. T. (2004). "Using diversity in cluster ensembles," in *IEEE International Conference on Systems, Man and Cybernetics, 2004* (The Hague: IEEE), 1214–1219. doi: 10.1109/ICSMC.2004.1399790

Lancichinetti, A., and Fortunato, S. (2009a). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E* 80:016118. doi: 10.1103/PhysRevE.80.016118

Lancichinetti, A., and Fortunato, S. (2009b). Community detection algorithms: a comparative analysis. *Phys. Rev. E* 80:056117. doi: 10.1103/PhysRevE.80.056117

Lancichinetti, A., and Fortunato, S. (2011). Limits of modularity maximization in community detection. *Phys. Rev. E* 84:066122. doi: 10.1103/PhysRevE.84.066122

Lancichinetti, A., and Fortunato, S. (2012). Consensus clustering in complex networks. *Sci. Rep*. 2, 1–7. doi: 10.1038/srep00336

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *N. J. Phys*. 11:033015. doi: 10.1088/1367-2630/11/3/033015

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* 78:046110. doi: 10.1103/PhysRevE.78.046110

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLoS ONE* 6:e18961. doi: 10.1371/journal.pone.0018961

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Lecca, P., and Re, A. (2015). Detecting modules in biological networks by edge weight clustering and entropy significance. *Front. Genet*. 6:265. doi: 10.3389/fgene.2015.00265

Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* 306, 1555–1558. doi: 10.1126/science.1099511

Li, Y., Liu, B., Li, J., and Li, G. (2019). Mimod: a new algorithm for mining biological network modules. *IEEE Access* 7, 49492–49503. doi: 10.1109/ACCESS.2019.2909946

Liu, J., Jing, L., and Tu, X. (2016). Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *BMC Cardiovasc. Disord*. 16:1. doi: 10.1186/s12872-016-0217-3

Long, J., Liu, Z., Wu, X., Xu, Y., and Ge, C. (2016). Screening for genes and subnetworks associated with pancreatic cancer based on the gene expression profile. *Mol. Med. Rep*. 13, 3779–3786. doi: 10.3892/mmr.2016.5007

Lorenz, D. M., Jeng, A., and Deem, M. W. (2011). The emergence of modularity in biological systems. *Phys. Life Rev*. 8, 129–160. doi: 10.1016/j.plrev.2011.02.003

Lucchetta, M., and Pellegrini, M. (2020). Finding disease modules for cancer and covid-19 in gene co-expression networks with the core&peel method. *Sci. Rep*. 10, 1–18. doi: 10.1038/s41598-020-74705-6

Lusseau, D. (2007). Evidence for social role in a dolphin social network. *Evol. Ecol*. 21, 357–366. doi: 10.1007/s10682-006-9105-0

Mall, R., Cerulo, L., Bensmail, H., Iavarone, A., and Ceccarelli, M. (2017a). Detection of statistically significant network changes in complex biological networks. *BMC Syst. Biol*. 11:32. doi: 10.1186/s12918-017-0412-6

Mall, R., Ullah, E., Kunji, K., D'Angelo, F., Bensmail, H., and Ceccarelli, M. (2017b). "Differential community detection in paired biological networks," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Boston, MA, 330–339. doi: 10.1145/3107411.3107418

Medus, A., and Dorso, C. (2009). Alternative approach to community detection in networks. *Phys. Rev. E* 79:066111. doi: 10.1103/PhysRevE.79.066111

Meilă, M. (2007). Comparing clusterings? An information based distance. *J. Multivariate Anal*. 98, 873–895. doi: 10.1016/j.jmva.2006.11.013

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Miecznikowski, J. C., Gaile, D. P., Chen, X., and Tritchler, D. L. (2016). Identification of consistent functional genetic modules. *Stat. Appl. Genet. Mol. Biol*. 15, 1–18. doi: 10.1515/sagmb-2015-0026

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* 298, 824–827. doi: 10.1126/science.298.5594.824

Mittal, R., and Bhatia, M. (2018). "Analyzing the structures of clusters in multi-layer biological networks," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)* (Jalandhar: IEEE), 502–507. doi: 10.1109/ICSCCC.2018.8703271

Monzón-Sandoval, J., Castillo-Morales, A., Urrutia, A. O., and Gutierrez, H. (2016). Modular reorganization of the global network of gene regulatory interactions during perinatal human brain development. *BMC Dev. Biol*. 16:1. doi: 10.1186/s12861-016-0111-3

Moreira-Filho, C. A., Bando, S. Y., Bertonha, F. B., Iamashita, P., Silva, F. N., da Fontoura Costa, L., et al. (2015). Community structure analysis of transcriptional networks reveals distinct molecular pathways for early-and late-onset temporal lobe epilepsy with childhood febrile seizures. *PLoS ONE* 10:e0128174. doi: 10.1371/journal.pone.0128174

Morohashi, M., Winn, A. E., Borisuk, M. T., Bolouri, H., Doyle, J., and Kitano, H. (2002). Robustness as a measure of plausibility in models of biochemical networks. *J. Theoret. Biol*. 216, 19–30. doi: 10.1006/jtbi.2002.2537

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328, 876–878. doi: 10.1126/science.1184819

Muraro, D., and Simmons, A. (2016). An integrative analysis of gene expression and molecular interaction data to identify dys-regulated sub-networks in inflammatory bowel disease. *BMC Bioinformatics* 17:1. doi: 10.1186/s12859-016-0886-z

Narula, J., Smith, A. M., Gottgens, B., and Igoshin, O. A. (2010). Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate. *PLoS Comput. Biol*. 6:e1000771. doi: 10.1371/journal.pcbi.1000771

Newman, M. (2010). *Networks: An Introduction*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199206650.003.0001

Newman, M. E. (2004a). Detecting community structure in networks. *Eur. Phys. J. B* 38, 321–330. doi: 10.1140/epjb/e2004-00124-y

Newman, M. E. (2004b). Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69:066133. doi: 10.1103/PhysRevE.69.066133

Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74:036104. doi: 10.1103/PhysRevE.74.036104

Newman, M. E. (2006b). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A*. 103, 8577–8582. doi: 10.1073/pnas.0601602103

Newman, M. E. (2012). Communities, modules and large-scale structure in networks. *Nat. Phys.* 8, 25–31. doi: 10.1038/nphys2162

Newman, M. E. (2013). Spectral methods for community detection and graph partitioning. *Phys. Rev. E* 88:042822. doi: 10.1103/PhysRevE.88.042822

Newman, M. E., and Girvan, M. (2003). "Mixing patterns and community structure in networks," in *Statistical Mechanics of Complex Networks* (Heidelberg: Springer), 66–87. doi: 10.1007/978-3-540-44943-0_5

Newman, M. E., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69:026113. doi: 10.1103/PhysRevE.69.026113

Oliveira, J. V., de Brito, A. F., Braconi, C. T., de Melo Freire, C. C., Iamarino, A., and de Andrade Zanotto, P. M. (2013). Modularity and evolutionary constraints in a baculovirus gene regulatory network. *BMC Syst. Biol.* 7:1. doi: 10.1186/1752-0509-7-87

Orman, G. K., Labatut, V., and Cherifi, H. (2012). Comparative evaluation of community detection algorithms: a topological approach. *J. Stat. Mech. Theory Exp*. 2012:P08001. doi: 10.1088/1742-5468/2012/08/P08001

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818. doi: 10.1038/nature03607

Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC Evol. Biol.* 7:169. doi: 10.1186/1471-2148-7-169

Peixoto, T. P. (2018). Nonparametric weighted stochastic block models. *Phys. Rev. E* 97:012306. doi: 10.1103/PhysRevE.97.012306

Pesantez-Cabrera, P., and Kalyanaraman, A. (2016). "Detecting communities in biological bipartite networks," in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Seattle, WA, 98–107. doi: 10.1145/2975167.2975177

Pilosof, S., Alcala-Corona, S. A., Wang, T., Kim, T., Maslov, S., Whitaker, R., et al. (2020). The network structure and eco-evolutionary dynamics of crispr-induced immune diversification. *Nat. Ecol. Evol.* 4, 1650–1660. doi: 10.1038/s41559-020-01312-z

Pons, P., and Latapy, M. (2005). "Computing communities in large networks using random walks," in *International Symposium on Computer and Information Sciences* (Poznan: Springer), 284–293. doi: 10.1007/11569596_31

Pons, P., and Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* 10, 191–218. doi: 10.7155/jgaa.00124

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices AMS* 56, 1082–1097.

Praneenararat, T., Takagi, T., and Iwasaki, W. (2011). Interactive, multiscale navigation of large and complicated biological networks. *Bioinformatics* 27, 1121–1127. doi: 10.1093/bioinformatics/btr083

Qi, Y., and Ge, H. (2006). Modularity and dynamics of cellular networks. *PLoS Comput. Biol.* 2:e174. doi: 10.1371/journal.pcbi.0020174

Qiu, X., Cheng, S.-H., Xu, F., Yin, J.-W., Wang, L.-Y., and Zhang, X.-Y. (2020). Weighted gene co-expression network analysis identified MYL9 and CNN1 are associated with recurrence in colorectal cancer. *J. Cancer* 11:2348. doi: 10.7150/jca.39723

Rahiminejad, S., Maurya, M. R., and Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics* 20:212. doi: 10.1186/s12859-019-2746-0

Ramadan, E., Naef, A., and Ahmed, M. (2016). Protein complexes predictions within protein interaction networks using genetic algorithms. *BMC Bioinformatics* 17:481. doi: 10.1186/s12859-016-1096-4

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374

Reichardt, J., and Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett*. 93:218701. doi: 10.1103/PhysRevLett.93.218701

Reichardt, J., and Bornholdt, S. (2006). Statistical mechanics of community detection. *Phys. Rev. E* 74:016110. doi: 10.1103/PhysRevE.74.016110

Reyna, M. A., Haan, D., Paczkowska, M., Verbeke, L. P., Vazquez, M., Kahraman, A., et al. (2020). Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* 11, 1–17. doi: 10.1038/s41467-020-14367-0

Ritchie, S. C., Watts, S., Fearnley, L. G., Holt, K. E., Abraham, G., and Inouye, M. (2016). A scalable permutation approach reveals replication and preservation patterns of network modules in large datasets. *Cell Syst*. 3, 71–82. doi: 10.1016/j.cels.2016.06.012

Rosvall, M., and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. U.S.A*. 104, 7327–7331. doi: 10.1073/pnas.0611034104

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A*. 105, 1118–1123. doi: 10.1073/pnas.0706851105

Rosvall, M., and Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS ONE* 6:e18209. doi: 10.1371/journal.pone.0018209

Sah, P., Singh, L. O., Clauset, A., and Bansal, S. (2014). Exploring community structure in biological networks with random graphs. *BMC Bioinformatics* 15:220. doi: 10.1186/1471-2105-15-220

Samal, A., Singh, S., Giri, V., Krishna, S., Raghuram, N., and Jain, S. (2006). Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* 7:118. doi: 10.1186/1471-2105-7-118

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., Euzenat, J., et al. (1999). Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. *Nucleic Acids Res*. 27, 89–94. doi: 10.1093/nar/27.1.89

Sardiu, M. E., Gilmore, J. M., Groppe, B., Florens, L., and Washburn, M. P. (2017). Identification of topological network modules in perturbed protein interaction networks. *Sci. Rep*. 7, 1–13. doi: 10.1038/srep43845

Schulz, S., Eckweiler, D., Bielecka, A., Nicolai, T., Franke, R., Dötsch, A., et al. (2015). Elucidation of sigma factor-associated networks in *Pseudomonas aeruginosa* reveals a modular architecture with limited and function-specific crosstalk. *PLoS Pathog*. 11:e1004744. doi: 10.1371/journal.ppat.1004744

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*. 34, 166–176. doi: 10.1038/ng1165

Şen, F., Wigand, R. T., Agarwal, N., Mete, M., and Kasprzyk, R. (2014). "Focal structure analysis in large biological networks," in *3rd International Conference on Environment, Energy and Biotechnology (ICEEB 2014)*, Bangkok.

Serban, M. (2020). Exploring modularity in biological networks. *Philos. Trans. R. Soc. B* 375:20190316. doi: 10.1098/rstb.2019.0316

Servis, M. J., and Clark, A. E. (2021). Cluster identification using modularity optimization to uncover chemical heterogeneity in complex solutions. *J. Phys. Chem. A* 125, 3986–3993. doi: 10.1021/acs.jpca.0c11320

Sevim, V., and Rikvold, P. A. (2008). Chaotic gene regulatory networks can be robust against mutations and noise. *J. Theoret. Biol*. 253, 323–332. doi: 10.1016/j.jtbi.2008.03.003

Shi, Z., Derow, C. K., and Zhang, B. (2010). Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst. Biol*. 4:1. doi: 10.1186/1752-0509-4-74

Shmulevich, I., Kauffman, S. A., and Aldana, M. (2005). Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proc. Natl. Acad. Sci. U.S.A*. 102, 13439–13444. doi: 10.1073/pnas.0506771102

Smith, E. P. (1985). Statistical comparison of weighted overlap measures. *Trans. Am. Fish. Soc*. 114, 250–257. doi: 10.1577/1548-8659(1985)114<250:SCOWOM>2.0.CO;2

Solé, R. V., Salazar-Ciudad, I., and Garcia-Fernández, J. (2002). Common pattern formation, modularity and phase transitions in a gene network model of morphogenesis. *Phys. A Stat. Mech. Appl*. 305, 640–654. doi: 10.1016/S0378-4371(01)00580-5

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A*. 98, 10869–10874. doi: 10.1073/pnas.191367098

Su, G., Kuchinsky, A., Morris, J. H., Meng, F., et al. (2010). Glay: community structure analysis of biological networks. *Bioinformatics* 26, 3135–3137. doi: 10.1093/bioinformatics/btq596

Tadaka, S., and Kinoshita, K. (2016). NCMine: core-peripheral based functional module detection using near-clique mining. *Bioinformatics* 32, btw488. doi: 10.1093/bioinformatics/btw488

Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2981–2986. doi: 10.1073/pnas.0308661100

Thieffry, D., and Romero, D. (1999). The modularity of biological regulatory networks. *Biosystems* 50, 49–59. doi: 10.1016/S0303-2647(98)00087-2

Torres-Sosa, C., Huang, S., and Aldana, M. (2012). Criticality is an emergent property of genetic networks that exhibit evolvability. *PLoS Comput. Biol.* 8:e1002669. doi: 10.1371/journal.pcbi.1002669

Tripathi, B., Parthasarathy, S., Sinha, H., Raman, K., and Ravindran, B. (2019). Adapting community detection algorithms for disease module identification in heterogeneous biological networks. *Front. Genet*. 10:164. doi: 10.3389/fgene.2019.00164

Tripathi, S., Moutari, S., Dehmer, M., and Emmert-Streib, F. (2016). Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC Bioinformatics* 17:1. doi: 10.1186/s12859-016-0979-8

Valverde, S. (2017). Breakdown of modularity in complex networks. *Front. Physiol.* 8:497. doi: 10.3389/fphys.2017.00497

van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Bacterial Mol. Netw. Methods Protoc*. 804, 281–295. doi: 10.1007/978-1-61779-361-5_15

Van Dongen, S. M. (2001). *Graph clustering by flow simulation* (Ph.D. thesis), Utrecht.

Verd, B., Monk, N. A., and Jaeger, J. (2019). Modularity, criticality, and evolvability of a developmental gene regulatory network. *eLife* 8:e42832. doi: 10.7554/eLife.42832

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput*. 17, 395–416. doi: 10.1007/s11222-007-9033-z

Wagner, G., Mezey, J., and Calabretta, R. (2001). *Modularity: Understanding the Development and Evolution of Complex Natural Systems. Natural Selection and the Origin of Modules*. Cambridge, MA: MIT Press.

Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat. Rev. Genet*. 8, 921–931. doi: 10.1038/nrg2267

Wang, H., Ye, M., Fu, Y., Dong, A., Zhang, M., Feng, L., et al. (2021). Modeling genome-wide by environment interactions through omnigenic interactome networks. *Cell Rep*. 35:109114. doi: 10.1016/j.celrep.2021.109114

Wang, J., Yi, Y., Chen, Y., Xiong, Y., and Zhang, W. (2020). Potential mechanism of rrm2 for promoting cervical cancer based on weighted gene co-expression network analysis. *Int. J. Med. Sci*. 17:2362. doi: 10.7150/ijms.47356

Wilkinson, D. M., and Huberman, B. A. (2004). A method for finding communities of related genes. *Proc. Natl. Acad. Sci. U.S.A*. 101(Suppl 1), 5241–5248. doi: 10.1073/pnas.0307740100

Xu, H., and Wang, S. (2010). "Research on functional modules of gene regulatory network," in *Advancing Computing, Communication, Control and Management* ed Q. Luo, (Heidelberg: Springer), 264–271. doi: 10.1007/978-3-642-05173-9_34

Xu, X., Zhou, Y., Miao, R., Chen, W., Qu, K., Pang, Q., et al. (2016). Transcriptional modules related to hepatocellular carcinoma survival: coexpression network analysis. *Front. Med*. 10, 183–190. doi: 10.1007/s11684-016-0440-4

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep*. 6, 1–18. doi: 10.1038/srep30750

Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enriquez, J. (2020). Gene expression and co-expression networks are strongly altered through stages in clear cell renal carcinoma. *Front. Genet*. 11:1232. doi: 10.3389/fgene.2020.578679

Zhan, M. (2007). Deciphering modular and dynamic behaviors of transcriptional networks. *Genomic Med*. 1, 19–28. doi: 10.1007/s11568-007-9004-7

Zhang, B., Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. 4:1128. doi: 10.2202/1544-6115.1128

Zhang, X., and Newman, M. (2015). Multiway spectral community detection in networks. *Phys. Rev. E* 92:052808. doi: 10.1103/PhysRevE.92.052808

Zheng, F., Zhang, S., Churas, C., Pratt, D., Bahar, I., and Ideker, T. (2020). Decoding of persistent multiscale structures in complex biological networks. *bioRxiv*. 92, 1–8. doi: 10.1186/s13059-020-02228-4

Zheng, F., Zhang, S., Churas, C., Pratt, D., Bahar, I., and Ideker, T. (2021). HiDeF: identifying persistent structures in multiscale 'omics data. *Genome Biol*. 22, 1–15.

Zhou, G., and Xia, J. (2018). OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res*. 46, W514–W522. doi: 10.1093/nar/gky510

Zhou, H. (2003). Network landscape from a Brownian particle's perspective. *Phys. Rev. E* 67:041908. doi: 10.1103/PhysRevE.67.041908

Zhou, H., and Lipowsky, R. (2004). "Network Brownian motion: a new method to measure vertex-vertex proximity and to identify communities and subcommunities," in *International Conference on Computational Science* (Krakow: Springer), 1062–1069. doi: 10.1007/978-3-540-24688-6_137

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet*. 40, 854–861. doi: 10.1038/ng.167

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 November 2021
doi: 10.3389/fgene.2021.746495

A Network-Centric Framework for the Evaluation of Mutual Exclusivity Tests on Cancer Drivers

Rafsan Ahmed[1]*, Cesim Erten*[2]*, Aissa Houdjedj*[2]*, Hilal Kazan*[2]* *and Cansu Yalcin*[2]

[1]*Electrical and Computer Engineering Graduate Program, Antalya Bilim University, Antalya, Turkey,* [2]*Department of Computer Engineering, Antalya Bilim University, Antalya, Turkey*

One of the key concepts employed in cancer driver gene identification is that of mutual exclusivity (ME); a driver mutation is less likely to occur in case of an earlier mutation that has common functionality in the same molecular pathway. Several ME tests have been proposed recently, however the current protocols to evaluate ME tests have two main limitations. Firstly the evaluations are mostly with respect to simulated data and secondly the evaluation metrics lack a network-centric view. The latter is especially crucial as the notion of common functionality can be achieved through searching for interaction patterns in relevant networks. We propose a network-centric framework to evaluate the pairwise significances found by statistical ME tests. It has three main components. The first component consists of metrics employed in the network-centric ME evaluations. Such metrics are designed so that network knowledge and the reference set of known cancer genes are incorporated in ME evaluations under a careful definition of proper control groups. The other two components are designed as further mechanisms to avoid confounders inherent in ME detection on top of the network-centric view. To this end, our second objective is to dissect the side effects caused by mutation load artifacts where mutations driving tumor subtypes with low mutation load might be incorrectly diagnosed as mutually exclusive. Finally, as part of the third main component, the confounding issue stemming from the use of nonspecific interaction networks generated as combinations of interactions from different tissues is resolved through the creation and use of tissue-specific networks in the proposed framework. The data, the source code and useful scripts are available at: https://github.com/abu-compbio/NetCentric.

Keywords: mutual exclusivity, network-centric mutual exclusivity evaluation, cancer drivers, cancer genomics, tumor mutation load

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Marco Mina,
HAYA Therapeutics, Switzerland
Vsevolod Jurievich Makeev,
Russian Academy of Sciences, Russia

****Correspondence:***
Hilal Kazan
hilal.kazan@antalya.edu.tr

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 July 2021*
Accepted: *27 October 2021*
Published: *26 November 2021*

Citation:
Ahmed R, Erten C, Houdjedj A, Kazan H and Yalcin C (2021) A Network-Centric Framework for the Evaluation of Mutual Exclusivity Tests on Cancer Drivers. Front. Genet. 12:746495. doi: 10.3389/fgene.2021.746495

1 INTRODUCTION

Cancer is a disease caused mostly due to a gradual accumulation of somatic alterations that give rise to pathway dysregulation through alterations in copy number, DNA methylation, gene expression, and molecular function. An important challenge in cancer genomics is to distinguish driver mutations from passenger mutations. The former are those determined to be causal for cancer progression, whereas the latter are characterized as those not leading to any selective advantage. Several computational methods have been proposed for the identification of cancer driver genes or driver modules of genes by integrating mutations data with various other types of genetic data; see Dimitrakopoulos and Beerenwinkel. (2017), Zhang and Zhang. (2018), Bailey et al., 2018, Tokheim et al., 2016 for recent comprehensive evaluations and surveys on the topic.

A phenomenon observed frequently in the data pertaining to the alterations that the tumors acquire is mutual exclusivity (ME); a driver mutation is less likely to occur in case of an earlier mutation that has common functionality in the same molecular pathway (Thomas et al., 2007; Yeang et al., 2008; Leiserson et al., 2016; van de Haar et al., 2019). Therefore several driver gene or module identification approaches employ ME detection as part of their problem definitions and optimization goals (Babur et al., 2015; Ciriello et al., 2012; Leiserson et al., 2013; Kim et al., 2015; Ahmed et al., 2019; Baali et al., 2020). Such a central role in driver gene and module identification has led to the design of many different approaches for defining and computing mutual exclusivity. Some of these approaches are based on combinatorial definitions of mutual exclusivity (Vandin et al., 2012; Leiserson et al., 2013; Sarto Basso et al., 2019; Ahmed et al., 2019; Song et al., 2020; Baali et al., 2020). In most cases the combinatorial definitions are incorporated and tested within a driver gene or module identification framework, rather than as stand-alone ME tests. On the other hand, the vast majority of the ME detection approaches are based on statistical tests (Ciriello et al., 2012; Szczurek and Beerenwinkel, 2014; Leiserson et al., 2015; Constantinescu et al., 2015; Hua et al., 2016; Canisius et al., 2016; Leiserson et al., 2016; Kim et al., 2017; Liu et al., 2020; Zhang et al., 2020) and in most cases for such approaches the specific goal is to provide ME significance results. Therefore the focus of the proposed framework is the evaluation of the latter set of approaches consisting of the statistical ME tests.

Among such approaches, MEMo builds a graph based on gene similarities and extracts cliques from this graph. To determine whether each clique has significant mutual exclusivity, it then proposes a null model generated by randomly permuting the set of genomic events, while preserving the overall distribution of observed alterations across both genes and samples, and introduces a Markov Chain Monte Carlo (MCMC) permutation strategy based on random network generation models (Ciriello et al., 2012). Szczurek and Beerenwinkel. (2014) propose a probabilistic, generative model of mutual exclusivity, explicitly taking coverage, impurity, and error rates into account. Based on such a model, they provide a statistical test of mutual exclusivity by comparing its likelihood to the null model that assumes independent gene alterations. Mutex defines the alteration of two genes to be mutually exclusive if their overlap in samples is significantly less than expected by chance, where the statistical significance of the overlaps are calculated using a hypergeometric test with the assumption of a uniform alteration frequency among samples (Constantinescu et al., 2015). This may not always be the case as in many data sources there are hyper-mutated samples. The problem is resolved partially by simply excluding such samples from the analysis. CoMEt (Leiserson et al., 2015) on the other hand provides an exact statistical test for mutual exclusivity conditional on the observed frequency of each alteration with the goal of introducing less bias towards high frequency alterations. Based on this it provides a tail enumeration procedure to compute the exact test, as well as a binomial approximation. DISCOVER provides a statistical independence test that makes no assumption of identical gene alteration probabilities across tumors (Canisius et al., 2016). The alteration probabilities are estimated by solving a constrained optimization problem guaranteeing the probabilities are consistent with both the observed number of alterations per gene and the observed number of alterations per tumor. The tumor-specific gene alteration probabilities are then used to compute the probability of concurrent alterations which in turn are used to decide whether the number of tumors altered in both genes deviates from the expectation through an analytical test based on the Poisson-binomial distribution. WeXT provides a weighted exact test that conditions simultaneously on the number of samples with a mutation and the per-event, per-sample mutation probabilities (Leiserson et al., 2016). A recursive formulation to compute p-values for this weighted test exactly and a saddle-point approximation of the test are proposed. WeSMe provides a permutation-based test and an approximation of significance through a weighted sampling technique that enables further improvements in running time spent for sampling and a way to obtain a better precision without increasing the computational time significantly (Kim et al., 2017). Mina et al. propose the SELECT method which uses a weighted version of mutual information to identify significant mutual exclusivity or co-occurrence patterns where significance is estimated by comparing against patterns observed in random permutations of the data (Mina et al., 2017). Two recently suggested ME tests are FSME (Zhang et al., 2020) and MEScan (Liu et al., 2020). The former proposes a seed-and-extend strategy to alleviate the computational cost of a permutation-based test. The seed pairs are constructed by a combinatorial formulation incorporating both ME and the coverage of the pair. The seeds are then grown with new genes by employing an independence test. MESCan provides a test statistic that incorporates a patient and gene-specific background mutation rate in the calculation to adjust for the background noise, and that includes a gene-specific weight to down-weigh genes with high mutation rates. Such a statistic is then employed in an MCMC algorithm followed by a false discovery rate control.

We propose a network-centric framework to evaluate the pairwise significances found by statistical ME tests. It is important to make a distinction between the network-centric view of the current study and that of the previous studies employing both network data and the concept of ME (Ciriello et al., 2012; Leiserson et al., 2013; Kim et al., 2015; Ahmed et al., 2019; Baali et al., 2020). The latter are network-centric in the sense that the proposed ME tests are applied on interacting pairs or subnetworks as part of a more general goal of identifying cancer driver genes/modules. Thus due to the nature of the set objectives their evaluations focus on the success of output genes/modules matching reference cancer-related drivers/pathways. The proposed study takes on an approach in the opposite direction; we assume the interaction network and the reference cancer-related drivers to be inputs to our framework which evaluates the success of various ME tests. The focus of the proposed framework is on pairwise significances since one of the major application areas where ME tests are commonly employed is knowledge-based cancer driver identification where pairwise ME significances are of major essence. In terms of the general objectives our work is most similar to that of Deng et al., 2017, where a framework for performance comparisons of statistical ME detection approaches is proposed and executed on six such tests. An important distinction is that the performance analysis of Deng et al. is based on experiments with simulated data and the framework does not suggest any mechanism to avoid confounders inherent in ME detection. One such confounder is due to the alterations specific to cancer subtypes (Deng et al., 2017; van de Haar et al., 2019). Alterations in different subtypes may be incorrectly diagnosed with ME, although the alterations are not due to any natural root causes of ME such as redundant functionality. Inspired by the observation that mutual exclusivity is enriched among physically interacting pairs of genes (Dao et al., 2017), our network-centric view aims to recognize such false positives by constructing reference sets based on known drivers gathered from neighborhoods of interaction networks. Furthermore, inspired by the mutation load confounding concept of van de Haar et al., 2019, we extend our network-centric framework to dissect side effects caused by mutation load artifacts; mutations that drive tumor subtypes with low mutation load might be incorrectly diagnosed as mutually exclusive. A possible drawback of the proposed network-centric evaluation framework would be due to the use of nonspecific interaction networks that are generated as combinations of interactions from different tissues and are thus suboptimal in resolving confounding issues of mutual exclusivity. In order to detect whether there exists such discrepancies or to limit their effect if they do, we therefore refine the network-centric approach by designing further tests on tissue-specific networks (TSN) we construct based on gene co-expression.

2 METHODS

The overall network-centric ME evaluations framework has three main components. The first one consists of definitions of the metrics employed in the network-centric ME evaluations. Such metrics are designed so that network knowledge and the reference set of known cancer genes are incorporated in ME evaluations under a careful definition of proper control groups. The second component detects whether the use of the interactome information provides similar advantages in ME corrections of pairwise mutual exclusivity findings as the subtype-stratification idea suggested by van de Haar et al., 2019. Finally, the third component extends our framework to incorporate tissue-specific networks with the aim of reducing the possible side effects of using nonspecific interaction networks.

2.1 Metrics for the Network-Centric Mutual Exclusivity Evaluations

Assuming that cancer driver genes in the same pathway are more likely to show mutually exclusive mutation profiles, we utilize the interactome to devise a strategy for evaluating the ME methods and the effects of the interactome information on quantifying ME. Let $\mathcal{G}, \mathcal{C}, \mathcal{T}, \mathcal{S}, p_t, c$ denote respectively the input Protein-Protein Interaction (PPI) network, the employed cohort, the statistical ME test undergoing the network-centric ME evaluations, the golden standard reference gene set of known cancer drivers, the p-value threshold for significance, and the type of the control group to be employed. Let $N_{\mathcal{S}}(g_i)$ denote the set of genes from $\mathcal{S}$ that are in the neighborhood of the node corresponding to gene g_i in the PPI network $\mathcal{G}$. For a gene $g_i \in \mathcal{S}$, corresponding to each neighbor $g_j \in N_{\mathcal{S}}(g_i)$, we randomly select a gene g_r from a control group $\mathcal{X}_c(g_i)$, and compute TP^{cur}, FP^{cur}, based on the $-$ log-transformed p-values $p_{i,j}$ and $p_{i,r}$ as computed by the ME test $\mathcal{T}$. Here $p_{i,j}$ denotes the significance of the mutual exclusivity of the pair g_i, g_j for $g_i \in \mathcal{S}$ and $g_j \in N_{\mathcal{S}}(g_i)$, and $p_{i,r}$ denotes the significance of the mutual exclusivity of the pair g_i, g_r for a random gene g_r from the control group. Based on the premise that cancer driver genes interacting in the PPI network are likely to exhibit ME, a pair g_i, g_j belongs to the set of True Positives if $p_{i,j}$ is significant and a pair g_i, g_r belongs to the set of False Positives if $p_{i,r}$ is significant.

To obtain robust results, the selection of the random genes from the control group is repeated robustness_iterations number of times, which is set to 100 in all the evaluations, except for those testing the robustness of the framework with respect to various parameter settings. Finally the medians of these 100 instances are summed over all genes $g_i \in \mathcal{S}$ to provide the necessary statistics TP, FP. Thus precision, sensitivity, and the F1 scores are computed based on these statistics. Precision is calculated as $|TP|/(|TP| + |FP|)$. Sensitivity is calculated with the formula $|TP|/|P|$ where P corresponds to condition positives which are defined as the gene pairs $g_i, g_j \in \mathcal{S}$ where g_i, g_j interact in $\mathcal{G}$.

We note that limiting our focus solely on these conventionally formed TP, FP classes may be misleading as each one considers the significance of $p_{i,j}$ and $p_{i,r}$ individually. A more detailed inspection with a simultaneous consideration of their values could prove more insightful in certain cases since they both involve a common gene g_i. Towards this aim we introduce the *strict* versions of these conventional classes. More specifically TP_{strict} consists of g_i, g_j pairs where $p_{i,j}$ is significant not only with respect to the given threshold but also as compared to the p-value of the control pair g_i, g_r. Similarly FP_{strict} consists of the control pairs g_i, g_r, where $p_{i,r}$ is more significant than both the threshold value and $p_{i,j}$. Based on these strict classes we can compute three metrics: precision$_{strict}$, sensitivity$_{strict}$, and F1$_{strict}$. Precision$_{strict}$ is defined as $|TP_{strict}|/(|TP_{strict}| + |FP_{strict}|)$ and sensitivity$_{strict}$ is defined as $|TP_{strict}|/|P|$. Such a consideration is especially convenient in reducing any potential bias inherent in genes like TP53 which have large mutation frequencies almost exclusively in tumors with small numbers of mutations; both $p_{i,j}$ and $p_{i,r}$ are likely to be significant in such a scenario giving rise to vagueness in the conventional F1 score. A comparison of F1$_{strict}$ values based on the two statistics simultaneous by their nature, TP_{strict} and FP_{strict} provides a more rigorous evaluation in such cases.

For the network-centric ME evaluations we employ two different definitions for the control groups. For the first one, the control group $\mathcal{X}_1(g_i)$ consists of genes in $\mathcal{S}$ that do not interact with g_i in the PPI network. For the second one, $\mathcal{X}_2(g_i)$

consists of neighbors of g_i in the PPI network that are not in $\mathcal{S}$. In the latter case only the genes $g_i \in \mathcal{S}$ for which the number of neighbors not in $\mathcal{S}$ is larger than or equal to the number of neighbors in $\mathcal{S}$ are taken into account.

2.2 Network-Centric Mutual Exclusivity Corrections in Relation to Mutation Load Association

Some statistical mutual exclusivity tests are based on the assumption that gene's alterations across tumors are identically distributed. Among the approaches considered in this study Fisher's Exact Test and MEGSA belong to this category. However, it has been observed that the number of alterations per tumor can vary quite considerably, even in tumors of the same type; colorectal tumors with microsatellite stability have a median of 66 non-synonymous mutations, but colorectal tumors with microsatellite instability have a median of 777 mutations (Vogelstein et al., 2013; Leiserson et al., 2016). It has been shown that under such settings the mutual exclusivity tests relying on identical alteration probabilities across tumors may lead to reduced sensitivity for mutual exclusivity analysis (Canisius et al., 2016). The effects of varying alteration probabilities on pairwise mutual exclusivity calculations have been formalized within the context of the so-called mutation load confounding (MLC) in a recent study by van de Haar et al., 2019. MLC is a correlation between the number of statistically significant mutual exclusivity findings and the mutation load association (MLA) of a gene. MLA of a gene is calculated by running a logistic regression where a gene's binary mutation status indicating whether the gene is mutated or not in a tumor is used as the only feature to predict the mutation load of that tumor. Mutation load is defined as the number of genes that are mutated in a tumor. Once the coefficient of the feature is obtained by fitting the logistic regression model, it is standardized by dividing by the standard error to make it comparable across the genes. This standardized coefficient value is defined as the MLA value. Note that negative MLA values correspond to higher mutation frequencies in tumors with low mutation loads, whereas positive values correspond to higher mutation frequencies in tumors with high mutation loads. Strong negative correlations between the MLA of a gene and the number of statistically significant pairwise mutual exclusivities have been observed, implicating the finding that the more negative a gene's MLA, the higher the number of other genes that show mutual exclusivity with that particular gene (van de Haar et al., 2019). However, such a negative correlation does not always imply true ME since a gene that exclusively shows large mutation frequency in tumors with low mutation loads, naturally has a better chance of forming mutually exclusive pairs with other genes. Thus extra sources of information are necessary to filter out the pairs with true ME relations among a set of statistically significant pairwise mutual exclusivities postulated by some exclusivity test. van de Haar et al., 2019 make use of the subtype information for such a purpose and show that MLC can be reduced by correcting via tumor subtype stratification. Such a correction greatly reduces the number of gene pairs reported to show mutual exclusivity, especially for pairs that include genes with low MLA. A major drawback is the absence of subtype information for many tumors. As part of our network-centric ME framework, we suggest that such a correction can be efficiently done with the interaction network data, rather than or better yet on top of the subtype information. For this purpose we calculate the correlation between the number of statistically significant pairwise ME findings and the MLA for two settings; one where pairwise mutual exclusivities are sought between a gene in $\mathcal{S}$ and all other genes in $\mathcal{S}$, and the other where a gene in $\mathcal{S}$ is checked against only its PPI neighbors that are in $\mathcal{S}$. The computations of the two settings are repeated with the subtype-stratified data as well, to see the added value of the network-centric ME corrections on top of the subtype-based corrections on statistically significant pairwise MEs.

2.3 Network-Centric Mutual Exclusivity Evaluations in Relation to Tissue-Specific Networks

Rather than using a common nonspecific network for all the cancer types, in this component of our evaluation framework we employ TSN based on the tissue in which the tumor develops. To construct the TSN for a particular tissue, we start with the original PPI network and remove the edges between the pairs of genes that are not co-expressed in the corresponding tissue. For this purpose, we download RNA-seq datasets from GTEX portal (GTEXConsortium, 2020). See **Supplementary Table S49** for the total number of available samples for each tissue. To determine the co-expressed genes, we follow the procedure described in Luck et al., 2020. For each pair of genes that have an edge in the original PPI network, we identify the number of samples where both genes have Transcripts Per Kilobase Million (TPM) values ≥1. We then divide this number with the total number of samples where either gene has a TPM value ≥1. The resulting value is called the co-expression ratio. Gene pairs interacting in the original network are included in the TSN_{cor} if the co-expression ratio is $\geq cor$, for a given threshold cor.

In addition to applying the network-centric metrics introduced in **Section 3.1** on the constructed TSNs, we also propose a more detailed evaluation in terms of ROC analysis based on tissue-specificity. For this purpose, we define the gene pairs with co-expression ratio value of 1 as tissue-specific gene pairs. Similarly, the gene pairs with co-expression ratio values ≤0.5 are called non-tissue-specific gene pairs. To test whether a specific ME test identifies stronger mutual exclusivities for the tissue-specific gene pairs in $\mathcal{S}$, we rank the gene pairs in $\mathcal{S}$ in increasing order of p-values. To construct the control group, we rank the same number of random samples of gene pairs not in $\mathcal{S}$ with respect to the p-values making sure that the sizes of the positive (or negative) sets of gene pairs not in $\mathcal{S}$ are exactly the same as those that are found for the gene pairs in $\mathcal{S}$. For both gene pairs in $\mathcal{S}$ and gene pairs not in $\mathcal{S}$, the set of positives consists of the tissue-specific gene pairs, whereas non-tissue-specific gene pairs are labelled as negatives. We then compute the True Positive Rate (TPR) and the False Positive Rate (FPR) for each case. Note that for robustness considerations the control group

TABLE 1 | Results of network-centric ME evaluation framework with control group $\mathcal{X}_1$ COADREAD t20 (498 samples, 196 CGC-CGC pairs).

Method	Precision	Sensitivity	F1 Score	Precision$_{\text{strict}}$	Sensitivity$_{\text{strict}}$	F1 Score$_{\text{strict}}$
DISCOVER	0.661	0.220	0.331	0.708	0.183	0.291
DISCOVER Strat	0.727	0.041	0.078	0.727	0.041	0.078
Fisher's Exact Test	0.500	0.031	0.058	0.500	0.031	0.058
MEGSA	0.611	0.056	0.103	0.588	0.051	0.094
MEMO	0.658	0.329	0.439	0.647	0.237	0.347
WExT	0.676	0.403	0.505	0.725	0.329	0.453

computations are repeated 100 times and the median TPR and FPR values are reported.

3 RESULTS

3.1 Input Data and Parameter Settings

The somatic mutation data from TCGA was preprocessed and provided by van de Haar et al., 2019. The 8 different cancer types and their corresponding tumor samples within the dataset is as follows: BLCA (411), BRCA (1026), COADREAD (498), LUAD (568), LUSC (485), SKCM (468), STAD (438) and UCEC (531). The preprocessing step involves the removal of all mutations with "variant_classification" of "Silent," "3'UTR," "Intron," "5'UTR," "RNA," "3'Flank" and "5'Flank" from the TCGA data. The input data is then further filtered by mutation frequency threshold, *t*, to include genes with $> t$ mutations across the cohort. More specifically, with $t = 20$ we include the genes that are mutated in more than 20 samples within the cancer type under study. Regarding subtypes, we download subtype information for BRCA from the cBioPortal (Cerami et al., 2012; Gao et al., 2013) and the CMS stratification for COADREAD from (Guinney et al., 2015). We use the COSMIC Cancer Gene Census database to compile the set of known cancer genes (Sondka et al., 2018).

For the results presented in the main document we employ the IntAct PPI network as it is a comprehensive and well-characterized database (Orchard et al., 2014). As a preprocessing step, we remove duplicate edges and edges below the confidence threshold of 0.35 from the network. The final network contains 15,079 nodes and 103,520 edges. For the gene expression data employed in the construction of TSNs, we download RNA-Seq data from the Genotype-Tissue Expression (GTEx) portal (GTEXConsortium, 2020) (05-06-2017).

For the comparative evaluations of our network-centric framework described in the previous section, we choose six popular statistical mutual exclusivity methods: DISCOVER (Canisius et al., 2016), DISCOVER Strat (Canisius et al., 2016; van de Haar et al., 2019), Fisher's Exact Test, WeXT (Leiserson et al., 2016), MEMo (Ciriello et al., 2012) and MEGSA (Hua et al., 2016). Among these, MEMo and MEGSA are originally designed to output *p*-values for a set of genes with size > 2. For MEMo, we re-implement the first part of the algorithm where pairwise ME *p*-values are estimated. We use $Q = 100$ and $N = 10,000$ as suggested by the original paper (Ciriello et al., 2012). For MEGSA, pairwise ME *p*-values are calculated by applying chi-square cumulative probability less than or equal to the value of the log likelihood calculated by the funestimate function. With regards to the parameter settings of our proposed framework, we employ the values of 5 and 20 for *t*.

3.2 Mutual Exclusivity Evaluations Based on Defined Metrics

Table 1 and **2** show the results of evaluating the 6 ME detection methods on COADREAD data where $t = 20$ and we use the data from 498 patients for which subtype information is available. We use $\mathcal{X}_1$ and $\mathcal{X}_2$ as the control group in **Table 1** and **2**, respectively. We first discuss the results of $\mathcal{X}_1$. We observe that DISCOVER Strat gives the highest precision and precision$_{strict}$ values. The ranking of the other methods from best to worst in terms of precision or precision$_{strict}$ is as follows: WeXT, DISCOVER, MEMo, MEGSA and Fisher's Exact Test. A comparison of the precision and precision$_{strict}$ values distinguishes two groups of ME methods; for DISCOVER, DISCOVER Strat, Fisher's Exact Test, and WexT the precision$_{strict}$ values are greater than or equal to the precision values, whereas the exact opposite is observed for MEGSA and MEMo. This suggests that the performance of the methods in the latter group gets worse when random control gene pair is considered simultaneously in the precision calculation, that is precision$_{strict}$. Compared to the precision, we observe much larger differences among the sensitivity or the sensitivity$_{strict}$ values output by the employed methods. We can group the methods into two where the first group contains WeXT, MEMo and DISCOVER, and the second group contains the rest of the methods. The first group of methods give much larger sensitivity or sensitivity$_{strict}$ values than the second. For instance, the sensitivity value obtained with WeXT is an order of magnitude larger than that of Fisher's Exact Test. This also shows that the second group of methods are more conservative than the first group of methods. WeXT is the least conservative approach based on its high sensitivity value. Even though WexT predicts many significant *p*-values, it still has a competitive precision$_{strict}$ value which is slightly lower than the maximum observed value (0.725 vs 0.727). Accordingly, WeXT obtains the best F1 score and F1$_{strict}$ score which is followed by MEMo and DISCOVER. The remaining three methods give much smaller F1 scores and they rank as follows from highest to lowest: MEGSA, DISCOVER Strat and Fisher's Exact Test. Comparing the conventional F1 score with the F1$_{strict}$ score of each ME method, the largest difference is observed for MEMo indicating that the consideration of the random pair as a control affects its performance dramatically. Another interesting observation is

TABLE 2 | Results of network-centric ME evaluation framework with control group $\mathcal{X}_2$ COADREAD t20 (498 samples, 107 CGC-CGC pairs).

Method	Precision	Sensitivity	F1 Score	Precision$_{strict}$	Sensitivity$_{strict}$	F1 Score$_{strict}$
DISCOVER	0.537	0.276	0.365	0.579	0.210	0.308
DISCOVER Strat	0.455	0.048	0.086	0.400	0.038	0.069
Fisher's Exact Test	0.444	0.038	0.069	0.375	0.028	0.052
MEGSA	0.571	0.075	0.133	0.538	0.066	0.118
MEMO	0.566	0.388	0.460	0.495	0.215	0.300
WExT	0.575	0.438	0.497	0.596	0.295	0.395

TABLE 3 | Results of network-centric ME evaluation framework with control group $\mathcal{X}_1$ COADREAD t5 (498 samples, 1748 CGC-CGC pairs).

Method	Precision	Sensitivity	F1 Score	Precision$_{strict}$	Sensitivity$_{strict}$	F1 Score$_{strict}$
DISCOVER	0.647	0.052	0.096	0.658	0.046	0.086
DISCOVER Strat	0.618	0.012	0.024	0.618	0.012	0.024
Fisher's Exact Test	0.583	0.008	0.016	0.565	0.007	0.014
WExT	0.645	0.121	0.203	0.668	0.102	0.177

TABLE 4 | Results of network-centric ME evaluation framework with control group $\mathcal{X}_2$ COADREAD t5 (498 samples, 1625 CGC-CGC pairs).

Method	Precision	Sensitivity	F1 Score	Precision$_{strict}$	Sensitivity$_{strict}$	F1 Score$_{strict}$
DISCOVER	0.721	0.052	0.097	0.746	0.048	0.090
DISCOVER Strat	0.641	0.013	0.025	0.641	0.013	0.025
Fisher's Exact Test	0.619	0.008	0.016	0.619	0.008	0.016
WExT	0.670	0.118	0.200	0.712	0.103	0.180

the lower performance of DISCOVER Strat compared to DISCOVER which suggests that the use of subtype information is not useful for COADREAD. **Table 2** shows the results where $\mathcal{X}_2$ is used as the control group. Since $\mathcal{X}_2(g_i)$ is defined as the non-CGC neighbors of g_i in the PPI network, we can only consider the CGC genes that have more non-CGC neighbors than CGC neighbors. As such, the number of pairs included in this analysis is much smaller than that of **Table 1** (107 vs 196). The ranking of the methods in **Table 2** with respect to F1 score and sensitivity remain the same as **Table 1**. However, there are differences in the ranking with respect to other metrics. For instance, WeXT ranks best in terms of precision whereas the best ranking method in **Table 1**, DISCOVER Strat, ranks the fifth. Compared to **Table 1**, the precision values of all the methods are smaller in **Table 2**. We see the opposite trend for sensitivity values. These changes are in parallel with the increase in percent significant p-values output by the methods. For instance, the percentage of significant p-values output by DISCOVER is 12% in **Table 1** and 18% in **Table 2**. We also observe differences between the conventional and the strict versions of the employed metrics. WeXT and DISCOVER have increased precision$_{strict}$ values compared to precision whereas we observe the opposite trend for the rest of the methods. Additionally, the ranking of the methods with respect to F1 score and F1$_{strict}$ score is different. Namely, MEMo's ranking decreases from second highest to third highest when we switch from F1 score to F1$_{strict}$ score. Accordingly, DISCOVER's ranking improves from third highest to second highest based on F1 score. This increases the confidence of DISCOVER results as F1$_{strict}$ requires a stricter definition of true and false positives. **Supplementary Table S1** shows the results with $\mathcal{X}_1$ control group and t = 20 filtering for the other cancer types. A detailed discussion of these results are available in the **Supplementary Material**.

Table 3 and **4** show the COADREAD results of t = 5 setting with $c = X_1$ and $c = X_2$, respectively. Using a lower value for t increases the number of gene pairs tested in our analysis. When we compare these results with the results we obtained when t = 20, we observe few differences. Though the number of tested gene pairs is larger, the percentage of significant p-values obtained by the methods decreases. For instance, the percentage of significant p-values output by WeXT for COADREAD data decreases from 42 to 14% when t is changed from 20 to 5. This is likely related to the larger inclusion of low mutation frequency genes when t = 5. An interesting observation for t = 5 results is the decrease in DISCOVER Strat's performance. For COADREAD, DISCOVER Strat's precision and precision$_{strict}$ value is the highest for t = 20 when $\mathcal{X}_1$ is used as the control group. However, when t = 5, we observe that it ranks after WeXT and DISCOVER in terms of precision/precision$_{strict}$ value. Similarly, for BRCA dataset, DISCOVER Strat ranks after WeXT for both control groups $\mathcal{X}_1$ and $\mathcal{X}_2$ (**Supplementary Table S25B**, **Supplementary Table S37B**).

Lastly, we investigate the robustness of our results with respect to robustness_iterations value, the p-value significance threshold value, the reference gene set and the employed PPI network. The results together with a discussion of these results are available in

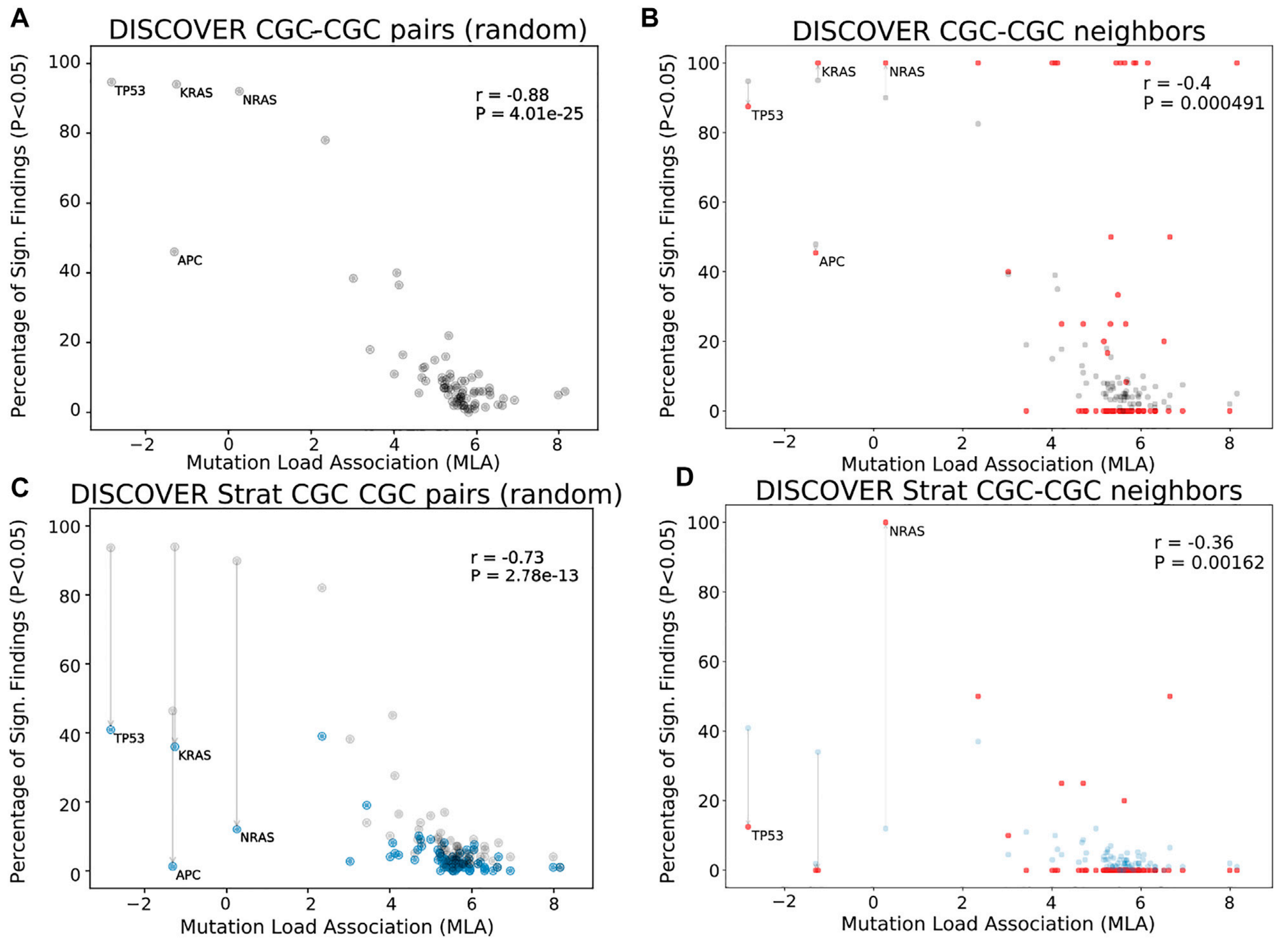


FIGURE 1 | Comparison of mutual exclusivity results of DISCOVER and DISCOVER Strat on TCGA COADREAD cohort (498 samples) **(A)** The scatterplot of percentage significance of ME runs (p-value < 0.05) of DISCOVER on COADREAD data where tests are performed between a CGC gene and a random subset of other CGC genes so that ME of a CGC gene of interest is checked with same sized group of genes in both A and B. **(B)** The scatter plot of percentage significance of ME runs of DISCOVER where tests are performed between a CGC gene and its PPI neighbors that are in CGC (red) compared with **(A)** in gray. **(C)** The scatterplot of percentage significance of mutual exclusivity runs of DISCOVER Strat where tests are performed between a CGC gene and a random subset of other CGC genes (blue) so that ME of a CGC gene of interest is checked with same sized group of genes in both C and D, results from **(A)** are shown in gray for comparison. **(D)** The scatterplot of percentage significance of mutual exclusivity runs of DISCOVER Strat where tests are performed between a CGC gene and its PPI neighbors that are in CGC (red) compared with **(C)** in blue.

Supplementary Table S2-S47. As a summary, our conclusions remain the same in these different settings and the largest differences are observed when the employed PPI network is changed.

3.3 Mutual Exclusivity Evaluations Based on Corrections via Mutation Load Association

Having compared the ME tests with respect to our novel network-centric evaluation framework, we now assess whether including network knowledge reduces the mutation load confounding (MLC) problem introduced by van de Haar et al., 2019. van de Haar et al. identified a strong negative correlation between the MLAs of genes and their percent significant findings in mutual exclusivity tests. In van de Haar et al., 2019, these statistics are computed for a set of 341 genes from an established cancer gene panel (Cheng et al., 2015) where, for each gene, mutual exclusivity tests are performed with all the other genes in the panel. Here, we first perform a similar analysis where we use the COSMIC CGC database (Forbes et al., 2017) to define the reference cancer gene set as it is more comprehensive and up to date.

Figures 1A shows the MLA of the reference cancer genes vs the percent significant findings in mutual exclusivity tests performed with DISCOVER for the TCGA COADREAD cohort (498 tumors). We observe a strong negative correlation between MLA values and percent significant findings in mutual exclusivity tests (Pearson correlation -0.88, p-value $4.0e-25$) similar to van de Haar et al., 2019. In **Figures 1B**, we take into account the PPI information to calculate percent significant findings. Namely, for each CGC gene, we perform mutual exclusivity tests only with its PPI neighbors that are also in CGC. Note that CGC genes which do not have any CGC

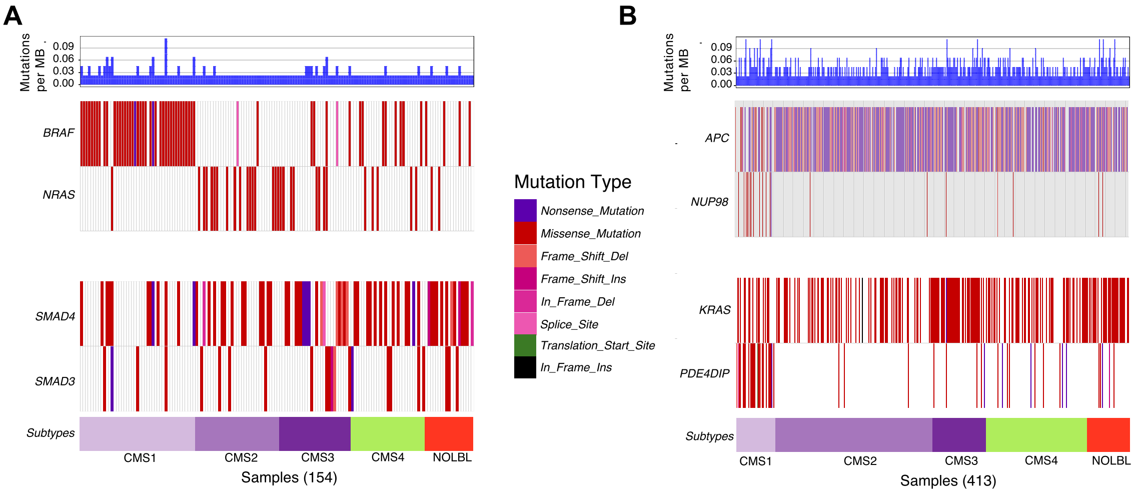


FIGURE 2 | Waterfall plots of the distribution of mutations for selected gene pairs. **(A)** Mutation distribution of two selected gene pairs (BRAF-NRAS and SMAD4-SMAD3) that are found to be significantly mutually exclusive based on both DISCOVER and DISCOVER-Strat estimations. **(B)** Mutation distribution of two selected gene pairs (APC-NUP98 and KRAS-PDE4DIP) that are found to be significantly mutually exclusive based on DISCOVER but not based on DISCOVER-Strat. Note that the set of samples included in each plot is determined by finding the set of patients that have a mutation in at least one of the listed genes. GenVisR R package is used to generate the waterfall plots (Skidmore et al., 2016). Subtype information is downloaded from (Guinney et al., 2015).

neighbors are excluded from this analysis. To make a fair comparison between **Figures 1A,B**, only the CGC genes that have CGC neighbors are shown in **Figures 1A**. We also ensure that the mutual exclusivity of a gene of interest is checked with same sized group of genes in both **Figures 1A,B**. To achieve this in **Figures 1A**, for each gene, we compute mutual exclusivity with a random subsample of the CGC reference set, the same size as the set of CGC neighbors of that gene. We repeat this random sampling 100 times and plot the mean percent significant findings value. For reference, **Supplementary Figure S3A, S3D** contains versions of **Figures 1A,C**, where all CGC genes (i.e., with and without CGC neighbors) are plotted and mutual exclusivities are checked between all CGC pairs, as it was done in van de Haar et al., 2019.

In **Figures 1B**, we observe a reduced correlation when network information is included (Pearson correlation -0.4, p-value $4.91e-4$). We also run DISCOVER Strat where stratification is based on CMS subtypes (Guinney et al., 2015). We plot these results in **Figures 1C** where we again ensure comparability with **Figures 1D** where both subtype and network information are considered. Comparing **Figures 1A** and **Figures 1C**, we verify the findings of van de Haar et al., although with less significance in correlation difference (Pearson correlation −0.73, p-value $2.8e-13$). It should be noted that the subtype stratification inherently causes an overall decrease in percent significant findings, not specific to genes with low MLA. On the contrary the idea of ME corrections through network incorporation, materialized in the comparison of **Figures 1A** and **Figures 1B**, inherently leads to an increase in percent significant findings. Most of the decreases occur in genes with small number of CGC neighbors. When we compare **Figures 1D** to **Figures 1B**, the decrease in correlation from −0.4 to −0.36 indicates that including subtype information is still useful when used on top of network-based corrections we propose.

Next, we utilize waterfall plots to compare the outputs of DISCOVER and DISCOVER-Strat to assess how MLA and subtype information can affect mutual exclusivity findings. **Figures 2A** shows two selected gene pairs that display significant mutual exclusivity based on both DISCOVER and DISCOVER-Strat estimations on TCGA COADREAD dataset. The mutual exclusivity between BRAF and NRAS, two members of the MAPK pathway, is well-known and has been detected in multiple cancer types including melanoma, myeloma and colorectal cancer (Samowitz et al., 2006; Roth et al., 2010; Popovici et al., 2012) BRAF is frequently mutated in patients from CMS1 subtype whereas NRAS shows almost no mutation across these patients. However, since BRAF and NRAS mutations are mutually exclusive across not only CMS1 subtype but also across the other subtypes, DISCOVER-Strat identifies this pair as significantly mutually exclusive. Similarly, SMAD3 and SMAD4 are two members of the TGF-β pathway and the mutual exclusivity between these two transcription factors is previously reported in colorectal cancer (Fleming et al., 2013). Mutations on SMAD3 and SMAD4 are distributed almost uniformly across the subtypes. As such, the mutual exclusivity between the mutations of these two genes is still significant when subtype information is incorporated. **Figures 2B** similarly shows two selected gene pairs that display significant mutual exclusivity based on DISCOVER but not based on DISCOVER Strat. For the first pair, we observe that NUP98 is mutated almost exclusively in patients from the CMS1 subtype which shows hypermutation due to microsatellite instability. On the other hand, there is a depletion of APC mutations among the patients from the CMS1 subtype which results in a low MLA value. As such, DISCOVER Strat fails to detect a significant ME between these two genes since it explores ME within each subtype separately. A similar observation can also be made for the KRAS-PDE4DIP pair where the former has a low MLA and the latter has a high MLA.

Supplementary Figure S2 compares the MLA of the reference cancer genes with the percent significant findings in mutual exclusivity tests for BRCA. Similar to the results that we obtain for COADREAD data, including network information reduces the correlation between MLA and ME detection rate (**Supplementary Figures S2B vs S2C**). The magnitude of reduction is even more significant than what we observe for COADREAD data (Pearson correlation −0.93 vs −0.27). Interestingly, including subtype information results in a very slight decrease in correlation coefficient (−0.93 to −0.91) (**Supplementary Figures S2B vs S2E**) as opposed to what we observe for COADREAD. We observe that including subtype information on top of network information results in no decrease in correlation (**Supplementary Figures S2C vs S2F**). This difference in the effect of including subtype information for BRCA and COADREAD datasets could be related to the average tumor mutation load of subtypes. BRCA subtypes have comparable average TML values (Her2: 146, LumA:65, LumB: 71, Normal: 55) whereas the CMS1 subtype in COADREAD has a dramatically larger average TML value compared to the other subtypes of COADREAD (CMS1: 1387, CMS2:93, CMS3: 272, CMS4: 212) We repeat the same analysis with the other ME detection methods as well as for other cancer types when t is set to 20 (**Supplementary Figures S1-S8**). We observe that the percent significant finding values can vary remarkably across the tumor types. Compared to other cancer types, we observe smaller percent significant findings for LUSC (**Supplementary Figures S5A,S5D, S5G**). Similarly, very few pairs have percent significance value ≥ 20 when we consider network information in LUSC (**Supplementary Figures S5C,S5F,S5I**). On the contrary, we observe many pairs with large percent significant values for CGC-CGC neighbors in UCEC data. This is particularly true for DISCOVER and WeXT results (**Supplementary Figures S8C–S8L**).

When we consider the correlation between MLA and percent significant values, we observe that adding network information decreases the correlation coefficient values for all cancer types and for all ME detection methods except for Fisher's Exact Test. Fisher's Exact Test results show an increased correlation with the addition of network information for LUSC and SKCM (**Supplementary Figures S5-S6** D vs F). Also, the correlation coefficient can not be computed for LUAD and STAD since Fisher's Exact Test gives a value of 0 for the percent significant findings of all considered genes (**Supplementary Figures S4D-S7F**). Another interesting observation is the variance in magnitude of decrease in correlation values across different tumor types. In particular, we observe a smaller decrease in correlation values for LUAD compared to other cancer types. The analogous results are also available for t = 5 setting (**Supplementary Figures S9-S16**). For all the cancer types, the correlation between MLA values and percent significant findings decreases and becomes non-significant for most cases.

We should also note that the majority of CGC genes have only one neighbor within the data setting of the cancer type under consideration. This leads to percentage significant findings of either 0 or 1 in many cases simply because these are the only possible values; for COADREAD see **Figures 1B** and **Figures 1D** where 41 out of 74 genes under study have only one CGC neighbor in the COADREAD data settings. To avoid any such possible biases, we repeat the same evaluations after filtering out those CGC genes with only one neighbor. The evaluations still provide significant decreases in correlation coefficient values analogous to the decreases observed in **Figures 1B** as compared to **Figures 1A** and **Figures 1D** as compared to **Figures 1C**. For detailed results, see **Supplementary Figures S17-S24** for t = 20 and **Supplementary Figures S25-S32** for t = 5.

Individual genes of interest are those that have increased percent significant findings when network neigborhood information is incorporated while at the same have significant number of CGC neighbors. More specifically, for the former constraint, we identify the CGC genes with at least 0.1 increase in percentage of significant findings value of WeXT, DISCOVER and MEMo when the network information is included as opposed to the scenario when it is not (e.g., for COADREAD, **Figures 1A** vs **Figures 1B**). We choose these 3 ME methods since they are top performers based on the defined metrics in **Section 3.1**. For STAD, SKCM and UCEC, since MEMo results are unavailable, we only consider WeXT and DISCOVER results. For the second constraint, we include the CGC genes with at least 3 CGC neighbors. For COADREAD, this selection procedure results in four genes: EP300, CREBBP, NCOA2 and NCOR2. Among these, EP300 is a well-known tumor suppressor in epithelial cancer types including COADREAD (Gayther et al., 2000). For BRCA, the only identified gene is PIK3R1. PIK3R1 is found to be significantly mutually exclusive with PIK3CA and SPEN based on both WeXT, DISCOVER and MEMo results. PIK3R1 and PIK3CA are members of the PI3K pathway and their mutual exclusivity has been previously established in the literature (Chen et al., 2018). For LUAD, PTPRB is the only identified gene and is found to be mutually exclusive with EGFR, a well-known oncogene in non-small cell lung cancer (Bethune et al., 2010). The set of identified genes for STAD are NCOA2, NCOR2 and CREBBP; all of which are found to be mutually exclusive with TP53. For SKCM, we identify ERBB4, RAC1, EP300 and ITK. ERBB4 is a well-known oncogene in skin cancer and found to be mutually exclusive with ERBB2 (Prickett et al., 2009; Nielsen et al., 2014). ERBB2 and ERBB4 indeed belong to the same family (i.e. ErbB family of receptor tyrosine kinases) and form a heterodimer receptor for Heparin-binding EGF-like growth factor (HB-EGF) (Iwamoto et al., 2017). RAC1 mutation P29S is an established driver in melanoma (Jiang et al., 2018). RAC1 is found to be mutually exclusive with MYH9, a tumor suppressor in melanoma (Singh et al., 2020). Lastly, ITK has been shown to be an oncogene in melanoma (Carson et al., 2015). For UCEC, we identify 33 genes in total. Among these, KIT and PTEN have established roles in UCEC cancer development (Chang et al., 2015; Wang et al., 2020). Moreover, PTEN is found to be strongly mutually exclusive with SPOP, whose mutations are also associated with endometrial cancer (Clark and Burleson,

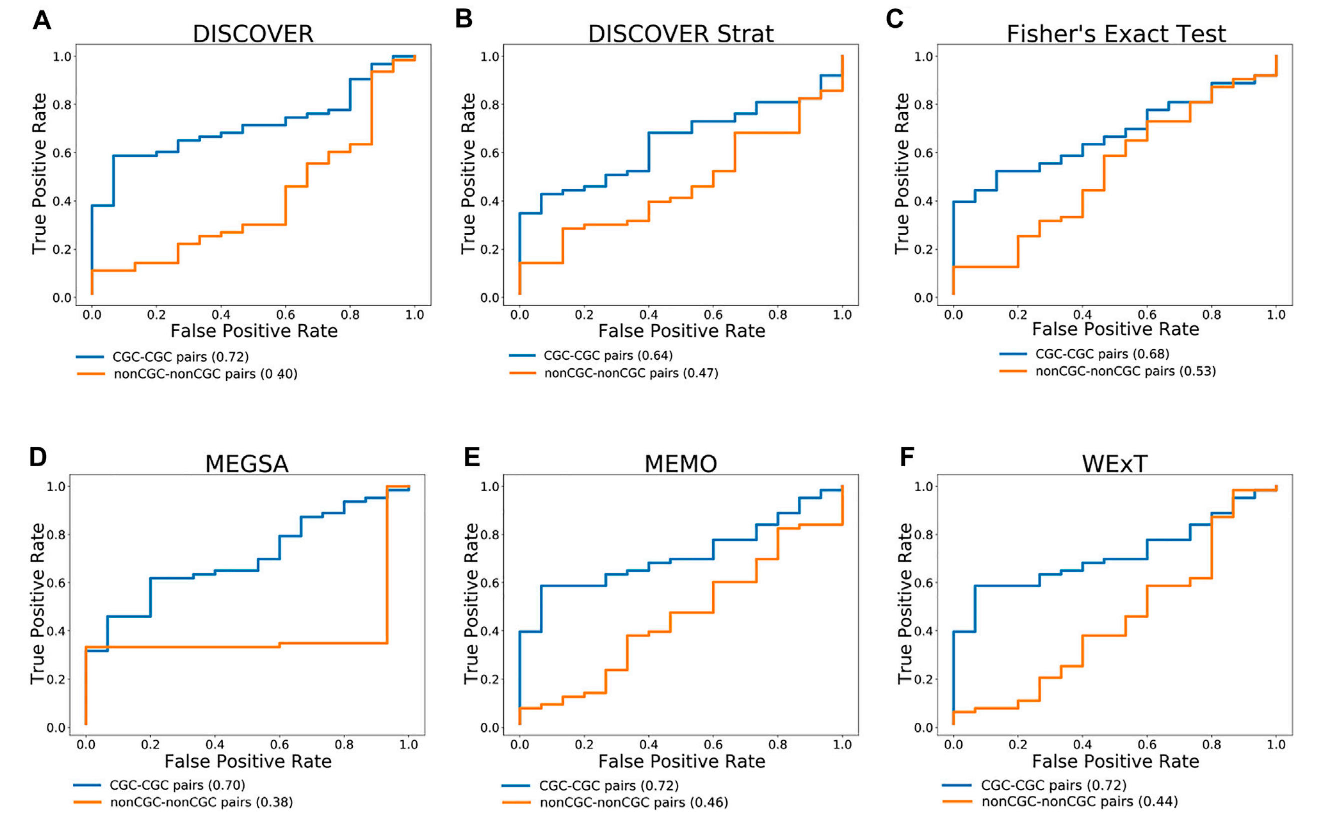


FIGURE 3 | Performance of selected ME tests in terms of discriminating TSN and non-TSN gene pairs based on estimated ME *p*-values on COADREAD data. Blue curve is plotted with CGC gene pairs and red curve is plotted with non-CGC gene pairs. Mutual exclusivities are estimated with **(A)** DISCOVER, **(B)** DISCOVER Strat, **(C)** Fisher's Exact Test, **(D)** MEGSA, **(E)** MEMo and **(G)** WeXT respectively.

2020). Lastly, for BLCA and LUSC, no gene satisfies the abovementioned criteria. Overall these results suggest that the CGC genes that show increased ME with network incorporation as well as their mutually exclusive partner genes often have established roles in the development of the particular cancer type.

3.4 Mutual Exclusivity Evaluations Based on Corrections via Tissue-Specific Networks

We first provide our ME evaluations with respect to the metrics defined in **Section 3.1** by replacing the non-specific networks with TSNs. We provide two types of comparisons; one where we compare $TSN_{0.5}$ with the original non-tissue specific Intact network and one where results of $TSN_{0.5}$ are compared against TSN_0. We do the latter to avoid artifacts that may be introduced due to the fact that some genes in the original Intact network might be simply missing from even TSN_0 since they may be nonexistent in the GTEX database. For the BLCA dataset, comparing the F1 scores of the ME methods under TSN_0 and $TSN_{0.5}$ settings, we observe that the scores of all methods are higher for the latter network. The largest percent increase of 10% is observed for WeXT when the control group is $\mathcal{X}_1$. Similarly, the largest percent increase of 12% is observed for MEMo when the control group is $\mathcal{X}_2$. On the other hand, when we compare the scores of TSN_0 against the original network, the differences are negligible. The next largest difference between the F1 scores obtained under $TSN_{0.5}$ as compared to TSN_0 is observed in STAD where we see a 7% increase in DISCOVER's score for $\mathcal{X}_1$, and a 10% increase in WeXT's score for $\mathcal{X}_2$. For the rest of the cancer types under study, for LUSC and UCEC we observe slight increase in performances of all the ME methods comparing the metrics under $TSN_{0.5}$ against TSN_0. For COADREAD, BRCA and SKCM we observe both increases and decreases in performances but the differences are almost negligible; see **Supplementary Tables S50-S81** for detailed results.

Figure 3 compares the ROC curves of CGC gene pairs and non-CGC gene pairs for COADREAD data where mutual exclusivities are estimated with DISCOVER, DISCOVER Strat, Fisher's Exact Test, MEGSA, MEMo and WeXT with t = 20. We observe that all the ME methods estimate stronger mutual exclusivities for tissue-specific CGC gene pairs compared to non-tissue-specific CGC gene pairs since AUROCs are greater than 0.5. Additionally, we observe much smaller AUROCs for the control group where we repeat the same analysis with non-CGC gene pairs. Analogous results are available for the other cancer types where both the positive and negative set contains at least 10 number of pairs when t is set to 20. (**Supplementary Figures S33–S35**). We observe a similar result for SKCM where CGC

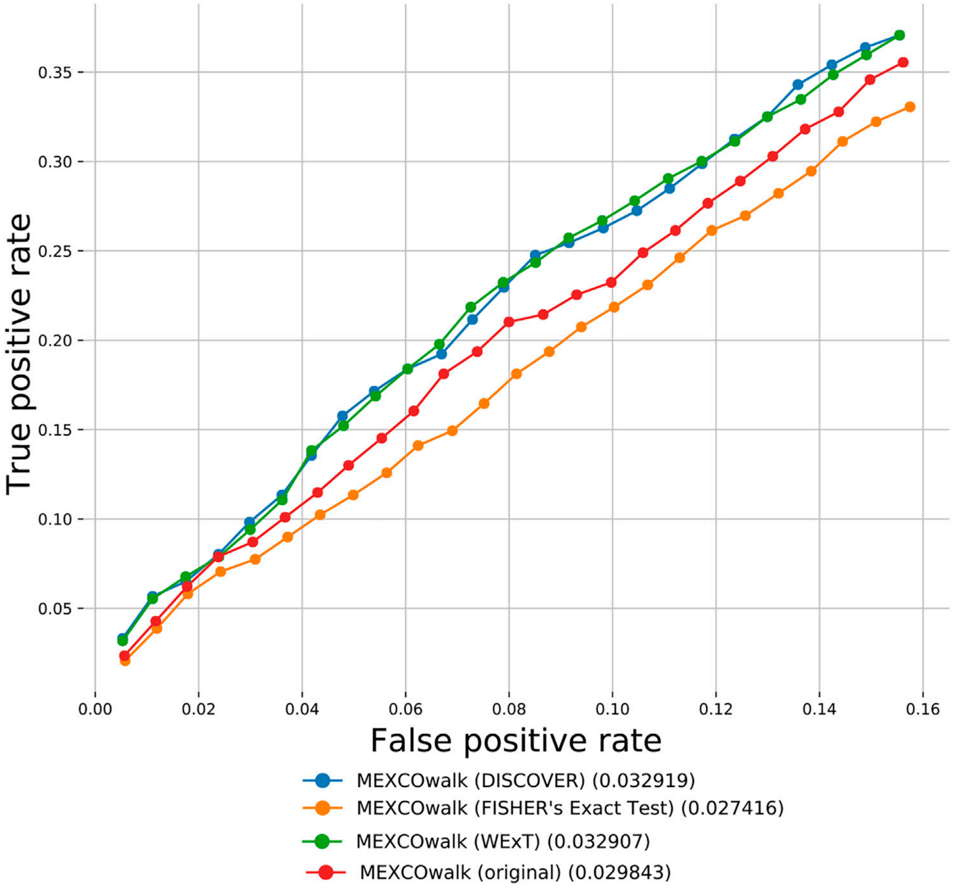


FIGURE 4 | The number of recovered CGC genes for the original MEXCOwalk as well as for its modified versions where mutual exclusivity values are estimated with DISCOVER, Fisher's Exact Test and WeXT. COADREAD dataset is used with $t = 5$ setting. The numbers in parentheses indicate the area under the ROC curve for the corresponding curve.

pairs result in larger AUROCs compared to non-CGC pairs for all ME methods (**Supplementary Figure S34**). We observe a steep increase in the ROC curves plotted for MEGSA results. This is due to the utilized likelihood ratio test that results in a *p*-value of 0.5 when the likelihood values are equal to each other. For UCEC, we see a significant difference between the ROC curves of CGC-pairs vs non-CGC pairs for Fisher's Exact Test and MEGSA; whereas the corresponding difference is negligible for DISCOVER and WeXT.

4 CASE STUDY

Apart from the defined network-centric ME evaluation framework, we discuss a case study where we assess whether mutual exclusivities estimated by the considered ME methods improve the performance of driver identification methods that utilize mutual exclusivity information. To this end, we compare the original version of MEXCOwalk with its alternatives where mutual exclusivity estimates are provided by the employed ME methods. Assuming that g_i and g_j genes are mutated in patient sets S_i and S_j, respectively; MEXCOWalk simply computes the mutual exclusivity between these two genes with the following formula: $|S_i \cup S_j|/(|S_i| + |S_j|)$. MEXCOWalk uses the estimated mutual exclusivity values as part of edge weights. As such, to utilize the *p*-values output by ME detection methods in MEXCOWalk, we first compute –log (*p*-value) and then convert the resulting values between 0 and 1. To this end, we replace all –log (*p*-value)'s larger than 10 with 1. We then find the maximum –log (*p*-value) less than 10 and divide all other –log (*p*-value)'s with this value. The reason why we set a threshold for finding the maximum is the large differences across the smallest *p*-values output by different ME methods. For instance, WeXT outputs a very large range of *p*-values and if we use the smallest *p*-value to scale, all other –log (*p*-value)s will be converted to values that are very close to 0. In the original MEXCOWalk study, a threshold of 0.7 is applied to ME values such that all values ≤0.7 are clamped to 0. This conversion is equivalent to removing those edges from the network since the edge weights include a multiplicative term for ME values. We find that the removal of these edges correspond to a 0.035 percent reduction in graph density. For the current analysis, we determine the threshold value for each ME detection method to achieve the same percent density reduction in the graph. **Figure 4** shows the number of recovered CGC genes for fixed output gene sizes from 100 to

2,500 as a ROC curve for original MEXCOwalk as well as for versions of MEXCOwalk where mutual exclusivity values are estimated with DISCOVER, Fisher's Exact Test and WeXT, respectively. We observe that MEXCOwalk with WeXT's ME values results in the best AUROC value for COADREAD. **Supplementary Figures S36** shows the analogous results for the other cancer types. For, LUSC, STAD and UCEC, MEXCOwalk with DISCOVER gives the best AUROC whereas for BLCA, LUAD and SKCM MEXCOwalk with Fisher's Exact Test performs the best. An important observation is the worse performance of MEXCOwalk with Fisher's Exact Test compared to the original MEXCOwalk for COADREAD, STAD and UCEC. As such, using Fisher's Exact Test in place of MEXCOwalk's original ME values does have the potential to decrease the performance whereas for the other ME methods we do not observe such a risk. Note that for these analysis we employ $t = 5$ since $t = 20$ filtering does not provide enough number of genes to be evaluated.

5 DISCUSSION

It is important to investigate whether the employment of an interaction network within our ME evaluation framework causes any ascertainment bias in the findings and to elaborate on how any such potential bias is mediated within the framework. It is established that known cancer genes have larger number of interactions compared to other genes in the network (Hou and Ma, 2014a). This implies a potential bias that needs to be resolved in cancer driver gene identification methods employing interaction network data. Such a bias is less of a problem for the current study, since our aim is not to identify novel cancer driver genes but to utilize the interaction network and known cancer genes to form a ground truth of mutually exclusive interactions for evaluating existing ME methods. On the contrary, the fact that most known cancer genes have well-characterized interactions in the network provides a benefit for our work as it supports the confidence of our true positive examples. Additionally, our framework makes use of not only genes from the reference set $\mathcal{S}$ but also genes not in $\mathcal{S}$ to create random controls. Nevertheless, the fact that some known cancer genes have significantly larger number of interactions compared to other known cancer genes could lead to a bias. For instance, for our analysis of the COADREAD data ($t = 20, \mathcal{S} = CGC$), there are 74 CGC genes among which five CGC genes have more than ten CGC neighbors whereas 41 have exactly one CGC neighbor. This could lead to a bias as CGC genes with large number of CGC neighbors contribute to the aggregate statistics and metrics much more than those CGC genes with small number of CGC neighbors. To mediate this bias, our framework includes additional results where all the statistics and the traditional measures such as the F1 score are calculated in a degree-normalized way for each gene and the gene-level results are then aggregated by taking an average across the genes. These results are available in the Supplementary Document; **Supplementary Tables S12, S24, S36, S48**. To summarize, the degree-normalized results are in agreement with those of the previous settings in almost all the cases in terms of ranking based on F1 score.

Another important point worth emphasizing is that apart from the aggregate statistics provided in the previous sections as part of the metrics for the network-centric ME evaluations, our proposed framework also provides analogous statistics at the gene-level as well. Such statistics may in fact be of more interest to cancer biologists than the aggregate statistics in certain cases. Several interesting observations can be made through an inspection of these gene-level evaluations, especially for the settings where the conventionally defined F1 score fails in quantifying ME. Genes with low MLA comprise an example setting, where TP53 is a leading member. Consider the case of TP53 in COADREAD evaluations for instance. With respect to the degree-normalized setting, the values of precision, sensitivity, precision$_{strict}$ and sensitivity$_{strict}$ for WeXT are respectively 0.5, 1, 0.25, 0.25 which gives rise to an F1 score of 0.66 and F1$_{strict}$ score of 0.25. On the other hand, MEMo provides the same precision, sensitivity and F1 scores as WeXT whereas its precision$_{strict}$, sensitivity$_{strict}$ and F1$_{strict}$ scores are all 0. To summarize, although the inspection of the F1 scores does not provide a distinction between the two results, an inspection of the F1$_{strict}$ scores establishes that MEMo is worse than WeXT in this setting. We note that the advantages of inspections based on the strict definitions of the metrics rather than the conventional ones are also apparent in the aggregate analysis as well. In addition to the COADREAD evaluations shown in **Table 2**, BRCA also contains an example instance where the conventional and the strict versions of the metrics provide different conclusions; see **Supplementary Table S7B**. In terms of the F1 scores, DISCOVER Strat ranks fourth, whereas comparing F1$_{strict}$ scores it ranks the second. Also, overall we observe that MEMo's performance gets severely affected when the strict versions of the metrics are employed.

Next, our robustness analysis results reveal some suggestions for potential users of our framework. We recommend using a p-value threshold smaller than 0.1 but larger than 0.05 as lower threshold values are too stringent and lead to too few predicted positives. Regarding robustness_iterations, we tested values both smaller than and higher than the default value of 100 for COADREAD evaluations: 5, 50, 100, 300 and 500. We repeated each experiment 20 times and calculated the standard deviation of the obtained set of F1 and F1$_{strict}$ scores. For the majority of the cases, we observe a large decrease in the standard deviation values when robustness_iterations is increased from 5 to 50. (**Supplementary Table S82**). This analysis suggests that the robustness_iterations should be set to a at least 50. Lastly, we observe that different PPI networks can lead to large differences in both the F1/F1$_{strict}$ scores and the ranking of the methods. As such, exploring different PPI sources would be beneficial.

To assess whether our findings extend to other datasets other than TCGA, we repeat our evaluations on somatic mutation data of 402 colon cancer patients within the Pancancer Analysis of Whole Genomes (PCAWG) study (Campbell, 2020). **Supplementary Tables S83-S86** shows the ME evaluations with respect to the metrics defined in section 4.2. We observe an overall decrease in F1$_{strict}$ scores of the methods. Compared to analogous results in TCGA data, WexT still performs the best in terms of F1$_{strict}$ score whereas the second best performing method is changed from MEMo to DISCOVER. The changes with respect to varying the *p*-value threshold, robustness_iterations value, input PPI, reference cancer gene set are consistent with the changes that we previously observe for the TCGA COADREAD dataset. When we switch from IntAct to its TSN version, we observe that all the ME methods estimate stronger mutual exclusivities for the tissue-specific CGC gene pairs compared to the non-tissue-specific CGC gene pairs as evident from AUROC values greater than 0.5; see **Supplementary Figures S37, S38** shows the results of MLA where we observe slightly smaller correlation values for DISCOVER (−0.84 vs −0.88) as compared to the results obtained from TCGA COADREAD dataset. We observe findings similar to those obtained from the TCGA COADREAD data in that the correlation values drop when the network information is incorporated. To summarize, our conclusions remain the same when we repeat our analyses on an entirely different cohort from the PCAWG study.

The majority of the somatic mutations observed in cancer genomics are passenger mutations. In the evaluations provided in the Results section we employ a simple filtering strategy where we remove silent mutations and mutations on non-coding regions of the genes. Additionally, we also assess the effects of employing a more elaborate mutation filtering procedure. To this end, we download the predictions of the Muiños et al. study on COREAD type (Muinos et al., 2021). This includes the classification of all possible mutations on 12 genes as driver or passenger mutations. Accordingly, we filter out the proposed passenger mutations from our mutation data and repeat all of our relevant analyses. We observe that the ranking of the methods according to the metrics proposed in section 4.2 remain the same where WExT, MEMO, and DISCOVER Strat show reduced $F1_{strict}$ scores, and DISCOVER and Fisher's Exact Test show higher $F1_{strict}$ scores (**Supplementary Tables S87-S90**). The TSN results and the MLA analysis results are also similar to our original results (**Supplementary Figures S39-S40**). Muiños et al. provides classifications of mutations on a subset of genes which have training data larger than a certain size. If such classifications become available for a larger set of genes in the future we can provide a better assessment regarding the filtering procedures employing these classifications.

Mutated genes in cancer prevalently exhibit a long tail phenomenon where few genes are mutated in many patients and large number of genes are mutated in few patients. To check whether assessing the mutual exclusivity of gene pairs with very different mutation frequencies bias the evaluations of the compared ME methods, we repeat our analyses after filtering out the genes with mutation frequencies $< 5\%$ and $> 30\%$. The results after this filtering step are available in **Supplementary Tables S91-S93**. We observe that the ranking of the methods remain the same where we see a significant increase in Precision/$\text{Precision}_{strict}$ values and a slight decrease in Sensitivity/$\text{Sensitivity}_{strict}$ values. When we look at the $\text{Precision}_{strict}$ values in more detail, we observe that the FP_{strict} values drop dramatically when we apply the filtering. This suggests that the control gene pairs that include genes with very low or very high mutation frequencies can have more significant *p*-values as compared to the *p*-values obtained for the corresponding CGC-CGC pair.

We also evaluate a more general ME detection method SELECT, which investigates both types of relationships among pairs, co-occurence and ME simultaneously. SELECT outputs ME associated scores to only a subset of the input gene pairs. Thus one strategy for comparing the results of SELECT against other methods is to focus only on such subsets. The relevant results where we use this strategy are available in **Supplementary Tables S94-S97**. We report evaluations on two subsets of TCGA COADREAD dataset: 1) the set of CGC-CGC pairs where SELECT results are available, 2) the set of CGC-CGC pairs where SELECT's version which uses subtype information (i.e., $\text{SELECT}_{\text{subtype}}$) are available. For the former, we observe that SELECT and $\text{SELECT}_{\text{subtype}}$ rank the fourth after WExT, MEMO, and DISCOVER. For the latter evaluation, $\text{SELECT}_{\text{subtype}}$ performs better than SELECT although both of them still rank the fourth among the other ME methods. Another strategy to fix this problem is to assign the worst ASC score to such pairs without specific ASC scores in the ME direction. We employ this approach as well and observe that it gives no significant difference in the comparisons.

Lastly, it is important to mention certain limitations of the proposed framework. Our framework is based on the presupposition that ME is likely to occur between interacting known cancer genes. Although rare, there may exist two different types of exceptions to this assumption; ME can be observed between non-interacting known drivers and the relationship between an interacting pair of known drivers can be that of co-occurrence rather than that of ME. These constitute respectively the false negative and the false positive events in our framework. An example instance of the former is the mutually exclusive mutations of APC and RNF43 observed in colorectal cancer [Mina et al., 2017] and example instance of the latter is the co-occurrence of CCNE1 and TP53 alterations [Zhang et al., 2014]. Both of these patterns are currently ignored by our framework and incorporation of mechanisms to dissect each such pattern to increase the performance of true ME detection is an important future step. Another limitation of the current framework is that it requires the availability of whole-genome or whole-exome sequencing data.

6 CONCLUSION

We propose a network-centric framework to evaluate pairwise mutual exclusivity findings reported by different ME algorithms. The first component of our framework consists of useful definitions of statistics employed in the network-centric ME evaluations. We observe that for the majority of the cancer types under study WeXT outperforms the other methods in terms of F1 score measured with respect to appropriately defined control groups. In half of the cancer types DISCOVER and in the other half MEMo perform as the second best methods. When comparing different cancer types we observe that BRCA and COADREAD are among the top two types leading to maximum F1 scores with at least one of the ME methods providing a score greater than 0.5. We note that DISCOVER Strat is only applicable in two cancer types among a total of eight since these are the only cancer types with well-defined subtypes. Furthermore, among these two cancer types, DISCOVER Strat outperforms original DISCOVER algorithm in BRCA, whereas it is the second worst method after Fisher's Exact Test in COADREAD. This is noteworthy since van de Haar et al. propose subtype stratification as employed by DISCOVER Strat as a way to emphasize true mutual exclusivity by reducing mutation load confounding (van de Haar et al., 2019). We also observe that Fisher's exact test and MEGSA are more conservative compared to DISCOVER and WeXT, where from the latter group, WeXT outputs notably larger number of significant p-values. The second component of our framework evaluates ME tests by comparing two types of measures obtained with and without network information. First measure is with respect to the percent significant findings of mutually exclusive gene pairs, whereas the second is based on MLC values. In most of the cancer types and for most of the genes we observe an increase with respect to the former whereas a decrease with respect to the latter measure. Finally, we repeat the same analysis by considering TSNs in the network-centric framework. Considerable improvements achieved due to the use of TSNs as opposed tissue nonspecific interaction network are only observed for BLCA and STAD datasets. A more detailed analysis in terms of comparing ROCs of CGC gene pairs and non-CGC gene pairs on cancer types with considerable number of tissue-specific gene pairs indicate the advantages of employing tissue specificity in detecting mutual exclusivity in COADREAD, SKCM, and UCEC. Finally we extend out network-centric evaluation framework to assess whether including network knowledge reduces the mutation load confounding problem.

As noted earlier the proposed framework is intended for the network-centric evaluations of mutual exclusivities of pairs of genes rather than groups of genes. Such a choice stems form the fact that the mutual exclusivities are commonly made use of in driver gene/module identification algorithms which mostly employ pairwise mutual exclusivities. Furthermore the extensive evaluation settings proposed, the number of ME methods under study and their own computational requirements, and the potentially exponential computational complexity inherent in handling groups of genes limits the scope of the current study to evaluations of pairwise ME scorings. Nonetheless most statistical ME methods are capable of providing ME results for groups of genes as well. Regarding the ME tests considered in this study, the main ME test provided by DISCOVER is based on a pairwise test definition but it also extends the definition for possible use in quantifying the ME of a group of genes, although the experiments involving the latter are based only on simulation data. The remaining tests MEGSA, MEMo, and WeXT are all ME tests specifically designed for groups of genes. An important direction for future work is to design a suitable extension of the proposed network-centric framework to evaluate the results of ME tests on groups of genes. Design choices relevant for such an extension would involve an appropriate and computationally efficient definition of the reference groups of genes analogous to a pair of interacting genes from the set $\mathcal{S}$ in the current setting and the definitions of control groups analogous to $\mathcal{X}_1$ and $\mathcal{X}_2$. Another future direction is to apply our network-centric framework on heterogeneous biological networks incorporating biological pathway information with PPI network data. Such incorporations have been successfully applied in other bioinformatics domains such as cancer driver identification (Hou and Ma, 2014b; Dinstag and Shamir, 2020).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Authors names are written in alphabetical order. CE and HK conceived the idea and supervised the study. RA implemented the code and performed the initial experiments. CY and AH repeated the experiments for other cancer types. All authors contributed to the preparation of the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by the Scientific and Technological Research Council of Turkey (grant number 117E879 to HK and CE.).

ACKNOWLEDGMENTS

We thank Joris van de Haar for sharing us the data used in van de Haar et al., 2019.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.746495/full#supplementary-material

REFERENCES

Ahmed, R., Baali, I., Erten, C., Hoxha, E., and Kazan, H. (2019). MEXCOwalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules. *Bioinformatics* 36 (3), 872–879. ISSN 1367-4803. doi:10.1093/bioinformatics/btz655

Baali, I., Erten, C., and Kazan, H. (2020). Driveways: A Method for Identifying Possibly Overlapping Driver Pathways in Cancer. *Sci. Rep.* 10. doi:10.1101/2020.04.01.015388

Babur, Ö., Gönen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., et al.B; ü; lent Arman Aksoy (2015). Systematic Identification of Cancer Driving Signaling Pathways Based on Mutual Exclusivity of Genomic Alterations. *Genome Biol.* 16 (1), 45. doi:10.1186/s13059-015-0612-6

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173, 371–e18. e18. doi:10.1016/j.cell.2018.02.060

Bethune, G., Bethune, D., Ridgway, N., and Xu, Z. (2010). Epidermal Growth Factor Receptor (Egfr) in Lung Cancer: an Overview and Update. *J. Thorac. Dis.* 2, 48–51.

Campbell, G. (2020). Pan-cancer Analysis of Whole Genomes. *Nature* 578, 82–93. doi:10.1038/s41586-020-1969-6

Canisius, S., Martens, J. W. M., Wessels, L. F. A., and Martens, M. (2016). A Novel independence Test for Somatic Alterations in Cancer Shows that Biology Drives Mutual Exclusivity but Chance Explains Most Co-occurrence. *Genome Biol.* 17 (261), 1–17. ISSN 1474-760X. doi:10.1186/s13059-016-1114-x

Carson, C. C., Moschos, S. J., Edmiston, S. N., Darr, D. B., Nikolaishvili-Feinberg, N., Groben, P. A., et al. (2015). Il2 Inducible T-Cell Kinase, a Novel Therapeutic Target in Melanoma. *Clin. Cancer Res.* 21 (9), 2167–2176. doi:10.1158/1078-0432.CCR-14-1826

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discov.* 2 (5), 401–404. ISSN 2159-8274. doi:10.1158/2159-8290.cd-12-0095

Chang, S.-W., Chao, W.-R., Ruan, A., Wang, P.-H., Lin, J.-C., and Han, C.-P. (2015). A Promising Hypothesis of C-KIT Methylation/Expression Paradox in C-KIT (+) Squamous Cell Carcinoma of Uterine Cervix ----- CTCF Transcriptional Repressor Regulates C-KIT Proto-Oncogene Expression. *Diagn. Pathol.* 10 (1). doi:10.1186/s13000-015-0438-2

Chen, L., Yang, L., Yao, L., Kuang, X.-Y., Zuo, W.-J., Li, S., et al. (2018). Characterization of Pik3ca and Pik3r1 Somatic Mutations in Chinese Breast Cancer Patients. *Nat. Commun.* 9 (1). doi:10.1038/s41467-018-03867-9

Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., et al. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* 17 (3), 251–264. doi:10.1016/j.jmoldx.2014.12.006

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual Exclusivity Analysis Identifies Oncogenic Network Modules. *Genome Res.* 22 (2), 398–406. doi:10.1101/gr.125567.111

Clark, A., and Burleson, M. (2020). Spop and Cancer: a Systematic Review. *Am. J. Cancer Res.* 10 (3), 704–726.

Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., and Beerenwinkel, N. (2015). TiMEx: a Waiting Time Model for Mutually Exclusive Cancer Alterations. *Bioinformatics* 32 (7), 968–975. doi:10.1093/bioinformatics/btv400

Dao, P., Kim, Y.-A., Wojtowicz, D., Madan, S., Sharan, R., and Przytycka, T. M. (2017). BeWith: A Between-Within Method to Discover Relationships between Cancer Modules via Integrated Analysis of Mutual Exclusivity, Co-occurrence and Functional Interactions. *Plos Comput. Biol.* 13, e1005695. doi:10.1371/journal.pcbi.1005695

Deng, Y., Luo, S., Deng, C., Luo, T., Yin, W., Zhang, H., et al. (2017). Identifying Mutual Exclusivity across Cancer Genomes: Computational Approaches to Discover Genetic Interaction and Reveal Tumor Vulnerability. *Brief in Bionform* 20, 254–266. doi:10.1093/bib/bbx109

Dimitrakopoulos, C. M., and Beerenwinkel, N. (2017). Computational Approaches for the Identification of Cancer Genes and Pathways. *Wires Syst. Biol. Med.* 9 (1), e1364, 2017 . ISSN 1939-5094. doi:10.1002/wsbm.1364

Dinstag, G., and Shamir, R. (2020). PRODIGY: Personalized Prioritization of Driver Genes. *Bioinformatics* 36 (6), 18311367–18394803. doi:10.1093/bioinformatics/btz815

Fleming, N. I., Jorissen, R. N., Mouradov, D., Christie, M., Sakthianandeswaren, A., Palmieri, M., et al. (2013). SMAD2, SMAD3 and SMAD4 Mutations in Colorectal Cancer. *Cancer Res.* 73 (2), 7251538–7357445. ISSN 0008-5472. doi:10.1158/0008-5472.CAN-12-2706

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al. (2017). Cosmic: Somatic Cancer Genetics at High-Resolution. *Nucleic Acids Res.* 45, D777–D783. doi:10.1093/nar/gkw1121

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the Cbioportal. *Sci. Signal.* 6 (269), pl1, 2013 . ISSN 1945-0877. doi:10.1126/scisignal.2004088

Gayther, S. A., Batley, S. J., Linger, L., Bannister, A., Thorpe, K., Chin, S.-F., et al. (2000). Mutations Truncating the Ep300 Acetylase in Human Cancers. *Nat. Genet.* 24, 300–303. doi:10.1038/73536

GTEX Consortium (2020). The Gtex Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* 369 (6509), 1318–1330. ISSN 0036-8075. doi:10.1126/science.aaz1776

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., et al. (2015). The Consensus Molecular Subtypes of Colorectal Cancer. *Nat. Med.* 21, 1350–1356. doi:10.1038/nm.3967

Hou, J. P., and Ma, J. (2014a). Dawnrank: Discovering Personalized Driver Genes in Cancer. *Genome Med.* 6 (56), 16. doi:10.1186/s13073-014-0056-8

Hou, J. P., and Ma, J. (2014b). DawnRank: Discovering Personalized Driver Genes in Cancer. *Genome Med.* 6 (7), 56, 2014b . ISSN 1756-994X. doi:10.1186/s13073-014-0056-8

Hua, X., Hyland, P. L., Huang, J., Song, L., Zhu, B., Caporaso, N. E., et al. (2016). Megsa: A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations. *Am. J. Hum. Genet.* 98, 442–455. doi:10.1016/j.ajhg.2015.12.021

Iwamoto, R., Mine, N., Mizushima, H., and Mekada, E. (2017). Erbb1 and Erbb4 Generate Opposing Signals Regulating Mesenchymal Cell Proliferation during Valvulogenesis. *Development* 144 (8), 2–e1. doi:10.1242/dev.152710

Jiang, Z. B., Ma, B. Q., Liu, S. G., Li, J., Yang, G. M., Hou, Y. B., et al. (2018). miR-365 Regulates Liver Cancer Stem Cells via RAC1 Pathway. *Mol. Carcinogenesis* 58 (1), 55–65. doi:10.1002/mc.22906

Kim, Y.-A., Cho, D.-Y., Dao, P., and Przytycka, T. M. (2015). MEMCover: Integrated Analysis of Mutual Exclusivity and Functional Network Reveals Dysregulated Pathways across Multiple Cancer Types. *Bioinformatics* 31 (12), i284–i292. doi:10.1093/bioinformatics/btv247

Kim, Y. A., Madan, S., and Przytycka, T. M. (2017). WeSME: Uncovering Mutual Exclusivity of Cancer Drivers and beyond. *Bioinformatics* 33 (6), 814–821. ISSN 1367-4803. doi:10.1093/bioinformatics/btw242

Leiserson, M. D. M., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous Identification of Multiple Driver Pathways in Cancer. *Plos Comput. Biol.* 9 (5), e1003054. doi:10.1371/journal.pcbi.1003054

Leiserson, M. D. M., Reyna, M. A., and Raphael, B. J. (2016). A Weighted Exact Test for Mutually Exclusive Mutations in Cancer. *Bioinformatics* 32, i736–i745. doi:10.1093/bioinformatics/btw462

Leiserson, M. D., Wu, H.-T., Vandin, F., and Raphael, B. J. (2015). Comet: A Statistical Approach to Identify Combinations of Mutually Exclusive Alterations in Cancer. *Genome Biol.* 16, 160, 2015 , ISSN 1474-7596. doi:10.1186/s13059-015-0700-7

Liu, S., Liu, J., Xie, Y., Zhai, T., Hinderer, E. W., Stromberg, A. J., et al. (2020). MEScan: a Powerful Statistical Framework for Genome-Scale Mutual Exclusivity Analysis of Cancer Mutations. *Bioinformatics* 11. doi:10.1093/bioinformatics/btaa957

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., et al. (2020). A Reference Map of the Human Binary Protein Interactome. *Nature* 580, 402–408. doi:10.1038/s41586-020-2188-x

Mina, M., Raynaud, F., Tavernari, D., Battistello, E., Sungalee, S., Saghafinia, S., et al. (2017). Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies. *Cancer Cell* 32 (2), 155–168. e6ISSN 1535-6108. doi:10.1016/j.ccell.2017.06.010

Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A., and Lopez-Bigas, N. (2021). In Silico saturation Mutagenesis of Cancer Genes. *Nature* 596, 428–432. doi:10.1038/s41586-021-03771-1

Nielsen, T. O., Poulsen, S. S., Journe, F., Ghanem, G., and Sorensen, B. S. (2014). Her4 and its Cytoplasmic Isoforms Are Associated with Progression-free Survival of Malignant Melanoma. *Melanoma Res.* 24 (1), 88–91. doi:10.1097/cmr.0000000000000040

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct Project-IntAct as a Common Curation Platform for 11 Molecular Interaction Databases. *Nucl. Acids Res.* 42 (D), D358–D363. ISSN 0305-1048. doi:10.1093/nar/gkt1115

Popovici, V., Budinska, E., Tejpar, S., Weinrich, S., Estrella, H., Hodgson, G., et al. (2012). Identification of a Poor-Prognosis BRAF-mutant-like Population of

Patients with colon Cancer. *Jco* 30 (12), 1288–1295. ISSN 1527-7755. doi:10.1200/JCO.2011.39.5814

Prickett, T. D., NeenaAgrawal, S., Agrawal, N. S., Wei, X., Yates, K. E., Lin, J. C., et al. (2009). Analysis of the Tyrosine Kinome in Melanoma Reveals Recurrent Mutations in Erbb4. *Nat. Genet.* 41 (10), 1127–1132. doi:10.1038/ng.438

Roth, A. D., Tejpar, S., Delorenzi, M., Yan, P., Fiocca, R., Klingbiel, D., et al. (2010). Prognostic Role of KRAS and BRAF in Stage II and III Resected Colon Cancer: Results of the Translational Study on the PETACC-3, EORTC 40993, SAKK 60-00 Trial. *Jco* 28 (3), 466–474. ISSN 1527-7755. doi:10.1200/JCO.2009.23.3452

Samowitz, W. S., Albertsen, H., Sweeney, C., Herrick, J., Caan, B. J., Anderson, K. E., et al. (2006). Association of smoking, CpG island methylator phenotype, and v600e BRAF mutations in colon cancer. *JNCI J. Natl. Cancer Inst.* 98 (23), 1731–1738. doi:10.1093/jnci/djj468

Sarto Basso, R., Hochbaum, D. S., and Vandin, F. (2019). Efficient Algorithms to Discover Alterations with Complementary Functional Association in Cancer. *PLOS Comput. Biol.* 15, e1006802. doi:10.1371/journal.pcbi.1006802

Singh, S. K., Sinha, S., Padhan, J., Jangde, N., Ray, R., and Rai, V. (2020). Myh9 Suppresses Melanoma Tumorigenesis, Metastasis and Regulates Tumor Microenvironment. *Med. Oncol.* 37 (10), 2020. doi:10.1007/s12032-020-01413-6

Skidmore, Z. L., Wagner, A. H., Lesurf, R., Campbell, K. M., Kunisaki, J., Griffith, O. L., et al. (2016). GenVisR: Genomic Visualizations in R. *Bioinformatics* 32 (19), 3012–3014. doi:10.1093/bioinformatics/btw325

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The Cosmic Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers. *Nat. Rev. Cancer* 18 (11), 696–705. doi:10.1038/s41568-018-0060-1

Song, J., Peng, W., and Wang, F. (2020). An Entropy-Based Method for Identifying Mutual Exclusive Driver Genes in Cancer. *Ieee/acm Trans. Comput. Biol. Bioinf.* 17 (3), 758–768. doi:10.1109/tcbb.2019.2897931

Szczurek, E., and Beerenwinkel, N. (2014). Modeling Mutual Exclusivity of Cancer Mutations. *Plos Comput. Biol.* 10 (3), e1003503–12. doi:10.1371/journal.pcbi.1003503

Thomas, R. K., Baker, A. C., Debiasi, R. M., Winckler, W., Laframboise, T., Lin, W. M., et al. (2007). High-throughput Oncogene Mutation Profiling in Human Cancer. *Nat. Genet.* 39, 347–351. doi:10.1038/ng1975

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2016). Evaluating the Evaluation of Cancer Driver Genes. *Proc. Natl. Acad. Sci. USA* 113, 14330–14335. doi:10.1073/pnas.1616440113

van de Haar, J., Canisius, S., Yu, M. K., Voest, E. E., Wessels, L. F. A., and Ideker, T. (2019). Identifying Epistasis in Cancer Genomes: A Delicate Affair. *Cell* 177, 1375–1383. doi:10.1016/j.cell.2019.05.005

Vandin, F., Upfal, E., and Raphael, B. J. (2012). De Novo discovery of Mutated Driver Pathways in Cancer. *Genome Res.* 22 (2), 375–385. doi:10.1101/gr.120477.111

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science* 339 (6127), 1546–1558. doi:10.1126/science.1235122

Wang, T., Ruan, S., Zhao, X., Shi, X., Teng, H., Zhong, J., et al. (2020). Oncovar: an Integrated Database and Analysis Platform for Oncogenic Driver Variants in Cancers. *NAR* 49 (D1), D1289–D1301. doi:10.1093/nar/gkaa1033

Yeang, C.-H., McCormick, F., and Levine, A. (2008). Combinatorial Patterns of Somatic Gene Mutations in Cancer. *FASEB j.* 22, 2605–2622. doi:10.1096/fj.08-108985

Zhang, J., Wu, L.-Y., Zhang, X.-S., and Zhang, S. (2014). Discovery of Co-occurring Driver Pathways in Cancer. *BMC bioinformatics* 15, 271. doi:10.1186/1471-2105-15-271

Zhang, J., and Zhang, S. (2018). The Discovery of Mutated Driver Pathways in Cancer: Models and Algorithms. *Ieee/acm Trans. Comput. Biol. Bioinf.* 15 (3), 988–998. doi:10.1109/TCBB.2016.264096310.1109/tcbb.2016.2640963

Zhang, Z., Yang, Y., Zhou, Y., Fang, H., Yuan, M., Sasser, K., et al. (2020). A Forward Selection Algorithm to Identify Mutually Exclusive Alterations in Cancer Studies. *J. Hum. Genet.* 66, 509–518. doi:10.1038/s10038-020-00870-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fgene.2021.795123

Human Immune System Diseasome Networks and Female Oviductal Microenvironment: New Horizons to be Discovered

Angela Taraschi[1,2], *Costanza Cimini*[1], *Alessia Colosimo*[1], *Marina Ramal-Sanchez*[1], *Fadl Moussa*[1,3], *Samia Mokh*[4], *Luca Valbonetti*[1,5], *Giulia Capacchietti*[1], *Israiel Tagaram*[1], *Nicola Bernabò*[1,5]* *and Barbara Barboni*[1]

[1]Faculty of Biosciences and Technology for Food, Agriculture and Environment, University of Teramo, Teramo, Italy, [2]Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale", Teramo, Italy, [3]Doctoral School of Science and Technology Lebanese University, Beirut, Lebanon, [4]National Council for Scientific Research (CNRS), Lebanese Atomic Energy Commission (LAEC), Laboratory for Analysis of Organic Compound (LACO), Beiru, Lebanon, [5]Institute of Biochemistry and Cell Biology (CNR-IBBC/EMMA/Infrafrontier/IMPC), National Research Council, Rome, Italy

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Malgorzata Kloc,
Houston Methodist Research Institute,
United States
Jose Artur Chies,
Federal University of Rio Grande do
Sul, Brazil

***Correspondence:**
Nicola Bernabò
nbernabo@unite.it

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 14 October 2021
Accepted: 14 December 2021
Published: 27 January 2022

Citation:
Taraschi A, Cimini C, Colosimo A,
Ramal-Sanchez M, Moussa F, Mokh S,
Valbonetti L, Capacchietti G,
Tagaram I, Bernabò N and Barboni B
(2022) Human Immune System
Diseasome Networks and Female
Oviductal Microenvironment: New
Horizons to be Discovered.
Front. Genet. 12:795123.
doi: 10.3389/fgene.2021.795123

Human hypofertility and infertility are two worldwide conditions experiencing nowadays an alarming increase due to a complex ensemble of events. The immune system has been suggested as one of the responsible for some of the etiopathogenic mechanisms involved in these conditions. To shed some light into the strong correlation between the reproductive and immune system, as can be inferred by the several and valuable manuscripts published to date, here we built a network using a useful bioinformatic tool (DisGeNET), in which the key genes involved in the sperm-oviduct interaction were linked. This constitutes an important event related with Human fertility since this interaction, and specially the spermatozoa, represents a not-self entity immunotolerated by the female. As a result, we discovered that some proteins involved in the sperm-oviduct interaction are implicated in several immune system diseases while, at the same time, some immune system diseases could interfere by using different pathways with the reproduction process. The data presented here could be of great importance to understand the involvement of the immune system in fertility reduction in Humans, setting the basis for potential immune therapeutic tools in the near future.

Keywords: diseasome, immune system, oviductal environment, human, biological network, immunological disease, rheumatoid arthritis, asthma

1 INTRODUCTION

Fertilization is a cell-cell recognition process that occurs naturally *in vivo* within the oviductal microenvironment of the female body. The successful interaction between the spermatozoa (male gametes) and the oocyte (female gamete) is supported by the presence of oviduct epithelial cells (OECs) and the oviductal fluid, that participates in this complex dialogue either by directly interacting with the gametes (OECs) and secreting (OECs) or carrying (oviductal fluid) different molecules necessary to achieve a successful fertilization.

The process initiates with the arrival of the ejaculated spermatozoa to the cervix, where only the healthiest spermatozoa are selected to advance towards the uterus (or are directly deposed within the uterus, depending on the species), cross the utero-tubal junction and reach the oviduct (reviewed in Suarez, 2016; Gadella, 2017; Li and Winuthayanon, 2017). Here, sperm cells are able to bind to the oviductal epithelium for an indefinite period of time, varying from hours to days (species-specific) and forming the so-called "functional sperm reservoir," before being released to continue their way towards the oocyte (Suarez and Pacey, 2006; Coy et al., 2012). As a result of this close interaction, it is originated a cross-talk between the OECs and the sperm cells, that is important to ensure the success of early reproductive events (Almiñana, 2015). With regard to the oviductal fluid, it is mainly composed of amino acids, energy metabolites, inorganic salts, glycosaminoglycans and numerous proteins (Ballester et al., 2014; Coy and Yanagimachi, 2015; Canha-Gouveia et al., 2019), that are either passively or actively transported over the epithelial barrier from the circulating blood or the interstitial tissue, or *de novo* secreted by the OECs (Saint-Dizier et al., 2020) and are able to sustain and drive the biochemical machinery of spermatozoa and embryos during their journey.

Thus, on the one hand, the oviduct and its secretions influence the physiology of the gametes (Avilés et al., 2010), while on the other one hand the reproductive cells are able to modulate the oviductal environment by activating a cell-type-specific signalling pathway leading directly to specific alterations in the tubal fluid composition (Georgiou et al., 2007).

Overall, the study of the interaction between the female counterpart with male gametes (firstly) and embryos (secondly) poses a fascinating and challenging questions involving all the hemostatic mechanisms of the body. If the role of neuro-endocrine system is evident, now new emerging evidences are highlighting the involvement of immune system. For instance, the spermatozoa are clearly not-self and the embryos are semi-allogenic, but instead to be attacked by the maternal immune system they are tolerated for days or even months (Zandieh et al., 2015), thus indicating the existence of a gamete recognition system (Georgiou et al., 2007), as will be explained in the discussion section. Moreover, the immune system is involved in the etiopathogenesis of reproductive diseases, as it happens in case of immune/immunological infertility. This condition is diagnosed when spontaneously produced antibodies bind to the antigens occurring on the male gametes, with the production of anti-sperm antibodies (ASA) (Bohring and Krause, 2003; Brazdova et al., 2016).

Ultimately, the involvement of immune system in determining the success of fertility, or its partial or total failure (hypo-fertility or infertility) is still far to be completely deciphered, and the molecules involved in linking reproductive function with immune response are still under investigation.

For this reason, here we carried out an innovative study to explore the possible involvement of genes encoding for proteins that participate to the functional dialogue existing between male gametes and female structures in immune pathologies. In particular, we used an approach based on the application of network theory to the study of biological complexity. By definition a network is a set of nodes (in our care the genes or the diseases) linked by edges (relationship between genes and diseases). The statistical study of network properties will lead to infer biologically relevant information, otherwise hidden by the complexity of the system.

To that, the work was carried as follow: I) retrieving in literature of the proteins involved in the sperm-oviduct interaction; II) creation of the list with the corresponding genes for those proteins; III) linking of the genes to the immune system disease in which it is involved, thus obtaining a bipartite network (a gene-disease network); IV) analysis of the network to infer biologically relevant information; and V) deep analysis of the relevance of this association in animal models of every human immune diseases, which constitutes one of the most valuable experimental approaches used in medical sciences.

The final aim was to suggest new players in the complex relationship between the reproductive function and immune pathology, to shed some light on how fertility could be compromised in immune system dysregulation.

2 MATERIALS AND METHODS

2.1 Data Collection

In order to recreate the microenvironment in which fertilization occurs, we collected the scientific literature published between 2005 and December 2020 in peer-reviewed international papers included in Scopus (https://www.scopus.com; accessed on 20/09/2021). In parallel, and as a quality control, two qualified researchers used the same key-words ("protein" AND "oviductal secretion" or "oviduct"), to carry out an independent search on the published manuscripts including information about the proteins found in the human oviduct. Then, the databases were compared, and a third qualified researcher verified the correctness of the record inserted, resolving eventual conflicts.

Data from each independent search was extracted to Excel spreadsheets (Microsoft Corporation, Albuquerque, USA), filling in and the following fields:

- *Species:* human;
- *Protein:* protein found in oviductal environment;
- *Gene:* protein-related gene;
- *Biological function:* physiological and/or pathological role of the protein;
- *Role in fertilization:* physiological and/or pathological role of the protein related to fertility;
- *OF/OEC/oviductal tissue:* protein identified within the oviductal fluid, on/in the oviductal epithelial cells or oviductal tissue;
- *References:* article reporting the above-mentioned data;
- *Phenotype ko mice:* existence of KO mouse and its relative phenotype;
- *Notes:* any further information useful for the study.

These data can be found in **Supplementary Material S1**.

2.2 Diseasome Creation and Visualization

Bioinformatics analysis was performed using Reactome, DisGeNET Cytoscape App, and Cytoscape 3.7.2.

First, we uploaded the gene list to Reactome (http://www.reactome.org/; accessed on 11/10/2021), a free, open-source, curated and peer-reviewed pathway database useful to visualize and analyse the biochemical pathways in which the genes are involved.

DisGeNET is a Cytoscape plugin designed to analyze human gene–disease association (GDA) networks, the diseasome. GDA is represented as a bipartite graph in which a set of nodes consists of diseases and the other one of disease-associated genes (Bauer-Mehren et al., 2010; Pavlopoulos et al., 2018). A disease and a gene are connected by a link only if the gene is implicated in the particular disease (Pavlopoulos et al., 2018). DisGeNET integrates information on human diseases and their genes from expert curated repositories, GWAS catalogues, animal models and the scientific literature discovered by text-mining approaches (Pinero et al., 20152015; Piñero et al., 2017; Piñero et al., 2020). Data are organized according to the type of source databases:

- CURATED: gene-disease association provided by expert curated resources, such as UniProt, ClinGen, Orphanet and CTD (human data), among others (Piñero et al., 2020);
- ANIMAL MODELS: gene-disease association provided by resources containing information about animal models (currently rat and mouse) of disease (RGD, MGD, and CTD) (Piñero et al., 2020);
- INFERRED: gene-disease association from the Human Phenotype Ontology and from VDAs reported by Clinvar, the GWAS catalogue and GWAS db (Piñero et al., 2020);
- ALL: gene-disease association from the previous sources and from LHGDN and BeFree (Piñero et al., 2020).

In addition, DisGeNET is able to classify the diseases according to the MeSH hierarchy and the genes according to the PANTHER Protein Class Ontology and Reactome top-level pathways (Pinero et al., 20152015). The gene-diseases associations are classified according to the DisGeNET association type ontology, that describes the different types of association between a gene and a disease, integrating information from the different databases (Bauer-Mehren et al., 2011). The GDA ontology is available at https://www.disgenet.org/dbinfo (accessed on 20/05/2021).

Using the DiGeNET Cytoscape App, we built two different networks for each gene in "Gene Disease Networks" tab, selecting "curated" or "animal models" as sources and "Immune System Diseases" as disease class. After merging the obtained networks on Cytoscape, we built two final diseasomes: the first curated (CURDi) and the second referred to animal models (AMDi). Both were then analysed using the plugin Network Analyzer.

2.3 Network Creation, Visualization, and Analysis

As previously stablished, the diseasome network was realized and analyzed using Cytoscape 3.7.2 and the specific plug-in Network Analyzer.

3 RESULTS AND DISCUSSION

3.1 Proteins Involved in the Sperm-Oviduct Interaction

The sperm-oviduct interaction and fertilization process can be considered as complex systems constituted by networks of heterogeneous elements interacting among them in a non-linearly way, giving rise to an emergent behavior. Thus, their properties cannot be explored or predicted simply by analysing their individual components, rather by putting their individual pieces togheter and building a network model. To this aim, a total of 145 proteins were identified through the literature search as proteins expressed within the oviduct and involved in the sperm-oviduct interaction in humans (see **Supplementary Material S1**, second sheet for the list of proteins and their corresponding genes, LOPaG). Here, we have used Reactome to investigate the pathways in which the identified proteins are involved. The analysis showed the 25 most relevant immunology pathways (see **Figure 1**; **Supplementary Material S2**), stressing the strong correlation between reproduction and immune system.

Then, by using the DisGeNET Cytoscape App and the genes list, we realized a bipartite network, i.e., a graph constituted by two families of nodes (genes and immune diseases) connected by edges and that represent the gene-disease association.

Depending on the data source (Curated or Animal Models Archives) we obtained two different diseasome networks: curated diseasome network (CURDi, see **Figure 2**) and animal model diseasome network (AMDi, see **Figure 3**).

3.2 CURDi Network and the Most Linked Genes

In CURDi network 54 of the 145 genes present in LOPaG were correlated with 124 immune system diseases.

As showed in **Figure 4**, the most linked genes in CURDi were *HLA-B, SERPING1* and *IFNG* (64, 52, and 42 links each one, respectively) (see **Supplementary Material S3**, sheet 1 for the complete list). These genes are well-studied for their key role in the immune response since their alteration may be responsible for several immune system diseases. Interestingly, there is growing evidence on the roles played by proteins encoded by the *HLA-B, SERPING1* and *IFNG* genes in several steps of the reproduction process.

3.2.1 HLA-B Gene

Among the 54 genes correlated with immune system diseases within the CURDi network, HLA-B stands out as the most linked one. This gene encodes for the human leukocyte antigen type B (HLA-B), one of the more than 200 genes belonging to the major histocompatibility complex (MHC) in humans. Located on chromosome 6p21.3, it comprises specific HLA class I (HLA-A and -B) and class II (HLA-DRB1, -DQA1, -DQB1, -DPA1 and -DPB1) genes that encode for cell-surface glycoproteins, whose main action is the induction and regulation of immune response (Leone et al., 2013; Wieczorek et al., 2017; Jongsma et al., 2019).

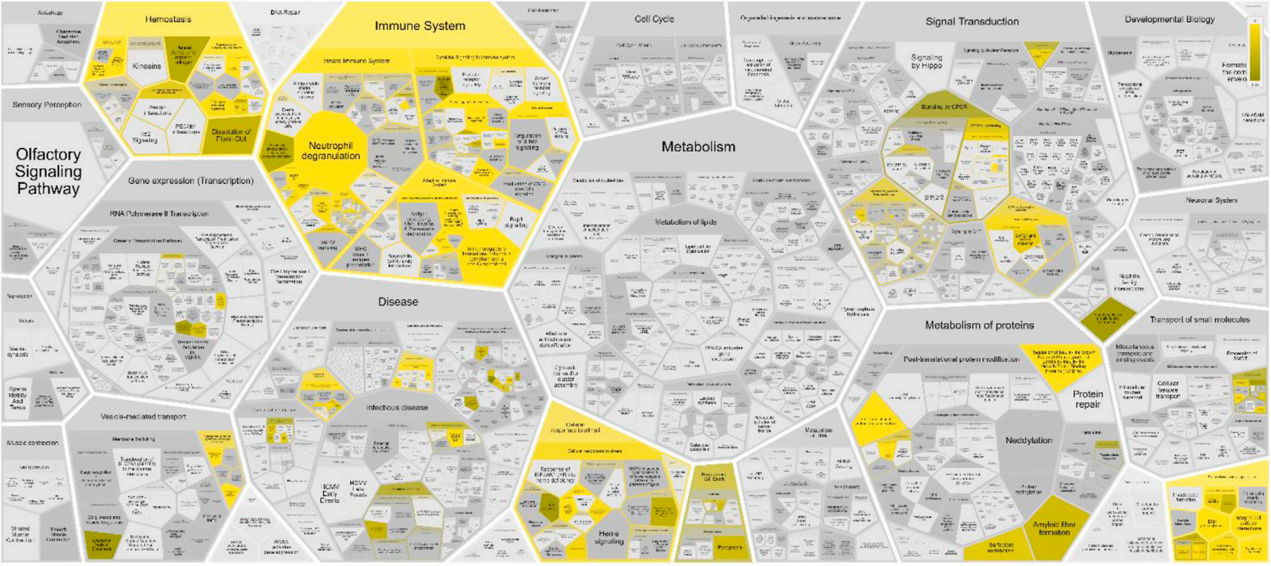


FIGURE 1 | Voronoi pathway visualization (Reacfoam) for the identified proteins in human oviduct. The color code denotes over-representation of that pathway in our input dataset. Light grey signifies pathways which are not significantly over-represented.

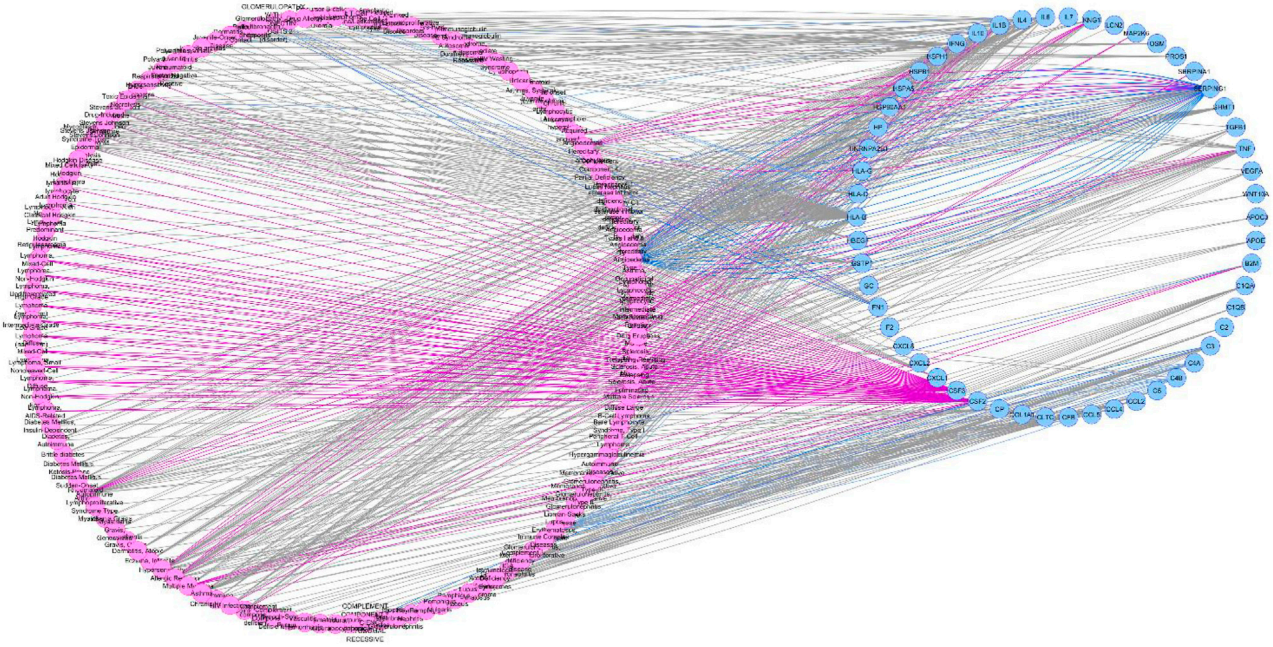


FIGURE 2 | Curated diseasome network (CURDi). CURDi forms the two node sets of bipartite networks with two types of nodes: diseases (pink circle) and gene (blue circle). Disease node and gene node are connected if the gene is implicated in the disorder.

The genes of the MHC are the most polymorphic of the human genome with a total of 13,023 HLA alleles (HLA class I: 9749; HLA class II: 3274) (Robinson et al., 2015). Interestingly, distinct HLA alleles have been associated with several human pathological conditions (Tersigni et al., 2020), while HLA proteins also own an important role in non-pathological conditions, such as lifespan and social behavior (Mosaad, 2015).

Regarding more Specifically, different alleles of the HLA-B gene have been associated with autoimmune diseases (such as HLA-B27 and its relationship with psoriatic arthritis and

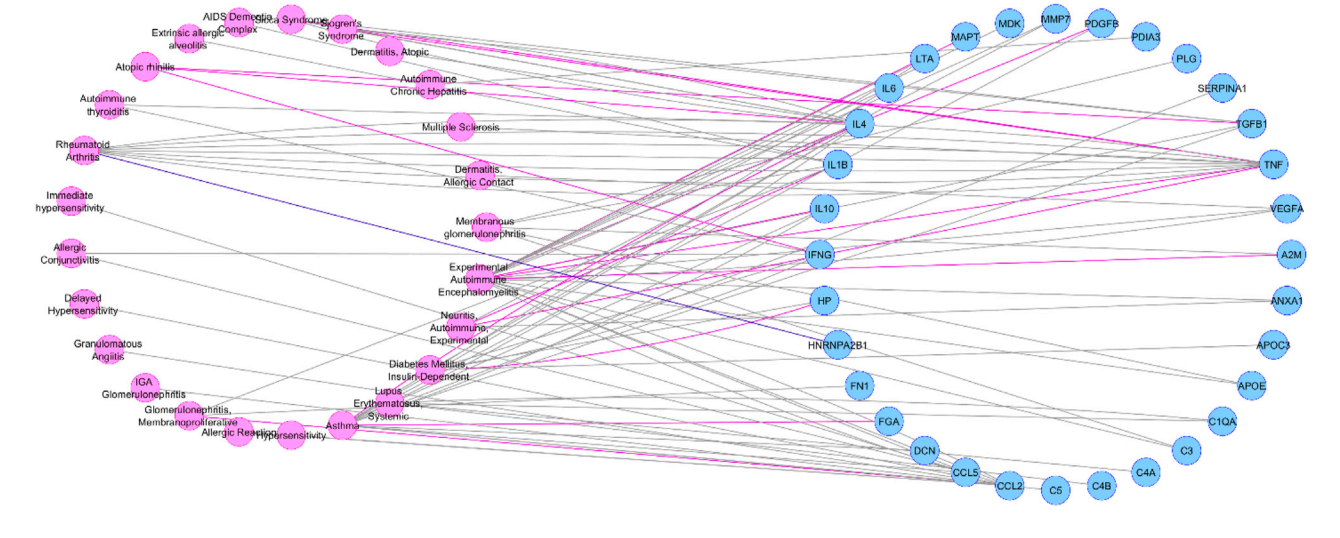


FIGURE 3 | Animal model diseasome network (AMDi)**.** AMDi form the two node sets of bipartite networks with two types of nodes: diseases (pink circle) and gene (blue circle). Disease node and gene node are connected if the gene is implicated in the disorder.

ankylosing spondylitis), inflammatory diseases (such as HLA-B*35 and systemic sclerosis, and HLA-B*52 and Takayasu arteritis), viral infections (such as HLA-B*35 phenotype and progression of Acquired Immune Deficiency Syndrome-AIDS) and tumor risks (such as HLA-B*52:01 and cervical cancer). In addition, it has been demonstrated an association of HLA-B alleles and severe drug hypersensitivity syndromes (such as HLA-B*57:01 and hypersensitivity to abacavir, and HLA-B*15:02 and use of carbamazepine) (Profaizer and Eckels, 2012).

In the reproductive field, the HLA antigens have been demonstrated to be crucial for the embryo-maternal tolerance and the achievement of a successful pregnancy (Chattopadhyay et al., 2014; Tersigni et al., 2020). For instance, some molecules as the high polymorphic HLA-C participate in the innate immune system by serving as a ligand for the inhibitory killer cell immunoglobulin-like receptors (KIRs) present on natural killer (NK) cells (Leone et al., 2013; Wilczyńska et al., 2020). HLA-C (along with the HLA-E, G and F ones) from both maternal and paternal origin is highly expressed by the extravillous trophoblasts invading the uterine tissues. While the paternal HLA-C protein represents a main target for maternal NK and T cells, an increased expression of foreign HLA-C (as in the case of oocyte donation) can be correlated with an incorrect placentation and further linked pathologies, thus requiring a tight regulation in the dual function of the protein (Papúchová et al., 2019). Despite the absence of evidence regarding the direct involvement between HLA-B and the immune response in the embryo, it might be possible to hypothesize that the close link between HLA-B and the encoding area of HLA-C could exert an indirect effect in the interaction between the NK cells from the uterus and the trophoblast HLA-C (Nielsen et al., 2017).

In addition, discordant results have been reported so far on the role of HLA polymorphisms on the susceptibility to pre-eclampsia (PE) (Emmery et al., 2016). This complex disease, exclusive to human pregnancy, shows clinical features as a new onset of hypertension and proteinuria after 20 weeks of gestation and is characterized by a systemic disproportionated inflammatory response, representing the main cause of maternal and perinatal morbidity and mortality with a prevalence of 3–8% in the total number of pregnancies worldwide and an increasing incidence. The four main potential causes underlying the pathophysiology of pre-eclampsia include: an immunological maladaptive tolerance between maternal, paternal, and fetal tissues; placental implantation with abnormal trophoblastic invasion; oxidative stress causing endothelial cell dysfunction; and genetic and epigenetic predisposing alterations (Agius et al., 2018). Regarding the immunological maladaption occurring between mothers and fetuses, few studies have focused on the role of HLA alleles in inducing pre-eclampsia. Wiktor and collaborators reported a significant increase of HLA-B13 allele frequency in patients with pre-eclampsia and of HLA-B22 allele in their male partners (Wiktor and Kozioł, 1998). A subsequent study of Zhang Z et al. in 119 Chinese pre-eclamptic patients showed a higher frequency of some HLA alleles shared by mothers and fetuses (HLA-A11, HLA-B13, HLA-B15, HLA-B22), and a lower frequency of a different protective allele (HLA-B14) (Zhang et al., 2009). On the contrary, a study carried out in 201 Danish couples of mothers and children reported no specific association with HLA-A, -B, and -DR alleles, denying the role of HLA antigens as risk factors for pre-eclampsia (Biggar et al., 2010). An association of HLA-G polymorphic alleles with pre-eclampsia has also been reported in several studies (Moreau et al., 2008; Tan et al., 2008; Persson et al., 2017).

Recently, a more comprehensive report of genome-wide association (GWAS), transcriptomics, proteomics and metabolomics studies identified inhibin as a potential preeclamptic biomarker (Benny et al., 2020).

Despite few studies have focused on the role of HLA alleles in inducing pre-eclampsia, further functional studies are necessary

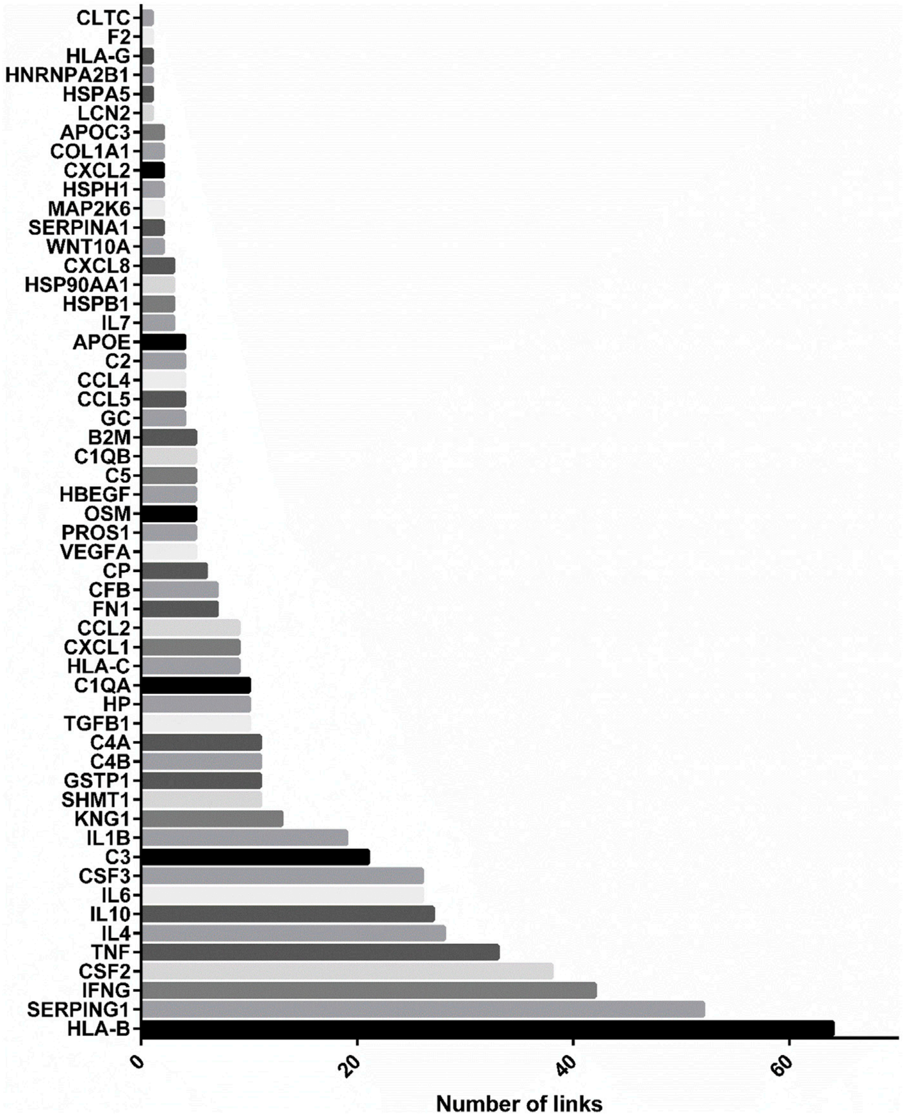


FIGURE 4 | Most linked genes in CURDi network. The histograms show the most linked genes to immune system diseases in CURDi: HLA-B, SERPING1 and IFNG.

to clarify an effective role of the classical HLA genes in its etiopathogenesis.

3.2.2 SERPING1 Gene

The second most linked gene, SERPING1, encodes for the plasma protease serine inhibitor (C1-INH), also known as SERPING1 or C1-inhibitor (Madsen et al., 2014). C1-INH regulates the activation of the classical and lectin complement pathways, coagulation and fibrinolysis cascades (López-Lera et al., 2014). Mutations in the SERPING1 gene are responsible for the largest cases of hereditary angioedema (HAE) (OMIM#106100), a rare autosomal dominant disorder that causes recurrent attacks of cutaneous angioedema, severe abdominal pain, and airway compromise (Santacroce et al., 2021). The disease course during pregnancy is unpredictable, with one study showing that seven Australian patients with HAE had reduced or absent attacks in the last two trimesters of pregnancy, while in the post-partum period they suffered from increased frequency and more severe attacks (Chinniah and Katelaris, 2009). However, fertility seems not to be impaired by HAE itself or by HAE medications (Yakaboski et al., 2020).

A network study by Sabetian and coll. (2014) built a sperm and oocyte protein interaction network and revealed new protein interactions. For example, the authors indicated that SERPINE1, also known as PAI-1 (plasminogen activator inhibitor), is located on the surface, in the tail and in the acrosome of mature spermatozoa, participating in the sperm-egg interaction by interacting with C1-INH of the oocyte (Sabetian et al., 2014). Thus, our results suggest that new studies could be useful to better clarify the interactions among the SERPING1 gene, immune diseases and fertility.

3.2.3 INFG Gene

The IFNG gene codifies for an extracellular proinflammatory cytokine (interferon γ, IFN-γ) that constitutes the main effector of cell-mediated immunity. Its main function is to recognize and eliminate pathogens by enhancing the antigen recognition through the antigen presenting cells and T cells, and is secreted by $CD4^+$, NK and NKT cells. It is able to intervene as the early host defense and autocrine regulation but also during the adaptative immune response (reviewed in (Schroder et al., 2004; Bhat et al., 2018; Kak et al., 2018)).

In reproduction, IFN-γ shows an important role on embryo implantation and pregnancy progression (Robertson et al., 2018). For instance, increased levels of IFN-γ have been associated with a reduced fertility (Carrasquel et al., 2014), as evidenced by the results of Carrasquel and coll. (2014). In that *in vitro* study, high concentrations of IFN-γ affected the intracellular calcium concentration, altering the sperm membrane permeability and thus impairing the sperm fertilizing ability (Carrasquel et al., 2014). Moreover, it has been demonstrated that an excess of the protein can also promote the generation of cytotoxic or $CD8^+$ cells during the embryo implantation that later drives to fetal loss (Robertson et al., 2018), thus supporting its fundamental involvement as a regulator of the maternal-fetal immune relationship.

Being secreted in the uterus during early pregnancy, IFN-γ plays a critical role in gestation, including remodeling of endometrial vasculature, angiogenesis at implantation sites, and maintenance of the decidual (maternal) component of the placenta. Alteration of INF-γ levels in the plasma of pregnant women may contribute to severe gestational pathologies, such as autoimmune disease, preterm labor, and preeclampsia (Sargent et al., 2006; Murphy et al., 2009; Yang et al., 2014).

One plausible mechanism could be the inability of the mother to switch from T helper cell type 1 (Th1) to Th2 cytokine profiles at the fetal-maternal interface, due to an altered expression of INF-γ and its receptors (IFN-γ R1 and IFN-γ R2) (Sargent et al., 2006).

3.2.4 Other Genes

In the list of most connected genes, CSF2 showed 38 links. This gene encodes for the granulocyte-macrophage colony-stimulating factor (GM-CSF), responsible for the growth and differentiation of hematopoietic precursor cells in granulocytes, macrophages, eosinophils and erythrocytes, among others. Interestingly, an important role has also been given to this protein during the fertilization process. Specifically, GM-CSF was found to mediate the maternal effects on embryonic development during preimplantation, probably by inducing the expression of IFN-γ (Loureiro et al., 2009). The presence of GM-CSF receptors has been also described in the midpiece and principal segment of the tail of mature spermatozoa in human and bovine species, while it was also demonstrated that GM-CSF was able to improve sperm motility when added to bovine sperm samples (Vilanova et al., 2003). In Csf2 null mutant mice, a deficiency in GM-CSF protein levels resulted in altered differentiation and maturation of junctional-zone trophoblast lineages, glycogen cells, and giant cells, thus suggesting the role of the Csf2 gene as a regulator of trophoblast differentiation and placental development (Sferruzzi-Perri et al., 2009).

Among the other most connected genes in the CURDi network stand out several genes codifying for cytokines, such as the tumor necrosis factor alpha (TNF-α), interleukins 4, 6 and 10 (IL-4, IL-6 and IL-10, respectively), and granulocyte colony-stimulating factor (G-CSF). TNF-α is a cytokine codified by the TNF gene and with a wide variety of functions. It is naturally produced by activated macrophages and monocytes, and its increased levels have been associated with infertility in humans (Eggert-Kruse et al., 2007; Yildizfer et al., 2015; Pinto-Bravo et al., 2017). Although few studies evaluated the role of TNF-α in the oviduct, evidence support that TNF-α may modulate the oviduct contraction necessary for transporting the gametes and embryo into the site of fertilization and the uterus, respectively (Wijayagunawardane et al., 2003; Parada-Bustamante et al., 2016). In addition, increased levels of TNF-α was detected in the tubal fluid of patients with hydrosalpinx and salpingitis due to chlamydial or gonococcal infection (Nasu et al., 2007). In these pathological conditions, TNF-α may induce the vascular endothelial growth factor (VEGF) production, which may further enhance the oviductal secretion by regulating vascular permeability (Nasu et al., 2007).

Interleukin-4 and -10 are pleiotropic anti-inflammatory cytokines that function mainly by suppressing the pro-inflammatory milieu (Chatterjee et al., 2014). For this reason, they play crucial roles in the success of pregnancy: progesterone induces the IL-4 and IL-10 production, which acts to inhibit Th1 responses during pregnancy, creating a tolerogenic environment in women (Chatterjee et al., 2014; Shahbazi et al., 2019). Indeed, while the trophoblastic cell implantation into endometrial cells is associated with an active Th1 pro-inflammatory response, the pregnancy maintenance is marked by an anti-inflammatory response, promoting fetal allograft tolerance and ensuring fetal development (Granot et al., 2012; Chatterjee et al., 2014).

Interleukin-6 is a pleiotropic cytokine involved in both acute and chronic inflammatory processes (Papathanasiou et al., 2008; Balasubramaniam et al., 2012). Papathanasiou and coll. (2008) showed that IL-6, in addition to act as an inflammatory marker, is capable *in vitro* to significantly reduce the ciliary beat function (CBF) causing a severe tubal damage, whereas the addition of anti-IL-6 restores the activity of CBF (Papathanasiou et al., 2008). IL-6 may also play a role in the pathophysiology of tubal ectopic gestation. Indeed, it was demonstrated that the expression of IL-6 is significantly increased near the implantation site in tubes with ectopic gestation, as compared with normal gestations (Balasubramaniam et al., 2012). On the other hand, IL-6 has been shown to affect sperm motility and to induce protein tyrosine phosphorylation in human spermatozoa (Laflamme et al., 2005).

Granulocyte-colony stimulating factor (G-CSF) is a pleiotropic cytokine belonging to the hematopoietic growth factor family that codifies by the CSF3 gene. Recent studies

TABLE 1 | Group of diseases and number of diseases included within each group.

Class	Disease class	Number of diseases
C15	Hemic and Lymphatic Diseases	45
C17	Skin and Connective Tissue Diseases	33
C04	Neoplasms	32
C16	Congenital, Hereditary and Neonatal Diseases and Abnormalities	14
C14	Cardiovascular Diseases	13
C12	Male Urogenital Diseases	11
C13	Female Urogenital Diseases and Pregnancy Complications	11
C05	Musculoskeletal Diseases	8
C25	Chemically-Induced Disorders	8
C10	Nervous System Diseases	7
C18	Nutritional and Metabolic Diseases	7
C23	Pathological Conditions Signs and Symptoms	7
C19	Endocrine System Diseases	6
C01	Infections	5
C07	Stomatognathic Diseases	5
C08	Respiratory Tract Diseases	3
C06	Digestive System Diseases	1
C11	Eye Diseases	1
C24	Occupational Diseases	1

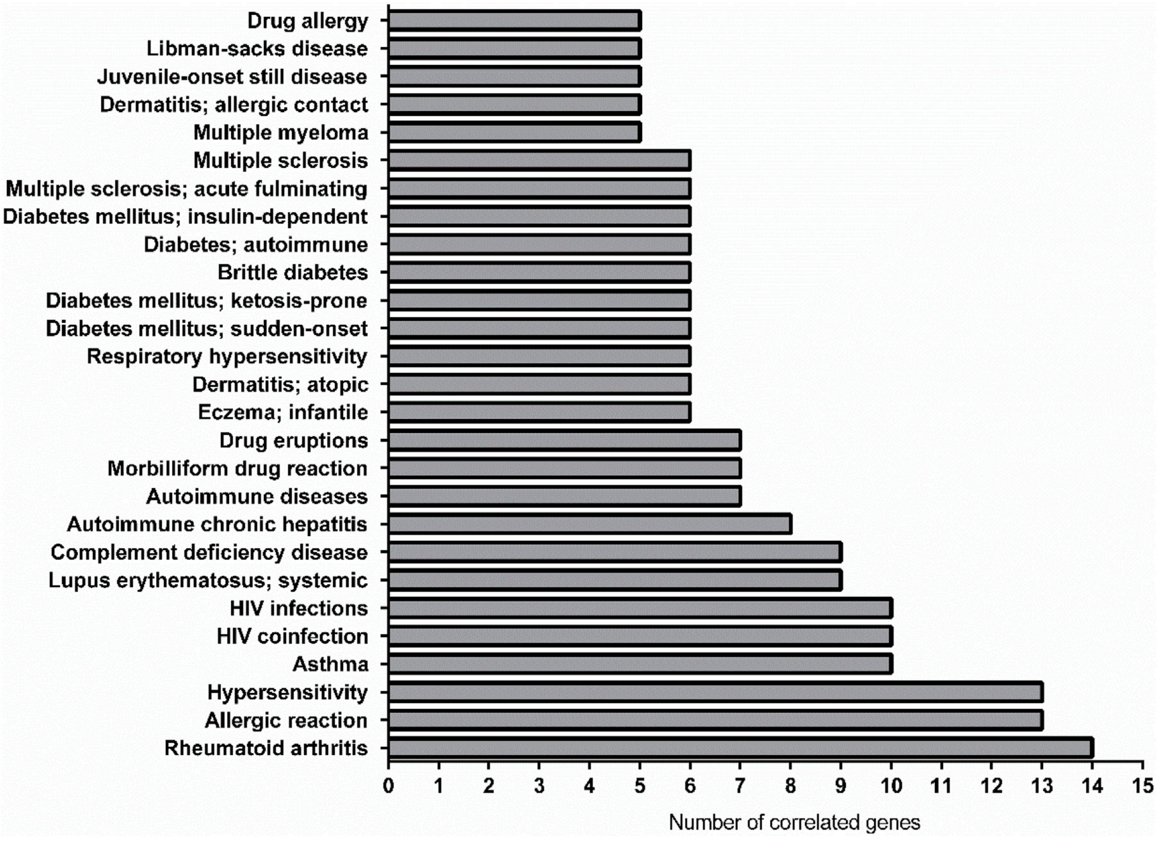


FIGURE 5 | Graphical representation of the most linked immune system diseases with the gene list in CURDi. The highest number of correlated genes are found in rheumatoid arthritis (14 linked genes), allergic reaction and hypersensitivity (13 linked genes) and asthma (10 linked genes).

has revealed granulocyte colony-stimulating factor (G-CSF) as a predictive biomarker of oocyte and embryo developmental competence in humans (Naghshineh et al., 2018; Cai et al., 2020), promoting endometrial thickening and improving the pathophysiology of endometriosis, which all fundamentally lead to preventing from the pregnancy loss (Cai et al., 2020).

3.3 CURDi Network and the Most Linked Immune Diseases

Analyzing the link of the selected gene set with diseases involving other organs and systems (different from the Immune System), we found that the largest number of pathologies were related to the following groups: "Hemic and Lymphatic diseases" (Chinniah and Katelaris, 2009), "Skin and Connective Tissues diseases" (Emmery et al., 2016) and "Neoplasms" (Nielsen et al., 2017) (**Table 1**; **Supplementary Material S3**, second sheet for the complete dataset).

As showed in **Figure 5** the most linked diseases to the list of genes from CURDi were rheumatoid arthritis (14 linked genes), allergic reaction and hypersensitivity (13 linked genes) and asthma (10 linked genes) (for the complete list of diseases and related information see **Supplementary Material S3**, sheet 3).

3.3.1 Rheumatoid Arthritis

Among the three most linked conditions, only rheumatoid arthritis (RA) has been related to fertility (Fattah et al., 2020) so far, maybe because the other two (i.e., allergic reaction and hypersensibility) show very high variability and multiple interconnected components. A recent review by Fattah and coll. (2020) provided several proofs regarding the relationship between women with RA and fertility, which seems declined and dependent on inflammatory milieu, mother age, hampered sexual activity and negative effects of non-steroidal anti-inflammatory drugs on ovarian function (Fattah et al., 2020). Indeed, it has been found that women with RA deliver fewer children when compared to healthy women (Fattah et al., 2020). The decreased fertility rate in women suffering from RA might be due to a reduced sexual activity (because of pain, fatigue, mental distress, functional limitations), treatment with antirheumatic medications hampering ovulation, as well as, to advanced maternal age, patients' choice, or a combination of all of these factors (Fattah et al., 2020). The results showed here demonstrate that at least 13 genes (CXCL8; CSF2; IL6; LCN2; TNF; VEGFA; IFNG; IL1B; IL10; CP; CXCL2; GC; F1) could be involved in this relationship.

3.3.2 Asthma

From the CURDi analysis, asthma showed 10 linked genes. The link between asthma and infertility was studied in a nationwide register-based twin study, in which a cohort of 15,250 twins living in Denmark participated in a questionnaire study including questions about the presence of asthma and fertility (Gade et al., 2014). Differences in time to pregnancy and pregnancy outcome were analysed in subjects affected with asthma and allergy and in healthy individuals, using multiple regression analysis. Results showed an association between asthma and an increased time to pregnancy, with a percentage of asthmatics with a time to pregnancy >1 year of 27% versus the 21.6% for the non-asthmatic individuals. Interestingly, the association remained significant after adjustment for age, age at menarche, body mass index and socioeconomic status and was more pronounced in those >30 years of age. In addition, untreated asthmatics had a significant increased risk of prolonged time to pregnancy compared to control individuals, while asthmatics receiving any kind of treatment for asthma tended to have a shorter time to pregnancy than untreated asthmatics (Gade et al., 2014). Thus, the authors concluded that asthma seems to be correlated with an alteration in fertility parameters, and that the negative effect of asthma on fertility increases with age and disease severity.

3.4 AMDi Network and the Most Linked Genes

Since the study of human diseases takes a huge advantage by the use of animal models as valuable resource for the investigation of pathogenesis, diagnostics, and therapeutics of human diseases, we realized the network representing the connections between the selected gene set and the immune diseases in animal models (AMDi). The most linked genes were IL4, TNF and CCL2, (12, 12 and 10 links, respectively) (see **Supplementary Material S4**, sheet 1 for the complete list).

The roles of IL4 and TNF, have been discussed before. The CCL2 gene codifies for the small chemokine CCL2, also referred to as monocyte chemotactic protein 1 (MCP1), which is secreted by endothelial, epithelial and stromal cells, monocytes and lymphocytes (Hess et al., 2013). It influences the innate immunity through its effects on monocytes, as well as the adaptive immunity through the control of T helper cell polarization (Hess et al., 2013). It was proposed that chemokines expressed by the oviductal epithelial cells contribute to normal physiological homoeostasis and protection from pathogens by activating the immune cells (Fahey et al., 2005). In addition to this protective function, chemokines, including CCL2, may protect these cells from malignant transformation, again suggesting that CCL2 may be involved in early tumour development (Wojnarowicz et al., 2012). It was also shown that a marked down-regulation of CCL2 may contribute to allogenic tolerance of the preimplantation embryo as it crosses the Fallopian tube (Hess et al., 2013).

Interestingly, an association between two CCL2 polymorphisms (rs1024611 and rs4586) and the development of gestational diabetes mellitus (GDM), the most common medical complication of human pregnancy, was demonstrated in 411 pregnant women (Teler et al., 2017). To this regard, a more recent study confirmed that blocking the CCL2/CCR2 pathway in a mouse GDM model, the inflammatory cytokines may be reduced, mitigating GDM symptoms and improving the reproductive outcomes in mice (Qi et al., 2021).

3.5 AMDi Network and the Most Linked Diseases

The AMDi network also provided very intriguing and useful information. For instance, the two pathologies related with the highest number of correlated genes are the Experimental Autoimmune Encephalomyelitis (EAE, 17 genes) and Asthma

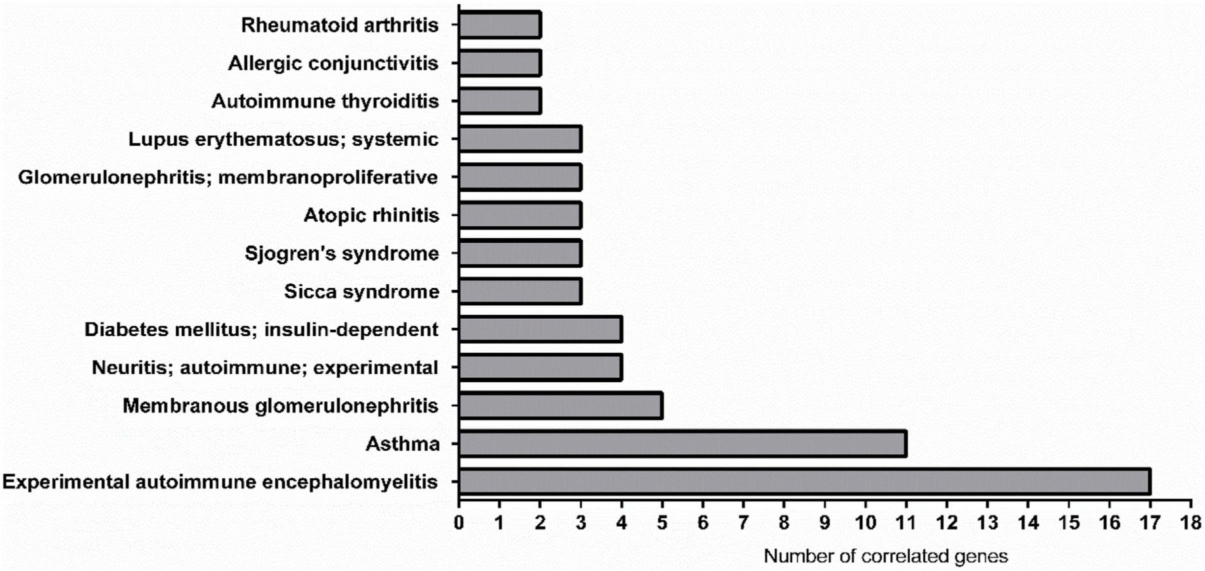


FIGURE 6 | Graphical representation of the most linked immune system diseases with the genes list in AMDi. The most linked diseases were experimental autoimmune encephalomyelitis (EAE, linked 17 genes) and asthma (linked 11 genes).

(11 genes), this last being already discussed above (for the complete list of diseases and related information, see **Figure 6**; **Supplementary Material S4**, second sheet).

3.5.1 Experimental Autoimmune Encephalomyelitis

EAE is an autoimmune encephalomyelitis commonly used as an experimental model for the human inflammatory demyelinating disease, multiple sclerosis (MS). It constitutes a complex condition in which the interaction between a variety of immunopathological and neuropathological mechanisms leads to the key pathological features of MS: inflammation, demyelination, axonal loss and gliosis (Constantinescu et al., 2011).

The exploration of the link between MS and infertility is very complex for several reasons. As discussed by Cavalla and coll. (2006), the frequency of childlessness in the female MS patients seems to be higher than in the general population (Cavalla et al., 2006). Rather than lowered fertility, this could reflect other issues related to this pathology, such as the fact that patients may choose to avoid or postpone pregnancy, mainly because of concern about taking care of the baby or about the risk of transmitting a genetic susceptibility to MS to their children (Cavalla et al., 2006). A recent study has shown that women affected with MS had lower live birth rates (LBR) compared to unaffected women (irrespective of their infertility diagnosis or treatment) (Houtchens et al., 2020). This statistically significant difference in LBRs was more evident in women in early (Bauer-Mehren et al., 2011; Profaizer and Eckels, 2012; Leone et al., 2013; Chattopadhyay et al., 2014; Mosaad, 2015; Pinero et al., 20152015; Robinson et al., 2015; Piñero et al., 2017; Wieczorek et al., 2017; Jongsma et al., 2019; Piñero et al., 2020; Tersigni et al., 2020; Wilczyńska et al., 2020) and middle (Emmery et al., 2016; Nielsen et al., 2017; Agius et al., 2018; Papúchová et al., 2019) childbearing years. The difference between women with and without MS disappeared after receiving infertility treatments, thus highlighting the importance of information regarding the efficacy of infertility treatments in women with autoimmune diseases (Houtchens et al., 2020).

Despite the fact that MS is three times more common in women than in men and that endocrine alteration commonly found in MS patients and immunosuppressive therapies could interfere with fertility, Glazer and co-workers evaluated the association of MS and male infertility in a register-based cohort study in Denmark between 1994 and 2015 (Glazer et al., 2017). A comparison was made between a group of 24,011 men diagnosed with male factor infertility and a control group of 27,052 normal males. Infertile men showed a higher risk of prevalent and incident MS when compared to the reference group, thus suggesting, for the first time, an association between male infertility and MS (Glazer et al., 2017).

Here we provided the evidence that in both EAE and asthma a common genetic background could explain, at least in part, the finding that a systemic inflammation can also involve the reproductive system.

From the results obtained in this study, we highlighted as immune system and reproductive function are closely linked. Indeed, as it was shown, some proteins involved in sperm-oviduct interaction could be involved in several immune system diseases, while, at the same time, some immune system diseases could interfere with the reproduction process, although their causal relationship is still unclear.

However, to better understand the cross-talk between the immune and the reproductive systems are needed further investigations, such as wider epidemiological studies and experimental research with the use of animal models.

In conclusion, our innovative approach fits well in the field of "reproductive immunology" that represents an active area of research aimed at understanding how the immune system contributes to human reproduction. In a clinical research scenario this comprehension might be fundamental in reducing implantation failure and recurrent miscarriage in assisted reproductive technologies (ARTs).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AT, CC, and NB conceptualized the work and performed the networks; all Authors curated the data and performed the formal analysis; AT, CC, and NB prepared the original draft; all Authors reviewed and edited the original draft and contributed to the writing; AT, CC, LV, and MR-S prepared the figures; NB and BB supervised and funded the research. All Authors reviewed and approved the publication of this manuscript in its current form.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.795123/full#supplementary-material

REFERENCES

Agius, A., Sultana, R., Camenzuli, C., Calleja-Agius, J., and Balzan, R. (2018). An Update on the Genetics of Pre-eclampsia. *Minerva Obstet. Gynecol.* 70, 465–479. doi:10.23736/S0026-4784.17.04150-8

Almiñana, C. (2015). Snooping on a Private Conversation between the Oviduct and Gametes/Embryos. *Anim. Reprod.* 12, 366–374.

Avilés, M., Gutiérrez-Adán, A., and Coy, P. (2010). Oviductal Secretions: Will They Be Key Factors for the Future ARTs? *Mol. Hum. Reprod.* 16, 896–906. doi:10.1093/molehr/gaq056

Balasubramaniam, E. S., Van Noorden, S., and El-Bahrawy, M. (2012). The Expression of Interleukin (IL)-6, IL-8, and Their Receptors in Fallopian Tubes with Ectopic Tubal Gestation. *Fertil. Sterility* 98, 898–904. doi:10.1016/j.fertnstert.2012.06.004

Ballester, L., Romero-Aguirregomezcorta, J., Soriano-Úbeda, C., Matás, C., Romar, R., and Coy, P. (2014). Timing of Oviductal Fluid Collection, Steroid Concentrations, and Sperm Preservation Method Affect Porcine *In Vitro* Fertilization Efficiency. *Fertil. Sterility* 102, 1762–1768.e1. doi:10.1016/j.fertnstert.2014.08.009

Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M. A., Sanz, F., and Furlong, L. I. (2011). Gene-disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases. *PLoS One* 6, e20284. doi:10.1371/journal.pone.0020284

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). DisGeNET: A Cytoscape Plugin to Visualize, Integrate, Search and Analyze Gene-Disease Networks. *Bioinformatics* 26, 2924–2926. doi:10.1093/bioinformatics/btq538

Benny, P. A., Alakwaa, F. M., Schlueter, R. J., Lassiter, C. B., and Garmire, L. X. (2020). A Review of Omics Approaches to Study Preeclampsia. *Placenta* 92, 17–27. doi:10.1016/J.PLACENTA.2020.01.008

Bhat, M. Y., Solanki, H. S., Advani, J., Khan, A. A., Keshava Prasad, T. S., Gowda, H., et al. (2018). Comprehensive Network Map of Interferon Gamma Signaling. *J. Cell Commun. Signal.* 12, 745–751. doi:10.1007/s12079-018-0486-y

Biggar, R. J., Poulsen, G., Ng, J., Melbye, M., and Boyd, H. A. (2010). HLA Antigen Sharing between Mother and Fetus as a Risk Factor for Eclampsia and Preeclampsia. *Hum. Immunol.* 71, 263–267. doi:10.1016/J.HUMIMM.2010.01.006

Bohring, C., and Krause, W. (2003). Immune Infertility: Towards a Better Understanding of Sperm (Auto)-immunity: The Value of Proteomic Analysis. *Hum. Reprod.* 18, 915–924. doi:10.1093/humrep/deg207

Brazdova, A., Senechal, H., Peltre, G., and Poncet, P. (2016). Immune Aspects of Female Infertility. *Int. J. Fertil. Steril* 10, 1–10. doi:10.22074/ijfs.2016.4762

Cai, L., Jeong, Y.-w., Jin, Y.-x., Lee, J.-y., Jeong, Y.-i., Hwang, K.-c., et al. (2020). Effects of Human Recombinant Granulocyte-colony Stimulating Factor Treatment during *In Vitro* Culture on Porcine Pre-implantation Embryos. *PLoS One* 15, e0230247. doi:10.1371/journal.pone.0230247

Canha-Gouveia, A., Paradela, A., Ramos-Fernández, A., Prieto-Sánchez, M. T., Sánchez-Ferrer, M. L., Corrales, F., et al. (2019). Which Low-Abundance Proteins Are Present in the Human Milieu of Gamete/embryo Maternal Interaction? *Ijms* 20, 5305. doi:10.3390/ijms20215305

Carrasquel, G., Camejo, M. I., Michelangeli, F., and Ruiz, M. C. (2014). IFN-gamma Alters the Human Sperm Membrane Permeability to Ca2+. *Syst. Biol. Reprod. Med.* 60, 21–27. doi:10.3109/19396368.2013.833658

Cavalla, P., Rovei, V., Masera, S., Vercellino, M., Massobrio, M., Mutani, R., et al. (2006). Fertility in Patients with Multiple Sclerosis: Current Knowledge and Future Perspectives. *Neurol. Sci.* 27, 231–239. doi:10.1007/s10072-006-0676-x

Chatterjee, P., Chiasson, V. L., Bounds, K. R., and Mitchell, B. M. (2014). Regulation of the Anti-inflammatory Cytokines Interleukin-4 and Interleukin-10 during Pregnancy. *Front. Immunol.* 5, 1. doi:10.3389/fimmu.2014.00253

Chattopadhyay, S., Chakraborty, N., Chattopadhyay, S., Pratheek, B., Nayak, T., Sahoo, S., et al. (2014). Mammalian Non-classical Major Histocompatibility Complex I and its Receptors: Important Contexts of Gene, Evolution, and Immunity. *Indian J. Hum. Genet.* 20, 129–141. doi:10.4103/0971-6866.142855

Chinniah, N., and Katelaris, C. H. (2009). Hereditary Angioedema and Pregnancy. *Aust. New Zeal J. Obstet. Gynaecol.* 49, 2–5. doi:10.1111/J.1479-828X.2008.00945.X

Constantinescu, C. S., Farooqi, N., O'Brien, K., and Gran, B. (2011). Experimental Autoimmune Encephalomyelitis (EAE) as a Model for Multiple Sclerosis (MS). *Br. J. Pharmacol.* 164, 1079–1106. doi:10.1111/j.1476-5381.2011.01302.x

Coy, P., García-Vázquez, F. A., Visconti, P. E., and Avilés, M. (2012). Roles of the Oviduct in Mammalian Fertilization. *Reproduction* 144, 649–660. doi:10.1530/REP-12-0279

Coy, P., and Yanagimachi, R. (2015). The Common and Species-specific Roles of Oviductal Proteins in Mammalian Fertilization and Embryo Development. *Bioscience* 65, 973–984. doi:10.1093/biosci/biv119

Eggert-Kruse, W., Kiefer, I., Beck, C., Demirakca, T., and Strowitzki, T. (2007). Role for Tumor Necrosis Factor Alpha (TNF-α) and Interleukin 1-beta (IL-1β) Determination in Seminal Plasma during Infertility Investigation. *Fertil. Sterility* 87, 810–823. doi:10.1016/j.fertnstert.2006.08.103

Emmery, J., Hachmon, R., Pyo, C. W., Nelson, W. C., Geraghty, D. E., Andersen, A. M. N., et al. (2016). Maternal and Fetal Human Leukocyte Antigen Class Ia and II Alleles in Severe Preeclampsia and Eclampsia. *Genes Immun.* 17, 251–260. doi:10.1038/gene.2016.20

Fahey, J. V., Schaefer, T. M., Channon, J. Y., and Wira, C. R. (2005). Secretion of Cytokines and Chemokines by Polarized Human Epithelial Cells from the Female Reproductive Tract. *Hum. Reprod.* 20, 1439–1446. doi:10.1093/humrep/deh806

Fattah, A., Asadi, A., Shayesteh, M. R. H., Hesari, F. H., Jamalzehi, S., Abbasi, M., et al. (2020). Fertility and Infertility Implications in Rheumatoid Arthritis; State of the Art. *Inflamm. Res.* 69, 721–729. doi:10.1007/s00011-020-01362-w

Gade, E. J., Thomsen, S. F., Lindenberg, S., Kyvik, K. O., Lieberoth, S., and Backer, V. (2014). Asthma Affects Time to Pregnancy and Fertility: A Register-Based Twin Study. *Eur. Respir. J.* 43, 1077–1085. doi:10.1183/09031936.00148713

Gadella, B. M. (2017). Reproductive Tract Modifications of the Boar Sperm Surface. *Mol. Reprod. Dev.* 84, 822–831. doi:10.1002/mrd.22821

Georgiou, A. S., Snijders, A. P. L., Sostaric, E., Aflatoonian, R., Vazquez, J. L., Vazquez, J. M., et al. (2007). Modulation of the Oviductal Environment by Gametes. *J. Proteome Res.* 6, 4656–4666. doi:10.1021/pr070349m

Glazer, C. H., Tøttenborg, S. S., Giwercman, A., Bräuner, E. V., Eisenberg, M. L., Vassard, D., et al. (2017). Male Factor Infertility and Risk of Multiple Sclerosis: A Register-Based Cohort Study. *Mult. Scler. MSJ J.* 9, 259–261. doi:10.1177/1352458517734069

Granot, I., Gnainsky, Y., and Dekel, N. (2012). Endometrial Inflammation and Effect on Implantation Improvement and Pregnancy Outcome. *Reproduction* 144, 661–668. doi:10.1530/REP-12-0217

Hess, A. P., Talbi, S., Hamilton, A. E., Baston-Buest, D. M., Nyegaard, M., Irwin, J. C., et al. (2013). The Human Oviduct Transcriptome Reveals an Anti-inflammatory, Anti-angiogenic, Secretory and Matrix-Stable Environment during Embryo Transit. *Reprod. BioMedicine Online* 27, 423–435. doi:10.1016/j.rbmo.2013.06.013

Houtchens, M. K., Edwards, N. C., Hayward, B., Mahony, M. C., and Phillips, A. L. (2020). Live Birth Rates, Infertility Diagnosis, and Infertility Treatment in Women with and without Multiple Sclerosis: Data from an Administrative Claims Database. *Mult. Scler. Relat. Disord.* 46, 102541. doi:10.1016/j.msard.2020.102541

Jongsma, M. L. M., Guarda, G., and Spaapen, R. M. (2019). The Regulatory Network behind MHC Class I Expression. *Mol. Immunol.* 113, 16–21. doi:10.1016/j.molimm.2017.12.005

Kak, G., Raza, M., and Tiwari, B. K. (2018). Interferon-gamma (IFN-γ): Exploring its Implications in Infectious Diseases. *Biomol. Concepts* 9, 64–79. doi:10.1515/bmc-2018-0007

Laflamme, J., Akoum, A., and Leclerc, P. (2005). Induction of Human Sperm Capacitation and Protein Tyrosine Phosphorylation by Endometrial Cells and Interleukin-6. *Mol. Hum. Reprod.* 11, 141–150. doi:10.1093/molehr/gah142

Leone, P., Shin, E.-C., Perosa, F., Vacca, A., Dammacco, F., and Racanelli, V. (2013). MHC Class I Antigen Processing and Presenting Machinery: Organization, Function, and Defects in Tumor Cells. *JNCI J. Natl. Cancer Inst.* 105, 1172–1187. doi:10.1093/jnci/djt184

Li, S., and Winuthayanon, W. (2017). Oviduct: Roles in Fertilization and Early Embryo Development. *J. Endocrinol.* 232, R1–R26. doi:10.1530/JOE-16-0302

López-Lera, A., Pernia, O., López-Trascasa, M., and Ibanez De Caceres, I. (2014). Expression of the SERPING1 Gene Is Not Regulated by Promoter Hypermethylation in Peripheral Blood Mononuclear Cells from Patients with Hereditary Angioedema Due to C1-Inhibitor Deficiency. *Orphanet J. Rare Dis.* 9, 1–5. doi:10.1186/s13023-014-0103-y

Loureiro, B., Bonilla, L., Block, J., Fear, J. M., Bonilla, A. Q. S., and Hansen, P. J. (2009). Colony-stimulating Factor 2 (CSF-2) Improves Development and Posttransfer Survival of Bovine Embryos Produced *In Vitro*. *Endocrinology* 150, 5046–5054. doi:10.1210/en.2009-0481

Madsen, D. E., Hansen, S., Gram, J., Bygum, A., Drouet, C., and Sidelmann, J. J. (2014). Presence of C1-Inhibitor Polymers in a Subset of Patients Suffering from Hereditary Angioedema. *PLoS One* 9, e112051–7. doi:10.1371/journal.pone.0112051

Moreau, P., Contu, L., Alba, F., Lai, S., Simoes, R., Orrù, S., et al. (2008). HLA-G Gene Polymorphism in Human Placentas: Possible Association of G*0106 Allele with Preeclampsia and Miscarriage. *Biol. Reprod.* 79, 459–467. doi:10.1095/BIOLREPROD.108.068874

Mosaad, Y. M. (2015). Clinical Role of Human Leukocyte Antigen in Health and Disease. *Scand. J. Immunol.* 82, 283–306. doi:10.1111/sji.12329

Murphy, S. P., Tayade, C., Ashkar, A. A., Hatta, K., Zhang, J., and Croy, B. A. (2009). Interferon Gamma in Successful Pregnancies1. *Biol. Reprod.* 80, 848–859. doi:10.1095/biolreprod.108.073353

Naghshineh, E., Eftekhar, M., and Khani, P. (2018). Role of Granulocyte colony-stimulating Factor in Human Reproduction. *J. Res. Med. Sci.* 23, 7. doi:10.4103/jrms.JRMS_628_17

Nasu, K., Itoh, H., Yuge, A., Nishida, M., Kawano, Y., and Narahara, H. (2007). Tumor Necrosis Factor-α Regulates Vascular Endothelial Growth Factor Secretion by Human Oviductal Epithelial Cells and Stromal Fibroblasts. *Fertil. Sterility* 87, 220–222. doi:10.1016/j.fertnstert.2006.05.082

Nielsen, H. S., and Hviid, T. V. (2017). "HLA Associations and Recurrent Pregnancy Loss," in *Early Pregnancy*. Editors R. G. Farquharson and M. D. Stephenson (Cambridge, UK: Cambridge Press).

Papathanasiou, A., Djahanbakhch, O., Saridogan, E., and Lyons, R. A. (2008). The Effect of Interleukin-6 on Ciliary Beat Frequency in the Human Fallopian Tube. *Fertil. Sterility* 90, 391–394. doi:10.1016/j.fertnstert.2007.07.1379

Papúchová, H., Meissner, T. B., Li, Q., Strominger, J. L., and Tilburgs, T. (2019). The Dual Role of HLA-C in Tolerance and Immunity at the Maternal-Fetal Interface. *Front. Immunol.* 10, 2730. doi:10.3389/fimmu.2019.02730

Parada-Bustamante, A., Oróstica, M. L., Reuquen, P., Zuñiga, L. M., Cardenas, H., and Orihuela, P. A. (2016). The Role of Mating in Oviduct Biology. *Mol. Reprod. Dev.* 83, 875–883. doi:10.1002/mrd.22674

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., and Bagos, P. G. (2018). Bipartite Graphs in Systems Biology and Medicine: A Survey of Methods and Applications. *Gigascience* 7, 1–31. doi:10.1093/gigascience/giy014

Persson, G., Melsted, W. N., Nilsson, L. L., and Hviid, T. V. F. (2017). HLA Class Ib in Pregnancy and Pregnancy-Related Disorders. *Immunogenetics* 69, 581–595. doi:10.1007/S00251-017-0988-4

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. *Nucleic Acids Res.* 45, D833–D839. doi:10.1093/nar/gkw943

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (20152015). DisGeNET: a Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. *Database* 2015, bav028. doi:10.1093/database/bav028

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Pinto-Bravo, P., Galvão, A., Rebordão, M. R., Amaral, A., Ramilo, D., Silva, E., et al. (2017). Ovarian Steroids, Oxytocin, and Tumor Necrosis Factor Modulate Equine Oviduct Function. *Domest. Anim. Endocrinol.* 61, 84–99. doi:10.1016/j.domaniend.2017.06.005

Profaizer, T., and Eckels, D. (2012). HLA Alleles and Drug Hypersensitivity Reactions. *Int. J. Immunogenet.* 39, 99–105. doi:10.1111/j.1744-313X.2011.01061.x

Qi, X., Xing, Y., and Wang, X. (2021). Blockade of CCL2/CCR2 Signaling Pathway Exerts Anti-inflammatory Effects and Attenuates Gestational Diabetes Mellitus in a Genetic Mice Model. *Horm. Metab. Res.* 53, 56–62. doi:10.1055/A-1250-8221

Robertson, S. A., Care, A. S., and Moldenhauer, L. M. (2018). Regulatory T Cells in Embryo Implantation and the Immune Response to Pregnancy. *J. Clin. Invest.* 128, 4224–4235. doi:10.1172/JCI122182

Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2015). The IPD and IMGT/HLA Database: Allele Variant Databases. *Nucleic Acids Res.* 43, D423–D431. doi:10.1093/nar/gku1161

Sabetian, S., Shamsir, M. S., and Naser, M. A. (2014). Functional Features and Protein Network of Human Sperm-Egg Interaction. *Syst. Biol. Reprod. Med.* 60, 329–337. doi:10.3109/19396368.2014.955896

Saint-Dizier, M., Schoen, J., Chen, S., Banliat, C., and Mermillod, P. (2020). Composing the Early Embryonic Microenvironment: Physiology and Regulation of Oviductal Secretions. *Ijms* 21, 1–21. doi:10.3390/ijms21010223

Santacroce, R., D'Andrea, G., Maffione, A. B., Margaglione, M., and d'Apolito, M. (2021). Clinical Medicine the Genetics of Hereditary Angioedema: A Review. *Jcm* 10, 2023. doi:10.3390/jcm10092023

Sargent, I. L., Borzychowski, A. M., and Redman, C. W. G. (2006). NK Cells and Human Pregnancy - an Inflammatory View. *Trends Immunol.* 27, 399–404. doi:10.1016/J.IT.2006.06.009

Schroder, K., Hertzog, P. J., Ravasi, T., and Hume, D. A. (2004). Interferon-γ: an Overview of Signals, Mechanisms and Functions. *J. Leukoc. Biol.* 75, 163–189. doi:10.1189/jlb.0603252

Sferruzzi-Perri, A. N., Macpherson, A. M., Roberts, C. T., and Robertson, S. A. (2009). Csf2 Null Mutation Alters Placental Gene Expression and Trophoblast

Glycogen Cell and Giant Cell Abundance in Mice1. *Biol. Reprod.* 81, 207–221. doi:10.1095/biolreprod.108.073312

Shahbazi, M., Ehsani, M., Mohammadnia-Afrouzi, M., Mirzakhani, M., and Esmaeilzadeh, S. (2019). Female Unexplained Infertility: A Disease with Imbalanced Adaptive Immunity. *J. Hum. Reprod. Sci.* 12, 274–282. doi:10.4103/jhrs.JHRS_30_19

Suarez, S. S. (2016). Mammalian Sperm Interactions with the Female Reproductive Tract. *Cell Tissue Res* 363, 185–194. doi:10.1007/S00441-015-2244-2

Suarez, S. S., and Pacey, A. A. (2006). Sperm Transport in the Female Reproductive Tract. *Hum. Reprod. Update* 12, 23–37. doi:10.1093/humupd/dmi047

Tan, C. Y., Ho, J. F. V., Chong, Y. S., Loganath, A., Chan, Y. H., Ravichandran, J., et al. (2008). Paternal Contribution of HLA-G*0106 Significantly Increases Risk for Pre-eclampsia in Multigravid Pregnancies. *Mol. Hum. Reprod.* 14, 317–324. doi:10.1093/MOLEHR/GAN013

Teler, J., Tarnowski, M., Safranow, K., Maciejewska, A., Sawczuk, M., Dziedziejko, V., et al. (2017). CCL2, CCL5, IL4 and IL15 Gene Polymorphisms in Women with Gestational Diabetes Mellitus. *Horm. Metab. Res.* 49, 10–15. doi:10.1055/S-0042-111436

Tersigni, C., Meli, F., Neri, C., Iacoangeli, A., Franco, R., Lanzone, A., et al. (2020). Role of Human Leukocyte Antigens at the Feto-Maternal Interface in normal and Pathological Pregnancy: An Update. *Ijms* 21, 1–13. doi:10.3390/ijms21134756

Vilanova, L. T., Rauch, M. C., Mansilla, A., Zambrano, A., Brito, M., Werner, E., et al. (2003). Expression of Granulocyte-Macrophage colony Stimulating Factor (GM-CSF) in Male Germ Cells: GM-CSF Enhances Sperm Motility. *Theriogenology* 60, 1083–1095. doi:10.1016/S0093-691X(03)00106-7

Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., et al. (2017). Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front. Immunol.* 8, 1–16. doi:10.3389/fimmu.2017.00292

Wijayagunawardane, M. P. B., Gabler, C., Killian, G., and Miyamoto, A. (2003). Tumor Necrosis Factor in the Bovine Oviduct during the Estrous Cycle: Messenger RNA Expression and Effect on Secretion of Prostaglandins, Endothelin-1, and Angiotensin II. *Biol. Reprod.* 69, 1341–1346. doi:10.1095/biolreprod.103.017327

Wiktor, H., and Kozioł, P. (1998). Histocompatibility Antigens in Pregnant Women with Preeclampsia and in Their Husbands. *Ginekol Pol.* 69, 937–942.

Wilczyńska, K., Radwan, P., Krasiński, R., Radwan, M., Wilczyński, J. R., Malinowski, A., et al. (2020). KIR and HLA-C Genes in Male Infertility. *J. Assist. Reprod. Genet.* 37, 2007–2017. doi:10.1007/s10815-020-01814-6

Wojnarowicz, P., Gambaro, K., De Ladurantaye, M., Quinn, M. C. J., Provencher, D., Mes-Masson, A.-M., et al. (2012). Overexpressing the CCL2 Chemokine in an Epithelial Ovarian Cancer Cell Line Results in Latency of *In Vivo* Tumourigenicity. *Oncogenesis* 1, e27. doi:10.1038/ONCSIS.2012.25

Yakaboski, E., Motazedi, T., and Banerji, A. (2020). Hereditary Angioedema: Special Considerations in Women. *Allergy Asthma Proc.* 41, S47–S50. doi:10.2500/AAP.2020.41.200077

Yang, Y., Su, X., Xu, W., and Zhou, R. (2014). Interleukin-18 and Interferon Gamma Levels in Preeclampsia: A Systematic Review and Meta-Analysis. *Am. J. Reprod. Immunol.* 72, 504–514. doi:10.1111/AJI.12298

Yildizfer, F., Donma, O., Yen, M., Ekmekci, O., Karatas Kul, Z. A., Keser, Z., et al. (2015). *In Vitro* fertilization, Levels of Pro-inflammatory Factors and Lipid Peroxidation. *Int. J. Fertil. Steril* 9, 277–284. doi:10.22074/ijfs.2015.4541

Zandieh, Z., Ashrafi, M., Jameie, B., Amanpour, S., Mosaffa, N., Salman Yazdi, R., et al. (2015). Evaluation of Immunological Interaction between Spermatozoa and Fallopian Tube Epithelial Cells. *Andrologia* 47, 1120–1130. doi:10.1111/and.12391

Zhang, Z., Jia, L. T., and Lin, Z. L. (2009). Polymorphism of HLA-A and HLA-B in Pre-eclampsia. *Beijing Da Xue Xue Bao Yi Xue Ban -Journal Peking Univ. Heal Sci.* 41, 418–425.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2022
doi: 10.3389/fgene.2021.814073

Predicting lncRNA–Protein Interactions by Heterogenous Network Embedding

Guoqing Zhao[1], Pengpai Li[1], Xu Qiao[1], Xianhua Han[2] and Zhi-Ping Liu[1]*

[1]Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, China, [2]Faculty of Science, Yamaguchi University, Yamaguchi, Japan

lncRNA–protein interactions play essential roles in a variety of cellular processes. However, the experimental methods for systematically mapping of lncRNA–protein interactions remain time-consuming and expensive. Therefore, it is urgent to develop reliable computational methods for predicting lncRNA–protein interactions. In this study, we propose a computational method called LncPNet to predict potential lncRNA–protein interactions by embedding an lncRNA–protein heterogenous network. The experimental results indicate that LncPNet achieves promising performance on benchmark datasets extracted from the NPInter database with an accuracy of 0.930 and area under ROC curve (AUC) of 0.971. In addition, we further compare our method with other eight state-of-the-art methods, and the results illustrate that our method achieves superior prediction performance. LncPNet provides an effective method via a new perspective of representing lncRNA–protein heterogenous network, which will greatly benefit the prediction of lncRNA–protein interactions.

Keywords: lncRNA–protein interaction, computational method, heterogenous network, network embedding, LncPNet

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Lei Deng,
Central South University, China
Wei Lan,
Guangxi University, China

****Correspondence:***
Zhi-Ping Liu
zpliu@sdu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *12 November 2021*
Accepted: *27 December 2021*
Published: *04 February 2022*

Citation:
Zhao G, Li P, Qiao X, Han X and Liu Z-P (2022) Predicting lncRNA–Protein Interactions by Heterogenous Network Embedding. Front. Genet. 12:814073. doi: 10.3389/fgene.2021.814073

1 INTRODUCTION

The non-coding RNA (ncRNA) plays important roles in biological processes, which can influence human health on various levels (Louro et al., 2009). Existing studies have shown that less than 2% of the human genome can be translated into proteins; while, over 80% of the genome has biochemical functions (Djebali et al., 2012). In addition, over 70% of ncRNAs are lncRNAs (Yang et al., 2014). It is demonstrated that lncRNAs play crucial roles in transcription, splicing gene expression (Ponting et al., 2009; Guttman and Rinn, 2012; Qu and Adelson, 2012; Zhu et al., 2013), and have a close relationship with complex diseases (Mercer et al., 2009; Yang et al., 2015). Therefore, lncRNA is of great importance for understanding the mechanisms of biological processes.

Most of the functions of lncRNA are still unknown. One of the mechanisms is lncRNAs usually function by binding to chaperone proteins (Mercer et al., 2009). Hence, the basis for understanding the functions of lncRNAs is to recognize the interactions between lncRNAs and proteins, which can help understand the mechanism of physiological processes. Experimental methods for identifying protein–RNA interactions include ChiRP, CHART, RIP, RIP-ChIP/Seq, and CLIP (Yang et al., 2015). Since these experimental methods are often time-consuming and expensive, an effective computational method is an alternative way for expanding our knowledge of lncRNA–protein interactions (Liu, 2021).

In recent years, some methods for predicting lncRNA–protein interactions have been developed. Muppirala et al. applied random forest (RF) (Breiman, 2001) and support vector machines (SVMs)

(Joachims, 1998) to classify an interaction only via the sequence information of lncRNA and protein (Muppirala et al., 2011). Lncpro was developed for predicting lncRNA–protein associations (Lu et al., 2013) by three types of features based on the Fisher linear discriminant approach, including classical protein secondary structures and hydrogen-bond and van der Waals propensities as well as six types of RNA secondary structures. In 2016, IPMiner was proposed to predict lncRNA–protein interactions from sequences, which employed deep learning and further improved the performance using stacked integration (Pan et al., 2016). Hu et al. introduced a method named HLPI-Ensemble specifically for human lncRNA–protein interactions (Hu et al., 2018). HLPI-Ensemble adopts three methods to extract the features of lncRNA and protein from sequences based on three mainstream machine learning algorithms of SVM, RF, and extreme gradient boosting (XGB) (Chen and Guestrin, 2016). Suresh et al. proposed an approach based on SVM classifiers by integrating sequence and structure features of the lncRNA and protein (Suresh et al., 2015). Zhang et al. combined multiple sequence-based features, lncRNA–lncRNA similarity and protein–protein similarity, and predicted lncRNA–protein interactions by RNA sequences and protein sequences as well as known lncRNA–protein interactions (Zhang et al., 2018b). Li et al. proposed a network-based computational method, which used a random walk with restart based on heterogenous network model (i.e., LPIHN), to infer the lncRNA–protein interactions (Li et al., 2015). Although LPIHN employs the method of network embedding, it does not consider the type of node. Moreover, these ordinary random walks cannot well retain the local and global information of the node from the network. LPLNP was developed for calculating the linear neighborhood similarity in the feature space and transferring it into the interaction space to predict unobserved interactions by a label propagation process (Zhang et al., 2018a). Yi et al. introduced a stacking ensemble-based computational model to predict lncRNA–protein interactions, called RPI-SE, which integrated XGB, SVM, and extremely randomized trees (ExtraTree) (Geurts et al., 2006) algorithms (Yi et al., 2020).

However, there are main drawbacks with the aforementioned methods. First, most of their extracted features for proteins as well as lncRNAs are hand-crafted, which consume much time and require strong domain knowledge. What is more, the previous studies attempt to construct a model to predict the lncRNA–protein interactions of all species. All these may lead to low robustness and overly optimistic predictions.

With the development of machine learning, network representation learning algorithm has become a pressing research task (Cui et al., 2019). In this study, we propose a new lncRNA–protein interactions prediction model called LncPNet based on heterogenous network embedding, which can solve the aforementioned problems in the existing methods. LncPNet is intentionally designed for predicting lncRNA–protein interactions in human, and thus it is trained by human lncRNA–protein interaction data. We apply network embedding to automatically generate features for proteins and lncRNAs. Specifically, a lncRNA–protein heterogenous network is constructed with lncRNA–lncRNA similarity, protein–protein similarity, and lncRNA–protein associations. Then, network embedding extracts, lncRNA features and protein features, are then fed into a SVM classifier to predict lncRNA–protein interactions. Moreover, we compare the performance of LncPNet with the previous models on the same benchmark database. The results demonstrate that LncPNet obtains predictive performance with higher accuracy and robustness.

2 MATERIALS AND METHODS

2.1 Framework of LncPNet

Figure 1 shows the schematic flowchart of our proposed LncPNet approach for predicting lncRNA–protein interactions based on heterogenous network embedding. The proposed method briefly includes three steps: 1) construction of a heterogenous network based on lncRNA–lncRNA similarity, protein–protein similarity, and known lncRNA–protein interactions; 2) the feature extraction for given lncRNA and protein using network embedding; and 3) training with SVM to predict novel lncRNA–protein associations. More detailed descriptions for each step are given below.

2.2 Datasets

In this study, we apply the known lncRNA–protein interaction data from NPInter v2.0 (Yuan et al., 2014) and lncRNA sequence data from NONCODE v6.0 (Zhao et al., 2016) as well as protein sequence data from UniProt (The UniProt Consortium, 2017). NPInter integrates experimentally verified functional interactions between ncRNAs (excluding tRNAs and rRNAs) and other biomolecules (proteins, RNAs, and genomic DNAs). NONCODE aims to present a complete collection and annotation of non-coding RNAs, especially long non-coding RNAs (lncRNAs). The UniProt knowledge base is a large resource of protein sequences and associated detailed annotation. First, we extract the human lncRNA–protein interactions from NPInter, which are filtered by restricting the organism, the type of lncRNAs, and the type of proteins to "Homo," "ncRNA," and "protein," respectively. After data cleaning, we obtain 7,523 experimentally validated human lncRNA–protein interactions, including 3,052 lncRNAs and 212 proteins. Then, we map these lncRNA IDs and protein IDs of NPInter into NONCODE IDs and UniProt IDs, respectively. From these lncRNAs and proteins that we have, we remove lncRNA and protein whose sequence information is unavailable. Finally, we obtain a dataset with 4,578 lncRNA–protein interactions between 2,009 lncRNAs and 78 proteins. In these datasets, only known lncRNA–protein associations (positive samples) are available. To train the classifier, we choose negative samples by a subcellular localization method with empirical tests of other alternatives. So, we randomly choose the same number of samples from all possible negative pairs. Meanwhile, the dataset is randomly divided into two parts, where one part is used for training set and the other is for testing. Among them, the quantity scale of the

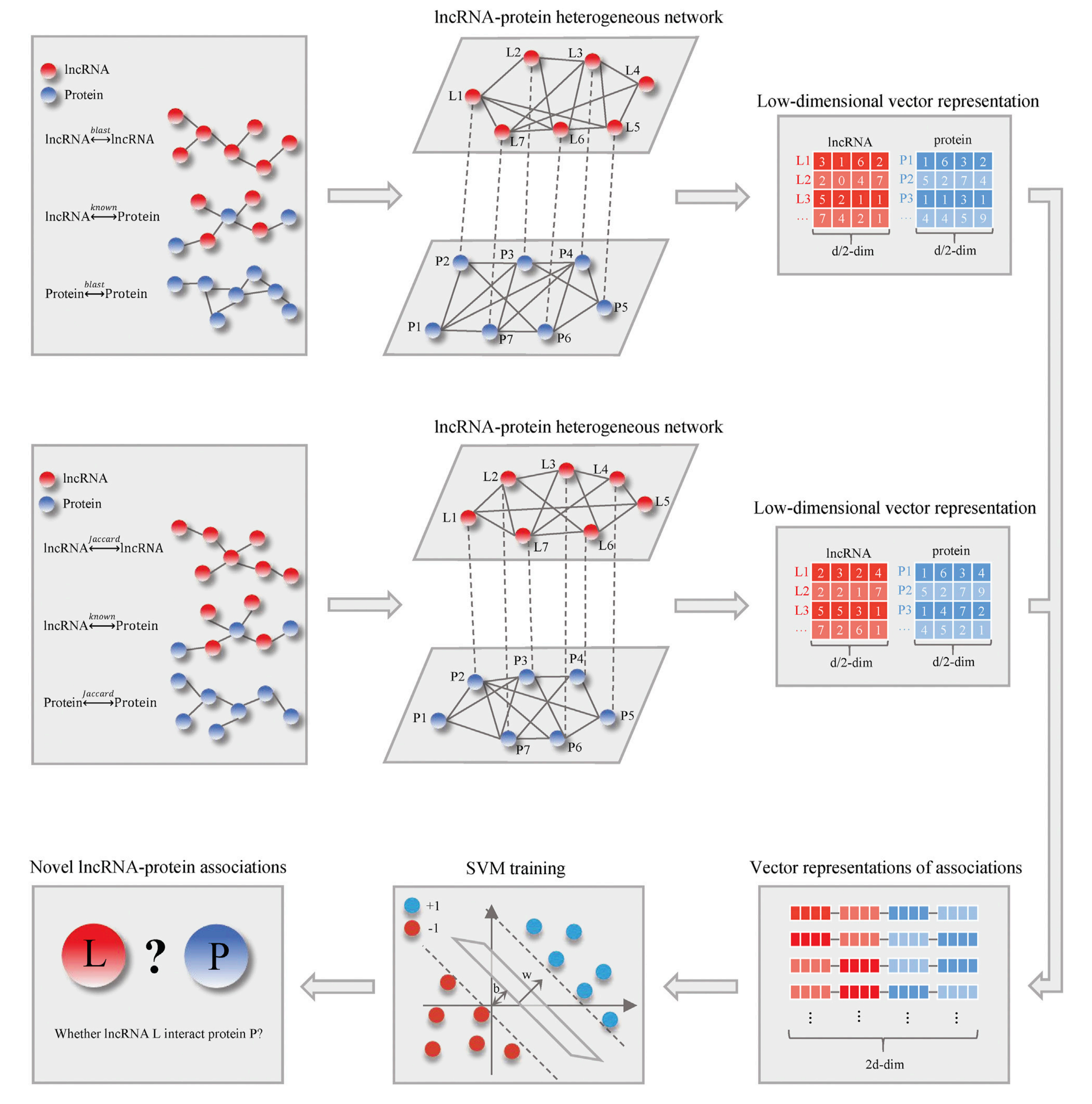


FIGURE 1 | Flowchart of LncPNet.

training set and test set is approximately 9:1, and the procedure is repeated three times.

2.3 Construction of a lncRNA–Protein Heterogenous Network

An lncRNA–protein heterogenous network is constructed with lncRNA–lncRNA similarity, protein–protein similarity, and known lncRNA–protein associations. lncRNA–lncRNA similarity and protein–protein similarity are both quantified in two different ways.

2.3.1 Jaccard Similarity

The Jaccard similarity (Bag et al., 2019) is an index used to measure the similarity of two sets. In this study, the Jaccard similarity is employed to calculate lncRNA–lncRNA similarities and protein–protein similarities. We define $L_i = \{p_1, p_2, \ldots, p_x\}$ and $P_j = \{l_1, l_2, \ldots, l_y\}$ as two sets of lncRNA i and protein j , which contain associated proteins of lncRNA i and associated lncRNAs of protein j, respectively. Given two lncRNAs, the similarity between two lncRNAs is defined as follows:

$$J\left(\mathrm{L_i}, \mathrm{L_j}\right) = \frac{\left|\mathrm{L_i} \cap L_j\right|}{\left|\mathrm{L_i} \cup L_j\right|}, \tag{1}$$

where L_i and L_j represent lncRNA i and lncRNA j associated proteins sets, respectively.

2.3.2 BLAST Similarity

BLAST is a fundamental and basic local alignment search tool for sequence similarity based on a local optimal alignment strategy (Ye et al., 2006). Essentially, BLAST is a heuristic algorithm. It first breaks the query sequence into sub-segments, called seed words. Furthermore, the seed is compared with the pre-indexed sequence, and the position with the higher continuous score of the seed is selected for further extension by the dynamic programming algorithm. The extension process will also be scored. When the score is below a certain limit, the extension process will be terminated and abandoned. Finally, a series of high-scored sequences are produced. In this study, we establish two local databases for lncRNA and protein. Then, the similarities between every two lncRNAs and every two proteins are calculated via BLAST.

2.3.3 The Heterogenous Network

The lncRNA–lncRNA Jaccard similarity network can be represented using a bipartite graph G_{11}, as follows:

$$G_{11} = (L, \mathrm{E}_{11}, J), \tag{2}$$

where $L = \{l_1, l_2, ...l_n\}$ represents the set of n lncRNAs, $E_{11} = \{e_1, e_2, ...e_m\}$ represents sets of edges between vertices, and l_i and l_j are connected if the Jaccard similarity is more than 0.5.

The lncRNA–lncRNA BLAST similarity network can be represented using a bipartite graph G_{12}, as follows:

$$G_{12} = (L, \mathrm{E}_{12}, B), \tag{3}$$

where $L = \{l_1, l_2, ...l_n\}$ represents the set of n lncRNAs, $E_{12} = \{e_1, e_2, ...e_m\}$ represents sets of edges between vertices, and l_i and l_j are connected if the BLAST similarity e-value is less than 0.001.

Similarly, two bipartite graphs G_{21} and G_{22} represent protein–protein similarities as follows:

$$G_{21} = (P, \mathrm{E}_{21}, J); \tag{4}$$

$$G_{22} = (P, \mathrm{E}_{22}, B), \tag{5}$$

where $P = \{p_1, p_2, ...p_n\}$ represents the set of n proteins, $E_{21} = \{e_1, e_2, ...e_m\}$ and $E_{22} = \{e_1, e_2, ...e_m\}$ represent sets of edges between vertices, and P_i and P_j are connected if their Jaccard similarity is more than 0 and the BLAST similarity e-value is less than 0.01.

Then, we construct two heterogenous networks. Among them, one is by known lncRNA–protein interactions, lncRNA–lncRNA similarities, and protein–protein similarities calculated with the Jaccard similarity. The other is by known lncRNA–protein interactions, lncRNA–lncRNA similarities, and protein-protein similarities calculated with BLAST similarity.

2.4 Heterogenous Network Embedding

Network embedding can use less information to represent nodes as dense- and low-dimensional vectors and has been rapidly developed and applied recently (Cao et al., 2016; Hamilton et al., 2018; Veličković et al., 2018; Zhang et al., 2020). According to the heterogenous network constructed previously, we employ network embedding to learn the low-dimensional latent representations based on the structural and semantic properties of the lncRNA–protein heterogenous network, which are able to characterize the lncRNA–protein associations. In LncPNet, we adopt the metapath2vec method (Dong et al., 2017) for network embedding because it takes better account of the type of nodes, which is suitable for representing the heterogenous network. Generally, metapath2vec can be divided into two steps. First, we employ meta-path-based random walks to generate paths that can capture both the semantic and structural correlations between different types of nodes and then facilitate the transformation of heterogenous network structures into metapath2vec$'$s skip-grams.

In detail, a meta-path scheme φ from V_1 to V_l is defined as the form of $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} ...V_t \xrightarrow{R_t} V_{t+1}... \xrightarrow{R_{l-1}} V_l$, where $R = R_1 \circ R_2 \circ ... \circ R_{l-1}$ is defined as the composite relations between node types V_1 and V_l. In this study, we define "LPLPL" and "LLPPLL" metapaths, in which "LPLPL" represents two lncRNAs interact via a protein and similarly for "LLPPLL". For the heterogenous network $G(V, E)$ and metapath $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} ...V_t \xrightarrow{R_t} V_{t+1}... \xrightarrow{R_{l-1}} V_l$, the transition probability at step i is defined as follows (Yang et al., 2019):

$$p\left(v^{i+1} \middle| v_k^i, \varphi\right) = \begin{cases} \dfrac{1}{\left|\mathrm{N}_j\left(v_k^i\right)\right|}, \left(v^{i+1}, v_k^i\right) \in E, \phi\left(v^{i+1}\right) = j \\ 0, otherwise, \end{cases}, \tag{6}$$

where v_j and v_k, respectively, denote the jth and kth node type in the path φ, $N_j(v_k)$ denotes the neighborhood of node v_k^j with respect to the jth node type, and $\phi(v)$ is a constraint function to make sure the node type of node v to be type j. In order to avoid the disclosure of the test set information, we remove the associations between lncRNA and protein in the test set when the metapath is generated. Then, skip-gram learns effective node representations for a heterogenous network $G(V, E)$ by maximizing the probability of having the heterogenous context.

LncPNet employs metapath2vec on the aforementioned two heterogenous networks to produce a 1 × 64 feature vector for every vertex. Moreover, we splice the two feature vectors of every lncRNA to obtain a 1 × 128 feature vector, which is the same to every protein encoded.

2.5 Prediction of lncRNA–Protein Interactions

With vector representations of lncRNA–protein associations as inputs, which of dimensionality is 1 × 256, SVM is trained to predict whether an lncRNA interacts with a protein. In particular, our training set and test set are pre-divided, and we conduct the

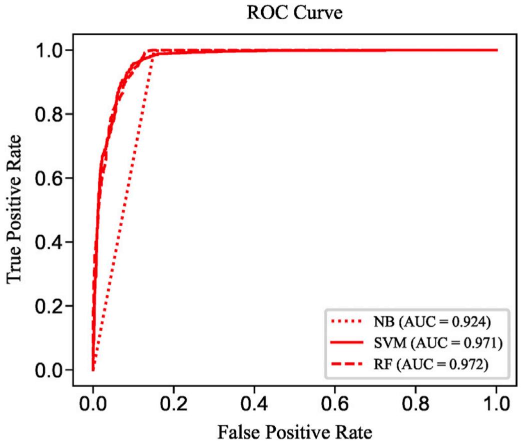


FIGURE 2 | ROC curves of SVM, RF, and NB.

procedure three times. What is more, we choose radial basic function (RBF) as the SVM kernel function.

2.6 Performance Evaluation

Precision (PRE), recall (REC), specificity (SPE), accuracy (ACC), Matthew's correlation coefficient (MCC), and F1-score are the most common classification model evaluation indicators. They can be defined as (Sokolova et al., 2006):

$$PRE = \frac{TP}{TP + FP}; \tag{7}$$

$$REC = \frac{TP}{TP + FN}; \tag{8}$$

$$SPE = \frac{TN}{FP + TN}; \tag{9}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}; \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}; \tag{11}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}, \tag{12}$$

where TP, FP, TN, and FN is the number of true positives, false positives, true negatives, and false negatives, respectively.

3 RESULTS AND DISCUSSION

3.1 Performance of LncPNet

To evaluate the prediction performance of LncPNet, we test RF (Breiman, 2001), naive Bayesian (NB) (Elkan, 1997), and SVM (Joachims, 1998) classifiers. As shown in **Figure 2**, SVM achieves the AUC of 0.971 on the NPInter v2.0 dataset. It increases by 4.7% over NB with the AUC of 0.924 and decreases by 0.1% over RF with the AUC of 0.972. But from **Figure 3**, SVM has comparable performance with RF. Thus, we choose SVM as our classifier implemented in LncPNet. What is more, we test different negative samples producing approaches on this model. Finally, LncPNet employs the SVM classifier to train the model and adopts the subcellular localization method to produce negative samples. For comparison study, we evaluate the performance of CF (Sarwar et al., 2001), RWR (Köhler et al., 2008), LPBNI (Ge et al., 2016), SFPEL-LPI (Zhang et al., 2018b), LPIHN (Li et al., 2015), LPLNP (Zhang et al., 2018a), RPI-SE (Yi et al., 2020), and IPMiner (Pan et al., 2016) on NPInter v2.0. Meanwhile, the performance of different sub-models has also been identified. In order to evaluate the performance of these methods comprehensively, we employ the ACC, PRE, REC, SPE, MCC, AUC, and F1 as the evaluation metrics. AUC (Huang and Ling, 2005) is the area under the ROC (Fawcett, 2006) curve, which is an evaluation dedicated to the classification model. In LncPNet, the average PRE, REC, SPE, ACC, MCC, F1, and AUC is 0.908, 0.957, 0.903, 0.930, 0.860, 0.932, and 0.971, respectively.

3.2 Comparisons With Sub-Models

In order to fully evaluate the performance, we compare LncPNet with three sub-models on NPInter v2.0. LncPNet model construction is mainly divided into three steps. Specifically, we construct a heterogenous network with lncRNA–lncRNA similarities, protein–protein similarities, and known lncRNA–protein interactions, where lncRNA–lncRNA similarities and protein–protein similarities are calculated by the Jaccard similarity and BLAST similarity, respectively. Then, a feature vector is generated from the heterogenous network with network embedding (metapath2vec) to characterize a pair of lncRNA and protein. Finally, with the feature vectors with class labels as inputs, SVM is trained to predict potential lncRNA–protein associations. The construction of heterogenous network contains four types of different strategies. In approach 1, only known lncRNA–protein interactions (KNet) are used to construct the network; in approach 2, known lncRNA–protein interactions and Jaccard similarity (KJNet) are used to construct the network; in approach 3, known lncRNA–protein interactions and BLAST similarity (KBNet) are used to construct the network; and in approach 4, known lncRNA–protein interactions, Jaccard similarity, and BLAST similarity (LncPNet) are used to construct the

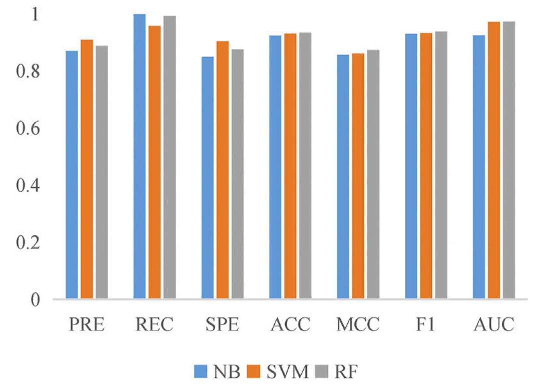


FIGURE 3 | Histogram of the six evaluation criteria achieved by SVM, RF, and NB models.

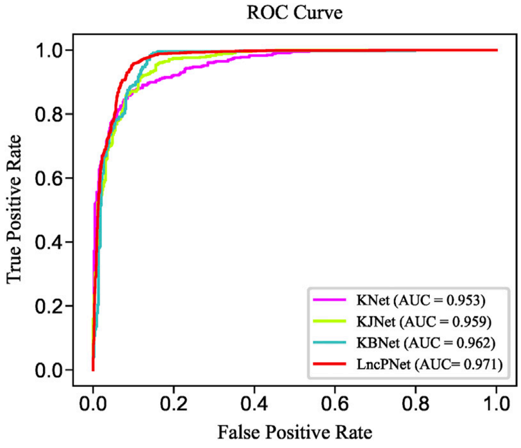


FIGURE 4 | ROC curves of LncPNet and sub-models.

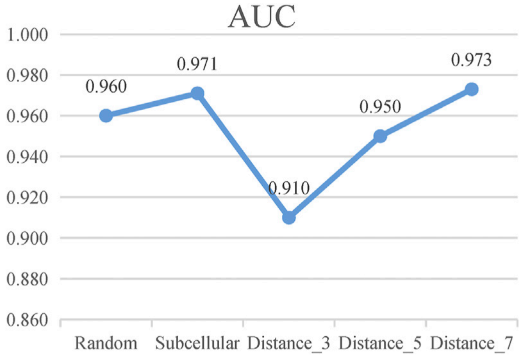


FIGURE 5 | AUC values of Random, Subcellular, "Distance_3," "Distance_5," and "Distance_7" (Random, random-pairing method; Subcellular, subcellular localization method).

TABLE 1 | Prediction results of LncPNet and sub-models.

Network	PRE	REC	SPE	ACC	MCC	F1	AUC
KNet	0.898	0.873	0.901	0.887	0.774	0.885	0.953
KJNet	0.887	0.914	0.884	0.899	0.799	0.900	0.959
KBNet	0.875	**0.982**	0.859	0.921	0.848	0.925	0.962
LncPNet	**0.908**	0.957	**0.903**	**0.930**	**0.860**	**0.932**	**0.971**

Every bold value means it corresponds to the highest value in the evaluation indicator.

TABLE 2 | Performance comparison of five negative sample models.

Method	PRE	REC	SPE	ACC	MCC	F1	AUC
Random	0.870	0.946	0.856	0.901	0.808	0.905	0.960
Subcellular	**0.908**	**0.957**	**0.903**	**0.930**	**0.860**	**0.932**	0.971
Distance_3	0.846	0.820	0.851	0.835	0.672	0.833	0.910
Distance_5	0.863	0.915	0.854	0.884	0.771	0.888	0.950
Distance_7	0.905	0.933	**0.903**	0.918	0.837	0.919	**0.973**

Every bold value means it corresponds to the highest value in the evaluation indicator.

network. **Figure 4** shows the ROC curve. **Table 1** illustrates the prediction results of different integration strategies on NPInter v2.0. From **Table 1**, we can find the experiments of LncPNet integrate the advantages of different branch models, achieving better performance than those of sub-models.

3.3 The Strategy of Negative Sampling

Missing negative samples has always been a problem in predicting molecular interactions, which leads to a wide variety of negative sample generation methods. However, few studies have proved how to generate negative samples is the most reliable. In this section, we summarize three commonly used negative sample construction methods. The first one, and also the most popular one, is the random pairing method. Negative samples are randomly sampled from the possible lncRNA–protein pairs except the positive samples. The second one is the method of subcellular localization, which is based on the assumption that the lncRNA and protein that are not in the same subcellular location would not interact with each other. Therefore, proteins and lncRNAs that are not in one organelle are regarded as negative sample pairs. The third one is the network distance method, which calculates the shortest-path distance between each lncRNA and protein in the prior interaction network, and treats the protein and lncRNA that are greater than a certain distance threshold, for e.g., six, as a negative sample pair.

According to these rules, we further categorize the distance method of selecting negative samples into three types of experiments: 1) "Distance_3": the negative samples with a distance equal to 3; 2) "Distance_5": the negative samples with a distance greater than 1 and less than or equal to 5; and 3) "Distance_7": the negative sample with a distance greater than 1 and less than or equal to 7. To avoid the imbalance problem when training the classifier, we choose negative samples with the same number of positive samples in the experiments. As presented in **Figure 5**, the subcellular localization method and "Distance_7" achieve a relatively higher value than the random pairing, "Distance_3" and "Distance_5" methods. Meanwhile, in the three distance-based methods, "Distance_3," "Distance_5," and "Distance_7", we find that as the distance of selecting negative sample increases, the AUC value becomes higher. This also validates the rationality of our proposed strategy and the former assumption in selecting negative samples. **Table 2** shows that the subcellular localization method achieves the best prediction performance according to the six evaluation metrics. This clearly shows that different negative samples have a concrete impact on the model, and more reliable negative samples will make LncPNet to achieve better prediction results. Thus, we employ the subcellular localization method as our negative sample generation method in LncPNet.

3.4 Comparison With Other State-Of-The-Art Models

In order to further demonstrate the reliability and robustness of prediction by the LncPNet method, we compare LncPNet with

TABLE 3 | Performance comparison of LncPNet and eight available methods.

Method	PRE	REC	SPE	ACC	MCC	F1	AUC
CF	0.583	0.894	0.361	0.627	0.301	0.706	0.761
RWR	0.739	0.798	0.717	0.757	0.517	0.767	0.830
LPBNI	0.740	0.840	0.698	0.769	0.548	0.785	0.859
SFPEL-LPI	0.769	0.920	0.724	0.822	0.657	0.838	0.916
LPIHN	0.807	0.966	0.769	0.867	0.750	0.879	0.938
LPLNP	0.832	0.943	0.810	0.876	0.761	0.884	0.944
RPI-SE	0.877	**0.974**	0.863	0.919	0.843	0.923	0.959
IPMiner	0.886	0.970	0.875	0.922	0.849	0.926	0.961
LncPNet	**0.908**	0.957	**0.903**	**0.930**	**0.860**	**0.932**	**0.971**

Every bold value means it corresponds to the highest value in the evaluation indicator.

the eight state-of-the-art methods, namely IPMiner, RPI-SE, LPLNP, RWR, CF, SFPEL-LPI, LPBNI, and LPIHN, on the same benchmark of NPInter v2.0. These methods are typical methods that have been proposed in recent years, and they can be divided into three categories:

(1) The first type of method is mainly based on sequence information, structural information, evolutionary knowledge, or physical and chemical properties to mine the distinguishing characteristics of the lncRNA and protein. For example, RPI-SE applied the position weight matrix combined with Legendre moments to obtain protein evolutionary information and k-mer sparse matrix to extract feature of lncRNA sequences. SFPEL-LPI used sequence information to build a feature projection ensemble-learning frame to predict lncRNA–protein interactions.
(2) The second type of method is mainly to use stacked autoencoders to extract high-level hidden features of proteins and lncRNAs. For example, IPMiner extracted raw sequence composition features from lncRNA and protein sequences, high-level features by applying stacked autoencoder, and fine-tuning features using label information, and then a training ensemble strategy such as RF classifier to robustly predict the interactions between lncRNAs and proteins.
(3) The third type of method mainly uses topological information to extract lncRNA and protein features. For example, LPLNP employed a linear neighborhood propagation method, to predict lncRNA–protein interactions. LPBNI used a bipartite network–based method for predicting lncRNA–protein interactions. RWR and CF are also the same type of methods. LPIHN constructed a lncRNA–protein heterogenous network and used a random walk with restart to infer novel lncRNA–protein interactions.

We replicate all these methods on the same dataset for fair comparisons. As shown in **Table 3**, LncPNet achieves a PRE of 0.908, SPE of 0.903, ACC of 0.930, MCC of 0.860, and F1 of 0.932, which outperform all the other methods. REC is a little worse than the best method, IPMiner. All these performance comparisons indicate that LncPNet has higher reliability in predicting lncRNA–protein interactions. **Figure 6** illustrates the ROC curves with AUCs of these methods. The results further demonstrate the effectiveness and advantage of our method, LncPNet. Although we use the heterogenous network with LPIHN, our metapath2vec method takes into account the node type and transition probability simultaneously, which makes it achieves better performance.

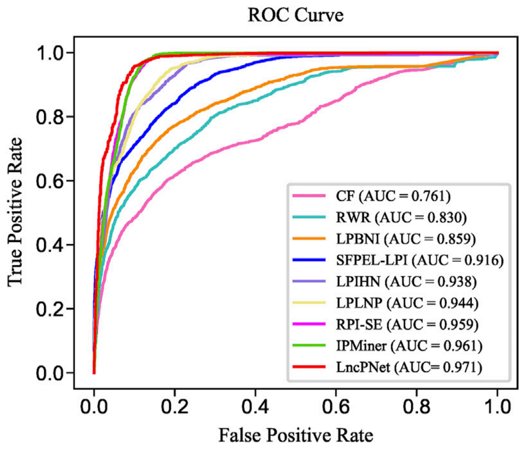


FIGURE 6 | ROC curves of LncPNet and eight comparing methods.

TABLE 4 | Top 10 novel interactions predicted by LncPNet.

Rank	lncRNA	Protein	Whether confirmed
1	NONHSAT032174.2	O00425	Yes
2	NONHSAT017141.2	O00425	Yes
3	NONHSAT125498.2	P61978	Yes
4	NONHSAT048327.2	Q01844	Yes
5	NONHSAT017141.2	O00571	No
6	NONHSAT125498.2	Q9NW64	Yes
7	NONHSAT017141.2	P78332	No
8	NONHSAT048327.2	Q9NW64	No
9	NONHSAT125498.2	Q8IYB8	Yes
10	NONHSAT067050.2	P70372	Yes

3.5 Case Study

In order to further evaluate the reliability of our prediction model, we propose a case study to verify its performance. As mentioned earlier, the dataset we used in LncPNet is NPInter v2.0, and currently NPInter has been updated to NPInter v4.0, which includes some novel lncRNA–protein interaction pairs. We test to predict the new lncRNA–protein interactions confirmed in NPInter v4.0 based on known interactions in NPInter v2.0. Specifically, we predict the 23 pairs of interactions newly discovered in NPInter v4.0 and the generated 23 pairs of negative samples and rank them according to the scores. As shown in **Table 4**, we list the top ten interactions predicted by LncPNet, in which seven novel interactions are confirmed in the new version of NPInter. **Figure 7** illustrates the constructed network diagram. The case study provides more evidence for the effectiveness,

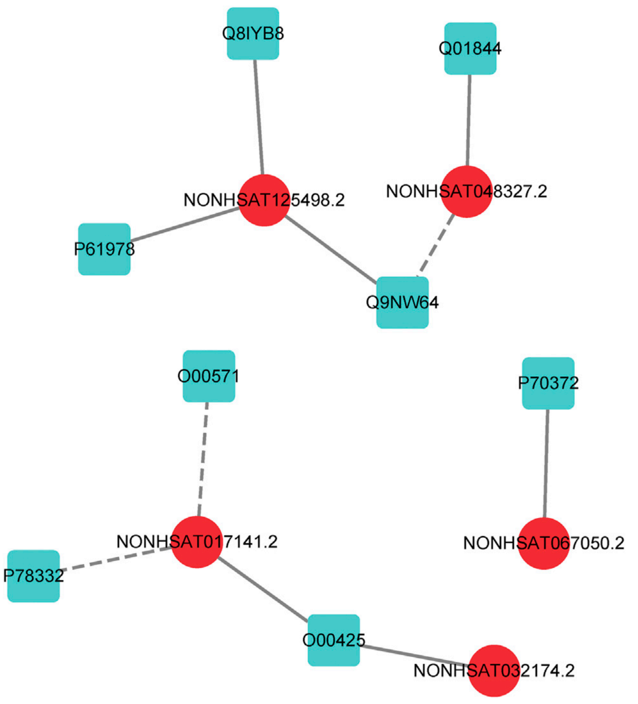


FIGURE 7 | lncRNA–protein network constructed by LncPNet. The lncRNA and proteins are shown in red (circular) and blue (square) nodes, respectively, while the correctly and wrongly predicted interactions are shown as solid and dotted lines, respectively.

flexibility, and extendibility in predicting lncRNA–protein interactions.

CONCLUSION

In this study, we proposed LncPNet based on a heterogeneous network embedding method for predicting lncRNA–protein interactions. The experimental results demonstrated that LncPNet achieves high prediction performance on our benchmark dataset and yields better results compared to other methods. As for the lncRNA–protein interaction predictive task is a nonnegative sample problem, we provided a new perspective into network embedding by comparing three kinds of methods for negative sampling. In addition, the case study results further demonstrated the effectiveness of LncPNet. The network embedding method is a general node representing method. The framework of LncPNet can be expanded to other interaction predictive task, such as miRNA–protein interaction prediction and lncRNA–disease interaction prediction.

DATA AVAILABILITY STATEMENT

The data and code in this study are available at: https://github.com/zpliulab/LncPNet.

AUTHOR CONTRIBUTIONS

GZ performed the experiments, analyzed the data, and wrote the manuscript. PL, XQ, and XH analyzed the data and wrote the manuscript. Z-PL conceived and designed the experiments and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was partially supported by the National Key Research and Development Program of China (No. 2020YFA0712402), the National Natural Science Foundation of China (No. 61973190), the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project 2019JZZY010423), the Natural Science Foundation of Shandong Province of China (ZR2020ZD25), the Innovation Method Fund of China (Ministry of Science and Technology of China, 2018IM020200), the Tang Scholar, and the Program of Qilu Young Scholar of Shandong University.

ACKNOWLEDGMENTS

We thank the members of our lab at Shandong University for their assistance.

REFERENCES

Bag, S., Kumar, S. K., and Tiwari, M. K. (2019). An Efficient Recommendation Generation Using Relevant Jaccard Similarity. *Inf. Sci.* 483, 53–64. doi:10.1016/j.ins.2019.01.023

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324

Cao, S., Lu, W., and Xu, Q. (2016). "Deep Neural Networks for Learning Graph Representations," in Proceedings of the AAAI Conference on Artificial Intelligence 30 (1), Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/10179.

Chen, T., and Guestrin, C. (2016). "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco California USA: Association for Computing Machinery, 785–794. doi:10.1145/2939672.2939785

Cui, P., Wang, X., Pei, J., and Zhu, W. (2019). A Survey on Network Embedding. *IEEE Trans. Knowl. Data Eng.* 31, 833–852. doi:10.1109/TKDE.2018.2849727

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of Transcription in Human Cells. *Nature* 489, 101–108. doi:10.1038/nature11233

Dong, Y., Chawla, N. V., and Swami, A. (2017). metapath2vecHalifax NS Can. ACM). in" Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, August 04, 2017, 135–144. doi:10.1145/3097983.3098036

Elkan, C. (1997). "Boosting And Naive Bayesian Learning," in Proceedings of the International Conference on Knowledge Discovery and Data Mining.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Lett.* 27, 861–874. doi:10.1016/j.patrec.2005.10.010

Ge, M., Li, A., and Wang, M. (2016). A Bipartite Network-Based Method for Prediction of Long Non-coding RNA-Protein Interactions. *Genomics Proteomics Bioinformatics* 14, 62–71. doi:10.1016/j.gpb.2016.01.004

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely Randomized Trees. *Mach. Learn.* 63, 3–42. doi:10.1007/s10994-006-6226-1

Guttman, M., and Rinn, J. L. (2012). Modular Regulatory Principles of Large Non-coding RNAs. *Nature* 482, 339–346. doi:10.1038/nature10887

Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Representation Learning on Graphs: Methods and Applications.*ArXiv170905584 Cs*. Available at: http://arxiv.org/abs/1709.05584 (Accessed September 10, 2021).

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-ensemble: Prediction of Human lncRNA-Protein Interactions Based on Ensemble Strategy. *RNA Biol.* 1, 1–10. doi:10.1080/15476286.2018.1457935

Jin Huang, Jin., and Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* 17, 299–310. doi:10.1109/TKDE.2005.50

Joachims, T. (1998). *Making Large-Scale SVM Learning Practical*. Dortmund: Technical Report, No. 1998,28, 18.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *Am. J. Hum. Genet.* 82, 949–958. doi:10.1016/j.ajhg.2008.02.013

Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting Long Noncoding RNA and Protein Interactions Using Heterogeneous Network Model. *Biomed. Res. Int.* 2015, 1–11. doi:10.1155/2015/671950

Liu, Z.-P. (2021). Predicting lncRNA-Protein Interactions by Machine Learning Methods: A Review. *Cbio* 15, 831–840. doi:10.2174/1574893615666200224095925

Louro, R., Smirnova, A. S., and Verjovski-Almeida, S. (2009). Long Intronic Noncoding RNA Transcription: Expression Noise or Expression Choice? *Genomics* 93, 291–298. doi:10.1016/j.ygeno.2008.11.009

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational Prediction of Associations between Long Non-coding RNAs and Proteins. *BMC Genomics* 14, 651. doi:10.1186/1471-2164-14-651

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10, 155–159. doi:10.1038/nrg2521

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-Protein Interactions Using Only Sequence Information. *BMC Bioinformatics* 12, 489. doi:10.1186/1471-2105-12-489

Pan, X., Fan, Y.-X., Yan, J., and Shen, H.-B. (2016). IPMiner: Hidden ncRNA-Protein Interaction Sequential Pattern Mining with Stacked Autoencoder for Accurate Computational Prediction. *BMC Genomics* 17, 582. doi:10.1186/s12864-016-2931-8

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell* 136, 629–641. doi:10.1016/j.cell.2009.02.006

Qu, Z., and Adelson, D. L. (2012). Evolutionary Conservation and Functional Roles of ncRNA. *Front. Gene* 3. doi:10.3389/fgene.2012.00205

Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. (2001). Item-based Collaborative Filtering Recommendation Algorithms. *Proc. ACM World Wide Web Conf.* 1, 285–295. doi:10.1145/371920.372071

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in Advances In Artificial Intelligence *Lecture Notes in Computer Science*. Editors A. Sattar and B. Kang (Berlin, Heidelberg: Springer Berlin Heidelberg), 1015–1021. doi:10.1007/11941439_114

Suresh, V., Liu, L., Adjeroh, D., and Zhou, X. (2015). RPI-pred: Predicting ncRNA-Protein Interaction Using Sequence and Structural Information. *Nucleic Acids Res.* 43, 1370–1379. doi:10.1093/nar/gkv020

The UniProt Consortium (2017). UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi:10.1093/nar/gkw1099

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph Attention Networks. ArXiv171010903 Cs Sta. .Available at: http://arxiv.org/abs/1710 (Accessed September 10, 2021)

Yang, K., Zhao, X., Waxman, D., and Zhao, X.-M. (2019). Predicting Drug-Disease Associations with Heterogeneous Network Embedding. *Chaos* 29, 123109. doi:10.1063/1.5121900

Yang, Q., Zhang, S., Liu, H., Wu, J., Xu, E., Peng, B., et al. (2014). Oncogenic Role of Long Noncoding RNA AF118081 in Anti-benzo[a]pyrene-trans-7,8-dihydrodiol-9,10-epoxide-transformed 16HBE Cells. *Toxicol. Lett.* 229, 430–439. doi:10.1016/j.toxlet.2014.07.004

Yang, Y., Wen, L., and Zhu, H. (2015). Unveiling the Hidden Function of Long Non-coding RNA by Identifying its Major Partner-Protein. *Cell Biosci* 5, 59. doi:10.1186/s13578-015-0050-x

Ye, J., McGinnis, S., and Madden, T. L. (2006). BLAST: Improvements for Better Sequence Analysis. *Nucleic Acids Res.* 34, W6–W9. doi:10.1093/nar/gkl164

Yi, H.-C., You, Z.-H., Wang, M.-N., Guo, Z.-H., Wang, Y.-B., and Zhou, J.-R. (2020). RPI-SE: a Stacking Ensemble Learning Framework for ncRNA-Protein Interactions Prediction Using Sequence Information. *BMC Bioinformatics* 21, 60. doi:10.1186/s12859-020-3406-0

Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2014). NPInter v2.0: an Updated Database of ncRNA Interactions. *Nucl. Acids Res.* 42, D104–D108. doi:10.1093/nar/gkt1057

Zhang, D., Yin, J., Zhu, X., and Zhang, C. (2020). Network Representation Learning: A Survey. *IEEE Trans. Big Data* 6, 3–28. doi:10.1109/TBDATA.2018.2850013

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018a). The Linear Neighborhood Propagation Method for Predicting Long Non-coding RNA-Protein Interactions. *Neurocomputing* 273, 526–534. doi:10.1016/j.neucom.2017.07.065

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018b). SFPEL-LPI: Sequence-Based Feature Projection Ensemble Learning for Predicting LncRNA-Protein Interactions. *PLoS Comput. Biol.* 14, e1006616. doi:10.1371/journal.pcbi.1006616

Zhao, Y., Li, H., Fang, S., Kang, Y., wu, W., Hao, Y., et al. (2016). NONCODE 2016: an Informative and Valuable Data Source of Long Non-coding RNAs. *Nucleic Acids Res.* 44, D203–D208. doi:10.1093/nar/gkv1252

Zhu, J., Fu, H., Wu, Y., and Zheng, X. (2013). Function of lncRNAs and Approaches to lncRNA-Protein Interactions. *Sci. China Life Sci.* 56, 876–885. doi:10.1007/s11427-013-4553-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 10 February 2022
doi: 10.3389/fgene.2022.815692

Generation of Realistic Gene Regulatory Networks by Enriching for Feed-Forward Loops

Erik K. Zhivkoplias[1†], Oleg Vavulov[2†], Thomas Hillerton[1] and Erik L. L. Sonnhammer[1]*

[1] Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Solna, Sweden, [2] Bioinformatics Institute, St. Petersburg, Russia

The regulatory relationships between genes and proteins in a cell form a gene regulatory network (GRN) that controls the cellular response to changes in the environment. A number of inference methods to reverse engineer the original GRN from large-scale expression data have recently been developed. However, the absence of ground-truth GRNs when evaluating the performance makes realistic simulations of GRNs necessary. One aspect of this is that local network motif analysis of real GRNs indicates that the feed-forward loop (FFL) is significantly enriched. To simulate this properly, we developed a novel motif-based preferential attachment algorithm, FFLatt, which outperformed the popular GeneNetWeaver network generation tool in reproducing the FFL motif occurrence observed in literature-based biological GRNs. It also preserves important topological properties such as scale-free topology, sparsity, and average in/out-degree per node. We conclude that FFLatt is well-suited as a network generation module for a benchmarking framework with the aim to provide fair and robust performance evaluation of GRN inference methods.

Keywords: network biology, gene regulatory networks, gene-gene interaction, network motif structure, network generation, network simulation, benchmarking

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Vitaly Gursky,
Ioffe Institute (RAS), Russia
Michael Mayo,
US Army Corps of Engineers,
United States

Correspondence:
Erik L. L. Sonnhammer
erik.sonnhammer@scilifelab.se

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 November 2021*
Accepted: *13 January 2022*
Published: *10 February 2022*

Citation:
Zhivkoplias EK, Vavulov O, Hillerton T and Sonnhammer ELL (2022) Generation of Realistic Gene Regulatory Networks by Enriching for Feed-Forward Loops. Front. Genet. 13:815692. doi: 10.3389/fgene.2022.815692

INTRODUCTION

Understanding large-scale biological relationships between genes and the proteins they encode remains a great challenge in systems biology. The wide availability of system-level expression datasets has given rise to a variety of reverse engineering methods that aim to reconstruct the hidden regulatory gene–gene and gene–protein relationships. Such relationships form a gene regulatory network (GRN) that regulates developmental processes in organisms and controls adaptation to changes in the environment (Davidson, 2010). By contrast with other networks in biological systems, GRNs are harder to validate as the interactions that occur between genes usually involve indirect interactions through biological molecules making the interaction hard to detect and quantify. The incompleteness and scarcity of ground-truth networks results in problems when evaluating the performance of methods that seek to infer GRNs from large-scale expression data (Emmert-Streib and Dehmer, 2018).

The problem of inferring a gene regulatory network from gene expression data has received significant attention. A variety of GRN inference methods are commonly used (Margolin et al., 2006; Faith et al., 2007; Friedman et al., 2010; Huynh-Thu et al., 2010; Zavlanos et al., 2011) to tackle this problem. It was also the focus of four separate Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges, with DREAM5 being the most recent one (Marbach et al., 2012).

Newer, more advanced algorithms require not only expression data but also utilize additional information such as experimentally validated interactions and Gene Ontology terms (Chouvardas et al., 2016), structures of genomic datasets and network topology (Siahpirani and Roy, 2017), DNA binding domains of transcription factors, and promoter sequences of its putative targets (Kang et al., 2018), or use the iterative kernel PCR model (Iglesias-Martinez et al., 2021). Despite this, for most methods the performance on real experimental datasets remains modest (Marbach et al., 2012; Chen and March 2018; Pratapa et al., 2020).

Regardless of the method used, it is important to fairly assess its performance with respect to other methods. As some methods can only predict Boolean networks, assessment should be done in terms of binary error classification such as the number of false positives and false negatives. In addition to this, experimental information about transcriptional interactions is usually only available in the binary form. Boolean networks can only be defined by their topology, which is why it is essential to understand the structure of GRN graphs. It is also worth pointing out that most GRN inference methods can only predict a static network structure, which implies that *in-silico* generated GRNs should also possess biological stability.

While the true structure of real GRNs is usually not known, they tend to share some topological features: the scale-free property (Barabasi and Albert, 1999), where the node degrees follow a power-law degree distribution, and often have the small world property (Watts and Strogatz, 1998), and where nodes form distinct clusters in which they are connected to each other in lattice rings. These properties are different from random graphs where node degrees are normal distributed across all nodes in the system. Some attempts to simulate GRNs have been made by implementing methods that generate random (Watts and Strogatz, 1998; Mendes et al., 2003) or scale-free (Barabasi and Albert, 1999) graphs with given sets of parameters, but eventually methods based on the idea of subnetwork-selection from biological networks gained more popularity (Van den Bulcke et al., 2006). One example of this is GeneNetWeaver (GNW) (Schaffter et al., 2011), which was used to generate *in silico* networks for the DREAM challenges.

The regulatory dynamics of GRNs is shaped by network patterns that are more frequent in GRNs than in other networks (Milo et al., 2002; Shen-Orr et al., 2002) and may carry information-processing functions. These local patterns, or motifs, and do not result in emergence of specific patterns in gene expression but rather determine dynamical boundaries of the phase space of the system (Ahnert and Fink, 2016). It was suggested that some motifs could be particularly important for network dynamics and therefore become overrepresented and drive the evolution of the networks (Prill et al., 2005). Examples of how feed-forward loops are involved in such dynamics are ample, including sign-sensitive delay elements (Mangan et al., 2003), bi-phase response generators (Kaplan et al., 2008), band-pass filters (Sohka et al., 2009), and decoders of oscillatory signals (Zhang et al., 2016). Due to this, simulating a network structure that preserves the overrepresentation of motifs is of utmost importance for capturing realistic dynamics of GRNs. The idea of building gene regulatory networks by using motifs as building blocks was first introduced by Abdelzaher et al. (2015a) that hypothesized that this could be important for the evolution of GRN topology in *E. coli*.

Network inference methods aim to solve the problem of finding regulatory interactions within a set of genes. This, however, doesn't imply that all edges in a reconstructed network represent physical binding between transcription factors and their respective targets. Gardner and Faith (2005) describe two groups of reverse-engineering algorithms. The first group seeks to identify regulators that directly control mRNA expression, and the second one is focused on identification of general regulatory interactions between different genes that may be indirect. Regardless of interaction type, simulated data should allow for exploring a wide range of network properties to evaluate inference algorithms performance. It was shown that FFLs are significantly overrepresented in experimentally validated transcriptional regulation databases (Lee et al., 2002; Milo et al., 2002). FFLs were also found to be significantly overrepresented in other databases of microRNAs and their predicted targets (Krek et al., 2005; Lewis et al., 2005) with Z-score range between 1.39 and 6.03 (Shalgi et al., 2007). Other TF-microRNA studies demonstrated that in the circuitry of gene regulation via intermediate microRNAs, in mouse and human, and the FFL motif is also enriched (Tsang et al., 2007). This suggests that FFL is an important signature of real GRNs that represent either direct or indirect interactions between genes.

In the present study the significance of 3-node motifs in four directed GRNs based on experimentally verified transcriptional interaction databases were evaluated. In agreement with previous studies (Lee et al., 2002; Milo et al., 2002; Boyer et al., 2005), it was found that the feed-forward loop (FFL) is the only motif that is overrepresented. This motivated us to develop a novel motif-based preferential attachment algorithm called FFLatt for simulating realistic structures of GRNs that are enriched with the FFL motif. The networks generated by FFLatt demonstrate structural properties that agree with biological GRNs, and have good robustness in stability analyses. Given their realistic properties, they are well suited for fair and robust evaluation of the performance of GRN inference algorithms.

METHODS

Transcriptional Interaction Databases

Three biological databases that contain information of experimentally validated transcriptional regulation were chosen as ground-truth networks: RegulonDB (Santos-Zavaleta et al., 2019) for *E. coli* (Balaji et al., 2006), for *S. cerevisiae*, and TRRUST v2 (Han et al., 2018) for *M. musculus* and *H. sapiens* transcription factor—target regulatory relationships.

Motif-Node Participation and Motif Enrichment

We chose to test for node-motif participation for all possible connected three-node motifs with no reciprocal links between them (**Figure 1**). Reciprocal links were not considered as they are

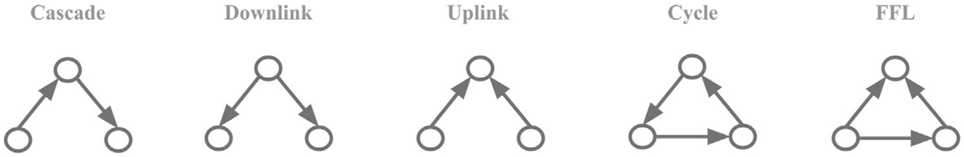


FIGURE 1 | Motif collection. The five possible three-node motifs with 2 or 3 unidirectional links.

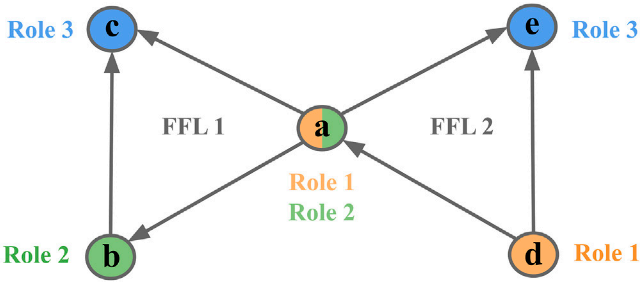


FIGURE 2 | Node participation in FFL motif. An example of 3-node motif counts given on an FFL motif. Node *a* plays different roles in two FFL motifs [(**a c)** and (**d, a, and e**) respectively]. Colors represent different roles.

very rare in the biological networks studied here. To calculate the motif-node counts, N_{real}, for every node in the network we calculated the presence of a given node in all different roles of a given motif, *N(i)*. and so for a set of nodes *{1 = 1, . . . , M}* in the network of size *M* it could be framed as:

$$N_{real} = \sum_{i=1}^{M} N_{role1}(i) + N_{role2}(i) + N_{role3}(i) \tag{1}$$

For example, node *a* could either participate in Role 1 (2 outgoing edges, 0 incoming), Role 2 (1 outgoing edge, 1 incoming), and Role 3 (0 outgoing edges, 2 incoming) of FFL motif 1 but at the same time participate in different role of other FFL motif 2 (**Figure 2**).

To test for motif enrichment, we calculated Z-score for every motif type:

$$\frac{N_{real} - \mu_{shuffled}}{\sigma_{shuffled}} \tag{2}$$

where N_{real} is the number of motif counts in the original network, $\mu_{shuffled}$ and $\sigma_{shuffled}$ are the mean and standard deviation of motif counts in the distribution of shuffled networks. Every network was shuffled with a preserved in/out-degree for all nodes until at least 80% of edges in the original network were swapped. To calculate the mean and standard deviation of motif counts in the shuffled networks every network was shuffled 10,000 times. To ensure that the same type of nodes stay connected after shuffling, we calculated the correlations between the degree of connected nodes as weighted average nearest-neighbors degrees (Barrat et al., 2004) in the original and shuffled networks.

Algorithm Description

The FFL-based generation algorithm starts with a nucleation step where an input network is used to find a subnetwork of predefined size (default 20 nodes) with all FFLs connected via shared nodes as in all analyzed networks, almost all FFL motifs share a common node with another FFL motif (**Table 1**). To avoid excessive parameters that could additionally control for in/out degree distribution, the *E. coli* GRN graph was used for the nucleation step. The degree distribution in the "FFL nucleus" sampled from a biological GRN was utilized by the preferential attachment rules as initial conditions to reconstruct a scale-free topology when attaching new edges and nodes to the growing network. The outline of the algorithm is presented graphically (**Figure 3**).

Once the substrate is selected the algorithm adds nodes and edges iteratively such that at every iteration, a candidate node is selected with a random uniform probability. Once selected, one of the four attachment rules (R1, R2, R3, and R4) is applied (**Figure 4**) based on four predetermined probabilities (*p1, p2, p3*, and *p4*) that add up to 1. The iterations are repeated until the required number of nodes in the network is reached.

If the random float number *r1* is less or equal to *p1* then R1 is picked. For the R1 rule we applied the modified preferential attachment algorithm from Abdelzaher et al. (2015a) with a power-law kernel:

$$P(g) = \frac{K_g^{\gamma}}{\sum_{i=1}^{n} K_i^{\gamma}} \tag{3}$$

where K_i denotes node-degree connectivity, *P(g)* is the probability that a new node will be connected to existing node *g*, and ɣ is a parameter that controls the shape of the out-degree distribution.

If *r1* is greater than p_1 then one of the motif-based preferential attachment rules (R2, R3 or R4) is applied, and so $1\text{-}p_1$ corresponds to the desired percentage of nodes that participate in FFL motifs. For R2-R4 rules, one of the already existing FFL motifs is picked based on it's connectivity with the others.

Once the candidate motif and rule are chosen, a new random float number, *r2*, is generated. If $0 < r2 \preccurlyeq p_2$, the R2 rule is applied. In that case, two new edges and one new node will be added to the existing node so the new FFL motif is formed. If $r2 > p_2$, one of the R3 or R4 rules is selected with equal probability. For the R3 rule, two edges are added to nodes in existing FFL motifs to create a new FFL motif. For the R4 rule, and one edge is added between nodes in two existing FFL motifs to create a new FFL motif. If R2 is applied, it creates an FFL motif where one node has

TABLE 1 | Biological GRNs' graph properties.

Organism	# Of nodes	% Of nodes that participate in FFL motifs	% Of FFL motifs sharing nodes with other FFLs	Sparsity	In-degree	Out-degree
E. coli	1,917	37.4	99.1	2.328	1.106	1.222
S. cerevisiae	4,441	27.0	100	2.899	1.421	1.477
M. musculus	2,862	31.5	99.7	2.643	1.274	1.369
H. sapiens	2,456	34.7	99.9	2.944	1.364	1.580

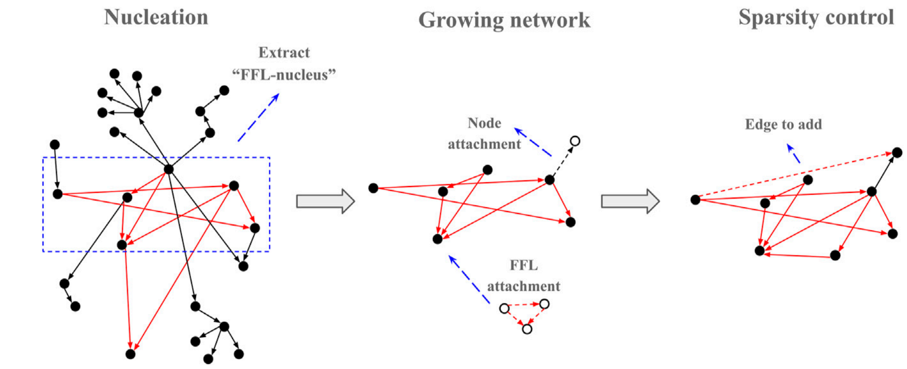


FIGURE 3 | Graphic outline of the FFLatt algorithm. It starts with selecting a seed from the input network, and then iteratively grows the nucleus until the required size is reached. Finally, the sparsity of the network is adjusted according to the sparsity level.

only incoming edges. If R2 or R3 is applied, it creates an FFL motif where all participating nodes have at least one incoming and one outgoing edge. See **Figure 4** for details.

All nodes have to have an out-degree smaller or equal to a threshold K_{max} after which no new outgoing edges are added. If the candidate motif doesn't satisfy the conditions for a chosen FFL attachment rule, another candidate motif picked and this is repeated until a motif is found that meets the rule conditions. If a new motif is created, the library with FFL motifs is updated.

When the desired network size is reached, the algorithm adjusts the sparsity (average number of connections per gene) until it reaches the set sparsity level in terms of average links per node. If the network is too dense, edges are selected for removal based on out-degree node connectivity so that an edge is proportionally more likely to be removed if it is attached to a node with a high out-degree. If the network is too sparse, edges are added to nodes selected proportionally to their out-degree connectivity, connecting them to randomly selected nodes. When network generation is completed, the network is saved as an unweighted directed graph.

Network Generation

For network simulation comparison five algorithms were chosen: FFLatt (developed in present study), GeneNetWeaver (GNW; Schaffter et al., 2011), NetworkX directed scale-free graph algorithm (NetworkX; Hagberg et al., 2008), and sparse uniformly distributed random matrix with and without allowing for feedback loops in the network (DAG and RandG; Guo and Amir, 2021). DAG and RandG matrices were binarized by setting all non-zero elements equal to 1. The NetworkX graph algorithm was modified to control for sparsity as the FFLatt algorithm does, i.e., edges are added to or removed from nodes proportionally to their out-degree node connectivity. For network generation of different sizes with FFLatt, the set of transcriptional interaction graph properties estimated from the *E. coli* transcriptional interaction network (**Table 1**) was used. For each organism, the number of nodes that participate in FFL motif was used to set *p1*, with *p2* equal to (1-*p1*)*0.9, and *p3*=*p4*=(1-*p1*)*0.05 respectively. For network generation of different sizes with other algorithms (except GNW), only network size and sparsity parameters were taken into account as only controllable parameters available. For network generation/subselection with GNW the following (default) parameters were used: *-random-seed*, *--greedy-selection*, *--keep-self-interactions* as well as the size of the subtracted network.

When mimicking the *E.coli* transcription network model, all three-node cycles were disrupted, by removal of one edge, as they are absent in the target network. The removal was done by deleting the outgoing edge of the node with the highest out-degree and an edge was instead attached to a random node with a probability based on the connectivity of each node.

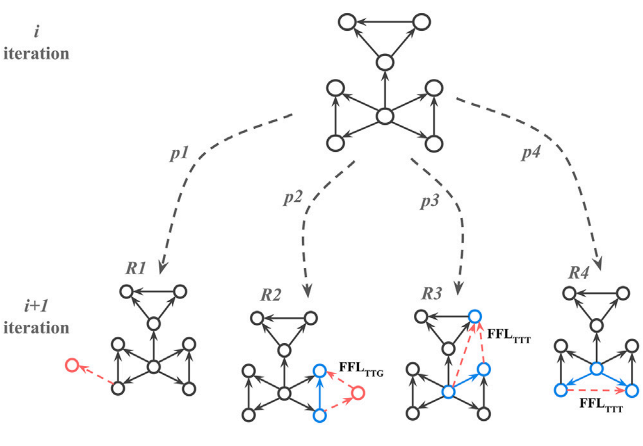


FIGURE 4 | Attachment rules that create FFL motif enriched network; *p1*, *p2*, *p3*, and *p4* correspond to probabilities for choosing rule *R* at the next iteration while growing network. FFL$_{TTG}$ and FFL$_{TTT}$ correspond to different FFL motif types, where G or T (Gene or Transcription factor) indicate whether a participating node has only incoming edges (G), or at least one outgoing edge (T). The red dotted arrows here show new edges added to the network and the solid blue arrows show edges participating in the new FFL motif with the new edges.

To mimic the complete three-node motif profile in biological GNRs in which non-FFL motifs are depleted, an optional motif depletion step can be executed. Here all three-node cycles are converted to FFL motifs by swapping the direction of one of the edges. In addition, up to one tenth of the cascades that do not share edges with FFL motifs were used to create new FFLs by adding an edge. The total number of edges that was used for motif conversion was taken into account when adjusting the network sparsity.

For stability analysis, self-loops (if any) were removed from network graphs generated with above mentioned algorithms before applying the stability analysis model.

Stability Analysis Model

To measure the stability of a network, i.e., how a network graph structure affects the dynamical stability of a gene regulatory interaction model, we utilized the model developed by (Guo and Amir, 2021) that explores how the dynamics of protein and mRNA concentrations control the transcriptional regulation. The model allows for multiple proteins acting on the same gene, and is defined by the authors as:

$$g_i\left(\vec{c}\right) = g_{i0} + \prod_j \left(1 + \gamma_{ij} f_{ij}\left(c_j\right)\right) \tag{4}$$

where g_i and g_{i0} is the effective gene copy number of gene i with and without input of other genes respectively, c_j is the concentration of transcription factor j, and γ_{ij} relates to the strength of the regulation of gene i by c_j. The functional relationship between the transcription factor and target gene, f_{ij}, is modelled as a sigmoid Hill function:

$$f_{ij}\left(c_j\right) = \frac{c_j^h}{K_{ij}^h + c_j^h} \tag{5}$$

where h is the saturation binding coefficient, i.e. the number of proteins required for saturation of binding to DNA, and K is the protein concentration threshold needed to produce a significant increase in mRNA.

The process of gene expression could be described as coupled dynamics of protein and mRNA concentrations. It was shown that in yeast (Zhurinsky et al., 2010) and mammalian cells (Schmidt and Schibler, 1995), the RNA polymerase concentration limits the transcription of mRNA, and the number of ribosomes limits the process of translation. The general transcription model (4) that connects transcription rate of gene i and the number of RNA polymerases can then be described as:

$$\frac{dCm_i}{dt} = k_m \phi_i\left(\vec{c}\right) n - C_{mi} k_p c_r - \frac{C_{mi}}{\tau} \tag{6}$$

$$\frac{dc_i}{dt} = k_p c_r \left(\frac{C_{mi}}{C_{mT}} - c_i\right) \tag{7}$$

where n is the total number of RNA polymerases, C_{mi} is the mRNA concentration of gene i, C_{mT} is the concentration of all mRNAs, ϕ is the gene allocation fraction of $g_i(\vec{c})$ controlled by RNA polymerases active on gene i, k_m is the transcription rate of RNA polymerase, k_p is the translation rate of the ribosome, c_r is the ribosomal concentration, and τ is the degradation rate difference between proteins and mRNA.

We assume that mRNAs degrade much faster than proteins, and as suggested by (Guo and Amir, 2021) we can set $\frac{dCm_i}{dt} \approx 0$ to neglect fast dynamics aiming to simplify the model. By substituting C_{mi} from **6** into **7**, the dynamics of transcription factors concentrations can be simplified as:

$$\frac{dc_i}{dt} \approx k_p c_r \left(\phi_i\left(\vec{c}\right) - c_i\right) \tag{8}$$

In such case, the stability of a steady-state in the dynamical model is dependent on the Jacobian matrix A of size NxN:

$$A = k_p c_r^{ss} \left(M - I\right) \tag{9}$$

where c_r^{ss} is the steady-state ribosomal concentration, M is the gene-gene interaction matrix that consists of γ_{ij} weights of the regulation, I is the identity matrix, and N is the number of genes in the system. The system is stable if the maximal real part of all eigenvalues of M, λ_M, is smaller than 1, i.e., the real part of all eigenvalues of A are negative. As the imaginary part of the eigenvalues is ignored, both oscillatory systems and systems without oscillations around the steady state are considered to be stable.

In contrast to random matrix theory (May 1972) or the generalized models (Gross and Feudel, 2006; Gross et al., 2010), the Jacobian matrix here is not a random matrix nor approximated through studying system bifurcations. In the Guo and Amir model it is derived by applying a knowledge-driven modelling approach which we find convenient for such a well-studied biological process like transcription. We applied this model to all network graphs simulated with different algorithms. Each graph, in a form of adjacency matrix, was supplied as a binary interaction matrix. For each replicate of a different size generated with a given algorithm, we repeated assigning the network graph with link strengths 10 times. To focus on the effect of the GRN structure and FFL content on stability, we forced the distribution of link strengths of all GRNs to be similar. This was done by randomly setting half of the links in the binary interaction matrix to be upregulated and the other half downregulated (setting max (γ_{ij}) and min (γ_{ij}) to 1.5 and −1.5 respectively as boundaries of a normal distribution). In every trial, we first numerically solved for the ribosomal concentration c_r^{ss} with which the system reaches its non-zero steady state with **Eq. 8**. Given c_r^{ss}, A was found such that it only has negative real part eigenvalues using **Eq. 9** by optimizing M, and the highest eigenvalue in λ_M from this solution was compared across networks of different sizes.

RESULTS

Feed-Forward Loop is the Only Enriched Three-Node Motif in Biological Gene Regulatory Networks

Of all possible 3-gene network motifs with 2 or 3 unidirectional links, we found a strong enrichment relative to shuffled networks of the FFL motif in the networks studied here, which are networks that mainly capture transcription factor to target interactions (**Supplementary Table S1**). This was previously shown for *E. coli* (Milo et al., 2002) and *S. cerevisiae* (Lee et al., 2002). We also found that the cascade, uplink, and downlink motifs were consistently and significantly (p-value < 0.05) depleted in all four target networks. To ensure that the shuffling procedure produced topologically similar networks, we verified that the distribution of correlations between the degree of connected nodes was similar for the original and shuffled networks (**Supplementary Figure S1**).

All depleted motifs are 3-node motifs with two edges (**Figure 1**), and these have previously been shown to be significantly depleted in other biological networks, for instance in a protein structure network and a human brain functional network (Mirzasoleiman and Jalili, 2011). However, how the depletion of these motifs contributes to the function of the gene circuitry, and how it relates to the evolution of gene regulatory networks, remains to be answered.

We found that FFL is the only enriched motif, and this was observed in all analyzed networks (**Supplementary Table S1**). Almost all FFL motifs share a common node with another FFL motif, as this fraction ranges from 99.1% in the *E. coli* GRN to 100% in *S. cerevisiae* (**Table 1**). The fraction of nodes that participate in FFL motifs ranges from 27 to 37.4%. This inspired us to develop a GRN generation algorithm that attaches nodes to form connected FFL motifs at a high rate. For each GRN we also calculated the average number of edges per node, here referred to as sparsity, and average in- and out-degrees, and these properties were also used as targets for the algorithm.

Each regulatory interaction in the FFL motif can be either positive or negative, i.e., activating or inhibiting, resulting in eight different types that can act as e.g. accelerators, delay-generators or pulsers (Mangan and Alon, 2003), resulting in different dynamics of gene circuits. Given the wide variety of FFL types and their importance to GRN dynamics, an unsigned *in silico* GRN graph needs a large number of FFLs to accommodate these. A combination of the eight signed types of FFL motifs will in turn reflect a realistic flow of GRN circuits.

We generated a set of GRNs of different sizes from 500 to 1,500 nodes, 10 replicates for each size, using five different algorithms: FFLatt, GNW, NetworkX graph, RandG, and DAG. For each algorithm we analyzed four properties of their GRNs: the number of nodes that participate in FFL motifs, network sparsity, average in- and out-degree within the network. We repeated these simulations for all four organisms, as they have different graph properties. The results for *E. coli* are shown in **Figure 5**, and for the other organisms in **Supplementary Figures S2, S3, and S4**. Each organism-related GRN was used to set the topological parameters in the GRN simulated by FFLatt as described in Methods.

To assess the accuracy of GRN inference algorithms, the topological parameters such as in- and out-degree distribution and sparsity should be controlled when simulating data for benchmark analysis. We found that sparsity as well as out-degree of artificial networks generated with the subnetwork selection based GNW algorithm deviates considerably from the target networks for *E. coli* in sizes 500 and 750 (**Figures 5B,D**), for *S. cerevisiae* in size 500 (**Supplementary Figures S2B, S2D**), and in all sizes for *M. musculus* and *H. sapiens* (**Supplementary Figures S3B, S3D, S4B, and S4D**). While this alone does not indicate a poor performance of the GNW algorithm, it does advocate for the necessity of network generation algorithms to control topological parameters.

More importantly, when subsetting networks from biological GRNs with the GNW algorithm, we obtained a significant underrepresentation of FFL motifs in sizes 500, 750, and 1,000 for *E. coli* (**Figure 5A**) in comparison with FFLatt networks. Similar results were obtained for GRNs of other organisms (**Supplementary Figures S2A, S3A, and S4A**). To confirm and extend these findings, we performed motif enrichment analysis on the simulated networks as well as on biological GRNs (**Figure 6**; **Supplementary Table S1**). This showed that FFL motifs are not significantly overrepresented in GNW networks, but they are highly significantly enriched in the *E. coli* GRN (Z-score 7.4). In networks generated with other algorithms, the FFL motif was also not significantly overrepresented, with the exception of FFLatt whose networks were significantly enriched with Z-scores between 2.95 and 4.98. By default, FFLatt does not deplete other 3-node motifs, and but this is possible with an optional motif depletion step. We explored how this step in combination with various parameter values can mimic the complete 3-node motif distribution profile with the FFL motif enriched, and all other motifs depleted (**Supplementary Table S2**).

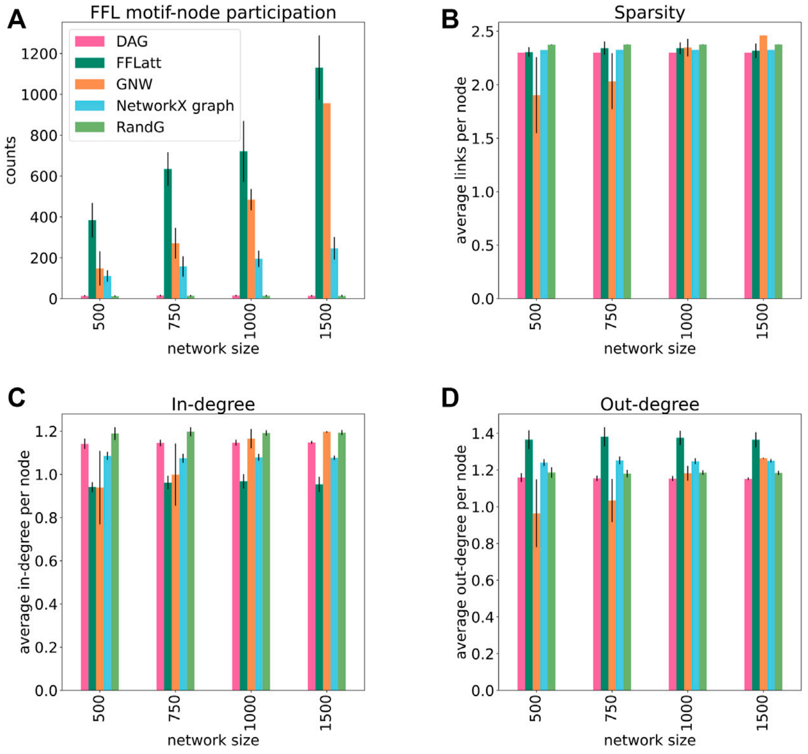


FIGURE 5 | Topological properties of simulated networks (*E. coli*). FFL motif node participation, average sparsity, in- and out-degree distribution in simulated networks. For FFL-motif node participation counts, up to three participations for each node were allowed (in different roles). Each data point was calculated as the average of ten different replicates of each network size. Error bars represent standard deviation.

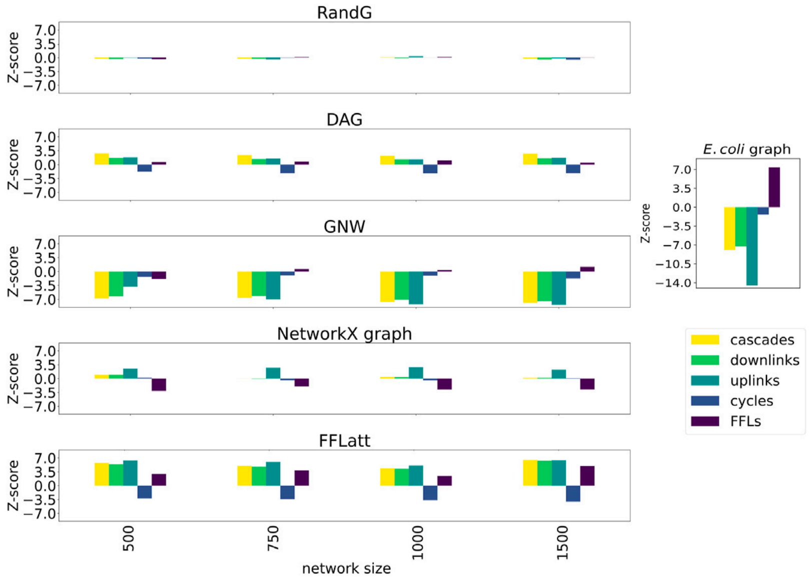


FIGURE 6 | Motif enrichment analysis of 3-node network motifs in simulated networks (*E. coli*). For networks generated with GNW, the *E. coli* RegulonDB (Santos-Zavaleta et al., 2019) database was used. For networks generated with FFLatt, we used the graph properties for *E. coli* specified in **Table 1**. RandG is a random assignment of links and DAG is the same with cycles removed. NetworkX graph GRNs are scale-free. For RandG, DAG, and NetworkX graph GRNs we used the *E. coli* network sparsity.

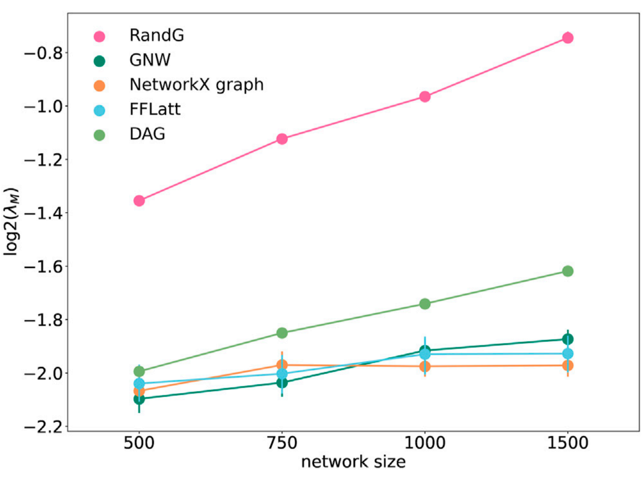


FIGURE 7 | Stability of randomly wired simulated network graphs. λ is the lowest eigenvalue of the interaction matrix M. Each data point was calculated as the average of ten different repeats of overlaying links chosen randomly with strengths from a standard distribution, with corresponding semi-transparent areas indicating the 95% confidence interval.

Topology, Motif Composition, and Network Stability

In biology, random matrix theory, that seeks to understand the properties of matrices with randomly drawn elements, is known from R. May's research on the stability of large biological systems (May 1972). He demonstrated that the stability of a large ecological system depends on satisfying the following inequality:

$$1 > \alpha\sqrt{nC} \tag{10}$$

where α is the average interaction strength, n is the number of species, and C is the density of interactions between them. Therefore, the larger a system gets the more unstable it becomes unless the sparsity and/or interaction strengths are scaled down accordingly. May's approach has been proven to be highly valuable to other biological networks (Aljadeff et al., 2015), including those that aim to describe gene regulations (Prill et al., 2005; Stone, 2018).

It was earlier suggested that motif composition contributes to fault-tolerance in transcriptional networks (Roy et al., 2020). To test if the structural composition is important for stability in artificially generated networks, we analysed the stability of the five network models using the method by Guo and Amir (2021). As expected, all GRNs with fixed sparsity and interaction strengths became more fragile when increasing in size. We found that GRNs with different motif profiles demonstrated different levels of network stability (**Figure 7**). The RandG GRNs that were neither enriched nor depleted with any 3-node motifs (**Figure 6**) were far less stable than the other ones. The DAG GRNs which are generated like RandG GRNs but without cyclic motifs were more stable but still considerably less stable than NetworkX, GNW, and FFLatt GRNs. We note that NetworkX, GNW, and FFLatt GRNs have different network motif abundances, such as either depleted or enriched FFL motifs, and yet they show similar stability. The abundance of the FFL motif alone therefore does not seem to be a major factor for network stability, which is congruent with previous findings about non-importance of the FFL motif to system robustness under random node failure test (Abdelzaher et al., 2015b).

We note that the two lines that represent size-dependent stability of DAG and RandG GRNs have a steeper slope than the other three. This means that as the GRN increases in size, DAG and RandG GRNs become less stable faster than the other three. To find a reason for this, we analyzed the degree distribution of the GRNs. Since RandG and DAG networks are sparse uniformly distributed random binary matrices, their degree distributions do not follow the power-law and therefore they are not scale-free (**Figure 8**). This suggests that a scale-free topology which has been previously found to be central for creating a robust system, protecting the GRN from random mutations (Greenbury et al., 2010), can in fact help gene regulatory systems to reach a stable state after perturbation.

DISCUSSION

Here we present a new algorithm, FFLatt, for generating realistic directed GRN graphs to enable more accurate and authentic performance evaluation of GRN inference methods. The novelty of the presented algorithm is that it generates networks with boosted FFL motifs, which are known to be important for network dynamics. Besides being enriched with the FFL motif, the resulting GRN graphs generated with FFLatt exhibit topological properties similar to experimentally validated biological GRNs.

We show that the motif profile and topological properties of FFLatt network graphs demonstrate a biological stability comparable with other models, such as the NetworkX and GNW algorithms. It is particularly important for network

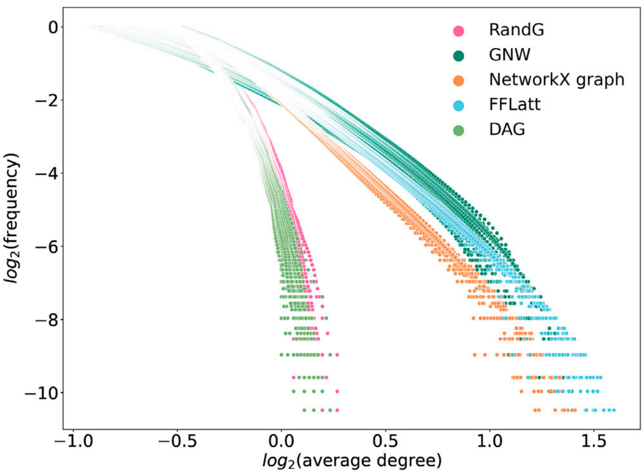


FIGURE 8 | Degree distributions in simulated networks generated by different algorithms. GRNs of sizes 500, 750, 1,000, and 1,500 were used, ten of each size. A power-law distribution should generate a straight line.

inference methods working with steady-state gene expression data as many of them, for instance Least-Squares with Cut-Off (LSCO; (Tjärnberg et al., 2013), LASSO (Tibshirani, 1996; Friedman et al., 2010), LASSO-VAR (Larvie et al., 2016), and GENIE3 (Huynh-Thu et al., 2010) aim to infer a stable static network from steady-state data. To summarize, the FFLatt graph generation algorithm provides an opportunity to simulate biologically meaningful network graphs that can be wired with realistic biological dynamics.

We also noted that the FFLatt networks were enriched with three other motifs: uplinks, downlinks and cascades whereas in GNW networks and biological GRNs these motifs are usually depleted. Sorrells and Jonhson (2015) suggested that in biological GRNs, FFL formation proceeds through a non-adaptive rewiring of gene regulatory regulation which could explain how the abundance of FFLs and the depletion of uplinks, downlinks, and cascades is coupled. The algorithm can be run to allow for depletion of other 3-node motifs while growing the network. However a reason that such depletions are important for network dynamics is yet to be found. A thorough search of the relevant literature did not yield in related articles. We also could not find evidence that different three-node motif profiles affect network stability. NetworkX, GNW, and FFLatt motif profiles are fairly different yet they demonstrated comparable stability across different sizes. While being out of scope for this study, it remains an interesting question how the composition of more complex and higher-order structures known to be present in GRNs (Benson et al., 2016; Gorochowski et al., 2018) could contribute to stability of the system.

In this article we focus on the proof of concept of the FFL attachment algorithm to demonstrate its necessity and feasibility. However, to increase model performance, it could be extended with other parameters. For example, to better capture "small world" (Watts and Strogatz, 1998) structural properties that are known to be present in biological networks, one parameter could be a desired number of biological modules so that within each module the connectivity is higher than in between them. The clustering algorithm should however be biologically motivated so that the connection between modular graph structure and expression dynamics is clear.

Despite a continued uncertainty of how structural properties and functional modularity of GRNs relate to each other, some patterns such as FFLs are known to be key signatures of transcriptional regulation networks. Here we developed a novel algorithm that generates biologically realistic structures of large artificial gene regulatory networks with controlled size, sparsity, topology, and number of FFLs. The implementation executes with reasonable runtimes (**Supplementary Figure S5**). FFLatt graphs are binary and can thus assume a wide range of dynamical structures with signed strengths. They could be used as input to already established tools based on Hill function kinetics such as GNW, which allows for knock-out and knock-down perturbation designs when generating expression data, and some control of the number of nodes, including the number of transcription factors, based on a user-defined input network. To generate expression data it utilizes a non-linear ordinary differential equations (ODE) model for gene expression, and stochastic differential equations (SDEs) for molecular noise generation. Potentially, they could also become a part of future deep learning frameworks that aim to model gene expression from DNA sequence (Zrimec et al., 2020; Avsec et al., 2021). In such frameworks, FFLatt networks could be used as a deep learning model constraint to incorporate prior knowledge of each node participation in FFL motifs. As a result, we believe that it will contribute to future development of benchmarking tools that could fairly and accurately evaluate the performance of GRN inference methods.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. The source code of the algorithm is available at https://bitbucket.org/sonnhammergrni/fflatt.

AUTHOR CONTRIBUTIONS

EZ and OV devised and implemented the algorithm. EZ and TH performed the calculations, analyzed the results, contributed to the discussion, designed the figures, and wrote the manuscript. ES participated in the design and coordination of the study, contributed to the discussion and design of figures, supervised and reviewed the writing of the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the Swedish Research Council Project 2019-04095.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.815692/full#supplementary-material

REFERENCES

Abdelzaher, A. F., Al-Musawi, A. F., Ghosh, P., Mayo, M. L., and Perkins, E. J. (2015b). Transcriptional Network Growing Models Using Motif-Based Preferential Attachment. *Front. Bioeng. Biotechnol.* 3, 157. doi:10.3389/fbioe.2015.00157

Abdelzaher, A. F., Mayo, M. L., Perkins, E. J., and Ghosh, P. (2015a). Contribution of Canonical Feed-Forward Loop Motifs on the Fault-Tolerance and Information Transport Efficiency of Transcriptional Regulatory Networks. *Nano Commun. Networks* 6, 133–144. doi:10.1016/j.nancom.2015.04.002

Ahnert, S. E., and Fink, T. M. A. (2016). Form and Function in Gene Regulatory Networks: The Structure of Network Motifs Determines Fundamental

Properties of Their Dynamical State Space. *J. R. Soc. Interf.* 13, 20160179. doi:10.1098/rsif.2016.0179

Aljadeff, J., Stern, M., and Sharpee, T. (2015). Transition to Chaos in Random Networks with Cell-type-Specific Connectivity. *Phys. Rev. Lett.* 114, 088101. doi:10.1103/PhysRevLett.114.088101

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., et al. (2021). Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions. *Nat. Methods* 18, 1196–1203. doi:10.1038/s41592-021-01252-x

Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M., and Aravind, L. (2006). Comprehensive Analysis of Combinatorial Regulation Using the Transcriptional Regulatory Network of Yeast. *J. Mol. Biol.* 360, 213–227. doi:10.1016/j.jmb.2006.04.029

Barabási, A., and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science* 286, 509–512. doi:10.1126/science.286.5439.509

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The Architecture of Complex Weighted Networks. *Proc. Natl. Acad. Sci.* 101, 3747–3752. doi:10.1073/pnas.0400087101

Benson, A. R., Gleich, D. F., and Leskovec, J. (2016). Higher-Order Organization of Complex Networks. *Science* 353, 163–166. doi:10.1126/science.aad9029

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* 122, 947–956. doi:10.1016/j.cell.2005.08.020

Chen, S., and Mar, J. C. (2018). Evaluating Methods of Inferring Gene Regulatory Networks Highlights Their Lack of Performance for Single Cell Gene Expression Data. *BMC Bioinformatics* 19, 232. doi:10.1186/s12859-018-2217-z

Chouvardas, P., Kollias, G., and Nikolaou, C. (2016). Inferring Active Regulatory Networks from Gene Expression Data Using a Combination of Prior Knowledge and Enrichment Analysis. *BMC Bioinformatics* 17, 181. doi:10.1186/s12859-016-1040-7

Davidson, E. H. (2010). Emerging Properties of Animal Gene Regulatory Networks. *Nature* 468, 911–920. doi:10.1038/nature09645

Emmert-Streib, F., and Dehmer, M. (2018). Inference of Genome-Scale Gene Regulatory Networks: Are There Differences in Biological and Clinical Validations? *Make* 1, 138–148. doi:10.3390/make1010008

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *Plos Biol.* 5, e8. doi:10.1371/journal.pbio.0050008

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Soft.* 33, 1–22. doi:10.18637/jss.v033.i01

Gardner, T., and Faith, J. (2005). Reverse-engineering Transcription Control Networks. *Phys. Life Rev.* 2, 65–88. doi:10.1016/j.plrev.2005.01.001

Gorochowski, T. E., Grierson, C. S., and di Bernardo, M. (2018). Organization of Feed-Forward Loop Motifs Reveals Architectural Principles in Natural and Engineered Networks. *Sci. Adv.* 4, 12. doi:10.1126/sciadv.aap9751

Greenbury, S. F., Johnston, I. G., Smith, M. A., Doye, J. P. K., and Louis, A. A. (2010). The Effect of Scale-Free Topology on the Robustness and Evolvability of Genetic Regulatory Networks. *J. Theor. Biol.* 267, 48–61. doi:10.1016/j.jtbi.2010.08.006

Gross, T., and Feudel, U. (2006). Generalized Models as a Universal Approach to the Analysis of Nonlinear Dynamical Systems. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 73, 016205. doi:10.1103/PhysRevE.73.016205

Gross, T., Stiefs, D., Rudolf, L., and Zumsande, M. (2010). Generalized Modeling of Heterogeneous Nonlinear Networks. *IEICE Proc. Ser.* 44, A2L–A1. doi:10.34385/proc.44.A2L-A1

Guo, Y., and Amir, A. (2021). Exploring the Effect of Network Topology, Mrna and Protein Dynamics on Gene Regulatory Network Stability. *Nat. Commun.* 12, 130. doi:10.1038/s41467-020-20472-x

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring Network Structure, Dynamics, and Function Using NetworkX," in *Proceedings of 7th Python in Science Conference (SciPy2008)*. Editors G. Varoquaux, T. Vaught, and J. Millman, 11–15.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). TRRUST V2: An Expanded Reference Database of Human and Mouse Transcriptional Regulatory Interactions. *Nucleic Acids Res.* 46, D380–D386. doi:10.1093/nar/gkx1013

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* 5, e12776. doi:10.1371/journal.pone.0012776

Iglesias-Martinez, L. F., De Kegel, B., and Kolch, W. (2021). KBoost: A New Method to Infer Gene Regulatory Networks from Gene Expression Data. *Sci. Rep.* 11, 15461. doi:10.1038/s41598-021-94919-6

Kang, Y., Liow, H.-H., Maier, E. J., and Brent, M. R. (2018). NetProphet 2.0: Mapping Transcription Factor Networks by Exploiting Scalable Data Resources. *Bioinformatics* 34, 249–257. doi:10.1093/bioinformatics/btx563

Kaplan, S., Bren, A., Dekel, E., and Alon, U. (2008). The Incoherent Feed-forward Loop Can Generate Non-monotonic Input Functions for Genes. *Mol. Syst. Biol.* 4, 203. doi:10.1038/msb.2008.43

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., et al. (2005). Combinatorial microRNA Target Predictions. *Nat. Genet.* 37, 495–500. doi:10.1038/ng1536

Larvie, J., Sefidmazgi, M., Homaifar, A., Harrison, S., Karimoddini, A., and Guiseppi-Elie, A. (2016). Stable Gene Regulatory Network Modeling from Steady-State Data. *Bioengineering* 3, 12. doi:10.3390/bioengineering3020012

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., et al. (2002). Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804. doi:10.1126/science.1075090

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes Are microRNA Targets. *Cell.* 120, 15–20. doi:10.1016/j.cell.2004.12.035

Mangan, S., and Alon, U. (2003). Structure and Function of the Feed-Forward Loop Network Motif. *Proc. Natl. Acad. Sci.* 100, 11980–11985. doi:10.1073/pnas.2133841100

Mangan, S., Zaslaver, A., and Alon, U. (2003). The Coherent Feedforward Loop Serves as a Sign-Sensitive Delay Element in Transcription Networks. *J. Mol. Biol.* 334, 197–204. doi:10.1016/j.jmb.2003.09.049

Marbach, D., Costello, J. C., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., et al. (2012). Wisdom of Crowds for Robust Gene Network Inference. *Nat. Methods* 9, 796–804. doi:10.1038/nmeth.2016

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., et al. (2006). ARACHNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* 7, S7. doi:10.1186/1471-2105-7-s1-s7

May, R. M. (1972). Will a Large Complex System Be Stable? *Nature* 238, 413–414. doi:10.1038/238413a0

Mendes, P., Sha, W., and Ye, K. (2003). Artificial Gene Networks for Objective Comparison of Analysis Algorithms. *Bioinformatics* 19, ii122–ii129. doi:10.1093/bioinformatics/btg1069

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 824–827. doi:10.1126/science.298.5594.824

Mirzasoleiman, B., and Jalili, M. (2011). Failure Tolerance of Motif Structure in Biological Networks. *PLoS ONE* 6, e20512. doi:10.1371/journal.pone.0020512

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data. *Nat. Methods* 17, 147–154. doi:10.1038/s41592-019-0690-6

Prill, R. J., Iglesias, P. A., and Levchenko, A. (2005). Dynamic Properties of Network Motifs Contribute to Biological Network Organization. *Plos Biol.* 3, e343. doi:10.1371/journal.pbio.0030343

Roy, S., Ghosh, P., Barua, D., and Das, S. K. (2020). Motifs Enable Communication Efficiency and Fault-Tolerance in Transcriptional Networks. *Sci. Rep.* 10, 9628. doi:10.1038/s41598-020-66573-x

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., et al. (2019). RegulonDB V 10.5: Tackling Challenges to Unify Classic and High Throughput Knowledge of Gene Regulation in *E. coli* K-12. *Nucleic Acids Res.* 47, D212–D220. doi:10.1093/nar/gky1077

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico Benchmark Generation and Performance Profiling of Network Inference Methods. *Bioinformatics* 27, 2263–2270. doi:10.1093/bioinformatics/btr373

Schmidt, E. E., and Schibler, U. (1995). Cell Size Regulation, a Mechanism that Controls Cellular RNA Accumulation: Consequences on Regulation of the

Ubiquitous Transcription Factors Oct1 and NF-Y and the Liver-Enriched Transcription Factor DBP. *J. Cel Biol.* 128, 467–483. doi:10.1083/jcb.128.4.467

Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y. (2007). Global and Local Architecture of the Mammalian microRNA-Transcription Factor Regulatory Network. *Plos Comput. Biol.* 3, e131. doi:10.1371/journal.pcbi.0030131

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network Motifs in the Transcriptional Regulation Network of *Escherichia coli*. *Nat. Genet.* 31, 64–68. doi:10.1038/ng881

Siahpirani, A. F., and Roy, S. (2017). A Prior-Based Integrative Framework for Functional Transcriptional Regulatory Network Inference. *Nucleic Acids Res.* 45, gkw963. doi:10.1093/nar/gkw963

Sohka, T., Heins, R. A., Phelan, R. M., Greisler, J. M., Townsend, C. A., and Ostermeier, M. (2009). An Externally Tunable Bacterial Band-Pass Filter. *Proc. Natl. Acad. Sci.* 106, 10135–10140. doi:10.1073/pnas.0901246106

Sorrells, T. R., and Johnson, A. D. (2015). Making Sense of Transcription Networks. *Cell* 161, 714–723. doi:10.1016/j.cell.2015.04.014

Stone, L. (2018). The Feasibility and Stability of Large Complex Biological Networks: A Random Matrix Approach. *Sci. Rep.* 8, 8246. doi:10.1038/s41598-018-26486-2

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tjärnberg, A., Nordling, T. E. M., Studham, M., and Sonnhammer, E. L. L. (2013). Optimal Sparsity Criteria for Network Inference. *J. Comput. Biol.* 20, 398–408. doi:10.1089/cmb.2012.0268

Tsang, J., Zhu, J., and van Oudenaarden, A. (2007). MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals. *Mol. Cel* 26, 753–767. doi:10.1016/j.molcel.2007.05.018

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., et al. (2006). Syntren: A Generator of Synthetic Gene Expression Data for Design and Analysis of Structure Learning Algorithms. *BMC Bioinformatics* 7, 43. doi:10.1186/1471-2105-7-43

Watts, D. J., and Strogatz, S. H. (1998). Collective Dynamics of 'small-world' Networks. *Nature* 393, 440–442. doi:10.1038/30918

Zavlanos, M. M., Julius, A. A., Boyd, S. P., and Pappas, G. J. (2011). Inferring Stable Genetic Networks from Steady-State Data. *Automatica* 47, 1113–1122. doi:10.1016/j.automatica.2011.02.006

Zhang, C., Tsoi, R., Wu, F., and You, L. (2016). Processing Oscillatory Signals by Incoherent Feedforward Loops. *Plos Comput. Biol.* 12, e1005101. doi:10.1371/journal.pcbi.1005101

Zhurinsky, J., Leonhard, K., Watt, S., Marguerat, S., Bähler, J., and Nurse, P. (2010). A Coordinated Global Control over Cellular Transcription. *Curr. Biol.* 20, 2010–2015. doi:10.1016/j.cub.2010.10.002

Zrimec, J., Börlin, C. S., Buric, F., Muhammad, A. S., Chen, R., Siewers, V., et al. (2020). Deep Learning Suggests that Gene Expression Is Encoded in All Parts of a Co-Evolving Interacting Gene Regulatory Structure. *Nat. Commun.* 11, 6141. doi:10.1038/s41467-020-19921-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 February 2022
doi: 10.3389/fgene.2022.815476

Totoro: Identifying Active Reactions During the Transient State for Metabolic Perturbations

Mariana Galvão Ferrarini[1,2], *Irene Ziska*[1,3], *Ricardo Andrade*[1,4], *Alice Julien-Laferrière*[5], *Louis Duchemin*[1], *Roberto Marcondes César Jr.*[4], *Arnaud Mary*[1,3], *Susana Vinga*[6] *and Marie-France Sagot*[1,3]*

[1]*Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, CNRS, Université de Lyon, Université Lyon 1, Villeurbanne, France,* [2]*Univ Lyon, INRAE, INSA-Lyon, BF2I, UMR 203, Villeurbanne, France,* [3]*INRIA Grenoble Rhône-Alpes, Villeurbanne, France,* [4]*Institute of Mathematics and Statistics (IME), University of São Paulo, São Paulo, Brazil,* [5]*Soladis GmbH, Basel, Switzerland,* [6]*INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal*

Motivation: The increasing availability of metabolomic data and their analysis are improving the understanding of cellular mechanisms and how biological systems respond to different perturbations. Currently, there is a need for novel computational methods that facilitate the analysis and integration of increasing volume of available data.

Results: In this paper, we present TOTORO a new constraint-based approach that integrates quantitative non-targeted metabolomic data of two different metabolic states into genome-wide metabolic models and predicts reactions that were most likely active during the transient state. We applied TOTORO to real data of three different growth experiments (pulses of glucose, pyruvate, succinate) from *Escherichia coli* and we were able to predict known active pathways and gather new insights on the different metabolisms related to each substrate. We used both the *E. coli* core and the iJO1366 models to demonstrate that our approach is applicable to both smaller and larger networks.

Availability: TOTORO is an open source method (available at https://gitlab.inria.fr/erable/totoro) suitable for any organism with an available metabolic model. It is implemented in C++ and depends on IBM CPLEX which is freely available for academic purposes.

Keywords: metabolomics, metabolic networks, transient state, metabolic perturbation, omics integration

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Philipp Wendering,
University of Potsdam, Germany
Samuel Seaver,
Argonne National Laboratory (DOE),
United States

****Correspondence:***
Marie-France Sagot
marie-france.sagot@inria.fr

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 November 2021*
Accepted: *24 January 2022*
Published: *21 February 2022*

Citation:
Galvão Ferrarini M, Ziska I, Andrade R, Julien-Laferrière A, Duchemin L, César RM, Mary A, Vinga S and Sagot M-F (2022) Totoro: Identifying Active Reactions During the Transient State for Metabolic Perturbations. Front. Genet. 13:815476. doi: 10.3389/fgene.2022.815476

1 INTRODUCTION

The increasing availability of metabolomic data and their analysis are currently enhancing our knowledge on diverse biological mechanisms and elucidating how cells and organisms respond to different perturbations (Sevin et al., 2015). Metabolomics can be used to obtain a metabolic profile that characterizes the physiological response of a cell, tissue or organism to a stress or to a general perturbation (Roessner and Bowne, 2009), and experiments ranging from shorter-term responses (such as stress response programs) to longer-term responses (such as acclimation) are broadly available for diverse species. Different network-based strategies for metabolomic data analysis have been recently reviewed in (Perez de Souza et al., 2020) and amongst others, such strategies can be used to establish associations between metabolites or to integrate them into metabolic pathways.

Metabolic profiles are often analyzed and interpreted with the help of bioinformatic software such as METEXPLORE (Cottret et al., 2018; Frainay et al., 2019), METABOANALYST (Xia et al., 2015; Chong et al., 2018) or 3OMICS (Kuo et al., 2013) that can identify the set of metabolites with a significant change in their concentration. The metabolomic data are projected on the annotated metabolic pathways in order to highlight the processes that may be linked to the observed changes. The

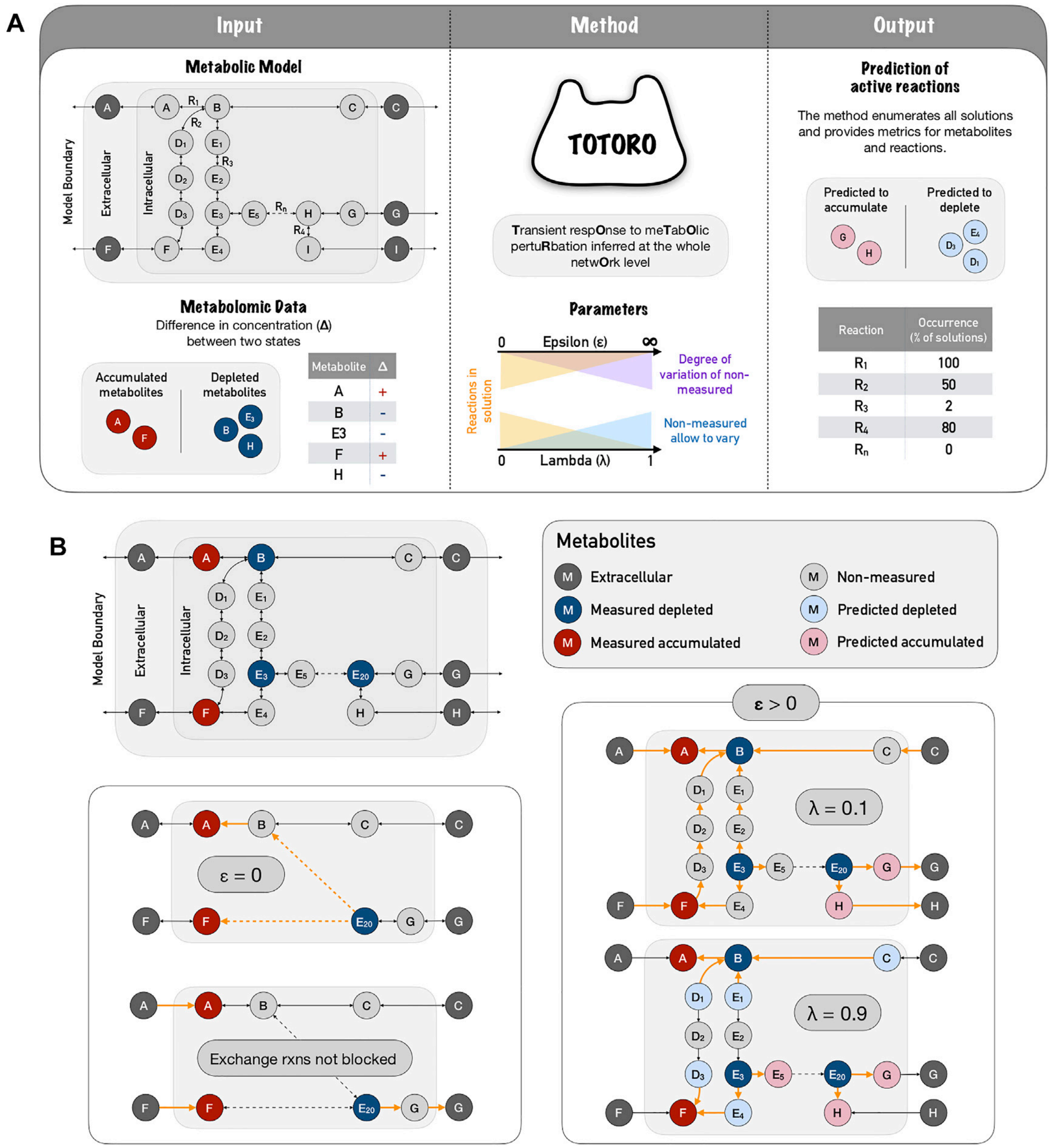


FIGURE 1 | TOTORO method explained. **(A)** TOTORO is able to integrate a metabolic model with metabolomic data in order to predict active reactions during the transient state between two conditions (or simply after a perturbation). The inputs of TOTORO are an SBML metabolic model, and a list of intervals for the difference in concentration (Δ) for each measured metabolite. In the metabolic model panel, grey circles depict metabolites and arrows depict reactions. In the metabolic data panel, accumulated metabolites are depicted in red circles, depleted metabolites are depicted in blue circles. The method TOTORO then requires two additional user-defined parameters to fine tune the results, namely λ and ϵ. TOTORO provides as output the predicted variation of metabolites and reactions that were most likely active between the two states in each enumerated solution as well as metric files grouping all enumerated solutions. In the figure, reaction occurrence is depicted as a percentage in all enumerated solutions. **(B)** The fine-tuning of parameters λ and ϵ are provided within a toy network, in which active reactions are showed in orange and dashed arrows indicate several reactions in a row. When we don't allow an accumulation of non-measured metabolites ($\epsilon = 0$), the method will try to connect the input deltas of distant and possibly unrelated metabolites; and in the case exchange reactions are not blocked, the method will most likely propagate the accumulation or depletion towards outside of the boundaries of the model. When accumulation is allowed ($\epsilon > 0$) a low lambda ($\lambda = 0.1$) will favor solutions in which fewer non-measured metabolites accumulate or deplete, and will include a larger number of reactions within the solutions. As we raise the parameter lambda ($\lambda = 0.9$), we favor local and smaller solutions.

aforementioned software also try to integrate different kinds of omic data (such as transcriptomic, metabolomic or proteomic data) in order to give a deeper understanding of the studied mechanisms (Cambiaghi et al., 2017). Different approaches were reviewed in (Rosato et al., 2018; Ivanisevic and Want, 2019; Stanstrup et al., 2019) and software for the enrichment analysis of metabolomic data were evaluated and their results compared in (Marco-Ramell et al., 2018). However, metabolic pathways have subjective definitions and can differ between databases (Ginsburg, 2009). Additionally, this kind of analysis can make it hard to identify the connections between metabolites since they can be part of many pathways and it is thus possible to miss paths which traverse several biological pathways.

Another approach is to use graph-based methods that allow to consider the whole metabolism as an integrated system focusing on the parts that are connecting the metabolites of interest. Usually, these methods rely mainly on the network structure, chemical information and on an input list of metabolites (Frainay and Jourdan, 2017). Another example can be seen in (Acuña et al., 2012; Milreu et al., 2014), with the enumeration of metabolic stories. A metabolic story is defined by the authors as the set of reactions that summarize the flow of matter from a set of source metabolites to a set of target metabolites and is characterized as a maximum directed acyclic subgraph connecting the metabolites of interest. One of the drawbacks of this approach is that a metabolic story is acyclic and thus, it is not possible to obtain sets of reactions that contain cycles. Nevertheless, cycles are common in metabolic networks and this assumption does not reflect reality. Additionally, the method does not take into account the stoichiometry of the reactions, which can lead to a set of unfeasible reactions in practice.

Metabolite concentrations have also been used to assess the responses to small perturbations in the context of constraint-based models (Palsson, 2000; Covert and Palsson, 2003; Klamt et al., 2014), and has been reviewed in detail by (Topfer et al., 2015). While standard flux balance analysis (FBA) tries to predict the flux distribution for one specific steady-state condition, dynamic FBA, as described in (Mahadevan et al., 2002), has been extensively used in smaller models to predict the evolution of the fluxes and of the metabolite concentrations over time. In (Reznik et al., 2013), the authors provide a method derived from the classical FBA framework, and showed that the variables of the dual problem (the so-called shadow prices, which correspond to the sensitivity of FBA to imbalances in the flux) can indicate if a metabolite is a growth-limiting metabolite in FBA. In (Bordbar et al., 2017) the authors describe the unsteady-FBA method (uFBA), created to integrate dynamic time-course metabolomics with a constraint-based metabolic model, allowing a bypass into the steady-state assumption for intracellular metabolites that are measured. In (Rohwer and Hofmeyr, 2008; Christensen et al., 2015), methods are presented to identify regulatory metabolites and paths by varying *in silico* their known concentrations in a measured steady-state using supply-demand analysis. Therefore, these methods are based on the response of an organism to a relatively small perturbation and on the influence of the metabolite concentrations on the reaction rates of the system to return to the original equilibrium.

In this paper, we focus not on the metabolite pools in one condition but on the difference of the obtained measurements between two conditions, which could be measured either within shorter or longer timeframes, depending on the biological question to be addressed. We also do not need neither comprehensive time-course datasets nor coupled data from the relative expression of genes or proteins, which are much harder to obtain. Our main hypothesis is that the difference of metabolite pools between two metabolic states can provide information on the transient state, that is, on the transition between the two measured conditions.

Similar problems have been studied in the literature. In (Sajitz-Hermstein et al., 2016), the authors provide a method (IREMETFLUX) to integrate relative metabolomic measurements in order to make predictions about differential fluxes. They use a constraint-based approach which minimizes the distance between the two flux vectors of the two different states based on the ratio between the measured metabolite concentrations in both conditions. For both states, steady-state is assumed for the flux vectors. However, the authors identify differential fluxes between the two conditions whereas we aim at finding reactions that are likely active during the transient state. In (Case et al., 2016), the authors investigated reachability problems in chemical reaction networks. Given two different states of the network, the goal is to identify a path that leads the network from the first state to the second one. They prove that this problem can be solved in polynomial time. However, they also discuss that a variant of this problem in which the maximum size of the path is fixed is more difficult to solve. Our approach overcomes this limitation at the same time that it minimizes the number of active reactions in the solutions, since we are interested in identifying only the parts of the network that are potentially active during the transient state. Even though other methods could be adapted to answer this problem, our objective is much simpler, requiring less computational complexity. By reformulating our problem in a simpler way we can also address larger genome-scale metabolic models, instead of focusing on smaller portions of the metabolism (e.g., core models).

We use constraint-based modeling to enumerate sets of reactions that explain the changes in concentrations for some measured metabolites, i.e., how the system moved from a state to another. We implemented our approach in a software we called TOTORO (for "Transient respOnse to meTabOlic pertuRbation inferred at the whole netwOrk level"), that is publicly available at https://gitlab.inria.fr/erable/totoro, along with the test datasets presented in this study. It is implemented in C++ and depends on IBM CPLEX which is freely available for academic purposes. We also tested our method with data from pulse experiments with different carbon sources (glucose, pyruvate and succinate) in *Escherichia coli*.

2 METHODS

A metabolic network can be represented as a weighted directed hypergraph $H(\mathcal{V}, \mathcal{R}, \mathcal{S})$ where $\mathcal{V}$ is the set of vertices, $\mathcal{R}$ the set of hyperarcs and $\mathcal{S}$ the stoichiometric matrix representing weights on the hyperarcs. Each $c \in \mathcal{V}$ represents a metabolite of the network and each hyperarc $r \in \mathcal{R}$ a reaction that connects two sets of disjoint metabolites $Subs_r$, $Prod_r$ with $Subs_r, Prod_r \subseteq \mathcal{V}$. To each hyperarc, a set of weights is associated representing the stoichiometric coefficients of the metabolites participating to the corresponding reaction. These weights are given by the stoichiometric matrix $\mathcal{S}$ which is a $m \times n$ matrix where each column represents a reaction and each row a different metabolite. It contains the stoichiometric coefficients which are positive if a metabolite is produced by a reaction and negative if it is consumed.

The set $X \subseteq \mathcal{V}$ contains all measured metabolites. The metabolomic data is given as a list which, for each measured metabolite in X, contains an interval. This interval describes by how much the internal metabolite concentration changed between two different states. Usually, small deviations for the measurements are available which can be used to calculate the minimum and the maximum possible difference between the internal metabolite concentrations. Furthermore, all reversible reactions of the network are split into forward and backward reactions.

We are interested in solving the following problem: Given a network H and a list containing the changes for some metabolite concentrations before and after a perturbation, we want to identify sets of reactions that were involved in diverting the system from the initial state before the perturbation to the state after the perturbation (**Figure 1A**). Here, we present a constraint-based approach to solve this problem where the change of concentrations (Δ) between two states is represented as an interval.

2.1 Core Method

The variation of the concentrations in time of the metabolites in X can be written as:

$$\frac{dX}{dt} = (\mathcal{S} \cdot v)_X. \tag{1}$$

In this equation, v is a flux vector and the $(\cdot)_X$ operator means that only the entries of the vector corresponding to the metabolites in X are taken into account. We use $[X]_t$ to denote the concentration for the metabolites in X at time point t. Considering two points t_0 and t_1 in time and $\Delta_X = [X]_{t_1} - [X]_{t_0}$, one can write:

$$\Delta_X = \mathcal{S} \cdot \varphi. \tag{2}$$

In this case, each entry of the vector φ can be interpreted as the overall number of moles that passed through the reaction j during the time interval $[t_0, t_f]$ which corresponds to the area under the reaction rate curve in this time interval:

$$\varphi_j = \int_{t_0}^{t_1} v_j(t) \cdot dt. \tag{3}$$

Due to biological and technical variability that can arise from different replicates of the same experiment, we assume that the measured variations in concentrations of the metabolites in X are represented by an interval rather than using a fixed number:

$$\Delta_X = [\Delta_X^{\min}, \ \Delta_X^{\max}]. \tag{4}$$

Furthermore, for the non-measured metabolites, we do not know if their concentration changed or not. Therefore, similarly to the approach of UFBA (Bordbar et al., 2017) and their 'node relaxation' to allow for changes in non-measured metabolites, we assume that a variation (ϵ) is possible for all non-measured metabolites $\bar{X} = \mathcal{V} \backslash X$:

$$\Delta_{\bar{X}} = [\epsilon^{\min}, \ \epsilon^{\max}]. \tag{5}$$

Based on these assumptions, we can model the production or consumption of metabolites between two states by the following constraints:

$$\begin{aligned} &\Delta^{\min} \le \mathcal{S} \cdot \varphi \le \Delta^{\max} \\ &0 \le \varphi_j \le u_j \qquad\qquad \forall j \in \mathcal{R}. \end{aligned} \tag{6}$$

All φ_j are positive and have an upper bound u_j. We have that $\Delta^{\min}$ is a vector composed of $\Delta_X^{\min}$ and $\epsilon^{\min}$ while $\Delta^{\max}$ is composed of $\Delta_X^{\max}$ and $\epsilon^{\max}$.

As showed above, in our formulation, the variable φ can only be zero or have a positive value. For this, we use an additional constraint as explained in **Section 2.2** in order to prevent both forward and reverse senses of reversible reactions from being picked in any given solution. However, this means that we do not know if the activity of the corresponding reaction was increased or decreased during the shift compared to the initial steady state. We only know that if φ_j is zero in the solution, reaction j is proposed as inactive during the shift while if φ_j has a non-zero value, reaction j is proposed as active during the shift. Hence, we are only interested in the reactions that have a non-zero φ because we want to identify the part of the metabolic network that was active during the metabolic shift. These reactions are represented by the support of the vector φ.

2.2 Minimizing the Number of Reactions and the Variation of the Concentrations for the Non-Measured Metabolites

Since the number of possible paths that can explain the measured metabolic shifts can be very large, we will focus on finding the smallest solutions with regard to the number of active reactions that still explain the metabolic shift. This corresponds to the parsimonious assumption that the fewest possible resources are used or the smallest changes are made. Thus, we are interested in identifying minimum sets of reactions that play a major role in the metabolic shift. For each reaction j, a binary variable y_j is then introduced that is set to zero if and only if the corresponding φ_j is zero and therefore, the reaction is not part of the solution. In this way, these variables will correspond to the support vector of φ and it will be sufficient to minimize their sum:

$$\begin{aligned} &y_j = 0 \leftrightarrow \varphi_j = 0 \quad\ \forall j \in \mathcal{R} \\ &y_j \in \{0, 1\}. \end{aligned} \tag{7}$$

Additionally, to prevent that both a reaction j and its reversible $\bar{j}$ can be picked at the same time for one solution, the following constraint is used:

$$y_j + y_{\bar{j}} \le 1 \quad \forall (j, \bar{j}) \in \mathcal{R}. \tag{8}$$

To minimize the number of reactions that are part of the solution, the objective function is written as:

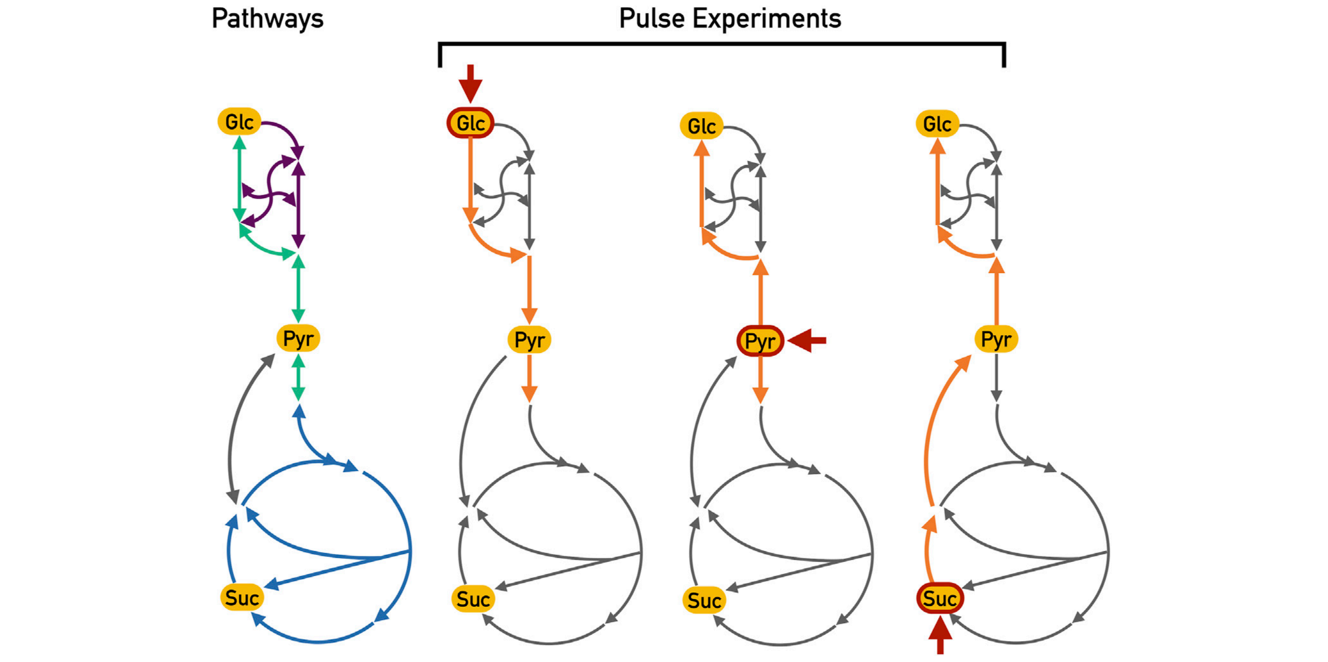


FIGURE 2 | Expected active reactions for different pulse experiments. These essential reactions along with their expected directions are highlighted in orange whereas other non-essential reactions (but which nonetheless could be chosen) are depicted in grey. Each pulse is indicated by the short red arrow (Glc: glucose; Pyr: pyruvate and Suc: Succinate). During the glucose pulse, the glycolysis reactions (depicted in green) should be active in order to generate ATP from the hydrolysis of glucose. On the other hand, the pyruvate and succinate pulse experiments should show gluconeogenesis activation (also depicted in green but in the opposite sense), generating glucose-6-phosphate from these two carbon sources. Furthermore, the TCA cycle (depicted in blue) can be fed from pyruvate during the pyruvate and glucose pulses. During the succinate pulse, the overflow in the TCA cycle should lead to the production of pyruvate with a subsequent activation of gluconeogenesis to produce biomass precursors. The pentose phosphate pathway (depicted in purple) is most likely active in all pulses in order to generate biomass precursors; however, since this pathway is a mere interconversion of carbohydrates, there is no particular expectation as to the actual direction of these reactions.

$$\min \sum_{j=1}^{m} y_j. \tag{9}$$

However, we are not only interested in minimizing the number of reactions in the solution but also in minimizing the variation in concentration for the non-measured metabolites $\bar{X}$. Since the measured compounds are usually the more important ones for analyzing the biological experiment, it is reasonable to aim for solutions where other compounds do not accumulate or deplete a lot. This leads to the following minimization:

$$\min \sum_{i\in\bar{X}} |(\mathcal{S}\cdot\varphi)_i|. \tag{10}$$

On the other hand, we are trying to explain as much change in the concentration as possible for the measured metabolites:

$$\max \sum_{i\in X} |(\mathcal{S}\cdot\varphi)_i|. \tag{11}$$

To combine both ideas in one objective function, a weight λ is used for both objectives:

$$\min \lambda \sum_{j=1}^{m} y_j + (1-\lambda)\sum_{i\in\bar{X}} |(\mathcal{S}\cdot\varphi)_i| - (1-\lambda)\sum_{i\in X} |(\mathcal{S}\cdot\varphi)_i|. \tag{12}$$

The value for λ should lie between 0 and 1. Finding a good balance between these two objectives can be challenging but necessary to identify meaningful biological solutions (for a schematic representation of TOTORO, see **Figure 1A**). A toy network example is provided in **Figure 1B** to show the influence of parameters λ and ϵ on the solutions. This will be further discussed in the following sections.

Summing up, the mixed-integer linear program (MILP) that is implemented in our software TOTORO is the following:

$$\begin{aligned}
\min_{\varphi,y}\ & \lambda \sum_{j=1}^{m} y_j + (1-\lambda)\sum_{i\in\bar{X}} |(\mathcal{S}\cdot\varphi)_i| - (1-\lambda)\sum_{i\in X} |(\mathcal{S}\cdot\varphi)_i| \\
s.t\ & \Delta^{\min} \le \mathcal{S}\cdot\varphi \le \Delta^{\max} \\
& 0 \le \varphi_j \le u_j \qquad \forall j \in \mathcal{R} \\
& y_j = 0 \leftrightarrow \varphi_j = 0 \qquad \forall j \in \mathcal{R} \\
& y_j + y_{\bar{j}} \le 1 \qquad \forall (j,\bar{j}) \in \mathcal{R} \\
& y_j \in \{0,1\}; \lambda \in (0,1); u_j, \varphi_j \in \mathbb{R}.
\end{aligned} \tag{13}$$

2.3 Enumerating Different Solutions

To enumerate different solutions, once a solution is found, it must be excluded for the next iteration. Two solutions are different if they do not contain the same reactions. We are using the following constraint where $y^{\star}$ is a previously found solution vector:

$$\sum_{j\in\mathcal{R}:\ y_j^{\star}=1} y_j \le \sum_{j=1}^{m} y_j^{\star} - 1. \tag{14}$$

This prevents that the exact same combination of reactions gets chosen again. Afterwards, we can solve the updated MILP again to compute a different solution. We repeat this process until no more new solutions can be found or until a desired number of solutions has been computed.

2.4 Dealing With Source/Sink Reactions and Non-Measured Metabolites

Source and sink reaction (i.e., reactions that have only products or only substrates) of the network should be blocked to avoid that changes in the concentration are just transferred outside of the network where they cannot be taken into account by the objective function. However, no information is lost if source and sink reactions are blocked. If the substrates of a sink reaction are accumulated or the products of a source reaction are depleted in a solution, this indicates that the corresponding source/sink reaction is active. Their use is limited by the chosen ϵ but it can be set to a very low or large value to imitate an infinite source or sink. Hence, specific sources or sinks can be added to the problem by specifying a large negative $\Delta^{\min}$ or a large positive $\Delta^{\max}$ for certain metabolites, but the method will remain robust to small variations, as long as the range of this parameter remains within a similar order of magnitude of the values of the measured metabolites.

However, if the minimization of the number of active reactions is prioritized ($\lambda \approx 1$) and the value of ϵ for the non-measured metabolites is higher than the one for the measured metabolites, the changes in concentration of the measured metabolites can simply be distributed to (accumulated on or taken from) the nearby non-measured metabolites (**Figure 1B**, $\epsilon > 0$, $\lambda = 0.9$) and prevents that larger sub-hypergraphs are chosen (which would instead connect several measured metabolites and explain how the depletion of one measured metabolite leads to the accumulation of another measured metabolite, or vice-versa). However, this can be addressed by decreasing the value of λ in the objective function and thereby giving more weight to the portion of the function that minimizes the accumulation in non-measured metabolites **Figure 1B**, $\epsilon > 0$, $\lambda = 0.1$). This should result in solutions that are larger but that connect the measured metabolites better than when only the number of reactions is minimized. Furthermore, based on other experimental data, the user might choose smaller values of ϵ, or constrain it to the highest measured metabolite to further restrict the accumulation/depletion of the non-measured metabolites.

3 RESULTS

To evaluate our approach, we used data from different pulse experiments with different carbon sources in *E. coli* as presented in (Taymaz-Nikerel et al., 2013). The authors measured the internal concentrations for several metabolites for a glucose baseline and for glucose, pyruvate and succinate pulse experiments. These data were used to apply the method on the *E. coli* core model (Orth et al., 2010) and the *E. coli* iJO1366 model (Orth et al., 2011) available from the BiGG database (King et al., 2015b). The *E. coli* core model consists of 72 metabolites and 95 reactions, the *E. coli* iJO1366 model of 1,805 metabolites and 2,583 reactions.

We were interested in the difference between the glucose baseline and the pseudo-steady state which was achieved in about 30–40s after each pulse. In (Taymaz-Nikerel et al., 2013), the authors provided the internal concentrations for the baseline, including the deviations for their measurements and the fold changes for the three different pseudo-steady states which we used to calculate the internal concentrations for each pseudo-steady state. In (Taymaz-Nikerel et al., 2013), deviations for the measured concentration of the glucose baseline are given that were derived from several replicates of the same experiment. We used them to be able to calculate the minimum difference $\Delta_X^{\min}$ and maximum difference $\Delta_X^{\max}$ in the concentrations between the glucose baseline and each pseudo-steady state. A detailed explanation can be found in the **Supplementary Material Section S1.1**. The calculated $\Delta_X^{\min}$ and $\Delta_X^{\max}$ for all three pulse experiments can be found in the **Supplementary Tables S1–S3**.

We used all measured metabolites that are present in the network and that had a significant change in their concentration as input. It should be noted that a change for each given metabolite must be either positive or negative. For further details, see the **Supplementary Material Section S1.1**.

Furthermore, source and sink reactions cannot be chosen as part of the solution and therefore glucose, pyruvate and succinate were added as sources for the corresponding pulse experiments. Oxygen was added as another source because in (Taymaz-Nikerel et al., 2013), the authors identified increased oxygen uptake rates during the pulse experiment. To allow unlimited growth, the biomass was added as sink.

The expected active reactions in the core metabolism of *E. coli* are displayed in **Figure 2** for each pulse experiment.

3.1 *E. coli* Core Model

At first, the method was applied using the *E. coli* core model. To better understand how the different parts of our model impacted the solutions, we did several runs with different values for λ (0.1, 0.5, and 0.9) and ϵ (5 and 10) for each pulse experiment. Although a single solution should be enough to identify some pathways responsible for the shift, we wanted to see if we could also obtain alternative pathways. Furthermore, we wanted to investigate how the solutions evolve when they are slightly suboptimal. For each different parameter setting, 100 different solutions were therefore enumerated. The results are displayed using *Escher* (King et al., 2015a) in the Supplementary Figures S1to S18.

In general, we could observe that solutions with $\lambda = 0.1$ were preferable since usually the goal is to have a final solution which is overall more connected. In this way, we were able to extract connected sub-hypergraphs that resemble complete biological pathways which played a role during the metabolic shifts. This was the case for all three pulse experiments. A higher λ led to solutions that were less connected since the optimizer prioritizes solutions with fewer active reactions, and depending on the case, it might be harder to interpret these solutions biologically. Nevertheless, the user is able to fine-tune the number of reactions in the final solution and the degree of connectivity (for instance, if the goal is to highlight only parts of the complete metabolic network instead of finding a connected sub-hypergraphs).

By adjusting the parameters λ and ϵ, TOTORO could propose connected sub-hypergraphs for all three pulse experiments. The

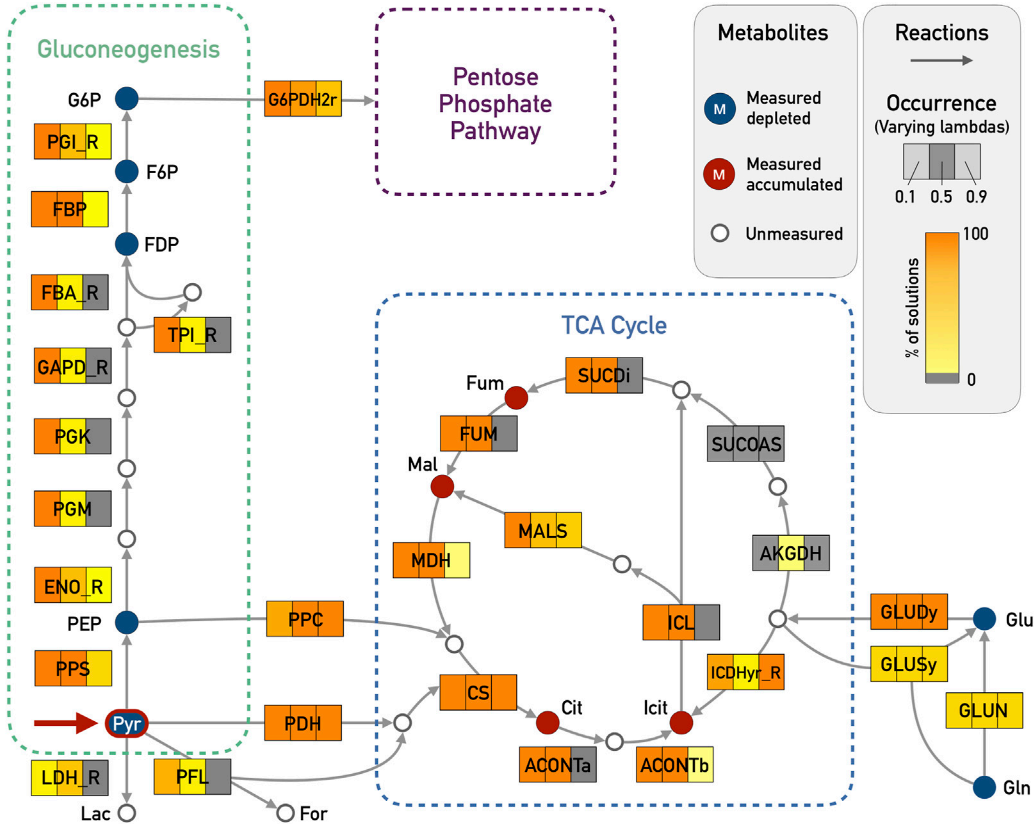


FIGURE 3 | *E. coli* core model - Results for Gluconeogenesis and TCA Cycle in the pyruvate pulse (red arrow in Pyr) with $\epsilon = 5$ and varying λ (0.1, 0.5, and 0.9). The metabolites that were given as input are highlighted in blue if the corresponding input deltas were below zero and red if they were above zero. Reactions that are highlighted in orange were chosen in almost all of the enumerated solutions, while light yellow corresponds to very few occurrences (less than 5%). Reactions in gray were not chosen in any solution. The expected reactions of the gluconeogenesis and part of the TCA cycle are active in all 100 solutions for $\lambda = 0.1$. The reversible reactions of the gluconeogenesis were chosen in the correct direction. For simplicity reasons, side compounds and cofactors were excluded from the figure. Abbreviations for metabolites: G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FDP, fructose-biphosphate; PEP, phosphoenolpyruvate; Pyr, pyruvate; Lac, lactate; For, formate; Mal, malate; Fum, fumarate; Cit, citrate; Icit, isocitrate; Glu, glutamate; Gln, glutamine; Abbreviations for reaction names (_R indicates the reverse direction of a reversible reaction within the original model): G6PDH2r, glucose 6-phosphate dehydrogenase; PGI, glucose-6-phosphate isomerase; FBP, fructose-bisphosphatase; FBA_R, fructose-bisphosphate aldolase; TPI, triose-phosphate isomerase; GAPD, glyceraldehyde-3-phosphate dehydrogenase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; ENO, enolase; PPS, phosphoenolpyruvate synthase; LDH, D-lactate dehydrogenase; PFL, pyruvate formate lyase; PPC, phosphoenolpyruvate carboxylase; PDH, pyruvate dehydrogenase; CS, citrate synthase; ACONTa, Aconitase (half reaction A); ACONTb, Aconitase (half reaction B); ICDHyr, Isocitrate dehydrogenase; AKGDH, 2-Oxoglutarate dehydrogenase; SUCOAS, Succinyl-CoA synthetase; SUCDi, Succinate dehydrogenase; FUM, fumarase; MDH, malate dehydrogenase; ICL, isocitrate lyase; MALS, malate synthase; GLUDy, glutamate dehydrogenase; GLUSy, glutamate synthase; GLUN, glutaminase.

predicted solutions did not use co-factors as shortcuts through the network. We therefore did not modify our method further to treat co-factors separately.

3.1.1 Pyruvate Pulse

For the pyruvate pulse, we expected that the activity of the TCA cycle would increase and that reactions for gluconeogenesis would be active (see **Figure 2**). Both observations could be reproduced with a $\lambda = 0.1$ (see **Figure 3** for a comparison of the values of λ), while higher values of lambda constrained the solutions locally around the measured metabolites. For $\lambda = 0.9$, neither the TCA cycle nor the gluconeogenesis pathway were proposed to be active. Setting λ to 0.5 already improved the results: the TCA cycle was proposed as active but the complete gluconeogenesis pathway was only recovered in less than 50% of the solutions.

The four measured metabolites citrate, isocitrate, L-malate and fumarate had positive input deltas and could thus be used as sinks. The results showed how the TCA cycle can be fed from pyruvate either by the phosphoenolpyruvate carboxylase (PPC) or by the combination of pyruvate dehydrogenase (PDH) and citrate synthase (CS). Furthermore, the pathway from pyruvate to glucose 6-phosphate (G6P) was active in 100% of solutions for $\lambda = 0.1$. The pathway from pyruvate to G6P contains nine reactions including seven reversible ones: glucose-6-phosphate isomerase (PGI), fructose-bisphosphate aldolase (FBA_R), triose isomerase (TPI), glyceraldehyde-3-phosphate dehydrogenase (GAPD), phosphoglycerate kinase (PGK), phosphoglycerate mutase (PGM) and enolase (ENO). Especially here, it is important to state that all these reversible reactions were predicted in the correct direction going from pyruvate towards G6P. The core

TABLE 1 | Comparison of different objective values for the best runs for each experiment. Since we are not fixing the objective value of the first solution in our optimization problem, the objective values for the subsequent solutions can be worse. In this table, we are comparing the difference in the objective values between the first solution and the 100th solution. In addition to the absolute differences, also the percentage of how much the objective value worsened compared to the first solution is displayed. The underlying optimization problem is a minimization problem. Therefore, smaller objective values are better.

Pulse experiment	1st sol.	100th sol.	Abs. diff.	%
Pyruvate ($\lambda = 0.1$, $\epsilon = 5$)	−32.1394	−30.6635	1.48	5.5
Glucose ($\lambda = 0.1$, $\epsilon = 1.2$)	5.3830	6.5582	1.18	21.8
Succinate ($\lambda = 0.1$, $\epsilon = 5$)	−158.1770	−157.5760	0.60	0.4

network results can be seen in **Supplementary Figures S1–S6**, with varying λ and ϵ. These figures were created using *Escher* (King et al., 2015a).

We do not fix the objective value in our optimization problem after obtaining the first solution but in every iteration, the minimization problem is solved again after excluding the newly found solution. This means that the next solution can have the same objective value but it is also possible that the objective value is worse than in the previous iteration. In this particular case, the 100th solution had an objective value that was only 5.5% worse than the objective value of the first solution (see **Table 1**) which shows that, as concerns optimality, all 100 solutions were very similar. They also had very similar active reactions. Comparing the 100 enumerated solutions for $\lambda = 0.1$ and $\epsilon = 5$, a total of 43 reactions with a specific direction were chosen in all solutions. Out of these 43 reactions, 24 were chosen in every solution (including reactions in the TCA cycle and the gluconeogenesis pathway). This means that certain core pathways were consistently picked also in slightly suboptimal solutions. Looking at only the ten best solutions, already 38 out of the 43 reactions were identified. The missing reactions were mostly part of the pentose phosphate pathway which also contains reactions that were part of the solution only in a few cases. Even with only ten solutions, we were able to obtain the alternative pathways feeding the TCA cycle (PPC/PDH). This indicates that it is not necessary to enumerate a large amount of solutions to get significant results and to identify alternative biological pathways.

To check the robustness of the method against small perturbations, we tested within the pyruvate pulse the results of TOTORO for the values of $\lambda = 0.1$ and 0.9, excluding one metabolite at a time, recomputing the results, and computing the distances to the results on the complete metabolite set for reaction occurrence (in terms of absolute difference of occurrences). Overall, the results for both $\lambda = 0.1$ and 0.9 differed from less than 5% to around 20%. In general, the results were robust ($< 10\%$ in average distance) for 70–80% of the metabolites tested (with $\lambda = 0.1$, and 0.9, respectively), but we noticed that excluding metabolites with a higher neighborhood connectivity (such as glutamate and glutamine) had a greater impact on the final results. These results show that even though the distances were small, the amount of information provided by different metabolites varied widely.

Moreover, we tested 10 random sets of measured metabolites (**Supplementary Tables S4, S5**), with a varying number of excluded metabolites to detect at which point the method would not behave as with the complete dataset. In accordance with the previous results for single exclusions, and within the tests with less than 50% of the measured metabolites excluded, the smallest distance ($\approx 10\%$) to the complete results came from a random dataset which included glutamate, glutamine and pyruvate (to ensure the carbon source uptake). As expected, when more than 50% of the measured metabolites were excluded, we detected a much higher difference ($\approx 40\%$) between the results from the complete dataset and those from the random datasets.

3.1.2 Glucose Pulse

For the glucose pulse, we expected that reactions that are part of the glycolysis pathway would be active as they convert glucose into pyruvate generating energy. Consequently, the TCA cycle should also be fed (see **Figure 2**). For $\lambda = 0.9$ and 0.5, the active reactions proposed by TOTORO were disconnected and it was not possible to identify active pathways. We believe that the results coming from this pulse were less insightful since the bacteria were already grown in glucose prior to the pulse, which in turn might be a reason why the changes in metabolites were not as informative as the other pulses. This was for the most part corrected if more metabolites were added as input to TOTORO when using the complete network as presented in **Section 3.2**. This also shows the importance of careful experimental design and how subtle perturbations may generate results that are not always homogeneous.

Even for $\lambda = 0.1$ and $\epsilon = 5$, only disconnected parts of the network were active (see **Supplementary Figure S9**). Since we were interested in testing the method to obtain more connected sub-hypergraphs, we decided to fine-tune the solutions by lowering the value of ϵ as much as possible. The result for $\epsilon = 2$ and 1.2 can be found in **Supplementary Figures S10, S11**, respectively. Lowering the value of ϵ to 1.1 rendered the underlying optimization problem infeasible. For $\epsilon = 1.2$, we got solutions that linked intermediate metabolites of the glycolysis pathway to the TCA cycle through the PPC reaction. In some solutions, the TCA cycle was also fed by PDH and CS to account for the accumulation of citrate. As previously mentioned, when the solutions are disconnected and this is unwanted, decreasing the value of ϵ can sometimes help to obtain more connected solutions. However, this should be used carefully in order to avoid linking unrelated and distant metabolites, which might not be meaningful biologically.

The 100 solutions were very similar ($\lambda = 0.1$, $\epsilon = 1.2$). They accounted for a total of 47 reactions (with distinct directions) and 30 of these appeared in all solutions. Similarly to the pyruvate pulse, the difference in these solutions were mostly based on a few reactions that are not part of the main pathways (glycolysis/TCA cycle). One critical observation is that the D-glucose transport reaction (GLCpts) was not part of every solution although glucose should be used as important source. As previously mentioned, the bacteria were already grown in glucose prior to the glucose pulse, which is possibly a reason why glucose was already internalized prior to the initial pulse. When comparing the objective values for these 100 solutions, the absolute difference between the first solution and the 100th one was similar to the one observed for the pyruvate pulse (see **Table 1**). However, proportionally this value was 21.8% worse than for the first solution. When we repeated the run for $\lambda = 0.1$ and $\epsilon = 1.2$ with 50 iterations, the D-glucose transport reaction was part of 42 solutions. For ten iterations, this reaction was picked in all ten solutions. Hence, the glucose transport reaction was active in solutions with the best objective values. This showed that although the solutions remained very similar, there was a decline in their quality. And similarly to the pyruvate pulse, we saw that it is not necessary to enumerate a large amount of solutions.

3.1.3 Succinate Pulse

After the succinate pulse, part of the TCA cycle should always be active. Furthermore, the gluconeogenesis pathway should be active to produce G3P and glucose-6-phosphate from succinate. Again, the results for $\lambda = 0.5$ and 0.9 led to smaller solutions that were more disconnected (see **Supplementary Figures S13–S16**). Therefore, we focused on the analysis of the results for $\lambda = 0.1$ (see **Supplementary Figures S17, S18**). For both $\epsilon = 5$ and 10, succinate entered the TCA cycle and turned into oxaloacetate. TOTORO proposed two possibilities to output the excess of the TCA cycle: Either phosphoenolpyruvate (PEP) was produced by PEP carboxykinase (PPCK) or by PEP synthase (PPS) using pyruvate as intermediate substrate. Subsequently, PEP was, as expected, transformed to G3P. The lower right part of the TCA cycle predicted as active can be explained by the fact that the concentration of L-glutamate decreased and the concentration of citrate increased. The active reaction in this part connected these two metabolites. Furthermore, reactions of the pentose phosphate pathway were proposed as active and the biomass precursors R5P, E4P, and G3P were produced.

The results for $\epsilon = 5$ and 10 were very similar. For example, one difference was that for $\epsilon = 10$, the reverse D-lactate dehydrogenase (LDH) was predicted to be active in 56 solutions which led to a small accumulation of D-lactate. It does make sense biologically because in general, D-lactate is one of the main products of the fermentation but we do not have the measurements for the concentration of D-lactate for this pulse experiment to actually verify this observation. However, in total, the differences were negligible and in contrast to the glucose pulse, the parameter ϵ had a lower impact on the outcome.

Again, the core reactions of all 100 solutions were very similar. In total, 41 reactions (with distinct directions) appeared in all 100 solutions (for $\lambda = 0.1$, $\epsilon = 5$). We observed that 22 of these were always active (mostly in the gluconeogenesis pathway and part of the TCA cycle). The objective values for all 100 solutions were extremely close (see **Table 1**).

3.2 *E. coli* iJO1366 Model

Based on the results for the *E. coli* core model, we only did runs with $\lambda = 0.1$ for the *E. coli* iJO1366 model. The inputs were updated because this network contains more metabolites and therefore, more measured metabolites could be added. The amount of iterations was decreased to ten because the runtime in the larger network is significantly higher and we had already established in the core model that it was not necessary to enumerate a larger amount of solutions. To decrease the runtime for each solution, CPLEX was configured differently. The relative MIP gap tolerance was set to 0.05 which means that the solver will stop an iteration if a solution is found that is within 5% of the optimal. This allows for a faster result and we could see in the core model that the first 100 solutions tended to be very similar. This means that even if we are enumerating slightly suboptimal solutions, we should be able to compute solutions that are very similar to the actual optimal solution. If the 5% limit is not reached after 48 h, the iteration is stopped. The memory usage of CPLEX was limited to 10 GB.

The runtime for the different pulse experiments differed a lot. The results for the pyruvate and glucose pulses were computed on a cluster. For the pyruvate pulse, the 5% limit was reached only in three iterations (see **Supplementary Table S6**). All other iterations were stopped after 48 h. However, all solutions obtained were within 7% of the optimum. Thus, we still took them into account when analyzing the predicted active reactions. In none of the iterations for the glucose pulse, the 5% limit was reached. The obtained solutions were within 8.5% of the optimal value (see **Supplementary Table S7**).

In contrast to the pyruvate and the glucose pulses, the 5% limit was reached in all iterations for the succinate pulse and computing all ten solutions took less than 5 min on a personal machine (2.90 GHz Intel i7-7820HQ CPU, 16 GB RAM). This shows that the constraints describing the input deltas in the MILP have a large influence on the difficulty of the optimization problem, and thus also on the runtime.

However, although the obtained solutions were suboptimal, the active reactions predicted by TOTORO for the core metabolism were similar to the best results of the *E. coli* core model for all three pulse experiments. For instance, in the pyruvate pulse results, out of 12 reactions in the TCA cycle within the large network, 8 were also present in the core model. In total, 5 were chosen in 100% of the solutions in the same direction in both core and large networks. The complete network was also able to correct the only inconsistency within the TCA cycle for the core network: the direction of the reaction ICDHyr, which shows the advantage of relying on complete networks whenever available. For the glycolysis/gluconeogenesis pathways, out of 12 reactions, 9 were also included in the core model. In total, 6 reactions were chosen in 100% and 1 in more than 80% of the solutions in the same direction in both networks. TOTORO predicted as active for pyruvate, glucose and succinate (in at least 1 solution) a total of 221, 284, and 189 reactions respectively. Moreover, 52% of the reactions were chosen across all iterations in the pyruvate pulse dataset, 81% in the succinate pulse dataset and 62% in the glucose pulse dataset.

The additional measurements that were added as input deltas for the large network were mostly amino acids (see **Supplementary Tables S1–S3**). In (Waschina et al., 2016), the authors show for the example of amino acid production in *E. coli* how the production cost for individual amino acids can depend on the available carbon source, and reactions close to the entry point of the carbon source might have considerably higher fluxes. A schematic representation of this is provided in **Figure 4A**. Indeed, from the experimental data, alanine and valine only accumulated during the pyruvate pulse, and were depleted

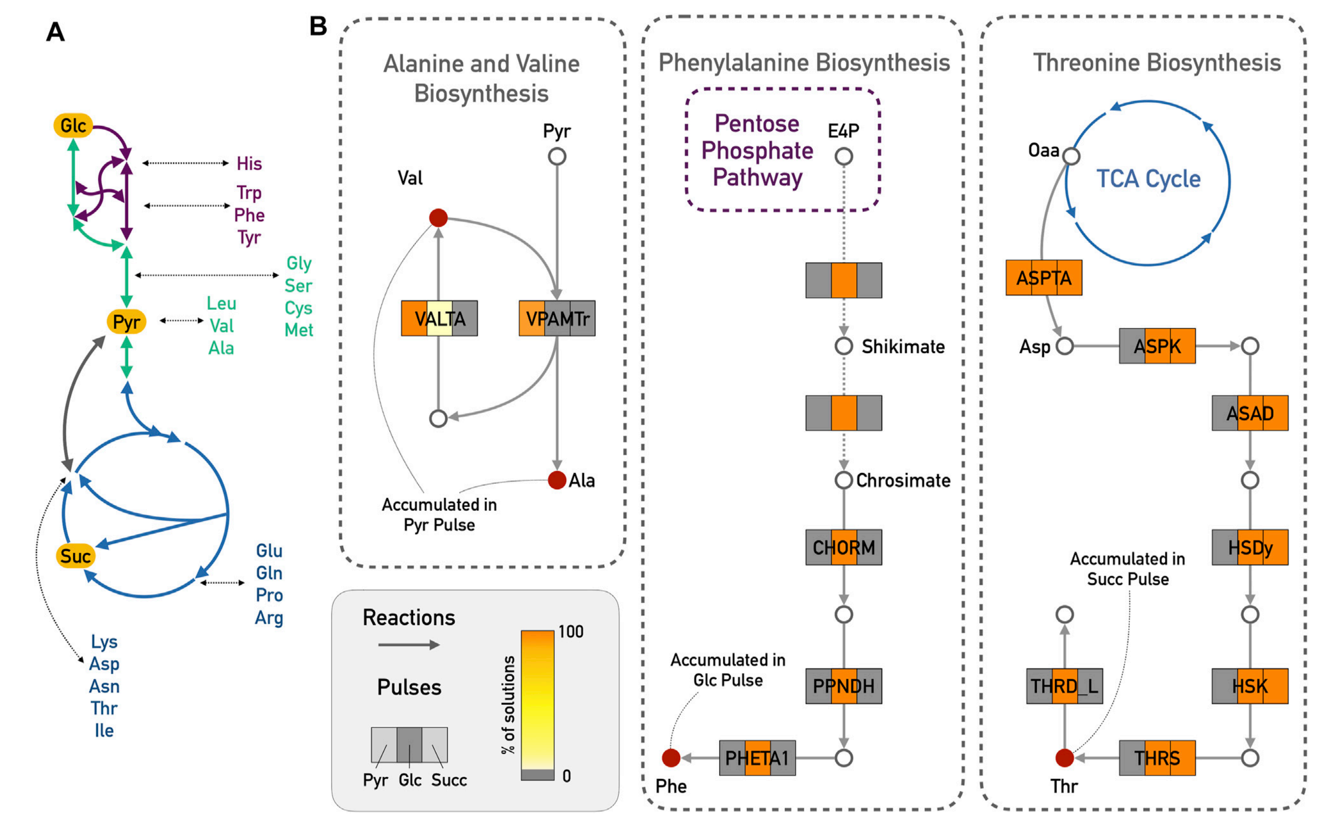


FIGURE 4 | Amino acid biosynthesis in the *E. coli* iJO1366 model. **(A)** Schematic representation of carbon sources with closely related amino acids. Glycolysis/ Gluconeogenesis in green; TCA cycle in blue and Pentose Phosphate Pathway in purple. **(B)** TOTORO results explaining the accumulation of valine (Val) and alanine (Ala) in the pyruvate (Pyr) pulse; accumulation of phenylalanine (Phe) from the glucose (Glc) pulse and accumulation of threonine (Thr) from the succinate (Succ) pulse. For simplicity reasons, side compounds and cofactors were excluded from the figure. Dashed arrows indicate several reactions from the shikimate and chorismate pathways. Abbreviations for reaction names are as follows: VALTA, valine transaminase; VPAMTr, valine-pyruvate aminotransferase; CHORM, chorismate mutase; PPNDH, prephenate dehydratase; PHETA1, phenylalanine transaminase; ASPTA, aspartate transaminase; ASPK, aspartate kinase; ASAD, aspartate-semialdehyde dehydrogenase; HSDy, homoserine dehydrogenase; HSK, homoserine kinase; THRS, threonine synthase; THRD_L, L-threonine deaminase.

with the other two carbon sources. Pyruvate is a direct precursor for valine production. We therefore expected that reactions of the alanine and valine biosynthesis should play a greater role in the predicted results for pyruvate compared to the other two pulses. TOTORO predicted an activation of the pathway from pyruvate to alanine and valine, which resulted in the accumulation of these amino acids (**Figure 4B**). In accordance with the predictions in (Waschina et al., 2016), another example is the accumulation of threonine during the succinate pulse. Threonine and succinate are closely connected, and TOTORO predicted active reactions leading to its biosynthesis and accumulation in the succinate pulse (**Figure 4B**). Compared to the results for succinate, TOTORO predicted more active reactions consuming threonine during the glucose pulse, and no reactions producing it in the pyruvate pulse, resulting in the depletion of this amino acid with those carbon sources. Moreover, only during the glucose pulse, phenylalanine was accumulated, and TOTORO proposed the complete pathway for the phenylalanine biosynthesis as active when compared to the pyruvate and succinate pulses (**Figure 4B**), in accordance with the predictions in (Waschina et al., 2016) of lower cost to produce this amino acid with glucose as carbon source.

4 DISCUSSION

TOTORO was able to predict expected pathways as active based on the differences in the measured concentrations for some internal metabolites for both the *E. coli* core and complete models. We show that in general, it is preferable to use smaller values of λ (e.g., $\lambda = 0.1$) though the method is not critically sensible to this setup, being robust to small perturbations. However, it is worth noting that a higher λ can lead to smaller solutions which might be biologically irrelevant. Here, we focused in extracting connected sub-hypergraphs that explained the changes in concentration between two different conditions. We also show that a reduction of ϵ can also be used to obtain more connected solutions. However, there might be situations where the user might be interested in only local changes around the measurements. In this context, it might be advantageous to choose higher values for λ and ϵ. We did not encounter problems specific to co-factors which is a known problem when looking for shortest paths in metabolic networks. This is probably due to the fact that we are not only minimizing the number of active reactions in the solutions but also focusing on the changes in the metabolite concentrations. By splitting reversible reactions, TOTORO was able to predict distinct directions for them.

Both in the core network and in the larger network, we were able to recover biologically meaningful pathways. Additionally, although the larger network contains more reactions and we added more input deltas, the predictions for the core metabolism of *E. coli* were fairly similar to the results for the core network. We also showed a particular case in which the perturbation was subtle, and the results from the complete model were more insightful than the ones from the core model. It must be however noted that the predictions do depend on the measured metabolites. If for large parts of the network, no metabolite concentrations are measured, TOTORO will likely not be able to find active pathways for these parts of the network.

Moreover, we could also see that it is not necessary to enumerate a high number of solutions which is especially important when larger networks are used and the runtime of TORORO increases. We enumerated 100 different solutions for the core network. However, in our case, the enumerated solutions were very similar and a large amount of reactions appeared in all 100 solutions. Therefore, already one (or few) solution(s) would have been sufficient to infer the most important reactions that were proposed to be active.

5 CONCLUSION

In this paper, we presented TOTORO, a method that identifies active reactions during the transient state based on the differences in the concentrations for some measured metabolites from two different conditions and we showed its prediction power on the example of different pulse experiments in *E. coli*. It is important to note that even though we provided several biologically trivial results, TOTORO only used metabolomic data as basis for these predictions, without any other source of bias such as defined metabolic pathways. Our method was also able to handle full networks which take into account model stoichiometry, and we did not perform any type of filtering for cycles, reversible reactions or co-factors.

With the current technologies, it gets more common to have different kinds of data available which creates a need for methods that combine, for instance, metabolomic, transcriptomic and proteomic data. We have recently developed a method for integration of metabolic networks and transcriptomic data (Pusa et al., 2019) and we intend in the future to adapt our approaches to be able to integrate multiple kinds of omic data, similarly to what was proposed in (Pandey et al., 2019) for thermodynamic, transcriptomic and metabolomic data, and in (Kleessen et al., 2015) for transcriptomic and metabolomic data. On a larger scale, it might be interesting also to consider whether some measures used in (hyper)graph theory such as connectivity or (hyper)path length might be related to the parameters used and thus provide an automatic and perhaps more reliable way of setting them. Notice that achieving this would be even more challenging in the case of hypergraphs for which such measures might have to be adapted.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

M-FS, AJ-L, RA, MF, SV: Conception of the work; IZ, RA, AJ-L, AM, LD, RC: Constructed the code; MF, IZ, RA: Analysis of datasets; IZ, MF, AJ-L, M-FS: Wrote the manuscript with input of other authors; All authors approved the last version of the manuscript.

FUNDING

This work was supported by a postdoctoral fellowship from the Agence Nationale de la Recherche (ANR-GREEN 17_CE20_0031_01) granted to MF; Fundação para a Ciẽncia e a Tecnologia (UIDB/50021/2020); Fundação de Amparo à Pesquisa do Estado de São Paulo (#15/13430-9, #15/22308-2) and was part of a PhD funded by Inria. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 951970 (OLISSIPO project).

ACKNOWLEDGMENTS

This work was performed using the computing facilities of the CC LBBE/PRABI.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.815476/full#supplementary-material

REFERENCES

Acuña, V., Birmelé, E., Cottret, L., Crescenzi, P., Jourdan, F., Lacroix, V., et al. (2012). Telling Stories: Enumerating Maximal Directed Acyclic Graphs with a Constrained Set of Sources and Targets. *Theor. Computer Sci.* 457, 1–9. doi:10.1016/j.tcs.2012.07.023

Bordbar, A., Yurkovich, J. T., Paglia, G., Rolfsson, O., Sigurjónsson, Ó. E., and Palsson, B. O. (2017). Elucidating Dynamic Metabolic Physiology through Network Integration of Quantitative Time-Course Metabolomics. *Sci. Rep.* 7, 46249. doi:10.1038/srep46249

Cambiaghi, A., Ferrario, M., and Masseroli, M. (2017). Analysis of Metabolomic Data: Tools, Current Strategies and Future Challenges for Omics Data Integration. *Brief Bioinform.* 18, bbw031–510. doi:10.1093/bib/bbw031

Case, A., Lutz, J. H., and Stull, D. M. (2016). "Reachability Problems for Continuous Chemical Reaction Networks," in International Conference on Unconventional Computation and Natural Computation (Springer), 1–10. doi:10.1007/978-3-319-41312-9_1

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., et al. (2018). MetaboAnalyst 4.0: towards More Transparent and Integrative Metabolomics Analysis. *Nucleic Acids Res.* 46, W486–W494. doi:10.1093/nar/gky310

Christensen, C. D., Hofmeyr, J.-H. S., and Rohwer, J. M. (2015). Tracing Regulatory Routes in Metabolism Using Generalised Supply-Demand Analysis. *BMC Syst. Biol.* 9, 89. doi:10.1186/s12918-015-0236-1

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., et al. (2018). Metexplore: Collaborative Edition and Exploration of Metabolic Networks. *Nucleic Acids Res.* 46, W495–W502. doi:10.1093/nar/gky301

Covert, M. W., and Palsson, B. O. (2003). Constraints-based Models: Regulation of Gene Expression Reduces the Steady-State Solution Space. *J. Theor. Biol.* 221, 309–325. doi:10.1006/jtbi.2003.3071

Frainay, C., Aros, S., Chazalviel, M., Garcia, T., Vinson, F., Weiss, N., et al. (2019). Metaborank: Network-Based Recommendation System to Interpret and Enrich Metabolomics Results. *Bioinformatics.* 35, 274–283. doi:10.1093/bioinformatics/bty577

Frainay, C., and Jourdan, F. (2017). Computational Methods to Identify Metabolic Sub-networks Based on Metabolomic Profiles. *Brief Bioinform.* 18, 43–56. doi:10.1093/bib/bbv115

Ginsburg, H. (2009). Caveat Emptor: Limitations of the Automated Reconstruction of Metabolic Pathways in Plasmodium. *Trends Parasitology.* 25, 37–43. doi:10.1016/j.pt.2008.08.012

Ivanisevic, J., and Want, E. J. (2019). From Samples to Insights into Metabolism: Uncovering Biologically Relevant Information in Lc-Hrms Metabolomics Data. *Metabolites.* 9, 308. doi:10.3390/metabo9120308

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015a). Escher: a Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *Plos Comput. Biol.* 11, e1004321. doi:10.1371/journal.pcbi.1004321

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., et al. (2015b). BiGG Models: A Platform for Integrating, Standardizing and Sharing Genome-Scale Models. *Nucleic Acids Res.* 44, D515–D522. doi:10.1093/nar/gkv1049

Klamt, S., Hädicke, O., and von Kamp, A. (2014). "Stoichiometric and Constraint-Based Analysis of Biochemical Reaction Networks," in Large-scale networks in engineering and life sciences (Springer), 263–316. doi:10.1007/978-3-319-08437-4_5

Kleessen, S., Irgang, S., Klie, S., Giavalisco, P., and Nikoloski, Z. (2015). Integration of Transcriptomics and Metabolomics Data Specifies the Metabolic Response of Chlamydomonas to Rapamycin Treatment. *Plant J.* 81, 822–835. doi:10.1111/tpj.12763

Kuo, T.-C., Tian, T.-F., and Tseng, Y. J. (2013). 3omics: a Web-Based Systems Biology Tool for Analysis, Integration and Visualization of Human Transcriptomic, Proteomic and Metabolomic Data. *BMC Syst. Biol.* 7, 64. doi:10.1186/1752-0509-7-64

Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002). Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*. *Biophysical J.* 83, 1331–1340. doi:10.1016/s0006-3495(02)73903-9

Marco-Ramell, A., Palau-Rodriguez, M., Alay, A., Tulipani, S., Urpi-Sarda, M., Sanchez-Pla, A., et al. (2018). Evaluation and Comparison of Bioinformatic Tools for the Enrichment Analysis of Metabolomics Data. *BMC bioinformatics.* 19, 1. doi:10.1186/s12859-017-2006-0

Milreu, P. V., Klein, C. C., Cottret, L., Acuña, V., Birmelé, E., Borassi, M., et al. (2014). Telling Metabolic Stories to Explore Metabolomics Data: a Case Study on the Yeast Response to Cadmium Exposure. *Bioinformatics.* 30, 61–70. doi:10.1093/bioinformatics/btt597

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., et al. (2011). A Comprehensive Genome-scale Reconstruction of *Escherichia coli* Metabolism-2011. *Mol. Syst. Biol.* 7, 535. doi:10.1038/msb.2011.65

Orth, J. D., Fleming, R. M., and Palsson, B. O. (2010). Reconstruction and Use of Microbial Metabolic Networks: the Core escherichia Coli Metabolic Model as an Educational Guide. *EcoSal plus.* 4 (1). doi:10.1128/ecosalplus.10.2.1

Palsson, B. (2000). The Challenges of In Silico Biology. *Nat. Biotechnol.* 18, 1147–1150. doi:10.1038/81125

Pandey, V., Hadadi, N., and Hatzimanikatis, V. (2019). Enhanced Flux Prediction by Integrating Relative Expression and Relative Metabolite Abundance into Thermodynamically Consistent Metabolic Models. *Plos Comput. Biol.* 15, e1007036. doi:10.1371/journal.pcbi.1007036

Perez de Souza, L., Alseekh, S., Brotman, Y., and Fernie, A. R. (2020). Network Based Strategies in Metabolomics Data Analysis and Interpretation: from Molecular Networking to Biological Interpretation. *Expert Rev. Proteomics* 17 (4), 243–255. doi:10.1080/14789450.2020.1766975

Pusa, T., Ferrarini, M. G., Andrade, R., Mary, A., Marchetti-Spaccamela, A., Stougie, L., et al. (2019). MOOMIN - Mathematical explOration of 'Omics Data on a MetabolIc Network. *Bioinformatics.* 36, 514–523. doi:10.1093/bioinformatics/btz584

Reznik, E., Mehta, P., and Segrè, D. (2013). Flux Imbalance Analysis and the Sensitivity of Cellular Growth to Changes in Metabolite Pools. *Plos Comput. Biol.* 9, e1003195–13. doi:10.1371/journal.pcbi.1003195

Roessner, U., and Bowne, J. (2009). What Is Metabolomics All about? *Biotechniques.* 46, 363–365. doi:10.2144/000113133

Rohwer, J. M., and Hofmeyr, J.-H. S. (2008). Identifying and Characterising Regulatory Metabolites with Generalised Supply-Demand Analysis. *J. Theor. Biol.* 252, 546–554. doi:10.1016/j.jtbi.2007.10.032

Rosato, A., Tenori, L., Cascante, M., De Atauri Carulla, P. R., Martins dos Santos, V. A. P., and Saccenti, E. (2018). From Correlation to Causation: Analysis of Metabolomics Data Using Systems Biology Approaches. *Metabolomics.* 14, 37. doi:10.1007/s11306-018-1335-y

Sajitz-Hermstein, M., Töpfer, N., Kleessen, S., Fernie, A. R., and Nikoloski, Z. (2016). Iremet-Flux: Constraint-Based Approach for Integrating Relative Metabolite Levels into a Stoichiometric Metabolic Models. *Bioinformatics.* 32, i755–i762. doi:10.1093/bioinformatics/btw465

Sévin, D. C., Kuehne, A., Zamboni, N., and Sauer, U. (2015). Biological Insights through Nontargeted Metabolomics. *Curr. Opin. Biotechnol.* 34, 1–8. doi:10.1016/j.copbio.2014.10.001

Stanstrup, J., Broeckling, C., Helmus, R., Hoffmann, N., Mathé, E., Naake, T., et al. (2019). The Metarbolomics Toolbox in Bioconductor and beyond. *Metabolites.* 9, 200. doi:10.3390/metabo9100200

Taymaz-Nikerel, H., De Mey, M., Baart, G., Maertens, J., Heijnen, J. J., and van Gulik, W. (2013). Changes in Substrate Availability in escherichia Coli lead to Rapid Metabolite, Flux and Growth Rate Responses. *Metab. Eng.* 16, 115–129. doi:10.1016/j.ymben.2013.01.004

Töpfer, N., Kleessen, S., and Nikoloski, Z. (2015). Integration of Metabolomics Data into Metabolic Networks. *Front. Plant Sci.* 6, 49. doi:10.3389/fpls.2015.00049

Waschina, S., D'Souza, G., Kost, C., and Kaleta, C. (2016). Metabolic Network Architecture and Carbon Source Determine Metabolite Production Costs. *Febs J.* 283, 2149–2163. doi:10.1111/febs.13727

Xia, J., Sinelnikov, I. V., Han, B., and Wishart, D. S. (2015). MetaboAnalyst 3.0-making Metabolomics More Meaningful. *Nucleic Acids Res.* 43, W251–W257. doi:10.1093/nar/gkv380

Conflict of Interest: Author AJ-L was employed by Soladis GmBH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 14 March 2022
doi: 10.3389/fgene.2022.814093

Knowledge Graphs for Indication Expansion: An Explainable Target-Disease Prediction Method

Ozge Gurbuz[1], Gregorio Alanis-Lobato[2], Sergio Picart-Armada[2], Miao Sun[2], Christian Haslinger[2], Nathan Lawless[2] and Francesc Fernandez-Albert[2]**

[1]Discovery Research Coordination Germany, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany, [2]Global Computational Biology and Data Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

OPEN ACCESS

Edited by:
Marco Pellegrini,
Institute of Computer Science and Telematics, Italy

Reviewed by:
Marco Brandizi,
Rothamsted Research, United Kingdom
Nasser Ghadiri,
Isfahan University of Technology, Iran

****Correspondence:***
Ozge Gurbuz
ozge.gurbuz@boehringer-ingelheim.com
Francesc Fernandez-Albert
francesc.fernandez-albert@boehringer-ingelheim.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *12 November 2021*
Accepted: *28 January 2022*
Published: *14 March 2022*

Citation:
Gurbuz O, Alanis-Lobato G, Picart-Armada S, Sun M, Haslinger C, Lawless N and Fernandez-Albert F (2022) Knowledge Graphs for Indication Expansion: An Explainable Target-Disease Prediction Method. Front. Genet. 13:814093. doi: 10.3389/fgene.2022.814093

Indication expansion aims to find new indications for existing targets in order to accelerate the process of launching a new drug for a disease on the market. The rapid increase in data types and data sources for computational drug discovery has fostered the use of semantic knowledge graphs (KGs) for indication expansion through target centric approaches, or in other words, target repositioning. Previously, we developed a novel method to construct a KG for indication expansion studies, with the aim of finding and justifying alternative indications for a target gene of interest. In contrast to other KGs, ours combines human-curated full-text literature and gene expression data from biomedical databases to encode relationships between genes, diseases, and tissues. Here, we assessed the suitability of our KG for explainable target-disease link prediction using a glass-box approach. To evaluate the predictive power of our KG, we applied shortest path with tissue information- and embedding-based prediction methods to a graph constructed with information published before or during 2010. We also obtained random baselines by applying the shortest path predictive methods to KGs with randomly shuffled node labels. Then, we evaluated the accuracy of the top predictions using gene-disease links reported after 2010. In addition, we investigated the contribution of the KG's tissue expression entity to the prediction performance. Our experiments showed that shortest path-based methods significantly outperform the random baselines and embedding-based methods outperform the shortest path predictions. Importantly, removing the tissue expression entity from the KG severely impacts the quality of the predictions, especially those produced by the embedding approaches. Finally, since the interpretability of the predictions is crucial in indication expansion, we highlight the advantages of our glass-box model through the examination of example candidate target-disease predictions.

Keywords: knowledge graphs, ontologies, drug discovery, target repurposing, target repositioning

INTRODUCTION

Indication expansion (IE) is an emerging subject in drug discovery that aims to find alternative therapeutic applications, or diseases (indications) for an existing drug target (Parisi et al., 2020). Considering the high cost and slow process of bringing a new drug into the market, *in silico* approaches for drug discovery and repurposing (Dudley et al., 2011; Picart-Armada et al., 2019; Sosa et al., 2020) became a popular subject in the bioinformatics community due to the increasing availability of both structured and unstructured data modalities. In fact, with the improvement in text mining technologies, literature mining has become an established and popular tool for indication expansion in drug discovery (Andronis et al., 2011; Lekka et al., 2012; Smalheiser, 2012; Sebastian et al., 2017; Sang et al., 2018; Sosa et al., 2020). One can search for all potential disease relations for a given drug in the literature via text mining techniques and expand the analysis to all targets of the drug to establish a more comprehensive search (Andronis et al., 2011; Lekka et al., 2012; Smalheiser, 2012). The outcome of this method is the direct disease-gene links (based on search criteria). On the other hand, analysis of biological data sources (such as molecular data, experimental data, gene expression data, etc.) are common approaches to search for novel target-disease links (Brown and Patel, 2017; Härtner et al., 2018; Picart-Armada et al., 2019).

A natural extension of these studies would be the integration of several data sources for a more comprehensive analysis. However, the heterogeneity of data formats and sources raises questions during their integration (Holzinger, 2018; Katsila and Matsoukas, 2018). The best way to undertake this data integration challenge, together with data contextualization, is the application of semantic web technologies: ontologies and knowledge graphs (Qu et al., 2009; Chen and Xie, 2010; Williams et al., 2012; Lin et al., 2017; Kanza and Frey, 2019; Zhu et al., 2020). The main ideas of ontologies and knowledge graphs (KGs) are that each resource has a unique identifier, and once each resource is defined with the identifier, regardless of where they are extracted from, they will be merged and the integration process will be effortless. Secondly, integrating the data sources brings up the topic of data governance, as data needs to be findable, accessible, interoperable, and reusable or, in other words, in alignment with the FAIR data principles (Wilkinson et al., 2016). For this, ontologies can also be very helpful (Williams et al., 2012) because all the data mapped using the same ontology will be already linked which makes it very easy to search, query, and reuse. Lastly, predictions from comprehensive and integrated data sets are often difficult to interpret (Holzinger, 2018; Lecue, 2020). This is a major challenge in the biological domain, which can be tackled by providing ontological perspective into the prediction process to incorporate human recognition and interpretation, thus making the methodology a “glass box” model (Holzinger et al., 2017). Due to the importance of semantic web technologies in addressing the above-mentioned challenges, many researchers have added a semantic layer and included KGs in their computational methods for drug discovery studies (Kanza and Frey, 2019).

In our previous work, we divided the studies that use KGs for drug discovery into two categories (Gurbuz et al., 2020): KGs built from biological data sources (Qu et al., 2009; Fu et al., 2016; Han et al., 2018; Celebi et al., 2019; Zhu et al., 2020) and KGs built from the literature (Sang et al., 2018, 2019; Sosa et al., 2020). Then, we distinguished between studies performing drug-disease predictions (Qu et al., 2009; Fu et al., 2016; Han et al., 2018; Sang et al., 2018, 2019; Sosa et al., 2020; Zhu et al., 2020) and those predicting drug-drug interactions (Herrero-Zazo et al., 2015; Celebi et al., 2019). The outcome of this review of the state-of-the-art was that the studies using structured biological data sources (BioGrid, StringDB, Human Protein Atlas, etc.) for building the KG applied several network analysis methods to predict either drug-disease relations or gene-disease associations. Even though the value of available biological sources cannot be denied, the outcome of such predictions based on statistical confidence scores may not be sufficiently persuasive to kick-off a full drug-development program (Holzinger, 2018). Literature support would be more convincing for further investigation. Therefore, the second group of studies constructed the KG from literature sources but did not implement KGs that combine both structured biological data sources and literature sources for a more comprehensive indication expansion or target repositioning approach. Additionally, all these studies did not truly benefit from semantic web technologies. Instead, they directly applied network analysis algorithms.

As a result, in the past we conducted an exploratory case study aimed at constructing a comprehensive KG to facilitate indication expansion (Gurbuz et al., 2020). We presented the methodology, defined the reasoning and inferencing on the KG, and successfully applied it to two randomly selected cases to predict the link between the target and disease. We ranked the paths connecting the target and disease based on the number of publications associated with its constituent edges. In addition, a path was considered more relevant when all the proteins in the path showed expression in the same tissue, either at the RNA or protein level. One limitation of the previous study was that we conducted the exploratory cases at a small scale with a target and a given candidate indication to find the mechanism of action. By contrast, in the current study we extend the identification of novel target-disease links to all available pairs, evaluate the performance of the inferred edges based on random baselines, and study the value of including RNA- and protein-level expression information in our predictions. Moreover, we show how the KG can be exploited to interpret candidate gene-disease associations through the examination of two examples.

RELATED WORK

In this section, we review the approaches that have resorted to the use of KGs for drug discovery regardless of whether the purpose was drug-disease, gene-disease, or drug-drug interaction prediction. **Table 1** shows a comprehensive overview of the reviewed methods.

Celebi et al. used KGs for drug-drug interaction identification and used a publicly available dataset called Bio2RDF to extract drug features (Celebi et al., 2019). After feature extraction via RDF2Vec, TransE, and TransD embeddings, they applied Logistic Regression, Naïve Bayes, and Random Forest models and evaluated which combination of embedding and machine learning models was better at predicting a reference set of drug-drug interactions. The best performance they achieved was using RDF2Vec together with a Random Forest model.

There are several studies which use KGs to predict drug-disease relations. Fu et al. (Fu et al., 2016) built a network from various biological and chemical data sources to

TABLE 1 | Overview of knowledge graph usage in drug discovery.

Study	Purpose	Method	Data source
Celebi et al. (2019)	Drug-Drug Interaction: evaluating the different embedding methods in various Cross Validation schemes	Embedding: RDF2Vec, CBOW, Skip Gram, TransE, TransD ML Model: Logistic Regression, Naive Bayes, Random Forest	Bio2RDF
Fu et al. (2016)	Drug-target interactions	Metapath + Random forest, SVM	Biological and chemical datasets
Han et al. (2018)	Drug target genes for Alzheimer's Disease	Inference + enrichment analysis	TTD, DrugBank, PharmGKB, AlzGene
Zhu et al. (2020)	Drug centric KG	Positive and Unlabeled Learning * SVM, Decision Tree and Random Forest	PharmGKB, TTD, KEGG DRUG, DrugBank, SIDER and DID
Sang et al. (2018)	Potential drugs for diseases	Logistic regression	Pubmed Abstracts
Sang et al. (2019)	Potential drugs for diseases	TransE embedding + LSTM	Pubmed Abstracts
Sosa et al. (2020)	FDA approved drugs for rare diseases	Network proximity	Pubmed Abstracts
Paliwal et al. (2020).	Predicting clinical failure	Tensor factorization + gene prioritization	20% is from biomedical literature and biological data sources
Nunes et al. (2020)	Predicting Gene-Disease links	Embeddings + Random Forest	Gene-Disease links from Disgenet
Geleta et al. (2021)	Knowledge Graph construction to support drug discovery like predicting Gene-Disease links and	Embeddings (RESCAL) + XGBoot	Gene and Disease nodes and edges from public databases
KG for IE	Target repurposing	Tissue based semantic inferencing + Embeddings & Random Forest	Human curated full text literature + biological database

predict drug-target relations using Random Forest and Support Vector machine algorithms. However, they only benefited from semantic web technologies at the stage of data integration and concentrated on drug-target relations. Han et al. (Han et al., 2018) integrates popular biological databases (DB) such as TTD, DrugBank, PharmGKB, and AlzGene to predict novel drug targets for Alzheimer's disease. Their novel strategy was to combine ontology inference together with enrichment analysis. However, their main goal was limited to finding genes for one specific disease, Alzheimer's in this case. Zhu et al. built a drug centric KG by integrating six drug data sources (PharmGKB, TTD, KEGG DRUG, DrugBank, SIDER, and DID) (Zhu et al., 2020). They implemented a machine learning approach on a path-based representation and an embedding-based representation, separately. To evaluate the effectiveness of the KG, the authors used positive samples and unlabeled samples (samples from diabetes mellitus only) and implemented positive and unlabeled learning (PU) with Decision Tree, Random forest, and SVM models. According to their performance evaluations, the best outcome came from SVM implemented on path-based representation (normalized path count). However, this study uses a drug centric KG to understand the drug-disease interaction.

On the other hand, there are two studies by Sang et al. (Sang et al., 2018; Sang et al., 2019) which used literature for building the KG including SemaTyp. This KG is built from PubMed abstracts by using a natural language processing (NLP) tool called SemaRep. In the first study (Sang et al., 2018), they applied logistic regression on the KG and outperformed the results obtained with a random walk method. Their aim was to predict drug-disease relations via drug-target-disease chains. Later, the authors published a continuation of their work called GrEDeL in which they used KG embedding methods for discovering drug-disease relations from literature (Sang et al., 2019). The authors again use SemaRep to extract associations from PubMed abstracts and build the KG. This time they claim that their previous work, which used logistic regression, couldn't reflect the order of the entities in the associations and couldn't show the detailed drug mechanism of action. Therefore, they first used the TransE embedding method and applied a Long Short-Term Memory (LSTM) based Recurrent Neural Network model to show that graph embeddings capture more information than logistic regression. However, they claimed that the limitation of both studies is that the effectiveness of the methods is dependent on the NLP tool. Likewise, Sosa et al. (Sosa et al., 2020) also constructed a KG from PubMed abstracts to repurpose FDA-approved drugs for rare diseases. They used graph embedding and network proximity for generating their hypothesis. The limitation of this study is that they missed important knowledge that is usually present in the full text but not in the abstract. Moreover, Nunes et al. (Nunes et al., 2020) implemented a KG using all curated gene-disease links extracted from DisGeNET[1]. They filtered out genes that did not have protein correspondence in Uniprot or annotations in the Gene ontology and genes and diseases that were not annotated in Human Phenotypes. Then, they created 3 different KGs based on this filtering and deployed several embedding strategies, noting that they achieved their best performance for predicting gene-disease links with OPA2Vec. However, they only included gene-disease relations in their KGs.

Furthermore, in another study by Paliwal et al. (Paliwal et al., 2020), the authors built a heterogeneous KG in which 20% of the

[1] https://www.disgenet.org/(accessed on 12.11.21).

data comes from biomedical literature databases and the rest from biological data sources. These sources consist of entities such as genes, proteins, diseases, gene ontology processes, pathways, and compounds (Paliwal et al., 2020). Although the aim of the study was to evaluate translatability of *in silico* predictions of clinical trial failure, they were able to predict therapeutic genes for diseases using gene prioritization algorithms. Note that, contrary to what we do in this work, Paliwal and colleagues searched for therapeutic genes related to a group of selected diseases. A similar study from Geleta et al. (Geleta et al., 2021) also presents a comprehensive knowledge graph built from internal data, external public databases such as ChEMBL and Ensembl, and information extracted from PubMed full-text using Natural Language Processing Techniques named Biological Insights Knowledge Graph (BIKG) to be used for knowledge discovery with machine learning. They use RESCAL for knowledge graph embeddings and XGBoost as machine learning method. They report their average F1 score as 88% for gene-disease link prediction where they reduce the size of the KG to Gene and Disease nodes. However, their focus lies on the creation of the KG, whereas our paper addresses the practical utility of KGs in the context of indication expansion in drug development. Furthermore, they depend on natural language parsers and the full details about the method for gene-disease prediction are unavailable, hindering their reproducibility and application to drug discovery. Likewise, Ochoa et al. (Ochoa et al., 2020) also present a comprehensive knowledge graph with characterization of targets, diseases, phenotypes, and drugs to support target identification and prioritization. This is part of an update within the Open Targets platform. While full text literature is a data stream within Open Targets, its use for drug discovery in indication expansion is not explored.

After analyzing these studies, we concluded that KGs are becoming mainstream for supporting drug discovery initiatives, but they have not benefited from semantic information and instead have relied directly on the application of network analysis. In consequence, we evaluated both tissue-based semantic inferencing and various embedding strategies. Additionally, most literature-based KGs were constructed with abstracts. However, the authors behind these studies have acknowledged that this is a limitation and that extracting information from full texts would increase the predictive power of KGs in general. Therefore, we set out to address these pitfalls and used full-text literature for building a KG for IE. Predictions based on this graph can be accompanied by the literature references supporting them, as well as the mechanisms of action. Furthermore, our KG takes tissue specificity information into consideration when inferencing and predicting target-disease links.

METHODOLOGY

Knowledge Graph Development for Indication Expansion

This section summarizes the methodology that improves upon the KG developed in our previous work to facilitate indication expansion studies. More details can be found in Gurbuz et al. (2020).

We start with the upper layer ontology, which defines the data and semantic layer of the KG. In the current study, we have improved the KG and included the following entities: Protein/Gene, RNA Tissue, Protein Tissue, Publication, and Disease. **Figure 1A** shows the updated upper layer ontology. We have used Python's RDFLib[2] for creating the ontologies and RDF graph. Since it was not possible to create edge properties with the RDF syntax and reification brings about efficiency problems, the new RDF* syntax can be used for creating a weight property on the edges (relations) with RDF4J [3]. Alternatively, a third entity can be created to store the references of these genes' connection information. In this study, we chose to create a third entity, named Publication, between gene and disease. This entity holds the information for PubMed IDs and the number of PubMed articles between the given gene and the disease as data properties.

After building the upper layer ontology, we populated it with Metabase[4], a commercial source for human curated full-text literature information. We only selected the *high confident* relations between gene-gene and gene-disease interactions provided by Metabase. We extracted the tissue-level gene expression from the Human Protein Atlas[5] (Uhlen et al., 2010). The pipeline for building and analyzing the KG is shown in **Supplementary Figure S1**. For data extraction and analysis, we used the R programming language and for ontology population and KG implementation we used Python's RDFlib. Both data extraction and ontology population processes were automated with R and Python scripts (see script KGbuild_toy.py, which can be used as a template for KG construction). Therefore, building the KG took less than 1 day. We used Ensembl IDs for gene/proteins and Mesh IDs for diseases as Unique Resource Identifiers (URI).

Characterization of the Knowledge Graph

We described the following topological features of the KG: in- and out-degree (i.e., number of directed links going in and out of a node, respectively), total degree (sum of in- and out-degree), edge density (ratio of the number of edges and the number of possible edges), value of the coefficient of the power-law distribution fitted to the degree distribution, and PageRank centrality (Page et al., 1999).

On the other hand, we explored the changes of the KG over the years, focusing on the largest (weakly) connected component consisting of only gene and disease nodes. To build the KG of a given year, we only kept the gene-gene and gene-disease edges whose first mention in literature was no later than that year. Then, we represented the evolution of the number of nodes, edges, the edge density, and the power law coefficient.

Tissue-Based Gene-Disease Link Prediction From the Knowledge Graph

Since there may be indirect links (Lekka et al., 2012) between a gene and a disease *via* secondary signaling cascades (modelled

[2] https://rdflib.readthedocs.io/en/stable/apidocs/rdflib.html (accessed on 12.11.21).
[3] https://rdf4j.org/documentation/programming/rdfstar/(accessed on 12.11.21).
[4] https://www.cortellislabs.com/page/?api=api-MB (accessed on 12.11.21).
[5] https://www.proteinatlas.org/(accessed on 12.11.21).

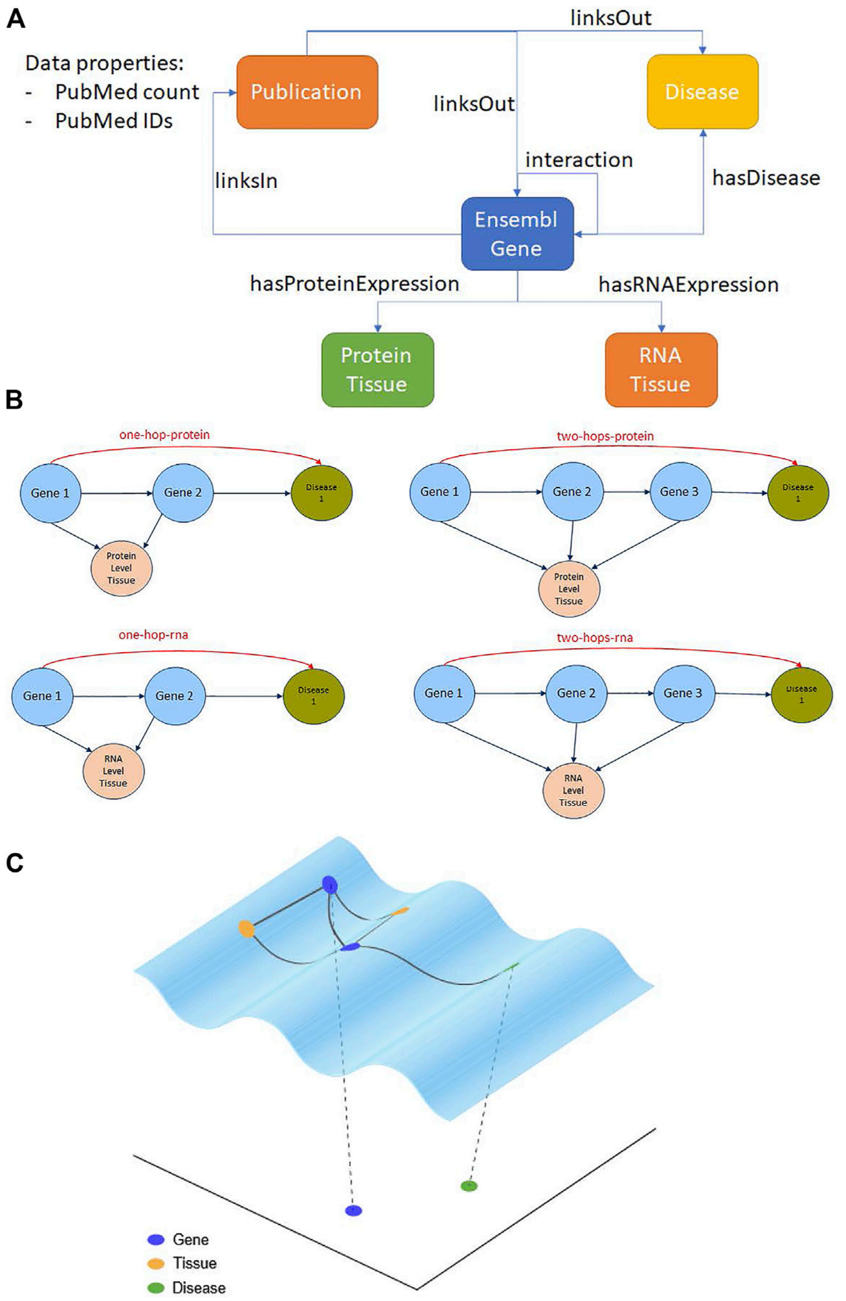


FIGURE 1 | Knowledge graph schema and gene-disease prediction strategies. **(A)** Upper layer ontology with the entities and relations defining the structure and content of our knowledge graph. **(B)** Hop-based prediction strategies to find novel gene-disease associations via intermediary genes expressed in the same tissue at the RNA or protein levels. **(C)** Embedding-based prediction strategy to find novel gene-disease associations via distances/similarities in a latent space.

as protein-protein interaction networks and pathways in our KG), we defined hop-based inferencing rules with RNA- and protein-level expression in tissues as key components (Gurbuz et al., 2020). For instance, say a protein/gene instance X interacts with another protein/gene instance Y and these two entities are expressed in the same tissues. Then, it is assumed that the disease D that Y is related to is one-hop-related to the instance X. Similarly, if protein/gene X interacts with Y, Y interacts with another protein/gene Z, all these entities are expressed in the same tissues and Z is associated with the disease D, we say that D is two-hop-related to the instance X (see **Figure 1B**). These candidate gene-disease links can be ranked according to the total number of publications in the X-Y-D or X-Y-Z-D path (i.e., the sum of the edge weights). A sample mock-up diagram can be found in **Supplementary Figure S2**.

TABLE 2 | Validation scheme based on the date when the interaction was first reported.

Node1	Node2	Interaction	First referenced	Graph	Type
Gi	Dk	hasDisease	≤2010	KG_Before2010	Train data
Gi	Gj	activates	≤2010	KG_Before2010	Train data
Gj	Dk	hasDisease	>2010	KG_After2010	Test data

For inferencing, we used five different strategies: one-hop links filtered by protein expression in the tissues, one-hop links filtered by RNA expression in the tissues, two-hop links filtered by protein expression in the tissues, two-hop links filtered by RNA expression in the tissues, and the union of all these types of predictions (see **Figure 1B**). We also evaluated the performance of the one- and two-hop strategies without the tissue filters.

Random Baselines

We created 100 random KGs to evaluate the performance of the hop-based predictions. To this end, we shuffled the identity of the protein/gene entities, which maintained the structure of the KG unchanged but affected the biology encoded by the gene-gene, gene-disease, and gene-tissue components of the graph.

In Silico Validation of Tissue-Based Gene-Disease Predictions

For each gene-gene and gene-disease link, we have the information of when the association was first reported (published) and what is the last record (publication) of such association. Accordingly, we used a prospective time-split validation scheme, where interactions and indications published before or in 2010 were eligible for the training data, whereas indications reported after 2010 were used to construct a gold standard or test set (see **Table 2**). It is important to note that the gold standard was constructed by making sure that only genes and diseases which also exists in KG_Before2010 were included, as these are the only cases that can be predicted. We further refined the gold standard by removing gene-disease pairs separated by more than two hops in the original KG. This led to fairer performance metrics because we considered a maximum of two hops of separation between genes and diseases in our predictions. Therefore, the final test set comprised 5,176 reference gene-disease associations.

Knowledge Graph Embeddings

We employed the Nunes et al. (2020) implementation of the most commonly used embedding methods for KGs, which are RDF2Vec[6], DistMult[7], TransE[8], TransH[9], and TransD[10] to embed KG_Before2010 into a low-dimensional space (see **Figure 1C**). We used a 200-dimensional space as recommended in Nunes et al. (2020). Therefore, we obtained 200-dimensional representations of all the gene and disease entities, which we used to calculate Euclidean distances and cosine similarities between gene-gene and gene-disease pairs. These distances/similarities were used to build a Random Forest model that we applied to the prediction of gene-disease links. We selected this machine learning approach based on the work of Celebi et al. (2019) and Nunes et al. (2020) who found that Random Forests outperformed other techniques in their studies for predicting gene-disease links from ontologies.

To train the Random Forest and evaluate its performance, we labeled all the gene-disease pairs separated by at most 2-hops and which did not take place in the train and test data (**Table 2**) as negative cases. This allowed to construct a training set (98,426 positive and 98,426 negative cases) and a test set (5,176 positive and 5,176 negative cases).

Performance Metrics

We evaluated the overall prediction accuracy of the inference strategies described above using the following definitions:

- True positive: Gene-disease link inferred from KG_Before2010 and that is listed in the KG_After2010 gold standard.
- False positive: Gene-disease link inferred from KG_Before2010 but that is not listed in the KG_After2010 gold standard.
- False negative: Gene-disease link not inferred from KG_Before2010 but that is listed in the KG_After2010 gold standard.

In addition, we constructed a table with all the possible gene-disease links that can be formed with the KG_Before2010 data (18,045 unique genes and 330 unique diseases for a total of 5,954,850 possible gene-disease associations). This list was further reduced to gene-disease pairs separated by at most two hops in the KG_Before2010 for a total of 458,640. Then, we determined which of those combinations were corroborated in the gold standard (positive cases) and scanned the list decreasingly based on the scores assigned to each pair by the hop-based prediction strategies (see **Figure 1B**). Gene-disease links not predicted by the hop-based methods were given a score of 0. This allowed us to construct Receiving Operating Characteristic (ROC) and Precision-Recall curves (Cannistraci et al., 2013) using the following definitions:

[6] https://github.com/IBCNServices/pyRDF2Vec (accessed on 11.10.21).

[7] https://github.com/thunlp/OpenKE (accessed on 11.10.21).

[8] https://github.com/thunlp/OpenKE (accessed on 11.10.21).

[9] https://github.com/liseda-lab/KGE_Predictions_GD (accessed on 11.10.21).

[10] https://github.com/liseda-lab/KGE_Predictions_GD (accessed on 11.10.21).

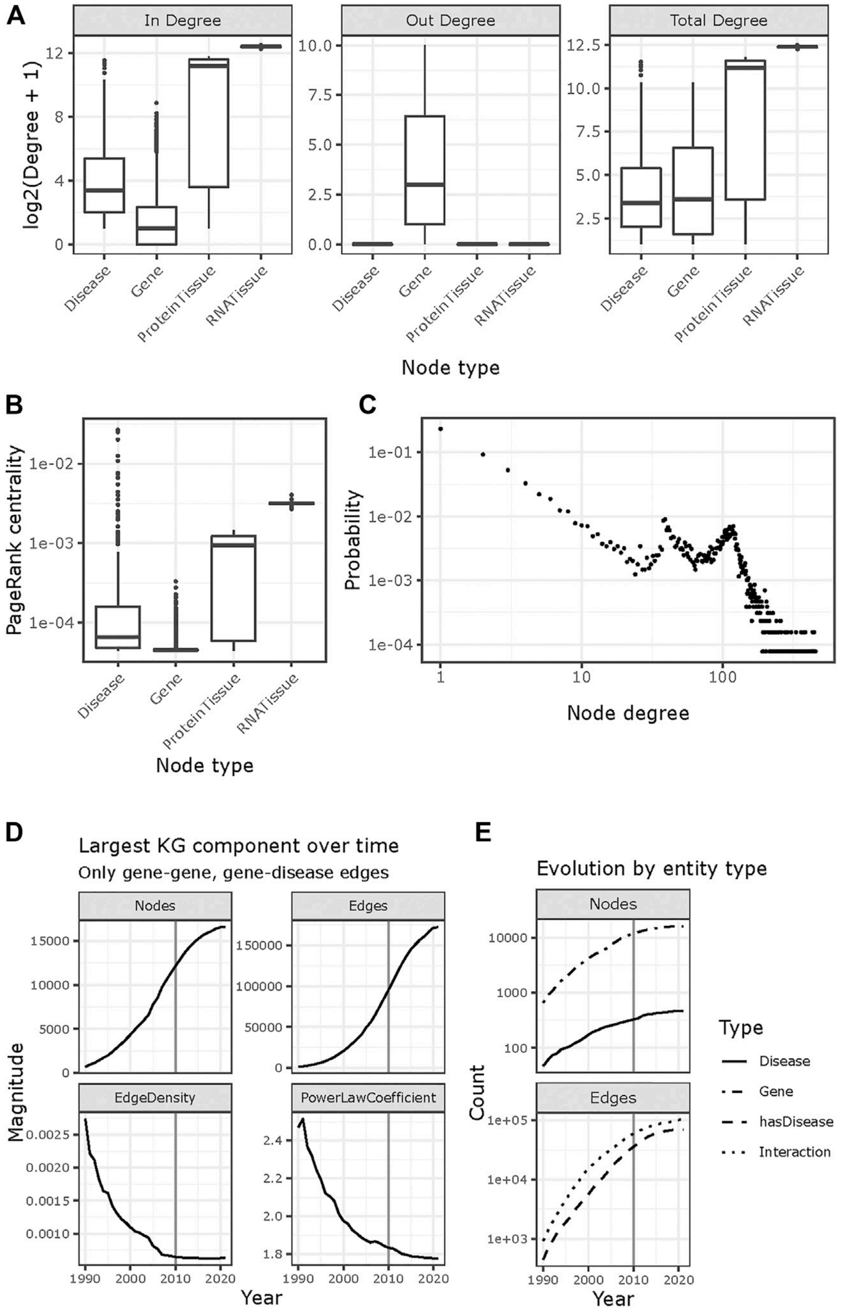


FIGURE 2 | Topological properties of KG_Before 2010. **(A)** In- and out-degree of the nodes in each node type. Also shown is the total degree, defined as the sum of the in- and out-degree. All node types have hubs with over 100 edges (log2 (101) ≈ 6.7). **(B)** PageRank centrality, by node type. **(C)** Probability of each node degree suggest a power law; both axes are log scaled. **(D, E)** Temporal evolution of gene-gene and gene-disease links between 1990 and 2021. Edges were filtered according to their first mention in the literature. **(D)** Evolution of the largest weakly connected component over time, in terms of node count, edge count, edge density and power law coefficient. **(E)** Details on the relative growth by node types (genes or diseases) and by edge types (gene-gene interactions and gene-disease annotations).

- True positive: Gene-disease link above current weight threshold that was reported after 2010.
- False positive: Gene-disease link above current weight threshold that was not reported after 2010.
- False negative: Gene-disease link below current weight threshold that was reported after 2010.
- True negative: Gene-disease link below current weight threshold that was not reported after 2010.

Data and Code Availability

Gene-gene links and gene-disease links were extracted from the commercial database Metabase[11], which prevents us from sharing these data. However, the code we used to define and populate our KG is available in the github link: https://github.com/bi-compbio/kg_for_ie and can be used with publicly available databases like StringDB[12] for gene-gene links and DisGeNET[13] for gene-disease links. Gene-tissue links for the resulting KG can be retrieved using the Human Protein Atlas R package[14] (Tran et al., 2019).

RESULTS

Characterization of the Graph

The KG was created from 18,790 unique nodes (464 diseases; 18,165 genes; 124 ProteinTissues; 37 RNATissues) and 669,900 edges (70,380 hasDisease; 263,106 hasProteinExpression; 234,294 hasRNAExpression; 102,120 Interaction). The graph was directed and contained no multi-edges. After imposing the publication date restriction, *KG_Before2010* had 12,906 nodes (330 diseases; 12,417 Ensembl genes; 122 Protein-Tissues; 37 RNA-Tissues) and 518,427 edges (34,201 hasDisease; 222,438 hasProteinExpression; 197,563 hasRNAExpression; 64,225 Interaction). Its edge density was 0.00311.

The topological properties of *KG_Before2010* suggest it follows a scale free architecture (power law coefficient of 2, **Figure 2C**). Regarding their in-degree, genes are the least central nodes, followed by diseases, Protein-Tissues, and RNA-Tissues (**Figure 2A**). The out-degree is only positive for genes, with a maximum of 1,025. The total degree shows trends like those in the in-degree, except for genes and disease being on par due to the addition of the out-degree of genes. Using PageRank as a centrality measure depicts a similar scenario to the in-degree (**Figure 2B**). All the node types show heavy tails and hubs with more than 100 connections (**Figures 2A,C**). Such properties are in line with those of molecular networks and KGs in the biomedical domain.

Temporal Evolution of Indications

To characterize the time dynamics of indication discovery, we started from the induced subgraph containing genes and diseases only and built year-specific subgraphs by removing the edges whose first mention in literature was posterior to the year under consideration (**Figure 2**). When accounting for all-time data (i.e., the 2021 network), the network encompassed 16,552 nodes and 172,118 edges (16,530 and 172,044 in the largest weakly connected component, respectively). In contrast, the largest connected component dating from 1990 consisted of 705 nodes and 1,360 edges, and the one from 2010 had 12,151 nodes and 95,375 edges. We observed a reduction of the increase rate in both the number of nodes and edges, more pronounced from 2015 onwards (**Figure 2D**), which might be explained by changes in the literature curation criteria or by the pace of data ingestion. Both edge density and the power law coefficient tend to decrease and plateau (**Figure 2D**), which might indicate the new addition of nodes over time that remain sparsely connected. The growth patterns in number of nodes and edges also hold for their sub-types (**Figure 2E**).

Performance Evaluation of Hop-Based Methods

Overall precision and recall values for the different strategies to predict indirect gene-disease links are shown in **Table 3**. The average performance metrics across 100 random KG are also reported, together with p-values from a one-sided z-test comparing the actual performance values and the distribution of random ones. In all cases, both precision and recall are higher than expected by chance with the one-hop with RNA tissue predictions producing the best precision-recall combination, followed by the one-hop with Protein tissue inferences (**Table 3** and **Figure 3A**). Interestingly, while removing the tissue expression entity from the KG does have an impact on precision, the sensitivity of the one-hop and two-hop strategies without tissue is higher. This responds to the fact that, in these cases, the intermediary nodes connecting the gene with its predicted associated disease (see **Figure 1B**) do not have to be expressed in the same tissue, resulting in many more predicted gene-disease links and a higher probability to identify pairs in the gold standard. This is also the case for the two-hop and the union of all predictions (**Figure 3A**). However, when precision and recall are summarized with the F1 statistic, it becomes evident that the best predictions come from the one-hop methods (**Figure 3A**). We believe that, even though the F1 metric from the one-hop no tissue approach is comparable to that of the predictions with tissue constraints, it is better to ensure tissue homogeneity.

Table 3 shows that each hop-based method predicts tens of thousands of gene-disease links, a number of associations that is unlikely to be validated by experimental means. Therefore, we assessed the performance of the hop-based approaches for early retrieval by looking at metrics for the top-100 predictions (see **Supplementary Table S1**). In particular, Precision@100 shows that one-hop with RNA tissue and one-hop without tissue constraints are the best approaches for early recognition.

To better understand whether predicted gene-disease links with high scores were corroborated in publications after 2010, we built performance curves by scanning a list of all possible gene-disease pairs in KG_Before2010 separated by at most 2 hops (see Methods). **Figure 3B** shows the receiver operating characteristic (ROC) and Precision-Recall curves of all the gene-disease inference strategies, while **Figure 3C** shows the areas under these curves. The plots corroborate that one-hop predictions are the best when it comes to early retrieval and the tails of the curves represent the random ranks for gene-disease pairs that were given artificial scores of 0 (see Methods).

[11] https://www.cortellislabs.com/page/?api=api-MB (accessed on: 09.11.2021).

[12] https://string-db.org/(accessed on: 12.11.2021).

[13] https://www.disgenet.org/(accessed on: 12.11.2021).

[14] https://bioconductor.org/packages/release/bioc/vignettes/hpar/inst/doc/hpar.html (Accessed on: 13.10.2021).

TABLE 3 | Types of inferencing and their overall performance scores based on a total of 5,176 reference gene-disease links reported after 2010. Average ± standard deviations are reported for the random predictions.

Type of inferencing	Predicted links	Precision	Precision at100	Precision (random)	*p*-value precision	Recall	Recall (random)	*p*-value recall
All the inferences	170,506	0.0296	0.23	0.0152 ± 0.0003	2.55E-284	0.9737	0.5449 ± 0.0223	1.50E-81
One-hop and protein tissue	33,633	0.0817	0.21	0.0227 ± 0.0006	0.00E+00	0.5307	0.2234 ± 0.0060	0.00E+00
One-hop and RNA tissue	45,664	0.0794	0.3	0.0227 ± 0.0006	0.00E+00	0.7007	0.2235 ± 0.0061	0.00E+00
Two-hop and protein tissue	120,319	0.0319	0.14	0.0158 ± 0.0003	0.00E+00	0.7417	0.5247 ± 0.0088	4.50E-127
Two-hop and RNA tissue	167,939	0.0295	0.23	0.0157 ± 0.0003	7.10E-286	0.9571	0.5286 ± 0.0088	0.00E+00
One-hop without tissue	47,734	0.0787	0.30	0.0227 ± 0.0006	0.00E+00	0.7262	0.2235 ± 0.0061	0.00E+00
Two-hops without tissue	174,305	0.0291	0.23	0.0157 ± 0.0003	7.10E-286	0.9795	0.5286 ± 0.0088	0.00E+00

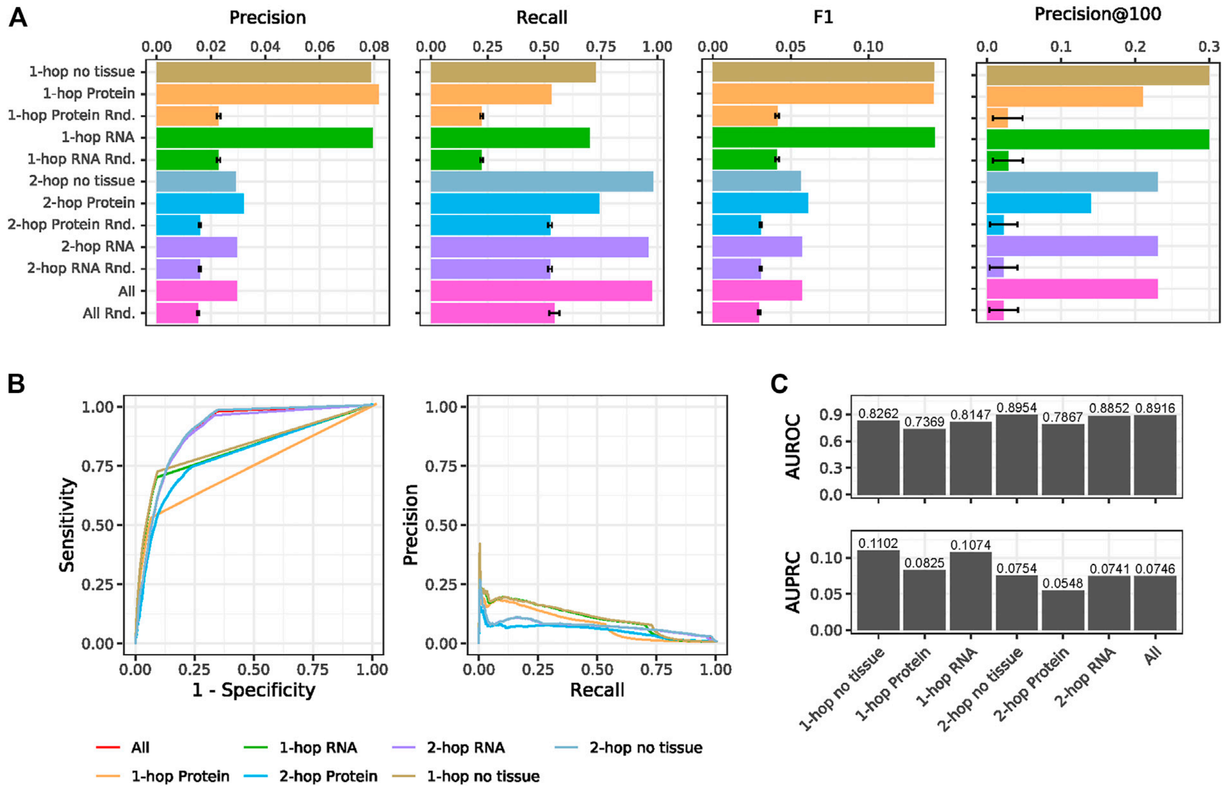


FIGURE 3 | Performance evaluation of the hop-based predictions. **(A)** Precision, Recall, F1 and Precision@100 metrics calculated from all the gene-disease links predicted by each hop-based approach. **(B)** ROC and Precision-Recall performance curves for all the hop-based prediction methods. **(C)** Area under the ROC (AUROC) and Precision-Recall (AUPRC) curves shown in **(B)**.

Performance Evaluation Based on Random Forest on Several Knowledge Graph Embeddings

We employed 5 different dimensionality reduction strategies[15] (Nunes et al., 2020) to embed the KG_Before2010 into a 200-dimensional space, obtain vector representations of genes and diseases, and use these vectors to build a machine learning model for gene-disease link prediction (see Methods and **Figure 1C**). Intuitively, a good embedding method should put gene-disease associations reported in the gold standard near each other in the latent space. We computed the distance between all gene-disease pairs, binned the distance range into 10 groups, and calculated the probability of finding gene-disease links reported after 2010 within each bin (**Supplementary Figures S3, S4**). This

[15] https://github.com/liseda-lab/KGE_Predictions_GD (Accessed on 13.10.2021).

TABLE 4 | Random Forest predictions on different embeddings.

With tissue				No tissue			
Category	**Precision**	**Recall**	**F1**	**Category**	**Precision**	**Recall**	**F1**
DistMult/Cos-sim/@all	0.5339	0.1609	0.247278	DistMult_notissue/Cos-sim/@all	0.3747	0.0326	0.059981
DistMult/Euclidean/@all	0.6758	0.2413	0.355622	DistMult_notissue/Euclidean/@all	0.4152	0.0917	0.150222
RDF2Vec/Cos-sim/@all	0.4765	0.2057	0.287353	RDF2Vec_notissue/Cos-sim/@all	0.5711	0.1636	0.25434
RDF2Vec/Euclidean/@all	0.412	0.242	0.304905	RDF2Vec_notissue/Euclidean/@all	0.4074	0.1246	0.190835
TransD/Cos-sim/@all	0.7356	0.3827	0.503468	TransD_notissue/Euclidean/@all	0.6038	0.3066	0.40669
TransD/Euclidean/@all	0.5312	0.3462	0.419196	TransD_notissue/Cos-sim/@all	0.6794	0.1027	0.178428
TransE/Cos-sim/@all	**0.6988**	**0.6854**	**0.692035**	TransE_notissue/Cos-sim/@all	0.6894	0.6049	0.644392
TransE/Euclidean/@all	0.6604	0.5085	0.57458	TransE_notissue/Euclidean/@all	0.5958	0.3098	0.407639
TransH/Cos-sim/@all	0.6922	0.5884	0.636093	TransH_notissue/Euclidean/@all	0.54	0.3021	0.387446
TransH/Euclidean/@all	0.6187	0.5818	0.599683	TransH_notissue/Cos-sim/@all	0.6601	0.6263	0.642756

Bold numbers show the highest performance.

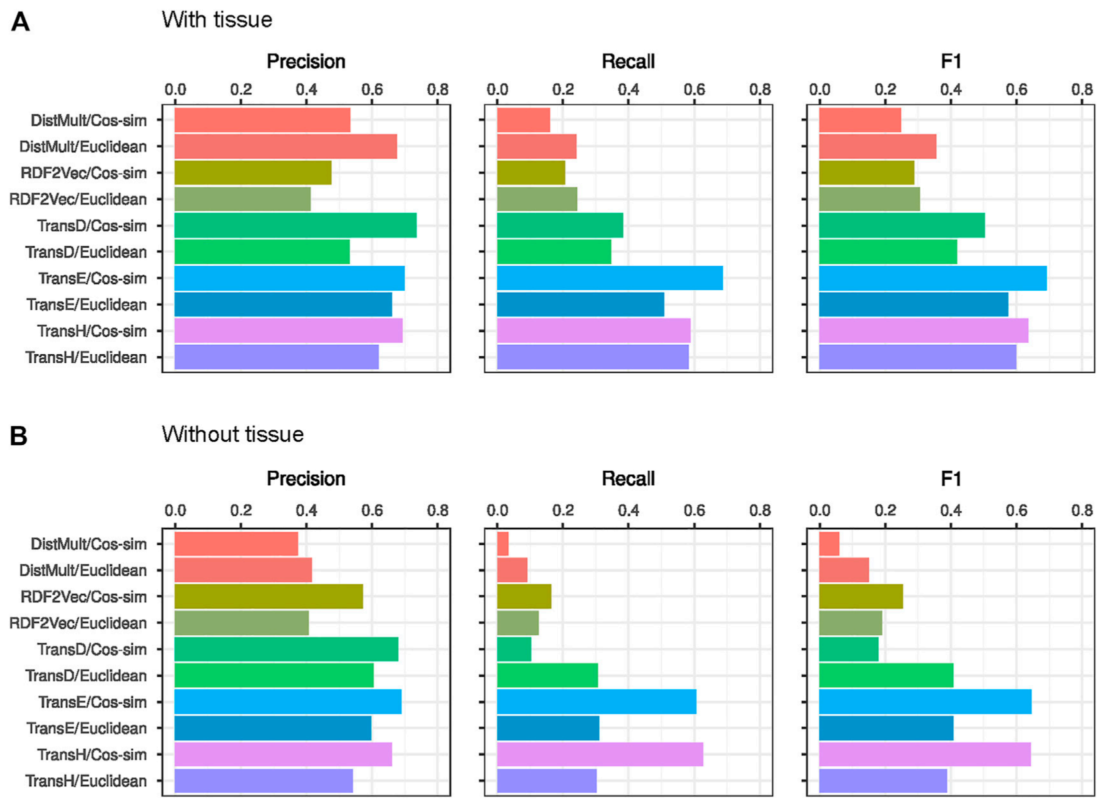


FIGURE 4 | Embedding based gene-disease prediction evaluation. **(A)** Embedding performances in which gene-tissue links were included in the knowledge graph. **(B)** Embedding performances in which gene-tissue links were not included in the knowledge graph.

analysis showed that TransD, TransE, and TransH were the approaches that produced the expected gene-disease proximity patterns. To confirm whether these methods would indeed produce good gene-disease link predictions, we computed the Euclidean distances and cosine similarities between genes and diseases in the five different 200-dimensional spaces and used these measures to train a Random Forest model whose performance was evaluated with the gold standard mentioned above (see Methods). **Table 4** and **Figure 4** show the performance of the Random Forest predictions refined with and without the tissue expression information. TransE embeddings using cosine similarity vector as the training data for Random Forest achieved the best performance overall. These results also show that embeddings from the KG that contains gene-tissue links outperform the embeddings that don't have this information, highlighting the importance of this entity for the embedding approaches (**Table 4**; **Figure 4** and **Supplementary Table S2**).

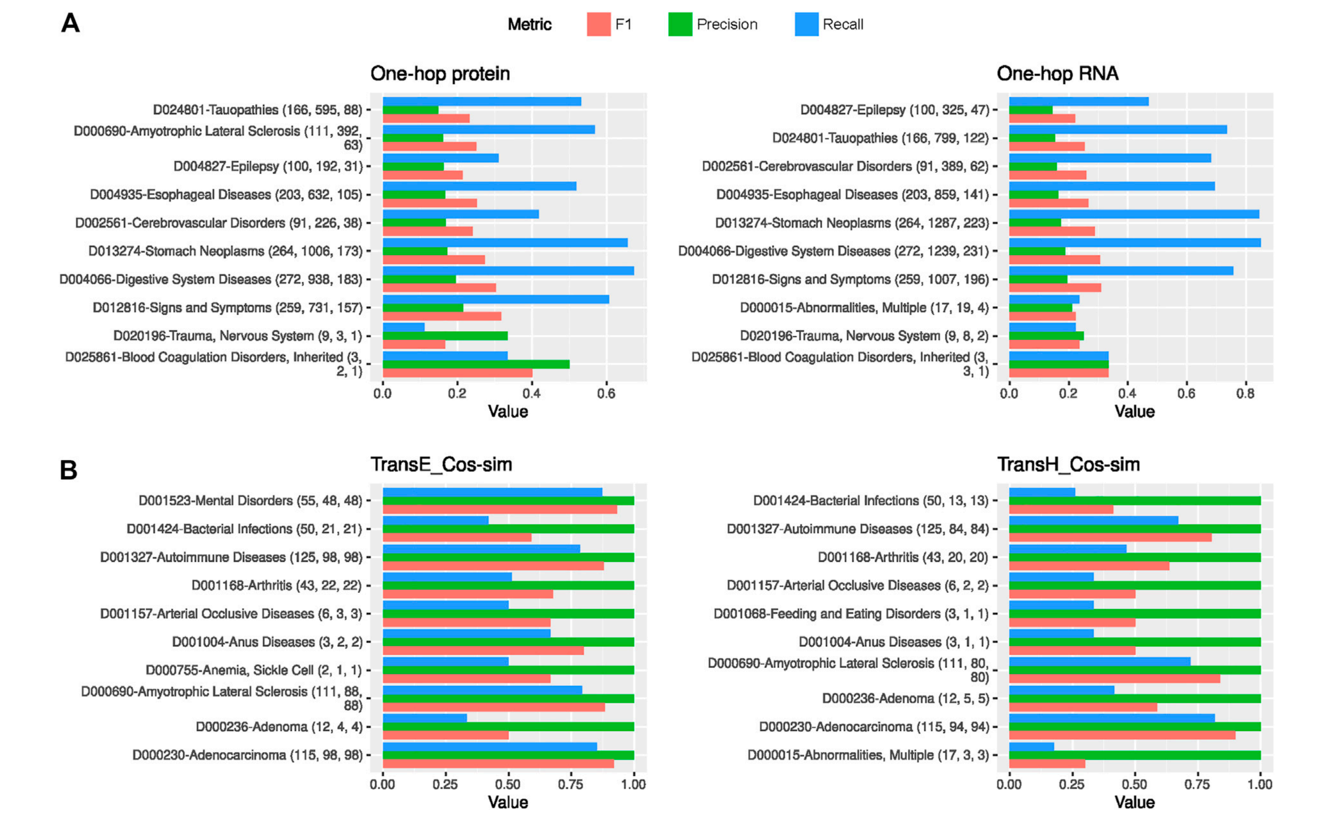


FIGURE 5 | Performance evaluation per disease. **(A)** Precision, recall and F1 metrics attained by each the top two best performing hop-based prediction methods. **(B)** Same as **(A)** but for the top two embedding methods. Only the top 10 diseases are shown based on the precision value. The numbers in parentheses indicate the total number of gene-disease links in the gold standard for that disease, the number of predicted gene-disease links and how many of those were positive, respectively.

Performance Evaluation Per Disease

Finally, we investigated the precision, recall, and F1 metrics for each disease separately to determine whether the biological knowledge encoded by the graph allows to make better predictions for certain diseases compared to others.

Figure 5A and **Supplementary Figure S5** show that the hop-based strategies tend to perform well in a common set of disorders like Tauopathies (D024801), Esophageal Diseases (D004935), Stomach Neoplasms (D013274), and Digestive System Diseases (D004066). A similar pattern is observed for the embedding methods, with Arthritis (D01168), Amyotrophic Lateral Sclerosis (D000690), Mental Disorders (D001523), and Bacterial Infections (D001424) among the top-10 diseases in at least four embedding approaches (**Figure 5B** and **Supplementary Figure S6**).

The hop-based strategies show a similar behavior across diseases: many predictions, which makes recall high but causes low precision (see numbers in brackets in **Supplementary Figure S5**). Yet, the one-hop strategies show more balance between precision and recall than the two-hop strategies, as reflected by higher F1 scores (**Supplementary Figure S5**). In contrast, the embedding methods produced, in general, less predictions for the top-10 diseases, which led to low recalls but very high precisions as most of them were true hits (**Supplementary Figure S6**). This corroborates the metrics reported in **Supplementary Table S2** and highlights that these prediction strategies are well suited for early retrieval tasks. Of note, when the embedding methods predicted more links for a disease (e.g., see Mental Disorders or ALS in TransE on **Figure 5B**), these were also mostly true hits, leading to high recall and F1 statistics. In the following section, we interpret some use cases for the diseases with the best local performance and showcase the interpretability of the predictions.

Use Cases

In order to showcase the potential of our approach, we identified the best performing disease areas as promising domains of application. Then, we demonstrate how both predictions with highest literature support and with highest prediction score yield sensible links that were confirmed after 2010.

Based on **Figure 4**, the best performing embedding method was TransE followed by a Random Forest prediction on the cosine-similarity of the gene and disease low-dimensional vectors. Performance evaluation per disease (**Figure 5B**) showed that this method attained its highest F1 score for Mental Disorders (D001523) and Amyotrophic Lateral Sclerosis (ALS) (D000690). There were 55 gene-Mental Disorders pairs that were published after 2010, and with the

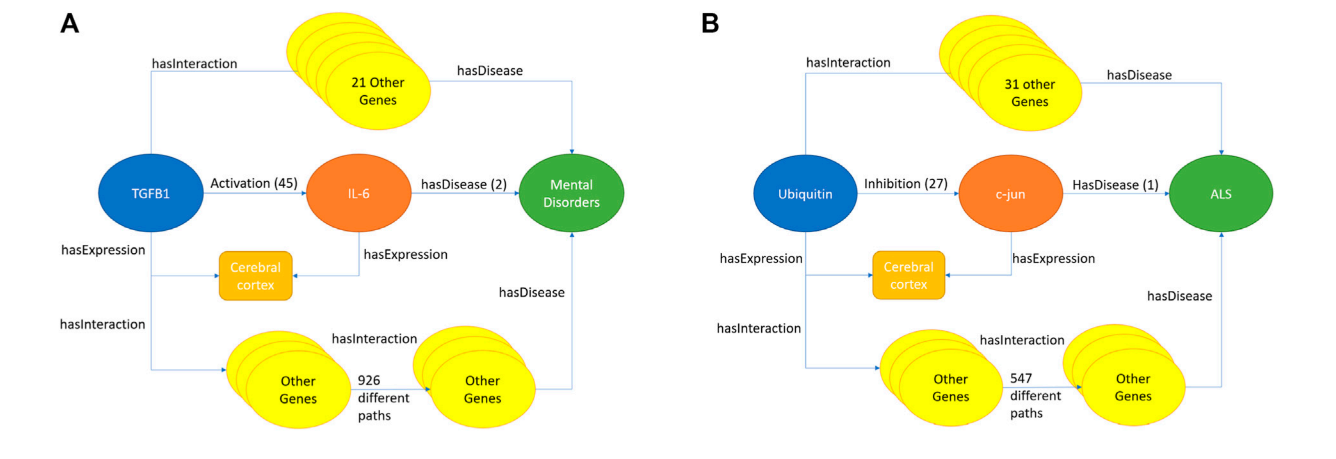


FIGURE 6 | Example prediction from the knowledge graph. **(A)** TGFB1 is connected to Mental Disorders via IL6. **(B)** Ubiquitin is connected to ALS via c-Jun. Both panels also show the number of alternative connections from the genes to the predicted disease via one-hop and two-hop links.

TransE embedding strategy 48 of them were correctly predicted. In the ALS case, TransE recovered 88 of the 111 gene-disease pairs, TransH recovered 80, and the one-hop with Protein tissue strategy covered 63 reported after 2010. To explain these predictions, one can go back to the KG and study their paths, literature, and tissue support.

For Mental Disorders, the path with the strongest literature backing (i.e., total number of publications) was the one linking TGFB1 with this disease group via IL-6, both genes co-expressed in the cerebral cortex (see **Figure 6A**). Before 2010, the activation of IL-6 by TGFB1 is endorsed by 45 publications, while the link between IL-6 and Mental Disorders is endorsed by 2 as shown in **Figure 6A**. Moreover, there are 22 different one-hop paths (TGFB1—gene X—Mental Disorders) and 926 different two-hops paths (TGFB1—gene Y—gene Z—Mental Disorders) between TGFB1 and Mental Disorders in which all genes are expressed in the cerebral cortex. The predicted TGFB1-Mental Disorders link, which later were published in López-González et al. (2019), supports the theory that dysfunction of the immune system plays an important role in the etiology of mental illnesses, such as schizophrenia and depression (Frydecka et al., 2013; Bialek et al., 2020). In fact, significantly higher serum levels of the IL-6 and TGFB1 cytokines have been reported in patients with schizophrenia compared to healthy controls (Ergün et al., 2017) and mutations in TGFB1 have been associated with the susceptibility and treatment response of schizophrenia (Frydecka et al., 2013) and major depressive disorder (Bialek et al., 2020).

For ALS, on the other hand, the top prediction from the embedding methods is Ubiquitin and ALS. There are 32 different one-hop paths (Ubiquitin—gene X—ALS) and 547 two-hops paths (Ubiquitin—gene Y—gene Z—ALS) in which all the genes in the paths were expressed in cerebral cortex as shown in **Figure 6B**. In this context, the strongest one-hop literature link (in terms of publication numbers) is Ubiquitin—c-Jun—ALS with 27 publications. The predicted Ubiquitin—ALS link is supported by the literature (Hasegawa and Arai, 2007; Watanabe et al., 2010; Keller et al., 2012) stating that Ubiquitin inclusions have been seen in ALS patients. JNK/c-Jun signaling has been found involved in the cell death caused by TDP-43, which is closely linked with ALS and ubiquitin inclusions (Suzuki and Matsuoka, 2013). It is important to note that the link between Ubiquitin and ALS has been discussed in the literature before 2010 (Hasegawa and Arai, 2007), but this was not considered a high-confidence association in Metabase and was therefore not known by our predictive model.

DISCUSSION

In this study, we presented the evaluation of the effectiveness of the methodology that we developed to build a comprehensive KG for target-repurposing (indication expansion) studies. We first evaluated the effectiveness of the constructed KG for target-disease prediction via semantic inferencing, i.e., by linking targets and diseases that are one or two hops away from each other passing through genes that are expressed in the same tissue as the target. In addition, we checked whether embedding our KG to a low dimensional space to then use the inferred gene and disease coordinates to generate dis-/similarity inputs for a machine learning model could lead to more reliable predictions. For these experiments, we divided the KG in two parts such that edges reported before 2010 were used as training data and edges reported after 2010 served as our gold standard. This splitting allowed us to have a reliable gold standard reference, supported by the literature.

Our experiments showed that the hop-based strategies using RNA- and Protein-level expression data significantly outperformed our random baselines and were more precise than hop-based predictors without tissue information. Also, the one-hop RNA prediction method outperformed the two-hop and the one-hop Protein strategies. This reflects the fact that there is much more available information about gene expression at the RNA level (and/or protein abundance data is still incomplete) and suggests that two-hop predictions incorporate too many false positives to be reliable, especially for early recognition. In addition, using Euclidean distances and cosine similarities between gene and disease vectors inferred by KG embeddings to train a Random Forest model led to much better gene-disease prediction results. In particular, the TransE and TransH embedding methods followed by the computation of cosine similarities between genes and diseases represented the best training platform for the constructed Random Forests. Our initial quality controls of the embeddings already hinted at this result, as the probability of finding gold standard gene-disease associations at short embedding distances was very high for these methods. Moreover, added value by gene-tissue links is more visible in the KG Embeddings strategies.

One of the limitations of this study is that when creating the training data set, the true negatives are usually unknown. We use as proxy gene-diseases for which no connection is known, but this does not imply that they are unrelated. This can also overestimate the number of false positives: even though a predicted link might have not yet discovered, we simply assumed that if the predicted link does not appear in the KG after 2010, then it is a false positive. Secondly, gene-gene interaction network is incomplete due to evolution of the network over time (which is continuous), and also it is technically challenging and costly to test each protein pairs' interaction in humans. Thirdly, we have the relations for tissue-specific expressions, but we cannot distinguish cell type-specific effects. And genes and diseases which are linked to low number of genes and diseases (in other words with less neighbors) are most likely result in worse predictions. Lastly, this study only focuses on the human data and other organisms are out of scope. However, this method can be applied on other organism data as well.

Although the explainability of the predictions, i.e., the glass-box property of the KG, is easier to see in the hop-based methods, it is also possible to query the KG in order to explain the predictions produced by embedding combined with machine learning approaches, as we did for our two use cases. In addition, it is possible to inspect the resulting Random Forest model to determine which features have a strong impact on a decision. This kind of analysis was outside of the scope of this study.

To the best of our knowledge, this work is the first one to apply inferencing constrained by tissue expression on a semantic KG. Moreover, our KG is built from full-text literature sources and not only abstracts, which means that the graph does not miss any important information and does not depend on NLP tools like other literature-based approaches. As future work, we plan to extend the data sources employed to construct our KG, explore other predictive modelling methods, as well as to make it a key component of our target identification pipelines.

DATA AVAILABILITY STATEMENT

The data analyzed in this study was obtained from Cortellis Metabase, the following licenses/restrictions apply: it is a commercial database. Code used throughout this study is available at https://github.com/bi-compbio/kg_for_ie. Any further requests to access these datasets should be directed to OG, ozge.gurbuz@boehringer-ingelheim.com

AUTHOR CONTRIBUTIONS

OG was major contributor for writing the manuscript and conducted the literature review, implemented the knowledge graph, inferencing and deploying embeddings and Random Forest. GA-L implemented the performance measures and constructed the gold standard. SP-A implemented the statistical characterization of the knowledge graph. MS, CH, NL, and FF-A supervised this project and provided the data. All of the authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank to Karlheinz Spenny, Andreas Leusch, Matthias Zentgraf, Jan-Nygaard Jensen, Udo Maier, Stephan Zahn, and Maksim Kolchin for their valuable support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.814093/full#supplementary-material

REFERENCES

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., and Persidis, A. (2011). Literature Mining, Ontologies and Information Visualization for Drug Repurposing. *Brief. Bioinform.* 12, 357–368. doi:10.1093/bib/bbr005

Bialek, K., Czarny, P., Watala, C., Wigner, P., Talarowska, M., Galecki, P., et al. (2020). Novel Association between TGFA, TGFB1, IRF1, PTGS2 and IKBKB Single-Nucleotide Polymorphisms and Occurrence, Severity and Treatment Response of Major Depressive Disorder. *Peerj* 8, e8676. doi:10.7717/peerj.8676

Brown, A. S., and Patel, C. J. (2017). A Standard Database for Drug Repositioning. *Sci. Data* 4, 170029. doi:10.1038/sdata.2017.29

Cannistraci, C. V., Alanis-Lobato, G., and Ravasi, T. (2013). From Link-Prediction in Brain Connectomes and Protein Interactomes to the Local-Community-Paradigm in Complex Networks. *Sci. Rep.* 3, 1613. doi:10.1038/srep01613

Celebi, R., Uyar, H., Yasar, E., Gumus, O., Dikenelli, O., and Dumontier, M. (2019). Evaluation of Knowledge Graph Embedding Approaches for Drug-Drug Interaction Prediction in Realistic Settings. *Bmc Bioinformatics* 20, 726. doi:10.1186/s12859-019-3284-5

Chen, H., and Xie, G. (2010). The Use of Web Ontology Languages and Other Semantic Web Tools in Drug Discovery. *Expert Opin. Drug Discov.* 5, 413–423. doi:10.1517/17460441003762709

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011). Exploiting Drug-Disease Relationships for Computational Drug Repositioning. *Brief. Bioinform.* 12, 303–311. doi:10.1093/bib/bbr013

Ergün, S., Yanartaş, Ö., Kandemir, G., Yaman, A., Yıldız, M., Haklar, G., et al. (2017). The Relationship between Psychopathology and Cognitive Functions with Cytokines in Clinically Stable Patients with Schizophrenia. *Psychiatry Clin. Psychopharmacol.* 28, 66–72. doi:10.1080/24750573.2017.1380920

Frydecka, D., Misiak, B., Beszlej, J. A., Karabon, L., Pawlak-Adamska, E., Tomkiewicz, A., et al. (2013). Genetic Variants in Transforming Growth Factor-β Gene (TGFB1) Affect Susceptibility to Schizophrenia. *Mol. Biol. Rep.* 40, 5607–5614. doi:10.1007/s11033-013-2662-8

Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., and Bolton, E. (2016). Predicting Drug Target Interactions Using Meta-Path-Based Semantic Network Analysis. *Bmc Bioinformatics* 17, 160. doi:10.1186/s12859-016-1005-x

Geleta, D., Nikolov, A., Edwards, G., Gogleva, A., Jackson, R., Jansson, E., et al. (2021). *Biological Insights Knowledge Graph: An Integrated Knowledge Graph to Support Drug Development.* Biorxiv. doi:10.28.46626210.1101/2021.10.28.466262

Gurbuz, O., Sun, M., and Lawless, N. (2020). "A Methodology to Develop Knowledge Graphs for Indication Expansion: An Exploratory Study," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea, 16-19 Dec. 2020 (IEEE), 1720–1727. doi:10.1109/bibm49941.2020.9313179

Han, Z. J., Xue, W. W., Tao, L., and Zhu, F. (2018). Identification of Novel Immune-relevant Drug Target Genes for Alzheimer's Disease by Combining Ontology Inference with Network Analysis. *Cns Neurosci. Ther.* 24, 1253–1263. doi:10.1111/cns.13051

Härtner, F., Andrade-Navarro, M. A., and Alanis-Lobato, G. (2018). Geometric Characterisation of Disease Modules. *Appl. Netw. Sci.* 3, 10. doi:10.1007/s41109-018-0066-3

Hasegawa, M., and Arai, T. (2007). Component of Ubiquitin-Positive Inclusions in ALS. *Brain Nerve* 59, 1171–1177.

Herrero-Zazo, M., Segura-Bedmar, I., Hastings, J., and Martínez, P. (2015). DINTO: Using OWL Ontologies and SWRL Rules to Infer Drug-Drug Interactions and Their Mechanisms. *J. Chem. Inf. Model.* 55, 1698–1707. doi:10.1021/acs.jcim.5b00119

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). *What Do We Need to Build Explainable AI Systems for the Medical Domain?* Arxiv.

Holzinger, A. (2018). "From Machine Learning to Explainable AI." in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Košice, Slovakia, 23-25 Aug. 2018 (IEEE), 55–66. doi:10.1109/disa.2018.8490530

Kanza, S., and Frey, J. G. (2019). A New Wave of Innovation in Semantic Web Tools for Drug Discovery. *Expert Opin. Drug Discov.* 14, 433–444. doi:10.1080/17460441.2019.1586880

Katsila, T., and Matsoukas, M.-T. (2018). How Far Have We Come with Contextual Data Integration in Drug Discovery? *Expert Opin. Drug Discov.* 13, 791–794. doi:10.1080/17460441.2018.1504767

Keller, B. A., Volkening, K., Droppelmann, C. A., Ang, L. C., Rademakers, R., and Strong, M. J. (2012). Co-aggregation of RNA Binding Proteins in ALS Spinal Motor Neurons: Evidence of a Common Pathogenic Mechanism. *Acta Neuropathol.* 124, 733–747. doi:10.1007/s00401-012-1035-z

Lecue, F. (2020). On the Role of Knowledge Graphs in Explainable AI. *Sw* 11, 41–51. doi:10.3233/sw-190374

Lekka, E., Deftereos, S. N., Persidis, A., Persidis, A., and Andronis, C. (2011). Literature Analysis for Systematic Drug Repurposing: a Case Study from Biovista. *Drug Discov. Today Ther. Strateg.* 8, 103–108. doi:10.1016/j.ddstr.2011.06.005

Lin, Y., Mehta, S., Küçük-McGinty, H., Turner, J. P., Vidovic, D., Forlin, M., et al. (2017). Drug Target Ontology to Classify and Integrate Drug Discovery Data. *J. Biomed. Semant* 8, 50. doi:10.1186/s13326-017-0161-x

López-González, I., Pinacho, R., Vila, È., Escanilla, A., Ferrer, I., and Ramos, B. (2019). Neuroinflammation in the Dorsolateral Prefrontal Cortex in Elderly Chronic Schizophrenia. *Eur. Neuropsychopharmacol.* 29, 384–396. doi:10.1016/j.euroneuro.2018.12.011

Nunes, S., Sousa, R. T., and Pesquita, C. (2020). Predicting Gene-Disease Associations with Knowledge Graph Embeddings over Multiple Ontologies. Available at: https://arxiv.org/ftp/arxiv/papers/2105/2105.04944.pdf.

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., et al. (2020). Open Targets Platform: Supporting Systematic Drug-Target Identification and Prioritisation. *Nucleic Acids Res.* 49, D1302–D1310. doi:10.1093/nar/gkaa1027

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab, Technical Report.

Paliwal, S., de Giorgio, A., Neil, D., Michel, J.-B., and Lacoste, A. M. (2020). Preclinical Validation of Therapeutic Targets Predicted by Tensor Factorization on Heterogeneous Graphs. *Sci. Rep.* 10, 18250. doi:10.1038/s41598-020-74922-z

Parisi, D., Adasme, M. F., Sveshnikova, A., Bolz, S. N., Moreau, Y., and Schroeder, M. (2020). Drug Repositioning or Target Repositioning: A Structural Perspective of Drug-Target-Indication Relationship for Available Repurposed Drugs. *Comput. Struct. Biotechnol. J.* 18, 1043–1055. doi:10.1016/j.csbj.2020.04.004

Picart-Armada, S., Barrett, S. J., Willé, D. R., Perera-Lluna, A., Gutteridge, A., and Dessailly, B. H. (2019). Benchmarking Network Propagation Methods for Disease Gene Identification. *Plos Comput. Biol.* 15, e1007276. doi:10.1371/journal.pcbi.1007276

Qu, X. A., Gudivada, R. C., Jegga, A. G., Neumann, E. K., and Aronow, B. J. (2009). Inferring Novel Disease Indications for Known Drugs by Semantically Linking Drug Action and Disease Mechanism Relationships. *Bmc Bioinformatics* 10, S4. doi:10.1186/1471-2105-10-s5-s4

Sang, S., Yang, Z., Liu, X., Wang, L., Lin, H., Wang, J., et al. (2019). GrEDeL: A Knowledge Graph Embedding Based Method for Drug Discovery from Biomedical Literatures. *Ieee Access* 7, 8404–8415. doi:10.1109/access.2018.2886311

Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H., and Wang, J. (2018). SemaTyP: a Knowledge Graph Based Literature Mining Method for Drug Discovery. *Bmc Bioinformatics* 19, 193. doi:10.1186/s12859-018-2167-5

Sebastian, Y., Siew, E.-G., and Orimaye, S. O. (2017). Learning the Heterogeneous Bibliographic Information Network for Literature-Based Discovery. *Knowledge-Based Syst.* 115, 66–79. doi:10.1016/j.knosys.2016.10.015

Smalheiser, N. R. (2012). Literature-based Discovery: Beyond the ABCs. *J. Am. Soc. Inf. Sci.* 63, 218–224. doi:10.1002/asi.21599

Sosa, D. N., Derry, A., Guo, M., Wei, E., Brinton, C., and Altman, R. B. (2020). A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Pac. Symp. Biocomput* 25, 463–474.

Suzuki, H., and Matsuoka, M. (2013). The JNK/c-Jun Signaling axis Contributes to the TDP-43-Induced Cell Death. *Mol. Cel Biochem* 372, 241–248. doi:10.1007/s11010-012-1465-x

Tran, A. N., Dussaq, A. M., Kennell, T., Willey, C. D., and Hjelmeland, A. B. (2019). HPAanalyze: an R Package that Facilitates the Retrieval and Analysis of the Human Protein Atlas Data. *Bmc Bioinformatics* 20, 463. doi:10.1186/s12859-019-3059-z

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a Knowledge-Based Human Protein Atlas. *Nat. Biotechnol.* 28, 1248–1250. doi:10.1038/nbt1210-1248

Watanabe, T., Okeda, Y., Yamano, T., and Ono, S. (2010). An Immunohistochemical Study of Ubiquitin in the Skin of Sporadic Amyotrophic Lateral Sclerosis. *J. Neurol. Sci.* 298, 52–56. doi:10.1016/j.jns.2010.08.026

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., et al. (2012). Open PHACTS: Semantic Interoperability for Drug Discovery. *Drug Discov. Today* 17, 1188–1198. doi:10.1016/j.drudis.2012.05.016

Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C., and Wang, F. (2020). Knowledge-driven Drug Repurposing Using a Comprehensive Drug Knowledge Graph. *Health Inform. J* 26, 2737–2750. doi:10.1177/1460458220937101

Conflict of Interest: OG, GA-L, SP-A, NL, and FF-A are employees of Boehringer Ingelheim Pharma GmbH & Co. KG. MS, and CH were employed by Boehringer Ingelheim Pharma GmbH & Co. KG at the time of the study.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, orclaim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

METHODS
published: 04 April 2022
doi: 10.3389/fgene.2022.839453

PPA-GCN: A Efficient GCN Framework for Prokaryotic Pathways Assignment

Yuntao Lu[1,2], *Qi Li*[1]* *and Tao Li*[1]*

[1]Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China, [2]College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing, China

With the rapid development of sequencing technology, completed genomes of microbes have explosively emerged. For a newly sequenced prokaryotic genome, gene functional annotation and metabolism pathway assignment are important foundations for all subsequent research work. However, the assignment rate for gene metabolism pathways is lower than 48% on the whole. It is even lower for newly sequenced prokaryotic genomes, which has become a bottleneck for subsequent research. Thus, the development of a high-precision metabolic pathway assignment framework is urgently needed. Here, we developed PPA-GCN, a prokaryotic pathways assignment framework based on graph convolutional network, to assist functional pathway assignments using KEGG information and genomic characteristics. In the framework, genomic gene synteny information was used to construct a network, and ideas of self-supervised learning were inspired to enhance the framework's learning ability. Our framework is applicable to the genera of microbe with sufficient whole genome sequences. To evaluate the assignment rate, genomes from three different genera (*Flavobacterium* (65 genomes) and *Pseudomonas* (100 genomes), *Staphylococcus* (500 genomes)) were used. The initial functional pathway assignment rate of the three test genera were 27.7% (*Flavobacterium*), 49.5% (*Pseudomonas*) and 30.1% (*Staphylococcus*). PPA-GCN achieved excellence performance of 84.8% (*Flavobacterium*), 77.0% (*Pseudomonas*) and 71.0% (*Staphylococcus*) for assignment rate. At the same time, PPA-GCN was proved to have strong fault tolerance. The framework provides novel insights into assignment for metabolism pathways and is likely to inform future deep learning applications for interpreting functional annotations and extends to all prokaryotic genera with sufficient genomes.

Keywords: graph convolution network, prokaryotic genome, metabolic pathway, deep learning, self supervised

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Massimo La Rosa,
National Research Council (CNR), Italy
Beifang Niu,
Chinese Academy of Sciences (CAS), China

****Correspondence:***
Qi Li
liqi@ihb.ac.cn
Tao Li
litao@ihb.ac.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *20 December 2021*
Accepted: *17 March 2022*
Published: *04 April 2022*

Citation:
Lu Y, Li Q and Li T (2022) PPA-GCN: A Efficient GCN Framework for Prokaryotic Pathways Assignment. Front. Genet. 13:839453. doi: 10.3389/fgene.2022.839453

INTRODUCTION

With the rapid development of sequencing technology, the number of newly released prokaryotic genomes has exploded, providing an important foundation for subsequent research work (Doerks et al., 2004). Functional annotation and pathway assignment are important components of understanding the details of metabolism. Accordingly, a series of reference genome databases and functional annotation platforms have been developed (Benson et al., 2012; Federhen, 2012; Keegan et al., 2016; Chen et al., 2019; Bazgir et al., 2020). The Kyoto Encyclopedia of Genes and Genomes (KEGG) is one of the most widely used and reliable functional platforms, and it provides three annotation software tools, namely, BlastKOALA, GhostKOALA, and KofamKOALA, for functional annotation (Suzuki et al., 2014;

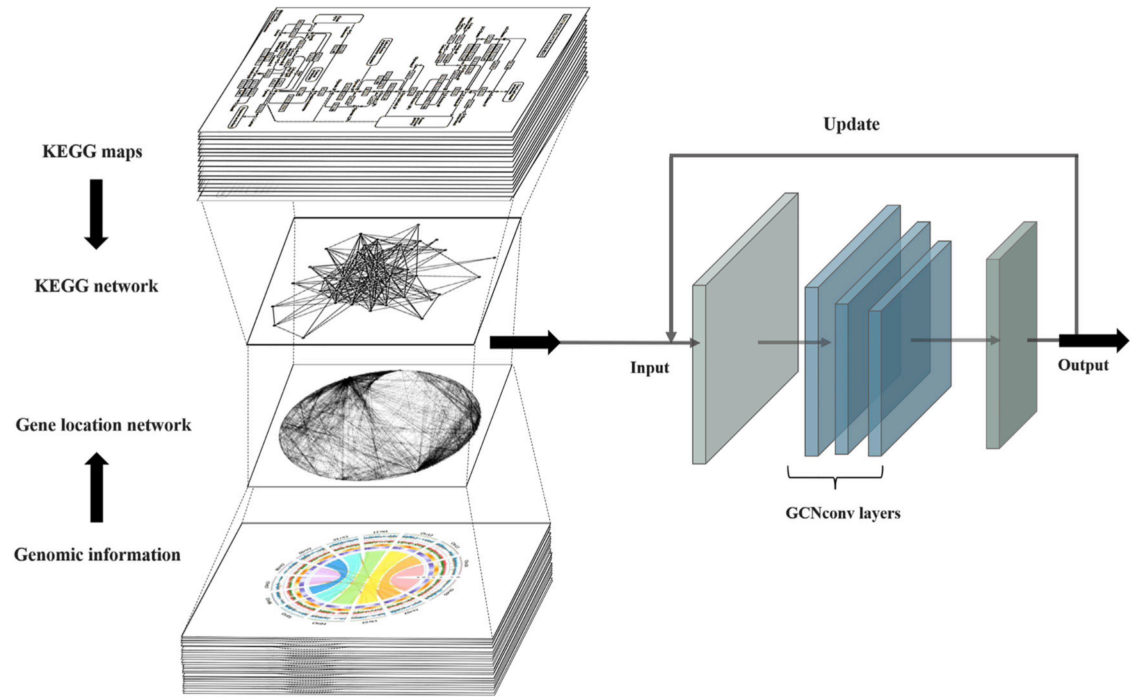


FIGURE 1 | PPA-GCN architecture. The input to the framework is the metabolic pathway network extracted from the KEGG metabolic pathways and the gene synteny network composed of the prokaryotic genomes. The graph convolutional layer attempts to construct a mapping relationship between the two input networks and iteratively uses the training results to update the input inspired by self-supervised learning until a steady state is reached and the final assignment output is obtained.

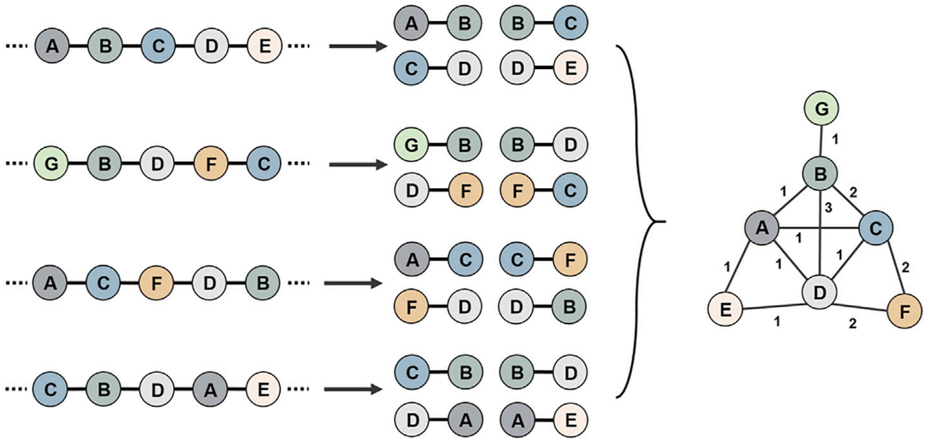


FIGURE 2 | Schematic diagram of the use of multiple genomes to construct a gene synteny network. First, all genomic genes are compared for sequence similarity, and genes that share high reciprocal similarity and cover ratios are assigned the same node id. Then, positional relationship pairs between two genes from each genome were constructed. Finally, all gene position relationship pairs are connected into a gene synteny network.

Kanehisa et al., 2016a; Kanehisa et al., 2016b; Aramaki et al., 2020). Currently, only 48% of the protein sequences are assigned to pathways in the KEGG GENES database (Aramaki et al., 2020). It is even lower for newly sequenced prokaryotic genomes, which has become a bottleneck for subsequent research (Suzuki et al., 2014). Thus, the development of a high-precision metabolic pathway assignment framework is urgently needed.

Here, we propose PPA-GCN, a framework based on graph convolutional network (GCN) that uses genomic gene synteny information within specific genus, from which the graph topological pattern and gene node characteristics can be learned, to disseminate node attributes in the network and provide assistance to the assignment of metabolic pathways. Synteny is defined as two or more pairs of homologous genes occupying the same

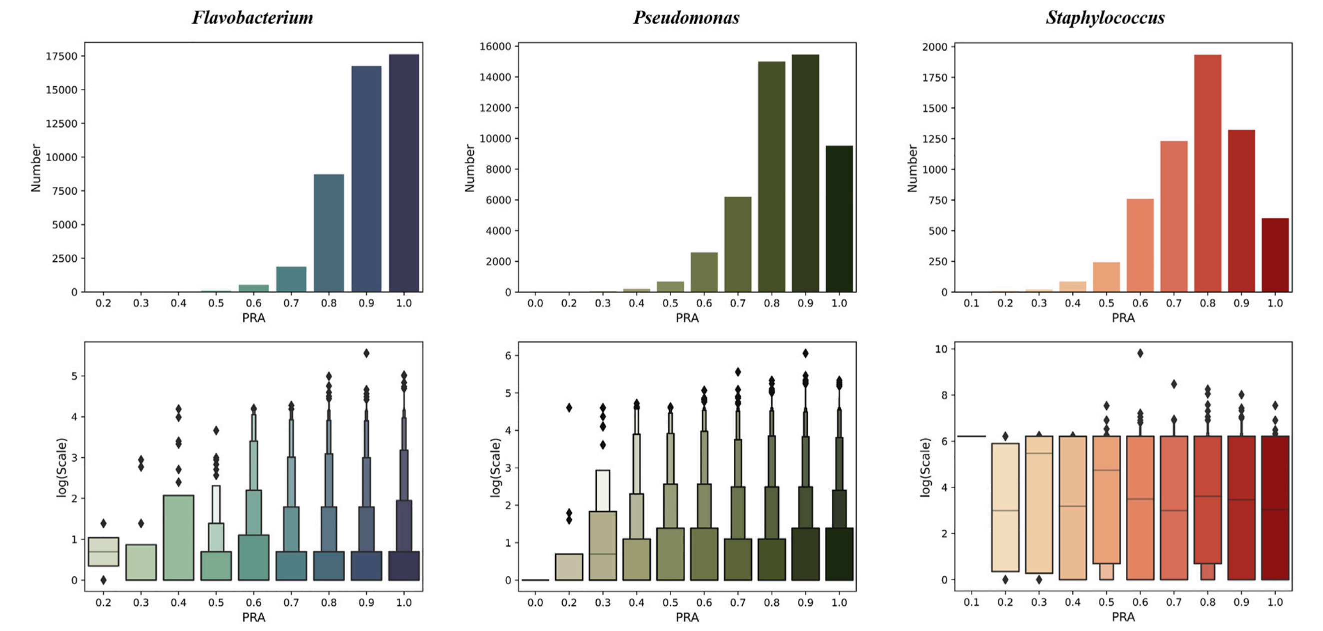


FIGURE 3 | The performance of PPA-GCN on three genera (in terms of the PRA) and the node scale distribution of the node set at each PRA level (10% as one level). From left to right are *Flavobacterium, Pseudomonas,* and *Staphylococcus*.

TABLE 1 | Performance under 5-fold cross-validation for the three genera.

Species	PRA	TLPR	WPRA	KC	HD	JS
Flavobacterium	0.848	0.846	0.829	0.842	0.008	0.751
Pseudomonas	0.770	0.728	0.736	0.721	0.014	0.609
Staphylococcus	0.710	0.691	0.698	0.689	0.008	0.651

chromosomal segment, where homologous loci are defined based on the similarity of function of the products of the corresponding genes (Nadeau and Taylor, 1984). Analyzing synteny can provide insight regarding the evolution and function of genes (Zhang et al., 2016). As an inherent biological attribute, bacteria of different genera have different synteny patterns. In general, bacterial genomes have two different pan-genome types. The pan-genome refers to all genes detected in a whole group of genomes (Wang L. et al., 2020). Some prokaryotes have genomes with highly conserved gene content (closed pan-genomes), while others are more flexible (open pan-genomes). Since the concept of a "pan-genome" was first proposed in 2005, pan-genome analysis has revealed the diversity and evolution of bacterial genomes (Tettelin et al., 2005). In present, there is currently no deep learning framework for direct assignment of functional pathways against KEGG database. To evaluate PPA-GCN, genome datasets of three different genus were used, and on all of them, the proposed framework had achieved excellent performance. PPA-GCN enables novel insights into assignment for functional pathways and is likely to inform future deep learning applications for interpreting functional annotations.

RELATED WORK

The study of gene location in the genome is one of the classic fields of genetics (Rogozin et al., 2004). In prokaryotes, genes encoding functional linked proteins are usually organized into gene clusters (Shmakov et al., 2019). There were methods assign protein function using neighborhood properties (Saha et al., 2012; Jun et al., 2017; Saha et al., 2018). It has been shown that the neighborhood milieu of genes in a network can assist in predicting the probable function of a gene for which no function is known (Hao et al., 2012). However, there is almost no method to assign KEGG pathways using gene neighborhood information.

In recent years, deep learning has been widely used in the field of life science, for example, for identifying and interpreting the contextual features of transcription factors (Zheng et al., 2021), generating functional protein sequences (Repecka et al., 2021), and identifying cell types (Lukassen et al., 2020; Wang M. et al., 2020). At present, the applications of graph neural networks in the medical and biology fields show strong representation and integration capabilities (Wu et al., 2020), including neuroimage analysis (Zhang et al., 2018), disease gene identification (Li et al., 2019; Schulte-Sasse et al., 2021), drug combination synergy prediction (Zitnik et al., 2018; Jiang et al., 2020; Manoochehri and Nourani, 2020), discovery of disease pathways (Agrawal et al., 2018), prediction of tissue cell function (Zitnik et al., 2017), pseudogene function prediction (Fan and Zhang, 2020), conducting taxonomic classification for phage contigs (Shang et al., 2021) and identifying missing protein–phenotype associations (Liu et al., 2021). The graph convolutional network (GCN) is a type of graph neural network that can learn the structure of a graph. This network model was originally proposed for semi-supervised classification (Kipf et al., 2016). A

TABLE 2 | Performance comparison under 5-fold cross-validation

Methods	PRA	TLPR	WPRA	KC	HD	JS
deepNF	0.562	0.365	0.339	0.511	0.273	0.379
Mashup	0.562	0.446	0.479	0.529	0.108	0.450
Pseudo2GO	0.578	0.470	0.466	0.513	0.051	0.433
SVM	0.483	0.304	0.319	0.506	0.118	0.414
DNN	0.402	0.365	0.339	0.501	0.063	0.429
PPA-GCN (without self-supervised learning)	0.607	0.570	0.539	0.522	0.034	0.402
PPA-GCN	0.710	0.691	0.698	0.689	0.008	0.651

TABLE 3 | Performance under the new data set.

Metrics	*Flavobacterium*	*Pseudomonas*	*Staphylococcus*
PRA	0.637	0.613	0.798
TLPR	0.606	0.538	0.723

GCN model can extensively integrate graph topological features and node information by defining each node as a computational graph and using neural networks to integrate neighbor node information.

MATERIALS AND METHODS

Problem Statement

Given an undirected graph G = (V_{tr}, V_{te}, E), where V_{tr} is the set of nodes that assigned function pathway, V_{te} is the set of nodes that unassigned function pathway, V = {V_{tr}, V_{te}}. E is the set of edges and the edge represents two genes belonging to different nodes are connected in the genome. A label set L = {l_1, l_2...l_k} is formed according to the KEGG secondary class. The relationship between the node set and the label set is represented by a matrix Y_{NxK}. Y_{ij} = 1, if there is a gene in node i has assigned to label j. Our goal is to assign the possible pathway labels to those nodes that have no labels.

Framework

PPA-GCN is a deep learning framework based on a graph convolutional model **(Figure 1)**. Gene synteny information from the selected genome is used to construct edges in a network, while genes sharing high sequence similarity and cover ratio are grouped into nodes. All node and edge information are used to construct the gene synteny network. PPA-GCN applies a three-layer graph convolutional architecture. Input features include node encoding, node scale and adjacency probability matrix. The KEGG metabolic pathway information of the secondary class is used as the node labels for initial training. Improve performance with inspiration from self-supervised learning. The final outputs are ranked in accordance with the stability of the assignment during the training process.

Graph Construction

Node Construction

Blast (Altschul et al., 1990) was used to compare the sequence similarity of all genome genes in one genus. In order to quickly and strictly find the similar genes, we directly adopted the reciprocal best hits comparison and controlled the identities and cover ratios to 65%. Taking *Flavobacterium* as an example, a total of 16,830 orthologs were obtained using OrthoFinder 2.0 (Emms and Kelly, 2019), and 51,247 nodes were obtained using our method, of which 50,998 nodes contained only one orthologs (99.5%). Therefore, our method is stricter than directly using orthologs. Node2vec algorithm (Grover and Leskovec, 2016) was used to generate graph embeddings for each node.

Edge Construction

Positional relationship pairs between two genes from each genome were constructed using the data of coding DNA sequence (CDS) (**Figure 2**). Through the correspondence between genes and nodes, all positional pairs were connected into a single gene synteny network, in which there could be more than one connection between two nodes. The adjacency matrix was constructed in accordance with the number of connections between nodes.

Construction of the Adjacency Probability Matrix

The adjacency probability is defined as the probability that two nodes form a certain number of connections in the network. First, the degree of each node in the gene synteny network (the number of connections by which a node is directly connected to surrounding nodes) was calculated. Then, the probability P_i that an edge is connected to a specific node *i* was calculated. Finally, the probability that there are *k* edges between node *i* and node *j* was defined as:

$$P_i = \frac{degree\,(i)}{\sum_{n=1}^{N} degree\,(n)} \tag{1}$$

$$P_{ij} = C^{k}_{degree\,(i)}\, P^{k}_{j}\left(1 - P_j\right)^{degree\,(i)-k} \tag{2}$$

where *N* is the total number of nodes in the gene synteny graph and *degree(i)* is the degree of node *i*, *C* is the combination symbol.

After the adjacency probabilities of all nodes had been formed into an *N**N* adjacency probability matrix, because there are no connections between most nodes, the node2vec algorithm was used to densify the adjacency probability matrix.

The GCN Model

Framework Architecture

Given an undirected graph with node feature matrix *X* and adjacency matrix *A*, the graph convolution operation (Kipf et al., 2016) is defined as:

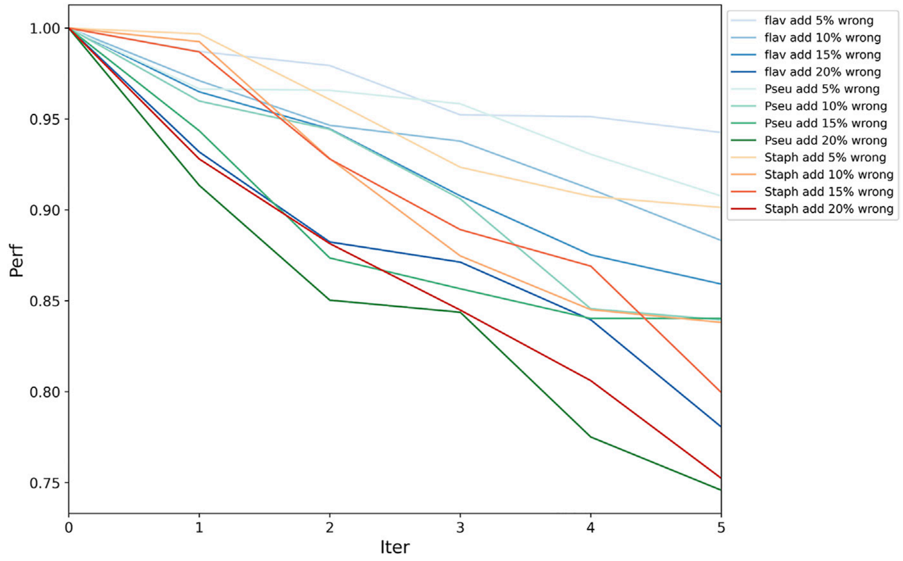


FIGURE 4 | Framework fault tolerance evaluation. On the three datasets, the performance was tested with the accumulation of 5–20% incorrectly labeled data in each epoch; the horizontal axis is the number of iteration, and the vertical axis is the performance indicator (current PRA/original PRA). This result shows that PPA-GCN has strong fault tolerance.

$$H = \sigma\left(D^{\frac{1}{2}}\hat{A}D^{-\frac{1}{2}}XW\right) \tag{3}$$

$$\hat{A} = A + I, \quad D_{ii} = \sum_{j}\widehat{A_{ij}} \tag{4}$$

where I is the identity matrix, W is the matrix of trainable weights in the neural network, X is the feature matrix before the update, H is the feature matrix after the update, and σ is the activation function (ReLU). The graph convolution operation iteratively calculates the weighted average of the node attributes of the neighbors of the current node to obtain the new feature matrix of the node. In this framework, the features of unlabeled nodes (nodes without assigned functional pathways) and the features of nearby labeled nodes (nodes with assigned functional pathways) are mixed to be propagated through the synteny network diagram. If two nodes have the same neighbor structure and neighbor features, their embedded feature matrix H will be exactly the same.

Python's PyTorch Geometric Module was used to implement PPA-GCN. Multiple graph convolutional layers can be stacked to enable learning on a larger domain structure. After testing, a three-layer stack was found to perform the best. The two-class cross entropy was used as the loss function because of the multilabel nature of the problem.

Self-Supervised Learning Inspiration

The original input was fed into the framework, and 50 epochs of random sampling verification training were performed with the test set. The nodes with an average cross-validation accuracy rate of less than 30% are removed from the training set, and nodes and labels with a assignment stability of 90% in the test set (that is, the same label is assigned more than 45 times) are added to the training set. After many iterations, when the number of nodes in the training set reached more than 90% of the total number of nodes in the gene synteny network, the training was considered to have reached a stable state, and the final assignment results were output.

Topological Analysis

Degree and Degree Distribution

The degree is defined as the number of all edge connections of a node in a graph, describing the first-order connection degree of the node. The degree distribution is an overall description of the nodes in a network, that is, the probability distribution or statistical distribution of the node degrees.

Clustering Coefficient

The clustering coefficient is used to describe the degree of clumping among the vertices of a network. Specifically, it is the degree of interconnection among the adjacent nodes of a node, describing the second-order connection degree of the node. For node i with degree k_i, the local clustering coefficient is defined as:

$$C_i = \frac{2L_i}{k_i\,(k_i - 1)} \tag{5}$$

where L_i is the number of connections among the k_i neighbors of node i. The overall aggregation coefficient of the network is characterized as the average value of the aggregation coefficients of all nodes.

RESULT

Data

All training genomes were downloaded from the National Center for Biotechnology Information (NCBI) database in June 2021

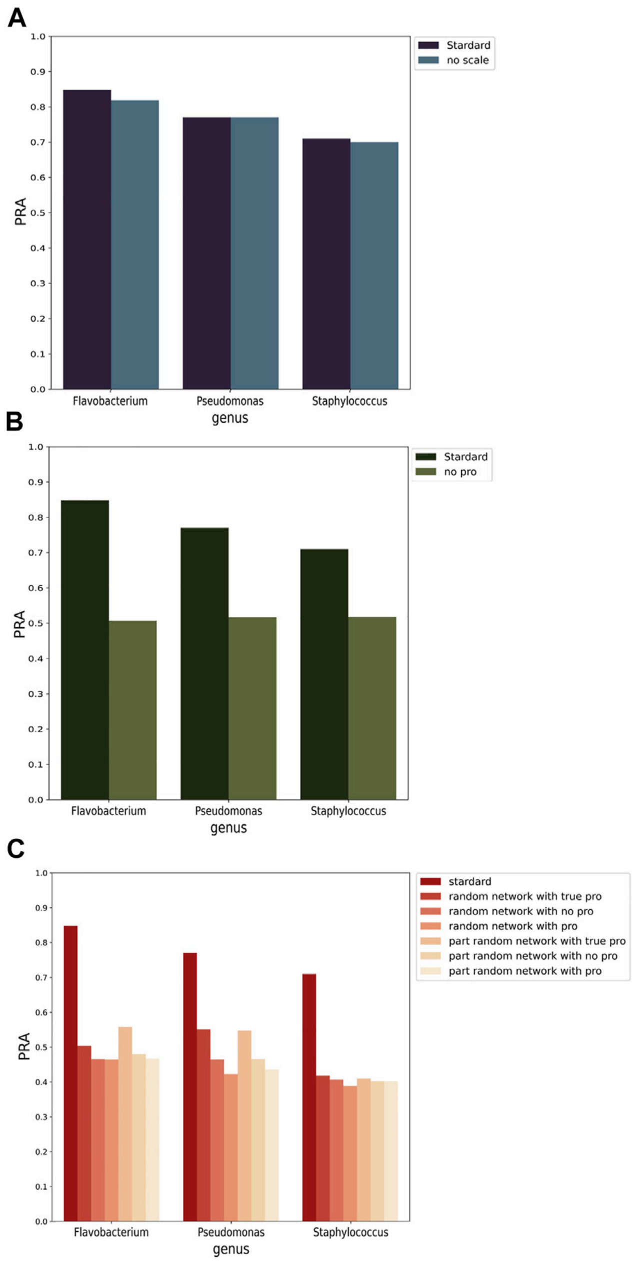


FIGURE 5 | (A) Feature importance assessment of the node scale. Comparison of performance changes before (standard) and after removing node scale (no scale). **(B)** Feature importance evaluation of the probability adjacency matrix. Comparison of performance changes before (standard) and after removing probability adjacency matrix (no pro). **(C)** Feature importance assessment of the gene synteny network. The horizontal axis represents standard training and training using random networks generated with three strategies: not including any adjacency probability matrix (with no pro), including the adjacency probability matrix of the newly generated network (with pro), and including the adjacency probability matrix of the real network (with true pro). The three graphs all use the prediction rate of assignment (PRA) as the evaluation index. The results show that node scale, adjacency probability matrix and network are very important features of PPA-GCN.

(https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/). The datasets include *Flavobacterium* (Gram-negative, 65 genomes), *Pseudomonas* (Gram-negative, 100 genomes) and *Staphylococcus* (Gram-positive, 500 genomes). *Staphylococcus* has a closed pan-genome. The 500 genomes selected for this study contain 1,332,382 genes grouped into a gene synteny network of 10,074 nodes. *Flavobacterium* and *Pseudomonas* have open pan-genomes. The 65 *Flavobacterium* genomes and 100 *Pseudomonas* genomes selected for this study contain 243,834 and 550,752 genes grouped into 51,247 and 79,941 nodes, respectively.

KEGG internal annotation tool KofamKOALA (version 100.0, updated October 1, 2021) was used to assign genes to functional pathways. The pathway labels belonging to the global and overview maps category were removed. *Staphylococcus* had 400,478 genes (1,324 nodes) assigned to metabolic pathways, *Flavobacterium* had 67,529 genes (3,694 nodes) assigned to metabolic pathways, and *Pseudomonas* had 272,388 genes (12,429 nodes) assigned to metabolic pathways (**Supplementary Table S1**). The original assignment rates for the three genera were 7.2% (*Flavobacterium*), 15.5% (*Pseudomonas*) and 13.1% (*Staphylococcus*).

In order to verify the performance of the model, the new genome data of the three genera were downloaded from the National Center for Biotechnology Information (NCBI) database in October 2021 (newly released genomes were downloaded first). The datasets include *Flavobacterium* (30 genomes), *Pseudomonas* (50 genomes) and *Staphylococcus* (200 genomes).

Evaluation Metrics

Pathway label assignment is essentially a multilabel classification problem. Hence, some commonly used evaluation indicators for binary classification problems are not suitable for PPA-GCN. We use six indicators to measure the effectiveness of the framework:

Prediction Rate of Assignment

PRA is the accuracy at the node level and is defined as the proportion of genes with at least one label assigned correctly.

Total Label Prediction Rate

The TLPR is the accuracy at the label level and is defined as the number of correctly assigned labels divided by the total number of labels.

Weighted Prediction Rate of Assignment

When a label is predicted for a node, we assign weights in accordance with the assignment probability, sum the WPRA of each label of a node to obtain the *WPRA* of that node, and divide by the total number of nodes to obtain the overall WPRA:

$$w_{prediction} = \frac{1}{N}\sum_{k\in T_i}\frac{2(I+1-k)}{I\,(I+1)} \qquad (6)$$

where N is the total number of nodes, I is the number of labels for node i, T_i is the order of the correct label probabilities assigned for

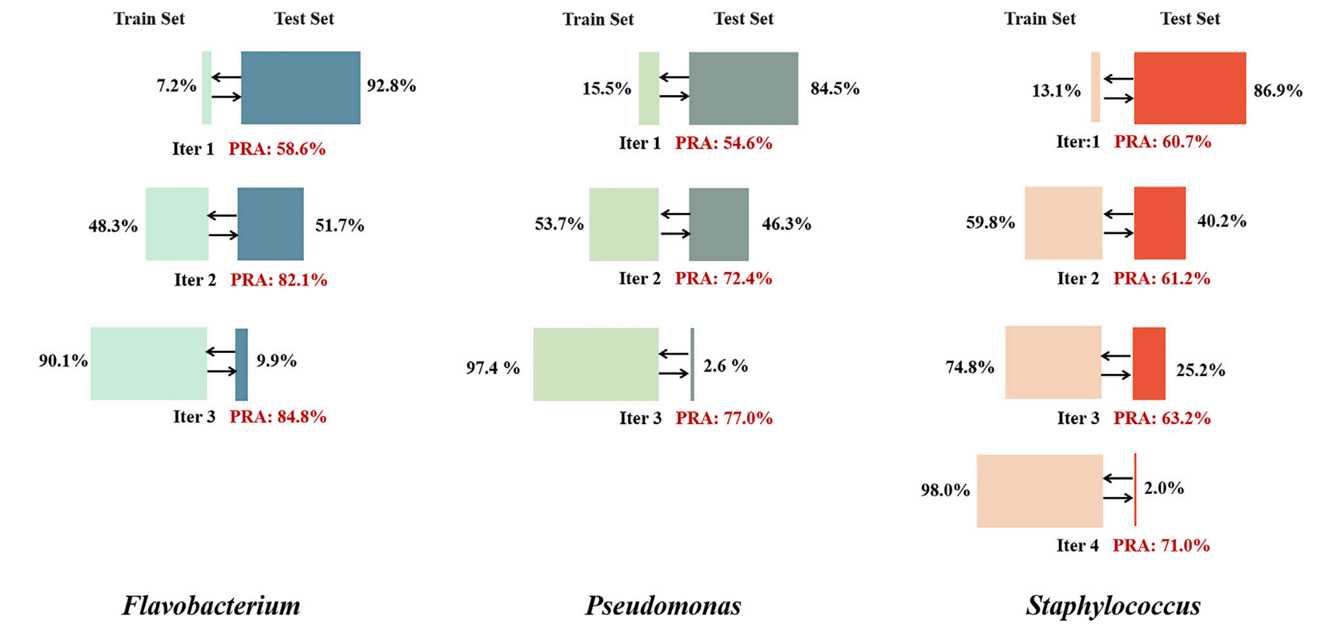


FIGURE 6 | Self-supervised learning iteration results for the three genera. In each iteration, training was performed for 50 epochs, nodes with an average PRA of less than 30% were removed from the training set, and nodes with stably assigned labels in the test set (with assignment consistency over more than 90% of epochs) were added to the training set. When the proportion of the total number of nodes included in the training set exceeded 90%, the final results were output.

node *i* (from large to small), and *k* is the k-th ranked probability label that was assigned correctly.

Kappa Coefficient

The kappa coefficient is often used for testing consistency, that is, whether the assignment effect of the model is consistent with the actual classification effect. Its value is between -1 and 1. When the value is greater than 0.6, it is considered substantial, and when it is greater than 0.8, it is considered almost perfect. The calculation of the kappa coefficient is based on the confusion matrix:

$$kappa = \frac{p_0 - p_e}{1 - p_e} \quad (7)$$

$$p_0 = \frac{\sum_i M_{ii}}{\sum_{ij} M_{ij}}, \quad p_e = \frac{\sum_i M_{i.} M_{.i}}{\left(\sum_{ij} M_{ij}\right)^2} \quad (8)$$

where M is the confusion matrix of the assignment results.

Hamming Distance

The Hamming distance is measure of the distance between the assigned and real labels, with a value between 0 and 1. A distance of 0 means that the assigned results are exactly the same as the real results, and a distance of 1 means that the model's results are completely opposite to the desired results. This indicator is calculated as the number of erroneously assigned labels divided by the total number of labels.

Jaccard Similarity Coefficient

This coefficient is an indicator for comparing the similarity of two finite sets, defined as the size of the intersection of two label sets (the true label set and the assigned label set) divided by the size of the union. When this coefficient is 1, the assigned results are completely consistent with the actual situation; when the coefficient is 0, the assigned results are completely inconsistent with the actual situation.

Results of Experiments

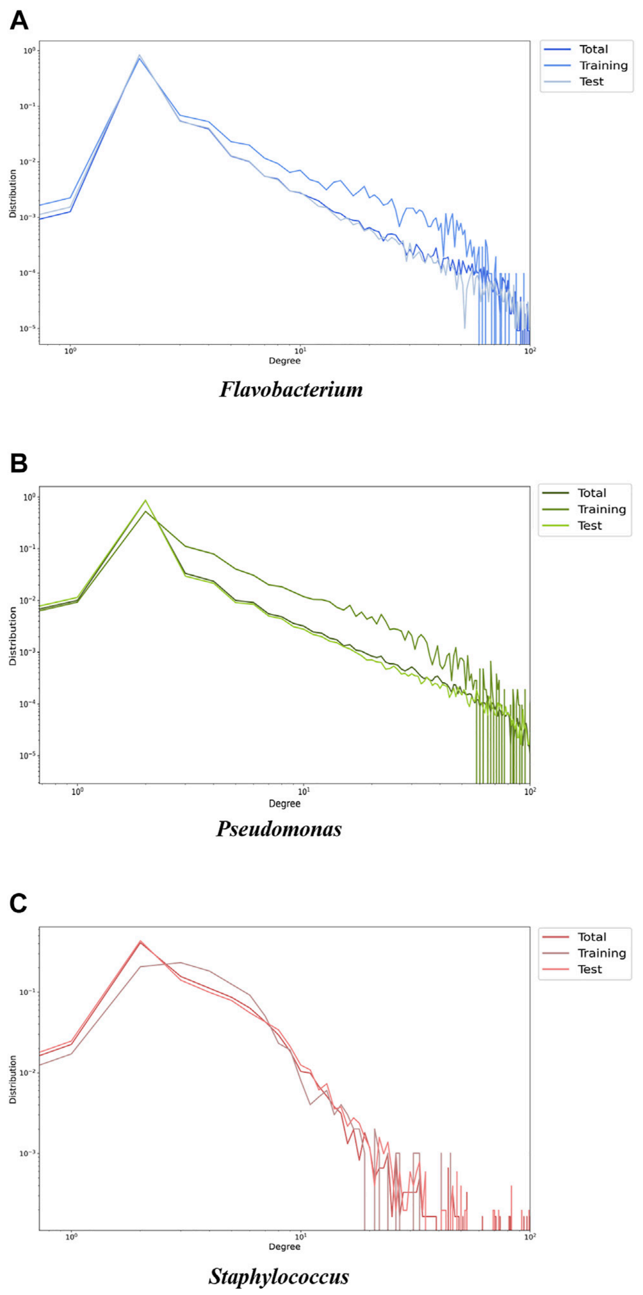
Results of Cross-Validation

We tested PPA-GCN with 5-fold cross-validation on three data sets. PPA-GCN achieved prediction rates of assignment (PRAs) of 84.8% (*Flavobacterium*), 77.0% (*Pseudomonas*) and 71.0% (*Staphylococcus*) on the three prokaryotic bacterial genera (**Figure 3**). According to the evaluation index results (**Table 1**), PPA-GCN is well adapted to all three genera.

In addition, we compared PPA-GCN with five other machine learning methods. deepNF (Gligorijevic et al., 2018), Mashup (Cho et al., 2016) and Pseudo2GO (Fan and Zhang, 2020) are three deep learning methods that use graph information for function prediction. Support vector machines (SVM) and deep neural networks (DNN) are two machine learning models that are not based on graph information. Using the *Staphylococcus* genome as the test data set, all methods use the same features in PPA-GCN as input, and use 5-fold cross-validation to test performance. The results (**Table 2**) show that, PPA-GCN achieves the best performance among all indicators.

Results of Test

In order to evaluate the adaptability of PPA-GCN to new data, the genes of the new genome were classified into network nodes. The test set node of the newly assigned functional path label in the



Flavobacterium

Pseudomonas

Staphylococcus

FIGURE 7 | Degree distribution curves describing the degree distributions of the overall network, the training set network and the test set network for each of the three genera [**(A)** *Flavobacterium,* **(B)** *Pseudomonas* **(C)** *Staphylococcus*]. The horizontal axis is the degree (truncated to 100), and the vertical axis is the probability distribution (The sum of the probabilities is 1). The results show that the degree distributions of the initial training set and the test set for the three genera are different, reflecting the genome characteristics of each genus to a certain extent.

network was used as the evaluation object, and the difference between the assigned output label and the real label is directly compared. The results are shown in **Table 3**, which proves that the results of PPA-GCN is reliable.

Fault Tolerance Evaluation

Because functional pathway assignment for bacterial genomes is still in the development stage, there will inevitably be some false pathway labels on the bacterial genes. Hence, we needed to test the fault tolerance of PPA-GCN. All assigned labels were assumed to be correct. In each epoch of training, some unlabeled nodes were given random labels to also participate in the training process. Two sets of experiments were conducted. In one, a certain percentage (5–20%) of incorrectly labeled samples were added in each epoch independently, and in the other, incorrectly labeled samples were added accumulatively. The PRA without the addition of incorrect labels was taken as the standard, and the PRA after the addition of incorrect labels was divided by the standard PRA to serve as the performance indicator. The results (**Figure 4**, **Supplementary Table S2**) show that PPA-GCN can still maintain more than 75% performance with the addition of incorrect labels at a rate of up to 100% (that is, the incorrectly labeled samples compose up to 50% of the training set). Because the distribution of wrong labels is random, and the distribution of correct labels is ordered, the influence of correct labels on the training results is greater than that of wrong labels, which enhances the fault tolerance of the framework. With an increasing proportion of incorrect labels, the efficiency of the framework did not drop sharply. This result shows that PPA-GCN has strong fault tolerance.

Feature Importance Test

A graph neural network can achieve excellent prediction accuracy, but it is difficult to give practical meaning to features. To evaluate the importance of the selected features, the PRAs before and after feature removal were compared (**Figure 5**, **Supplementary Table S3**). There are three important features in the PPA-GCN input: the node scale, the adjacency probability matrix and the gene synteny network.

The node scale is defined as the number of genes grouped into one node. The node scale was selected as an input feature because it can reflect the characteristics of a group of genomes. *Staphylococcus* has a closed pan-genome with an average node scale of 132.3, that is, an average of approximately 132 genes grouped into one node. *Flavobacterium* and *Pseudomonas* have open pan-genomes with average node scales of only 4.8 and 6.9, respectively. The node scale was one of the major observed differences between the labeled (training set) and unlabeled (test set) node sets in the gene synteny network. PPA-GCN showed no significant difference in performance when the node scale information was removed from the input (**Figure 5A**). The node scale has no effect on framework training, and this is beneficial for the applicability of the framework to unlabeled nodes.

The locations of genes in genomes are often specific, and the gene synteny network extracted from the same genus could reflect the intrinsic properties of the genus. The adjacency probability matrix is defined as the probability that two specific nodes can achieve a certain number of connections in a specific genome synteny network. Adding the adjacency probability matrix to the

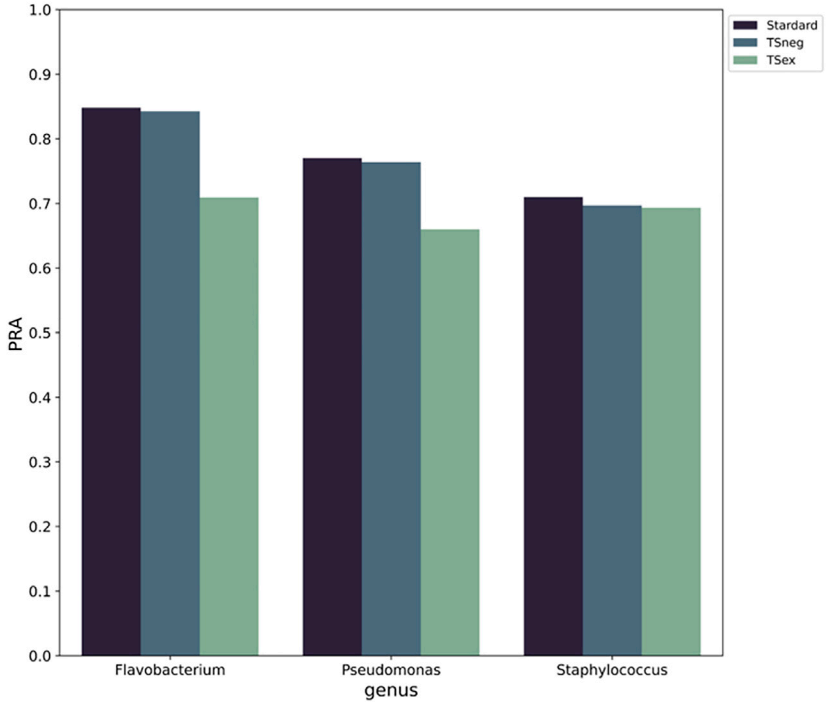


FIGURE 8 | The impact of MGEs on PPA-GCN performance. The horizontal axis represents standard training for the three genera (left), training with the MGEs as negative samples (middle) and training with the MGEs removed from the gene synteny network (right). The vertical axis uses PRA as an evaluation index. The results show that when the MGE nodes are removed from the networks, the performance of PPA-GCN is significantly reduced. When they are used as negative samples, the performance of the framework is only slightly reduced.

input was found to greatly improve the performance of the framework (**Figure 5B**). The adjacency probability matrix provides PPA-GCN with an information dissemination pattern for a specific bacterial genus in the gene synteny network.

Since the adjacency probability matrix can be used to extract synteny information patterns for specific microbial species, we wished to verify whether the gene synteny network could be replaced. Two types of random networks were designed while keeping the degree distribution constant. In one case, the arrangement of the gene positions in each sample genome was disrupted, and in the other, the positional relationships of all genomes were disrupted. Three strategies were considered for feature selection: not including any adjacency probability matrix, including the adjacency probability matrix of the newly generated network, and including the adjacency probability matrix of the real network. The training results show that (**Figure 5C**), regardless of which random network was used, the training performance when using a random network was much lower than that achieved using the real network. Interestingly, the true probability adjacency matrix can improve the framework training performance, while including the matrix of a random network actually impairs performance. This further shows that the adjacency probability matrix can capture specific information patterns of bacterial genomes. The gene synteny network and the adjacency probability matrix can provide the framework with different information patterns, and neither can replace the other.

Effectiveness of Self-Supervised Learning Inspiration

Currently, the assignment rate for gene metabolism pathways is lower than 50% in the KEGG GENES database. For the tested genera of three prokaryotes, the assignment rate for metabolic pathways is less than 20% of all nodes in the network, which greatly limits the training performance. The inspiration of self-supervised learning was adopted to extend the training set. Nodes with low PRAs in the validation set were temporarily excluded from the training set, and nodes with highly stable assigned labels in the test set were temporarily added to the training set. After several iterations, the performance eventually stabilized and showed a great improvement over the initial performance (**Figure 6**).

We speculate that PPA-GCN's performance could be significantly improved because labeled nodes spread node attributes in a certain pattern, ultimately causing the entire gene synteny network to present a genus-specific information pattern. The question of whether this kind of propagation can be universally applied to different types of gene synteny networks or is suitable only for network structures with a more "uniform" topology should be considered. Labeled and unlabeled nodes were extracted to construct training and test networks, respectively, and the topological structures of the two new networks were compared. Because PPA-GCN iteratively extracts information from the first- and second-order neighbors of nodes, the tightness of the first- and second-order connections in the network, as measured in terms of the degree distribution and clustering coefficient, need to be considered. The results (**Figure 7**, **Supplementary Table S4**) show that the degree distributions of the initial training set and the test set for the three genera are different, reflecting the genome characteristics of each genus to a certain extent. The degree distribution curves and clustering coefficients for the closed pan-genome (*Staphylococcus*) are not significantly different between the initial training set and the test set; in contrast, the initial training set networks of the open pan-genomes (*Flavobacterium* and *Pseudomonas*) are more closely connected than the test set networks, and the overall networks exhibit some level of inhomogeneity. These findings show that the self-supervised inspiration can effectively adapt to gene synteny networks with different topologies.

The Impact of Different Types of Genomes on Training

Synteny has been used to filter, organize and process local similarities between genome sequences of related organisms to build a coherent global chromosomal context (Deb et al., 2020). Each genus of prokaryotes possesses characteristic genomic gene synteny information, and its patterns are broadly associated with many bacterial functional traits (Brbić et al., 2016). Integrating gene synteny data from one genus can provide assistance to the functional pathway assignments of all genes.

Whether different types of genomes would affect training results should be considered. In addition to the node scale, the run number of self-supervised iterations needed to reach convergence can also reflect differences between different types of genomes. *Staphylococcus* requires more iterations to reach a steady state than *Flavobacterium* or *Pseudomonas*. This suggests that the information pattern of a closed pan-genome is relatively conservative and cannot be easily extended, while the information pattern of an open pan-genome is easier to spread. PPA-GCN could provide insights for judging genome types in accordance with the number of iterations needed for self-supervised learning when analyzing the genome of an unknown species.

The Role of Hyperlink Nodes in the Gene Synteny Network

There are several nodes with a “super connection number” in the gene synteny network of each genus. Further analysis revealed that these hyperlinked nodes have certain similarities in function. A large proportion of such nodes is assigned to mobile genetic elements (MGEs), which have the potential to disrupt the synteny of the involved genomes and are considered to cause gradual changes (sometimes mutations) in biological genes and promote biological evolution (Muszewska et al., 2019; Richards et al., 2019).

We investigated whether the insertion of MGEs into the genomes is random and has an impact on the pattern of functional labels. Two sets of experiments were designed. In the first set, all MGE nodes were removed from the gene synteny network to verify whether the insertion of the MGEs disrupted the information pattern of the original gene synteny networks. In the second set, all MGE nodes were added to the training set as negative samples to verify whether the intervention of the MGEs affected the distribution of functional labels. The results show (**Figure 8**) that when the MGE nodes are removed from the networks, the performance of PPA-GCN is significantly reduced. When they are used as negative samples, the performance of the framework is only slightly reduced. This indicates that from the perspective of gene location, MGEs may constitute an important part of the gene synteny network of a specific genus, and removing them will destroy the information pattern of the existing gene synteny network. Moreover, MGEs do not interfere with the distribution pattern of gene function.

DISCUSSION

In present, PPA-GCN is the first deep learning framework that uses genomic structure information to directly assist metabolic pathway assignments of prokaryotic genomes against KEGG information. Datasets representing three genera (*Flavobacterium*, *Pseudomonas and Staphylococcus*) were used to evaluate the assignment rate of the framework, and on all of them, good performance and strong fault tolerance were achieved. These results support the broad application of PPA-GCN to prokaryotic genomic research. For example, it can provide support for the mechanism research of pathogenic bacteria and the design of synthetic biology elements, modules and pathways.

Although all bacterial genome had been fragmented and shuffled by the endless genomic reconstruction and horizontal gene transfer, the localized genome structure was conserved within specific genus of bacteria. Gene synteny structure is intrinsic and stable under genus level and PPA-GCN relies on it. PPA-GCN captures the graph structure and node attributes from the gene synteny information through a graph convolutional network. To maximize the given pathway information of genomes of a genus, PPA-GCN obtains and mines as many possibilities for label assignment through the network as possible. Then PPA-GCN constructs the adjacency probability matrix to evaluate all possibilities, improving the certainty of all assigned labels. The idea of self-supervised learning is adopted to expand the training set and reinforce the training process.

PPA-GCN has the potential for further improvement. The runtime and memory usage of PPA-GCN will be optimized (**Supplementary Table S5**). At present, only one kind of graph information (the gene synteny network) is used to make assignments. In the future, some other information networks could be incorporated to improve the performance of PPA-GCN, potentially providing the perfect complement to the existing framework, such as a protein-protein interaction network and gene co-expression network.

PPA-GCN exhibits good performance and shows promise to help guide experimental verification and provide considerable additional space for downstream analysis. PPA-GCN could be applied to more genera of prokaryotes with sufficient whole genome sequences and used to build a database of consensus sequences from the perspective of functional pathway assignment, that could describe the differences in prokaryotes of various genera. In short, we present a deep learning framework with great potential to explain the relationship between gene synteny and KEGG pathway information in prokaryotes, which can provide novel insights into functional pathways assignments and is likely to inform future deep learning applications for interpreting functional annotations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

YL developed the framework, TL conceived and supervised the project. YL and QL wrote the manuscript.

FUNDING

This research was supported by the National Key R&D Program of China (Grant No. 2020YFA0907402), the National Natural Science Foundation of China (Grant No. 91851118 and No.31900096), and the Science and Technology Basic Resources Investigation Program of China (Grant No. 2017FY100300).

ACKNOWLEDGMENTS

The authors would like to thank the support from Yan Lin.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.839453/full#supplementary-material

REFERENCES

Agrawal, M., Zitnik, M., and Leskovec, J. (2018). "Large-scale Analysis of Disease Pathways in the Human Interactome," in PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium, 111–122. doi:10.1142/9789813235533_0011

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/s0022-2836(05)80360-2

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., et al. (2020). KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold. *Bioinformatics* 36 (7), 2251–2252. doi:10.1093/bioinformatics/btz859

Bazgir, O., Zhang, R., Dhruba, S. R., Rahman, R., Ghosh, S., and Pal, R. (2020). Representation of Features as Images with Neighborhood Dependencies for Compatibility with Convolutional Neural Networks. *Nat. Commun.* 11 (1), 4391. doi:10.1038/s41467-020-18197-y

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Genbank. Nucleic Acids Research* 41 (D1), D36–D42. doi:10.1093/nar/gks1195

Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., and Supek, F. (2016). The Landscape of Microbial Phenotypic Traits and Associated Genes. *Nucleic Acids Res.* 44, 10074–10090. doi:10.1093/nar/gkw964

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., et al. (2019). IMG/M v.5.0: an Integrated Data Management and Comparative Analysis System for Microbial Genomes and Microbiomes. *Nucleic Acids Res.* 47 (D1), D666–D677. doi:10.1093/nar/gky901

Cho, H., Berger, B., and Peng, J. (2016). Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cel Syst.* 3 (6), 540–548. doi:10.1016/j.cels.2016.10.017

Deb, S., Jayaprasad, S., Ravi, S., Rao, K. R., Whadgar, S., Hariharan, N., et al. (2020). Classification of Grain Amaranths Using Chromosome-Level Genome Assembly of Ramdana, A. Hypochondriacus. *Front. Plant Sci.* 11, 579529. doi:10.3389/fpls.2020.579529

Doerks, T., Von Mering, C., and Bork, P. (2004). Functional Clues for Hypothetical Proteins Based on Genomic Context Analysis in Prokaryotes. *Nucleic Acids Res.* 32 (21), 6321–6326. doi:10.1093/nar/gkh973

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol.* 20 (1), 238. doi:10.1186/s13059-019-1832-y

Eslami Manoochehri, H., and Nourani, M. (2020). Drug-target Interaction Prediction Using Semi-bipartite Graph Model and Deep Learning. *BMC bioinformatics* 21 (4), 248. doi:10.1186/s12859-020-3518-6

Fan, K., and Zhang, Y. (2020). Pseudo2GO: a Graph-Based Deep Learning Method for Pseudogene Function Prediction by Borrowing Information from Coding Genes. *Front. Genet.* 11, 807. doi:10.3389/fgene.2020.00807

Federhen, S. (2012). The NCBI Taxonomy Database. *Nucleic Acids Res.* 40 (D1), D136–D143. doi:10.1093/nar/gkr1178

Gligorijević, V., Barot, M., and Bonneau, R. (2018). deepNF: Deep Network Fusion for Protein Function Prediction. *Bioinformatics* 34 (22), 3873–3881.

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable Feature Learning for Networks," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 855–864. doi:10.1145/2939672.2939754*KDD*2016

Hao, K., Bossé, Y., Nickle, D. C., Paré, P. D., Postma, D. S., Laviolette, M., et al. (2012). Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. *Plos Genet.* 8 (11), e1003029. doi:10.1371/journal.pgen.1003029

Jiang, P., Huang, S., Fu, Z., Sun, Z., Lakowski, T. M., and Hu, P. (2020). Deep Graph Embedding for Prioritizing Synergistic Anticancer Drug Combinations. *Comput. Struct. Biotechnol. J.* 18, 427–438. doi:10.1016/j.csbj.2020.02.006

Jun, S. R., Nookaew, I., Hauser, L., and Gorin, A. (2017). Assessment of Genome Annotation Using Gene Function Similarity within the Gene Neighborhood. *BMC bioinformatics* 18 (1), 345. doi:10.1186/s12859-017-1761-2

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016b). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44 (D1), D457–D462. doi:10.1093/nar/gkv1070

Kanehisa, M., Sato, Y., and Morishima, K. (2016a). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and

Metagenome Sequences. *J. Mol. Biol.* 428 (4), 726–731. doi:10.1016/j.jmb.2015.11.006

Keegan, K. P., Glass, E. M., and Meyer, F. (2016). "MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function," in *Microbial Environmental Genomics (MEG)* (New York, NY: Humana Press), 207–233. doi:10.1007/978-1-4939-3369-3_13

Kipf, T. N., and Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

Li, Y., Kuwahara, H., Yang, P., Song, L., and Gao, X. (2019). PGCN: Disease Gene Prioritization by Disease and Gene Embedding through Graph Convolutional Neural Networks. *bioRxiv* 2019, 532226.

Liu, L., Mamitsuka, H., and Zhu, S. (2021). HPOFiller: Identifying Missing Protein-Phenotype Associations by Graph Convolutional Network. *Bioinformatics* 2021, btab224. doi:10.1093/bioinformatics/btab224

Lukassen, S., Ten, F. W., Adam, L., Eils, R., and Conrad, C. (2020). Gene Set Inference from Single-Cell Sequencing Data Using a Hybrid of Matrix Factorization and Variational Autoencoders. *Nat. Mach Intell.* 2 (12), 800–809. doi:10.1038/s42256-020-00269-9

Muszewska, A., Steczkiewicz, K., Stepniewska-Dziubinska, M., and Ginalski, K. (2019). Transposable Elements Contribute to Fungal Genes and Impact Fungal Lifestyle. *Sci. Rep.* 9 (1), 4307–4310. doi:10.1038/s41598-019-40965-0

Nadeau, J. H., and Taylor, B. A. (1984). Lengths of Chromosomal Segments Conserved since Divergence of Man and Mouse. *Proc. Natl. Acad. Sci. U.S.A.* 81 (3), 814–818. doi:10.1073/pnas.81.3.814

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., et al. (2021). Expanding Functional Protein Sequence Spaces Using Generative Adversarial Networks. *Nat. Mach Intell.* 3 (4), 324–333. doi:10.1038/s42256-021-00310-5

Richards, V. P., Velsko, I. M., Alam, M. T., Zadoks, R. N., Manning, S. D., Pavinski Bitar, P. D., et al. (2019). Population Gene Introgression and High Genome Plasticity for the Zoonotic Pathogen Streptococcus Agalactiae. *Mol. Biol. Evol.* 36 (11), 2572–2590. doi:10.1093/molbev/msz169

Rogozin, I. B., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2004). Computational Approaches for the Analysis of Gene Neighbourhoods in Prokaryotic Genomes. *Brief. Bioinformatics* 5 (2), 131–149. doi:10.1093/bib/5.2.131

Saha, S., Chatterjee, P., Basu, S., Kundu, M., and Nasipuri, M. (2012). "Improving Prediction of Protein Function from Protein Interaction Network Using Intelligent Neighborhood Approach," in 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS) (Kolkata, India: IEEE), 584–587. doi:10.1109/codis.2012.6422270

Saha, S., Prasad, A., Chatterjee, P., Basu, S., and Nasipuri, M. (2018). Protein Function Prediction from Protein-Protein Interaction Network Using Gene Ontology Based Neighborhood Analysis and Physico-Chemical Features. *J. Bioinform. Comput. Biol.* 16 (06), 1850025. doi:10.1142/s0219720018500257

Schulte-Sasse, R., Budach, S., Hnisz, D., and Marsico, A. (2021). Integration of Multiomics Data with Graph Convolutional Networks to Identify New Cancer Genes and Their Associated Molecular Mechanisms. *Nat. Mach Intell.* 3 (6), 513–526. doi:10.1038/s42256-021-00325-y

Shang, J., Jiang, J., and Sun, Y. (2021). Bacteriophage Classification for Assembled Contigs Using Graph Convolutional Network. *arXiv preprint arXiv:2102.03746*. doi:10.1093/bioinformatics/btab293

Shmakov, S. A., Faure, G., Makarova, K. S., Wolf, Y. I., Severinov, K. V., and Koonin, E. V. (2019). Systematic Prediction of Functionally Linked Genes in Bacterial and Archaeal Genomes. *Nat. Protoc.* 14 (10), 3013–3031. doi:10.1038/s41596-019-0211-1

Suzuki, S., Kakuta, M., Ishida, T., and Akiyama, Y. (2014). GHOSTX: an Improved Sequence Homology Search Algorithm Using a Query Suffix Array and a Database Suffix Array. *PloS one* 9 (8), e103833. doi:10.1371/journal.pone.0103833

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae : Implications for the Microbial "Pan-Genome". *Proc. Natl. Acad. Sci. U.S.A.* 102 (39), 13950–13955. doi:10.1073/pnas.0506758102

Wang, L., Nie, R., Yu, Z., Xin, R., Zheng, C., Zhang, Z., et al. (2020). An Interpretable Deep-Learning Architecture of Capsule Networks for Identifying Cell-type Gene Expression Programs from Single-Cell RNA-Sequencing Data. *Nat. Mach Intell.* 2 (11), 693–703. doi:10.1038/s42256-020-00244-4

Wang M., M., Zhu, H., Kong, Z., Li, T., Ma, L., Liu, D., et al. (2020). Pan-Genome Analyses of Geobacillus Spp. Reveal Genetic Characteristics and Composting Potential. *Ijms* 21 (9), 3393. doi:10.3390/ijms21093393

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32 (1), 4–24. doi:10.1109/TNNLS.2020.2978386

Zhang, K., Yue, D., Wei, W., Hu, Y., Feng, J., and Zou, Z. (2016). Characterization and Functional Analysis of Calmodulin and Calmodulin-like Genes in Fragaria Vesca. *Front. Plant Sci.* 7, 1820. doi:10.3389/fpls.2016.01820

Zhang, X., He, L., Chen, K., Luo, Y., Zhou, J., and Wang, F. (2018). Multi-View Graph Convolutional Network and its Applications on Neuroimage Analysis for Parkinson's Disease. *AMIA Annu. Symp. Proc.* 2018, 1147–1156.

Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H., and Gymrek, M. (2021). Deep Neural Networks Identify Sequence Context Features Predictive of Transcription Factor Binding. *Nat. Mach Intell.* 3 (2), 172–180. doi:10.1038/s42256-020-00282-y

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics* 34 (13), i457–i466. doi:10.1093/bioinformatics/bty294

Zitnik, M., and Leskovec, J. (2017). Predicting Multicellular Function through Multi-Layer Tissue Networks. *Bioinformatics* 33 (14), i190–i198. doi:10.1093/bioinformatics/btx252

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 28 April 2022
doi: 10.3389/fgene.2022.855739

BioTAGME: A Comprehensive Platform for Biological Knowledge Network Analysis

Antonio Di Maria[1], *Salvatore Alaimo*[1], *Lorenzo Bellomo*[2], *Fabrizio Billeci*[3], *Paolo Ferragina*[4], *Alfredo Ferro*[1] *and Alfredo Pulvirenti*[1]*

[1]Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy, [2]Scuola Normale Superiore, Pisa, Italy, [3]Department of Maths and Computer Science, University of Catania, Catania, Italy, [4]Department of Computer Science, University of Pisa, Pisa, Italy

The inference of novel knowledge and new hypotheses from the current literature analysis is crucial in making new scientific discoveries. In bio-medicine, given the enormous amount of literature and knowledge bases available, the automatic gain of knowledge concerning relationships among biological elements, in the form of semantically related terms (or entities), is rising novel research challenges and corresponding applications. In this regard, we propose BioTAGME, a system that combines an entity-annotation framework based on Wikipedia corpus (i.e., TAGME tool) with a network-based inference methodology (i.e., DT-Hybrid). This integration aims to create an extensive Knowledge Graph modeling relations among biological terms and phrases extracted from titles and abstracts of papers available in PubMed. The framework consists of a back-end and a front-end. The back-end is entirely implemented in Scala and runs on top of a Spark cluster that distributes the computing effort among several machines. The front-end is released through the Laravel framework, connected with the Neo4j graph database to store the knowledge graph.

Keywords: knowledge graph, text mining, annotation tools, TAGME, wikipedia, DT-hybrid

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Mourad Oussalah,
University of Oulu, Finland
Marco Brandizi,
Rothamsted Research,
United Kingdom

****Correspondence:***
Alfredo Pulvirenti
alfredo.pulvirenti@unict.it

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 January 2022*
Accepted: *24 March 2022*
Published: *28 April 2022*

Citation:
Di Maria A, Alaimo S, Bellomo L, Billeci F, Ferragina P, Ferro A and Pulvirenti A (2022) BioTAGME: A Comprehensive Platform for Biological Knowledge Network Analysis. Front. Genet. 13:855739. doi: 10.3389/fgene.2022.855739

1 INTRODUCTION

The increasing amount of scientific literature is raising new challenges for scientists. For example, identifying the proper set of articles dealing with a specific topic could be a not straightforward task. Thus, the possibility of missing essential references and relevant research is high nowadays. In particular, in research areas such as Biology or Bio-Medicine, thanks to fast-track publication journals, the number of published papers increases significantly fast, thus making it very difficult for scientists to keep track of literature evolution.

Furthermore, network analysis has become a key enabling technology to help the understanding of life mechanisms, living organisms and, in general, and uncover the underlying fundamental biological processes. Examples of applications include 1) analyzing disease networks for identifying disease-causing genes and pathways Barabási et al. (2010); 2) discovering the functional interdependence among molecular mechanisms through functional network querying (Xiaoke and Lin (2012)); 3) deriving network-based inferences for drug repurposing (Himmelstein et al. (2017)).

The large number of publicly available ontologies, which hold entities and their relations (Lambrix et al. (2007)), and the repositories of open-access articles such as PubMed Central (Beck (2010)), arXiv, and bioarXiv, are driving the academic community to rely on text mining tools and machine learning algorithms for extracting *semantic knowledge* from documents such as understanding how proteins interact each other, which gene mutations are involved in a disease, etc. In this context, the Biological Expression Language (BEL) (Hoyt et al. (2018)) or the Resource Description Framework (RDF) (McBride (2004)) are widely employed to represent this *knowledge* as triplets having the following structure: < subject, predicate, object > . The subject and the object represent biological elements, whereas the predicate represents a (logical or physical) relationship.

Since the implementation of biological text mining methodologies requires skills in natural language processing (NLP) that usually end-users do not have, several tools have been made available to scientists: 1) PubAnnotation (Kim et al. (2019)) is based on the "Agile text mining" concept, and it is a public resource for sharing annotated biomedical texts; 2) PubTator (PTC, Wei et al. (2019)) is a web service for viewing and retrieving bio-concept annotations (for genes/proteins, genetic variants, diseases, chemicals, species, and cell lines) from all PubMed abstracts and more than three million PubMed full-texts. These annotations are downloadable in multiple formats (XML, JSON, and tab-delimited) via the online interface, a RESTful web service, and bulk FTP. PTC is synchronized with PubMed and PubMed Central, adding new articles daily.

The literature also offers many frameworks for building functional networks. **STRING** (Szklarczyk et al. (2016)) is a database that collects known and predicted functional protein-protein associations for many organisms. Each protein-protein association is given a score (between zero and one) which summarizes the biological reliability of the interaction, its specificity, and the supporting evidence. Another significant contribution of these interactions is the so-called "interolog" transfer, based on the observation that orthologs of interacting proteins in one organism are often also interacting in another organism. The STRING resource is available online[1]. **Hetionet** (Himmelstein et al. (2017) is a heterogeneous network of biomedical knowledge constructed over genes, diseases, and compounds, extracted from the processing of a collection of 29 publicly available databases and millions of publications. It was created as part of Project Rephetio to predict new uses for existing drugs. In the last few years, it has been modified for working over a wider variety of purposes: such as drug repurposing and prioritizing disease-associated Genes. Hetionet is available at[2] **Reactome** (Croft et al. (2010) is a peer-reviewed knowledge base of biomolecular pathways that contains a detailed representation of cellular processes interconnecting terms to form a graph modeling biological knowledge. Reactome adopts Neo4j as a graph database to improve the graph traversal performance and knowledge discovery. Reactome is also available online[3]. **SemRep** (Rindflesch and Fiszman (2003)) is an NLP advanced information management application, which extracts relationships from biomedical sentences in PubMed titles and abstracts by mapping textual content to an ontology representing its meaning. To establish the binding relation, SemRep relies on internal rules (called "indicator rules"), which map syntactic elements, such as verbs, prepositions, and nominalization, to predicates in the Semantic Network. It is available at[4] **Kindred** (Lever and Jones (2017)) is a Python package built on top of the Stanford CoreNLP framework and the scikit-learn library. It performs relation extraction in biomedical texts, where relation candidates are created by finding every possible pair of entities within each sentence. Next, it exploits an SVM classifier to rank and select the most promising candidates. In **NetME** (Muscolino et al. (2022)), authors propose a tool that allows to query PUBMED and build knowledge networks synthesizing the concepts described through the selected papers. In the context of clinical Text Analysis and Knowledge Extraction, **cTAKES** (Savova et al. (2010)) is a system for information extraction from electronic medical record free-text. The pipeline comprises several modules, such as sentence boundary detector, tokenizer, normalizer, part-of-speech tagger, Shallow parser, and named entity recognizer. Other relevant work include **CKG** (Santos et al. (2022)). CkG is an open-source knowledge-graph platform, which includes 20 million nodes and 220 million relationships that represent relevant experimental data, public databases and literature. CKG incorporates statistical and machine learning algorithms to accelerate the analysis and interpretation of common proteomics workflows.

This paper introduces BioTAGME, a knowledge graph inferred from more than 33 million titles and abstracts in the PubMed database (Williamson and Minter (2019)), and downloadable as XML files via third-party applications.

BioTAGME uses two well-known tools to generate the Knowledge Graph. First, entities are extracted from each abstract using the TAGME annotation system (Ferragina and Scaiella (2010)). TAGME is a tool that analyzes short texts and extracts entities related to its content. It makes use of Wikipedia to perform the annotation. All the entities extracted from the abstracts are treated as nodes of the knowledge graph. Next, the DT-Hybrid (Alaimo et al. (2013)) recommendation system is applied to predict possible relationships among entities coming from different abstracts. These relationships form the edges of the knowledge graph. Finally, such predicted relationships are enriched with those from publicly available databases (the complete list is provided in **Section 2**) to generate a comprehensive Knowledge Graph, stored in the Neo4j database and made available to users via our web app. Such a knowledge graph consists of more than 161 thousand nodes and 40 million edges. Moreover, there are three different types of edges: 1) Literature edge: indicates a piece of biological evidence resulting from laboratory experiments, biological and biophysical processes; 2) STRING edge: represents STRING predicted protein-protein associations; finally 3) BioTAGME edge: are edges predicted by the combination of TAGME relatedness

[1] http://string-db.org/.

[2] https://neo4j.het.io/browser.

[3] https://reactome.org.

[4] https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html.

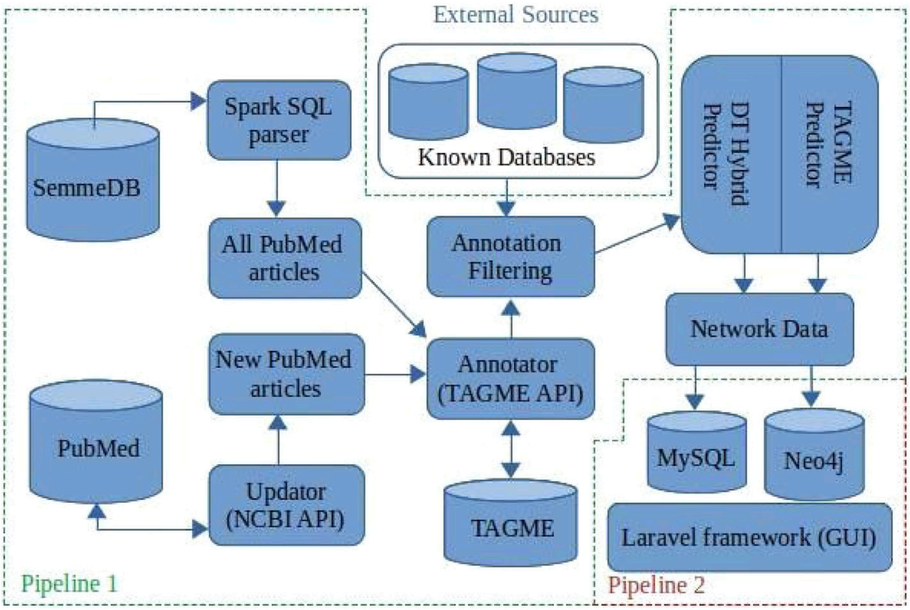


FIGURE 1 | BioTAGME pipelines: The first one pipeline is the core of the project. It transforms the whole set of PubMed abstracts within nodes and edges of the knowledge graph. It has been implemented in Scala e Spark. The seconds pipeline, allows a user to extract information from the graph. It has been implemented using Laravel and Reactstrap.

and BioTAGME one. Both BioTAGME edges and STRING ones are marked with the corresponding score value to indicate the interaction's likelihood. Biotagme is available at: https://biotagme.eu/[5]

The paper is organized as follows. In the **Section 2**, we introduce the back-end of our tool. Next, we introduce the web app to browse and query the system. Moreover, we show a BSG-Diseases network that reports literature evidence and BioTAGME prediction. Finally, in section conclusions, we explain future work about our tool.

2 MATERIALS AND METHODS

BioTAGME is a framework backed by two different pipelines (**Figure 1**) for building a biological knowledge graph from PubMed documents' titles and abstracts. It integrates two different learning algorithms, DT-Hybrid (Alaimo et al. (2013)) and TagME (Ferragina and Scaiella (2010)).

The first pipeline is built on top of the Apache SPARK analytic engine and Hadoop Distributed File System (HDFS). This implementation guarantees large-scale data processing through cluster managers (Apache Meson, YARN, Stand Alone, and Kubernetes). The pipeline collects results into DataFrames (Apache-Spark (2016)) the data coming from several freely available online databases as shown in **Table 1**. In addition, the complete set of PubMed titles and abstracts in order to build a life science knowledge graph using the Spark SQL language. DataFrame and SQL language provide a common way to access various data files, including Hive, Avro, Parquet, CSV, TSV, and JSON.

The major functionalities provided by the first pipeline are 1) Download and import, 2) SQL to JSON parser, 3) Integrating databases, 4) Annotation, 5) Prediction, 6) Network generation, and 7) Updating.

The second pipeline is built on top of the Laravel framework and consists of the following components: 1) MySQL for storing names, aliases, BioTAGME IDs, and Wikipedia pages IDs; 2) Neo4j for storing the knowledge graph, and allow querying the network (i.e., compute the shortest path between two user-specified biological entities (nodes)); 3) the User Interface (GUI), based on Laravel and React, used for wrapping the Neo4j queries and making them more accessible and more intuitive. Queries can be: 1) Search on the graph; 2) Shortest path. (Detailed information are in **Section 2.2**).

Data processing is done in PHP and bash to achieve high performance. In addition, all the GUI modules have been realized in react-native.

2.1 Pipeline One: Data Loader and Network Synthesis

This section describes all components and functionalities of the first pipeline underling BioTAGME.

2.1.1 Download and Import Module

This module allows importing the external databases into Hadoop Distributed File System (HDFS) through a custom bash script, which consists of three main sections:

[5] if the url does not work, more information about a possible new url is reported within the readme of the repository: https://github.com/Anto188bas/biotagme_docker.git.

TABLE 1 | Ontologies.

Source name	Citation	Data type
DisGeNET	Piñero et al. (2019)	human gene-disease association
DiseaseOntology (DO)	Schriml et al. (2018)	human disease
DiseaseEnhancer	Zhang et al. (2017)	human disease-associated enhancer
DrugBank	Wishart et al. (2007)	drug and drug target
PharmGKB	Thorn et al. (2013)	human-genetic variation on drug resp
HGNC	Daugherty et al. (2012)	human gene
ENSEMBL	Birney et al. (2004)	vertebrates genomic information
LNCipedia	Volders et al. (2012)	human long non-coding RNAs
miRcode	Jeggari et al. (2012)	human microRNA-target predictions
miRBase	Kozomara et al. (2018)	microRNA sequences
miRTarBase	Huang et al. (2019)	microRNA-target interactions
miRCancer	Xie et al. (2013)	microRNA expression profile in cancer
Reactome	Fabregat et al. (2017)	pathway
PathBank	Wishart et al. (2019)	pathway
UniProt	The UniProt Consortium (2016)	protein sequence
STRING	Szklarczyk et al. (2018)	protein–protein interaction
BRENDA	Chang et al. (2020)	enzyme

- PubMed section: it downloads titles and abstracts of PubMed articles through SemmedDB SENTENCE table (Kilicoglu et al. (2012)). Such table contains all the sentences related to the articles' title and abstract in PubMed.
- Literature databases section: it downloads the external databases which are used for i) filtering of noisy annotation entities caused by disambiguation and high generality of the Wikipedia corpus; ii) building literature edges, a biological evidence resulting from laboratory experiments, biological and biophysical processes. These edges allow us to evaluate the quality of BioTAGME prediction. Note that some databases, such as DrugBank (Wishart et al. (2007)), PharmGKB (Thorn et al. (2013)), Brenda (Chang et al. (2020)), require free registration or authorization to be downloaded. Therefore, such a procedure is left to the user.
- The import section transfers the downloaded databases from the local file system to the Hadoop FileSystem (HDFS).

2.1.2 SQL to JSON Parser Module

Although SemmedDB guarantees faster downloads than NCBI Entrez APIs, it has two main issues: the 1) title and abstract of each PMID (Document identifier in PubMed) are divided into sentences, and 2) the SENTENCE table is in a SQL format, which is not natively supported by the Spark engine.

To solve these issues, we implemented a new Spark module, named SQL2Json parser, that extracts headers, and every data row from a table by applying Spark SQL Window methodology. Each row is then aggregated to form the complete title and abstract through Spark built-in collect_list, concat_ws, and group-by functions. Finally, the parsed data is converted into JSON format and stored within the Hadoop FileSystem.

2.1.3 External Databases Integration Module

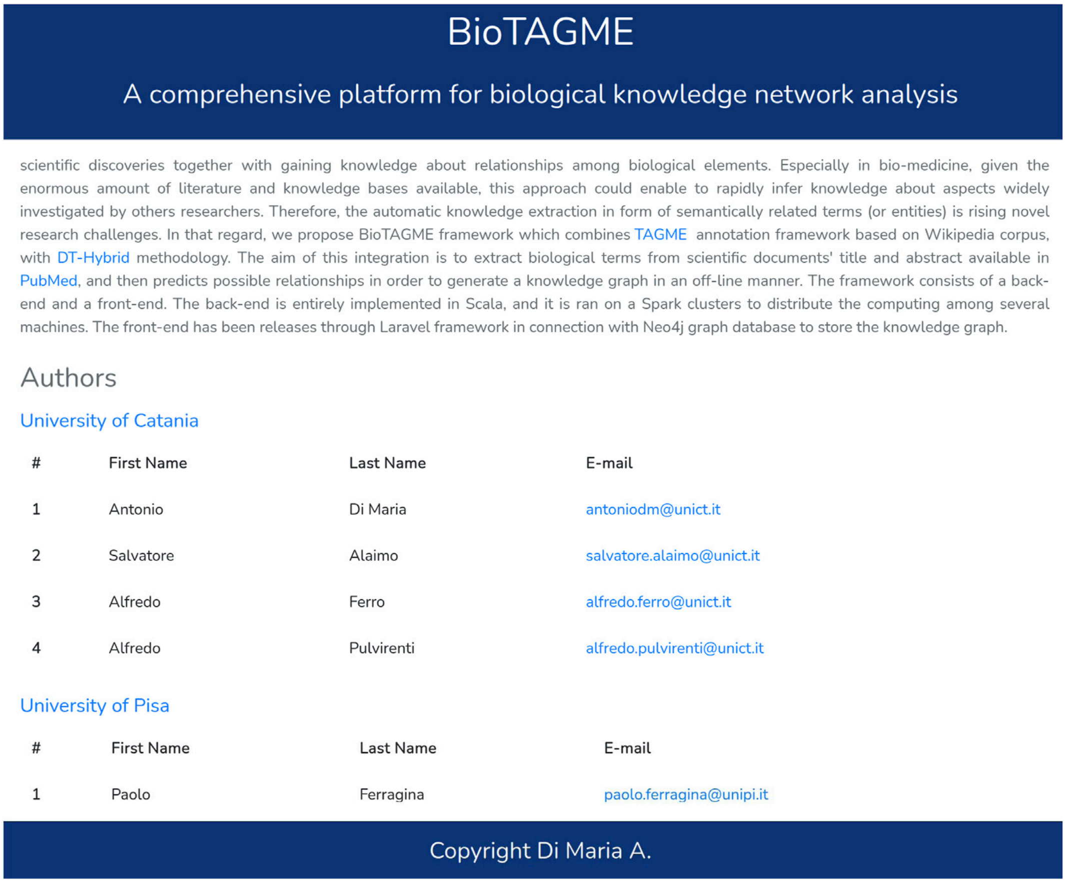
As previously mentioned, several databases are integrated into our pipeline. However, there are a few issues to consider: 1) Different databases often use different words to describe the same entity (synonyms). For example, DisGenNET uses "Colorectal cancer, hereditary nonpolyposis, type 1", while DiseaseOntology (DO) uses "Lynch syndrome 1" to refer to the same disease. 2) Equivalent attributes have different names in different databases. For example, a database might use the attribute name "mirna_nr", while another database might use "id". 3) Different databases might use different files formats, such as JSON, XML, TXT, CSV, TAB, OBO, GTF, FASTQ, and SQL, etc.

We implemented an integration module that executes the following tasks to tackle such issues. First, all databases are loaded into Spark DataFrames. We use the built-in Spark functions for CSV (read.csv), Tab-delimited and TXT (read.txt), and JSON (read.json) files. To import OBO, GFT, SQL, and FASTQ files, we implemented custom spark modules that convert such formats into DataFrames. The Databricks Spark-XML (Databricks (2021)) library is used for XML files. Then, each DataFrame is processed and subjected to a schema redefinition by using external databases metadata, synonyms list, and references (toward other external databases) list to harmonize the contents of the different data sources. This module is a fundamental intermediate layer that transforms all external databases into new ones having the same schema, attributes, format, and nomenclature.

2.1.4 Annotation Module

This module transforms documents' titles and abstracts into a list of annotation entities. Thus, for each document "t_i", a tuple (TI_AB, TAGME parameters map)$_i$ is generated and sent to the TAGME API through an HTTP POST request. We use TI_AB to represent the union of Document$_i$ Title and Abstract.

TAGME removes all stop-words and punctuation symbols from the TI_AB text at first. Then, a list of "annotation entities" is extracted and returned in response to the request, where each entity can be one or more words. Each annotation entity contains entity text, Wikipedia page title, Wikipedia page categories, and Wikipedia page ID. Each entity will be a node of the knowledge graph.



BioTAGME

A comprehensive platform for biological knowledge network analysis

scientific discoveries together with gaining knowledge about relationships among biological elements. Especially in bio-medicine, given the enormous amount of literature and knowledge bases available, this approach could enable to rapidly infer knowledge about aspects widely investigated by others researchers. Therefore, the automatic knowledge extraction in form of semantically related terms (or entities) is rising novel research challenges. In that regard, we propose BioTAGME framework which combines TAGME annotation framework based on Wikipedia corpus, with DT-Hybrid methodology. The aim of this integration is to extract biological terms from scientific documents' title and abstract available in PubMed, and then predicts possible relationships in order to generate a knowledge graph in an off-line manner. The framework consists of a back-end and a front-end. The back-end is entirely implemented in Scala, and it is ran on a Spark clusters to distribute the computing among several machines. The front-end has been releases through Laravel framework in connection with Neo4j graph database to store the knowledge graph.

Authors

University of Catania

#	First Name	Last Name	E-mail
1	Antonio	Di Maria	antoniodm@unict.it
2	Salvatore	Alaimo	salvatore.alaimo@unict.it
3	Alfredo	Ferro	alfredo.ferro@unict.it
4	Alfredo	Pulvirenti	alfredo.pulvirenti@unict.it

University of Pisa

#	First Name	Last Name	E-mail
1	Paolo	Ferragina	paolo.ferragina@unipi.it

Copyright Di Maria A.

FIGURE 2 | BioTAGME homepage.

TAGME annotations are not entirely accurate. The authors provide an estimate F_1 measure of 0.78, where F_1 is the harmonic mean between the precision and the recall of the annotation process. However, this does not considers any improvement due to 1) more up-to-date Wikipedia dumps and 2) pages filtering to obtain only Wikipedia pages relevant to the Biological field. Indeed, we properly pruned the Wikipedia network using the main biological categories[6] to 1) perform annotation only on Biological entities, and 2) mitigate the disambiguation problem.

Finally, the documents with their annotation entities are sent to the prediction module to generate the relationships.

2.1.5 Prediction Module

Our methodology aims to predict a potential relationship between i-th entity and j-th entity based on the BioTAGME score value ($BioTG_{i,j}$). This score is defined as the product between the DT-Hybrid score $s_{i,j}$ (Alaimo et al. (2013)) and the TAGME relatedness one $r_{i,j}$ (Ferragina and Scaiella (2010)). The higher is the score value, the higher is the meaningfulness of the predicted relationship.

The domain tuned-hybrid (DT-Hybrid) tool (Alaimo et al. (2013)) defines a recommendation method based on a bipartite network projection technique that implements the concept of resources transfer within the network to predict the robustness of the relationship between a pair of entities.

The DT-Hybrid score is computed by using a DT-Hybrid version running on Spark; the TAGME relatedness is computed through the online TAGME service available at[7]. The relatedness value is in the range [0,1] and expresses how much two entities are semantically related within the Wikipedia corpus. The value zero means no relationships between them; the value one means equivalence between two entities.

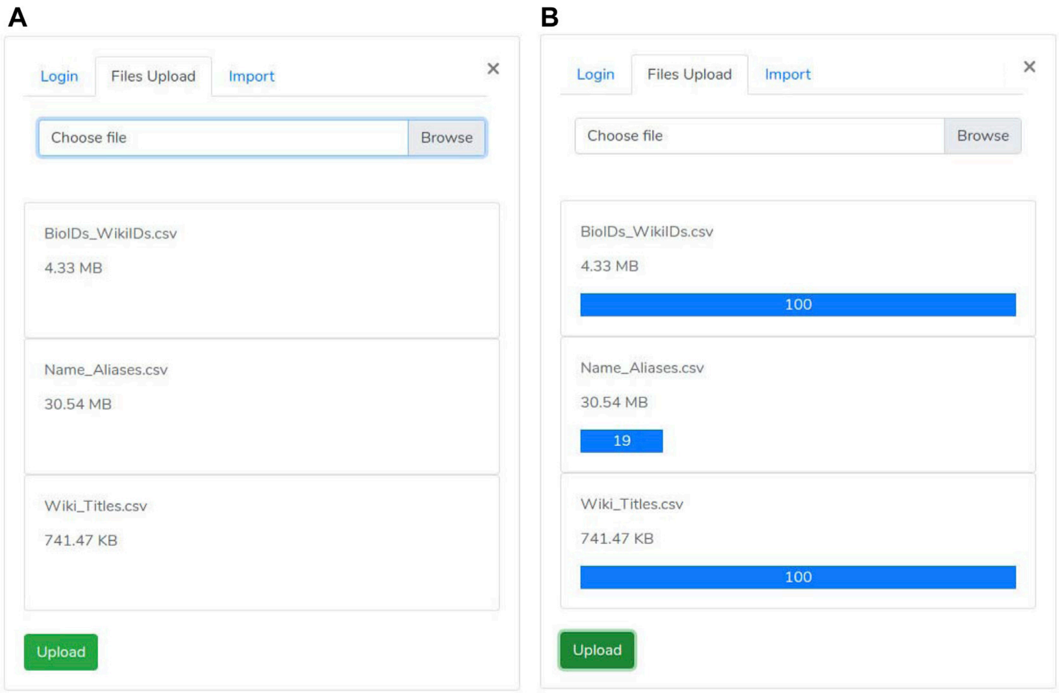
The output of this step is a set of relations between entities. These relations are then integrated during the network-construction phase with others coming from the external databases.

2.1.6 Network Construction

As soon as the documents have been annotated and the prediction procedure has been completed, the last step of the pipeline is to build the Knowledge Graph containing logical or physical relationships among biological elements. Physical

[6] https://en.wikipedia.org/wiki/Portal:Biology.

[7] https://tagme.d4science.org/tagme/.



A

BioIDs_WikiIDs.csv
4.33 MB

Name_Aliases.csv
30.54 MB

Wiki_Titles.csv
741.47 KB

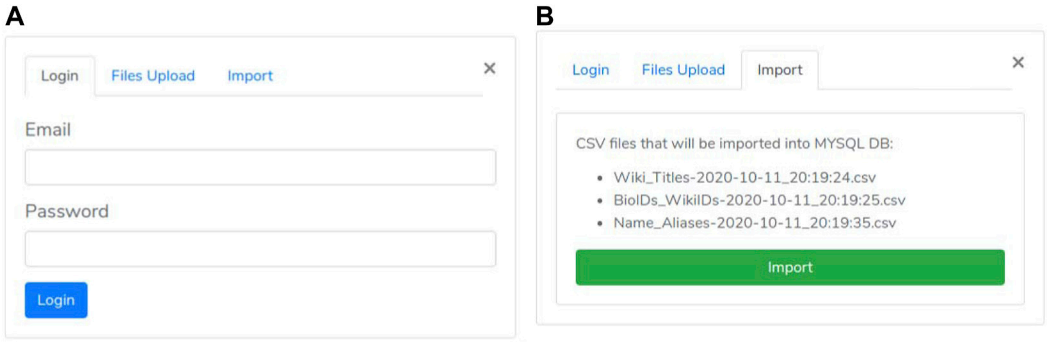
B

BioIDs_WikiIDs.csv
4.33 MB

Name_Aliases.csv
30.54 MB

Wiki_Titles.csv
741.47 KB

FIGURE 3 | Upload Panel: Such a panel allows a manager user of the site to load the graph and metadata. **(A)** File upload panel; **(B)** File upload panel with progress bar displaing the upload status.



CSV files that will be imported into MYSQL DB:

- Wiki_Titles-2020-10-11_20:19:24.csv
- BioIDs_WikiIDs-2020-10-11_20:19:25.csv
- Name_Aliases-2020-10-11_20:19:35.csv

FIGURE 4 | Authentication and Import panel. **(A)** Authentication panel; **(B)** Import panel.

relationships represent the real connection between biological entities. Instead, the logical one represents the effect that a biological entity (i.e., Drug) could have on another one (i.e., Disease or Gene).

For every $Entity_i$–$Entity_j$ association obtained during the prediction procedure, our system creates three different edges types:

- Literature: indicates an interaction derived from a publication, describing a biological evidence resulting from laboratory experiments, biological, and biophysical processes, etc.
- STRING: represents the predicted protein-protein associations stored in the STRING database. We report this information because our system integrates STRING *Homo sapiens* protein-protein interactions.
- BioTAGME: the edges predicted by our tool.

Both BioTAGME edges and STRING edges are marked with the corresponding score value to indicate the interaction's likelihood. More information about the plotting of the network, motif search, and shortest path computations are reported in the following **Section 2.2**.

We publicly release our network on Zenodo. The link is provided in the Supplementary Data section. Data is fully compliant with *FAIR* principles (Wilkinson et al. (2016)).

Supplemental Data

The networks data (nodes, edges, and other files) are available at: https://doi.org/10.5281/zenodo.6325345360.

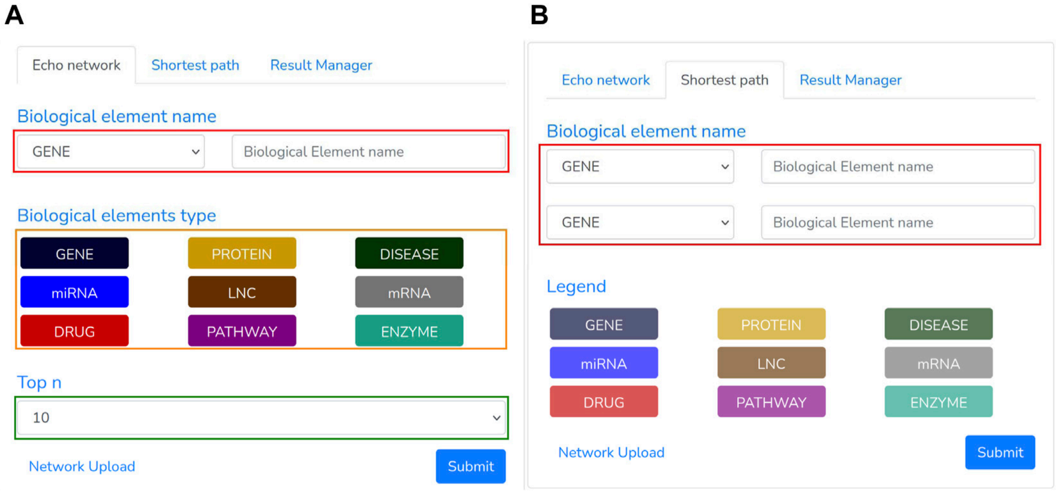


FIGURE 5 | Echo Network and Shortest path panel: The first **(A)** is used to extract the neighborhood of a given node and type. **(B)** The second one, instead, returns the shortest path among two specified biological entities.

The pipeline one code is available at: https://github.com/Anto188bas/biotagme pipeline.git361

The pipeline two code is available at: https://github.com/Anto188bas/biotagme laravel.git362

The docker-compose.yml file is available at: https://github.com/Anto188bas/biotagme docker.git363

2.1.7 Updating Procedure

BioTAGME pipeline annotates Pubmed documents' titles and abstracts to predict the relationships among their corresponding biological entities. A periodical update is needed since many new documents are submitted daily to the Pubmed database.

Our pipeline carries out the following steps to achieve this purpose. First, it downloads all the PMIDs (Documents' identifier in PubMed) within an established data range [mindate, maxdate] through an NCBI esearch POST request. "Mindate" usually refers to the last updating date; whereas "maxdate" is usually set to the actual date.

Once the PMIDs list has been obtained, the updating module downloads the title and abstract of these PMIDs using the NCBI efetch API. For performance reasons, the PMIDs list is partitioned into chunks of proper size, and then several chunk-based NCBI efetch post requests are generated and sent to the Pubmed server to obtain the required data. NCBI does not impose a maximum on the number of requests to be submitted, especially when a POST request is used. However, we suggest keeping this value under 10,000 to reduce the computational burden of our job.

Once the documents' titles and abstracts have been downloaded, the annotation, prediction, and network construction procedures are executed to update the Knowledge Graph's edges and nodes.

The update procedure is incremental. It does not require the entire PubMed abstracts corpus. It runs on a subset of abstracts within a date range ([start_date, end_date]), and then generate a knowledge graph only on those abstracts. Therefore, this procedure could be used to produce a temporal knowledge graphs over a certain topic of interest.

2.2 Pipeline Two: Network Deployment and Query Interface

The second pipeline has been implemented for importing the Knowledge Graph into the Neo4j database and querying the network to get the neighborhood of a biological element or compute the shortest path between two nodes. The interface module for network querying is crucial to exploit such graphs and infer putative novel biological knowledge. This pipeline employs the Laravel model-view-controller and the React Native framework to implement the back-end and web-pages components. In this section, we will describe such modules (**Figure 2**).

2.2.1 Network Import Module

A user may access the upload section through the "biological element search" panel by clicking on the "network files upload" link. Such section includes three consecutive phases:

- the first one is the "authentication phase" ensuring that only authorized users may execute the import procedure (**Figure 4A**).
- then, the "files selection phase" is enabled (**Figure 3**). During this phase, the user selects both "nodes.csv" and "edges.csv" files containing the network components and the "Name_Aliases.csv" file about biological elements aliases. Since the size of the files is large (GB), our system uses the "Pion" library (Pion (2021)) to split the file into small chunks (client-side) and re-assemble them as soon as these are correctly received (server-side).
- As all files are successfully received, the "import phase" is enabled. It shows a summary (**Figure 4B**) of the uploaded files to check for file selection mistakes. If everything is

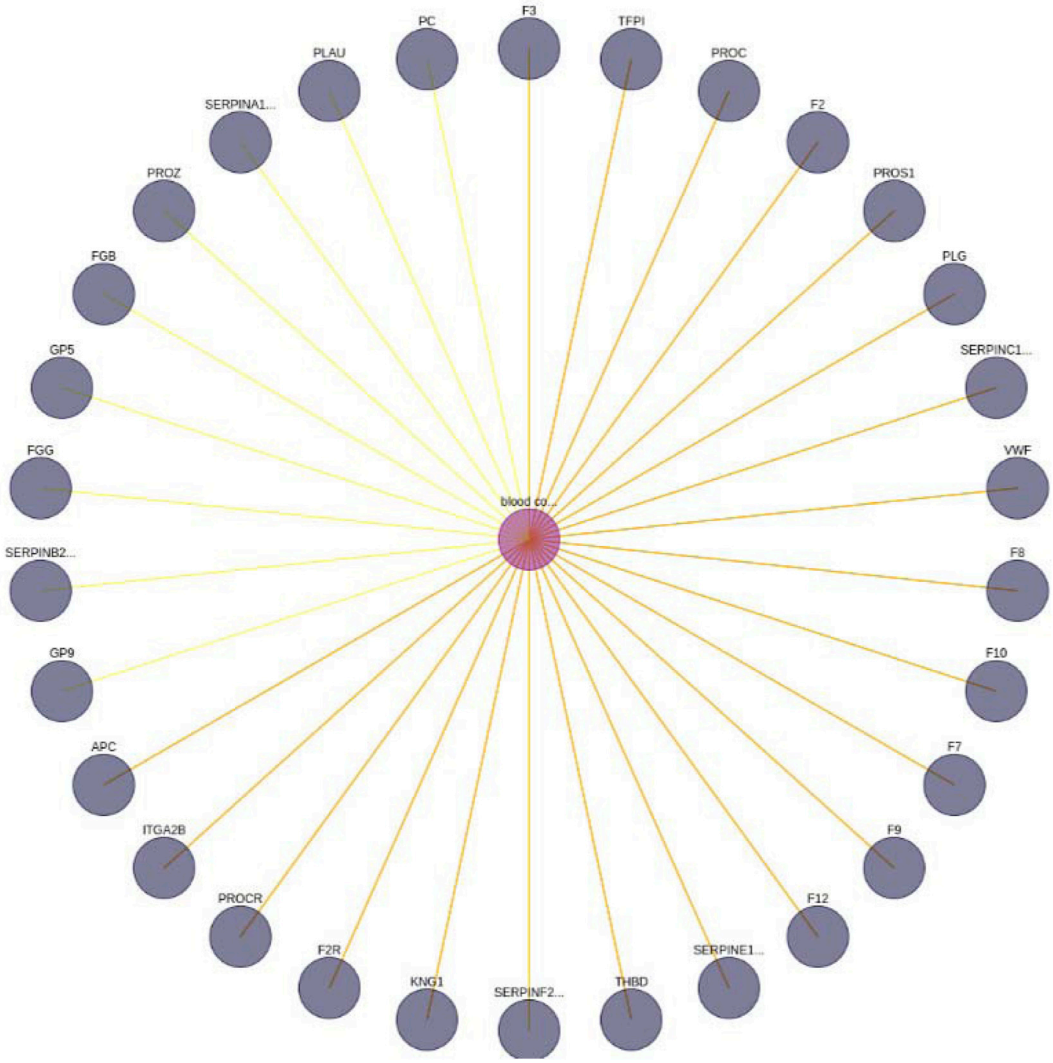


FIGURE 6 | Blood coagulation—gene interaction network. A limit of 30 has been set. In addition, the yellow edges represent a set of BioTagME unpredicted edges (extracted by external databases). Instead, the orange ones (yellow + red) are edges both predicted (by BioTagME) and extracted from the external databases.

correct, the user can trigger network loading on Neo4j by clicking the import button.

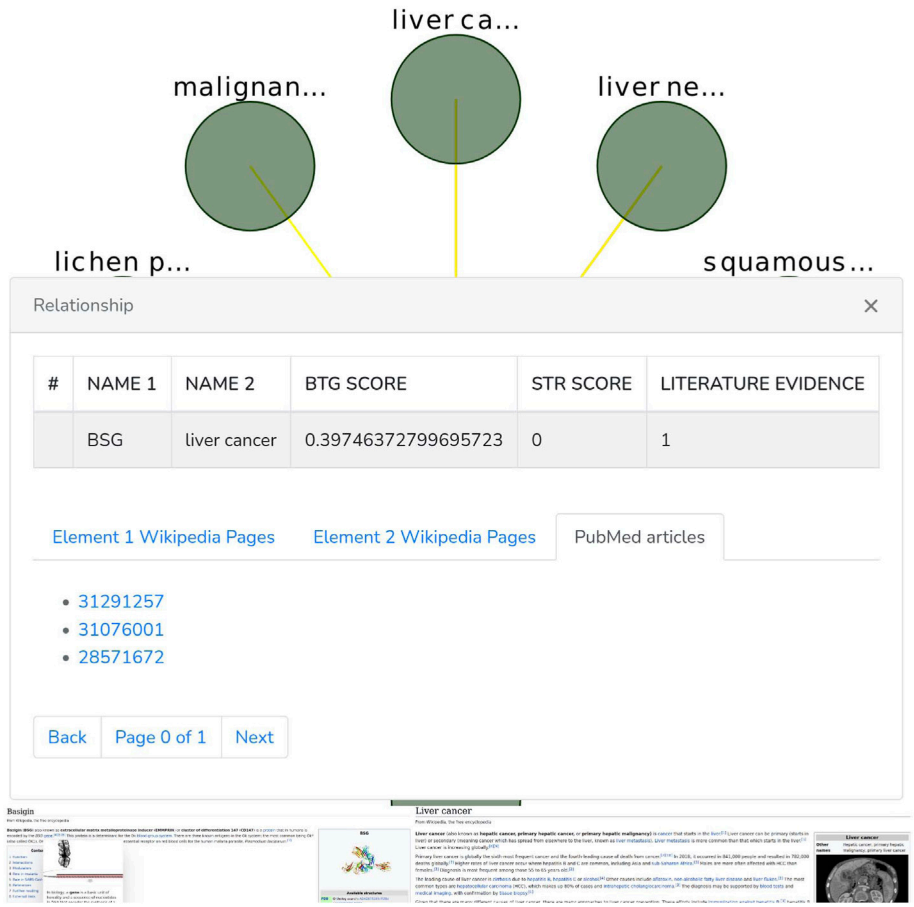
2.2.2 Searching Module

Once the network has been imported, a user may execute several queries through our "GUI", composed of the following panels: Searching panel (**Figure 5**) and Graph panel (**Figure 6**).

The Searching Panel is used for setting the query parameters based on the selected menu: 1) Echo network or 2) shortest path.

- When the Echo Network option is selected, a user may search the Echo Network of a biological entity "be_i". Therefore, he should provide the type and name of the biological entity to be analyzed (**Figure 5A**, red rectangle) and the type of the other entities (**Figure 5A**, orange rectangle) to include within the echo network. To avoid building a large graph, a maximum number of entities has to be supplied (ranging from 10 to 200 nodes) through the "Top n" section (**Figure 5A**, green rectangle). Once all the required parameters have been filled, the search process can be triggered by clicking the Submit button. This process transforms the specified parameters in a "Cypher query"[8] that looks for the "Top n" nodes having one or more links from/to "be_i".
- When the Shortest Panel option is selected (**Figure 5B**), a user looks for the shortest path between two biological entities. First, the user specifies the type and name of the source "*el_src*" and destination "*el_dst*" entities (**Figure 5B**, red rectangle), and then BioTAGME transforms all these parameters into a proper "Cypher query" which is mainly based on a Neo4j shortest path computation.

[8] Cypher is Neo4j′s query language to retrieve data from the graph, and was inspired by SQL.



Relationship ×

#	NAME 1	NAME 2	BTG SCORE	STR SCORE	LITERATURE EVIDENCE
	BSG	liver cancer	0.39746372799695723	0	1

Element 1 Wikipedia Pages | Element 2 Wikipedia Pages | PubMed articles

- 31291257
- 31076001
- 28571672

Back | Page 0 of 1 | Next

FIGURE 7 | Relationship window.

The Graph phanel is used to plot [by using the CytoscapeJS library (Franz et al. (2015))] the sub-graph (**Figure 6**) corresponding to a user-submitted query. The edges of such sub-graph are interactive. Thus, if a user clicks on them, then a relationship window (**Figure 7**) containing the following data is shown:

- A table containing the name of the source and destination nodes as well as the BioTagME and STRING scores. In addition, the last column of the table also reports the literature evidence (1 if the relationship is reported in at least one of the literature databases, 0 otherwise).
- A navigation panel with three different options. The first two (Element 1 Wikipedia Pages and Element 2 Wikipedia Pages) show several links among Wikipedia pages and source or destination nodes, respectively. The last one (PubMed articles) shows all the links to PubMed articles containing the selected relationship.

3 EXPERIMENTAL ANALYSIS

We analyzed the the reliability of BioTagMe on two case studies. The first one aims at determining preduction quality by evaluating our ability to extract “Basigin” relationships. The results were compared with STRING (Szklarczyk et al. (2018)). The second case study focuses on the construction of a “blood coagulation” network. Such a network is then compared against a literature one (generated by the links among the external databases employed in BioTagME, **Table 1**).

3.1 Case study 1

Many tools and computational models (Alaimo et al. (2020)) rely on existing network databases, such as KEGG (Kanehisa and Goto (2000)) and Reactome (Fabregat et al. (2017)). However, despite the enormous amount of available data, these databases are still incomplete and therefore have partial information.

In this case study, we have chosen *Basigin* (BSG), also known as CD147 or EMMPRIN, as a starting point to construct a protein-protein functional network. This gene represents an example of a biological element that should be supplemented to the KEGG network since it is not currently described in their pathways. BSG is a transmembrane glycoprotein of the immunoglobulin superfamily, expressed in many tissues and cells. It is known to participate in several highgly relevant biological and clinical processes. Furthermore, BSG is a crucial molecule in the pathogenesis of several human diseases (Xiong et al. (2014)).

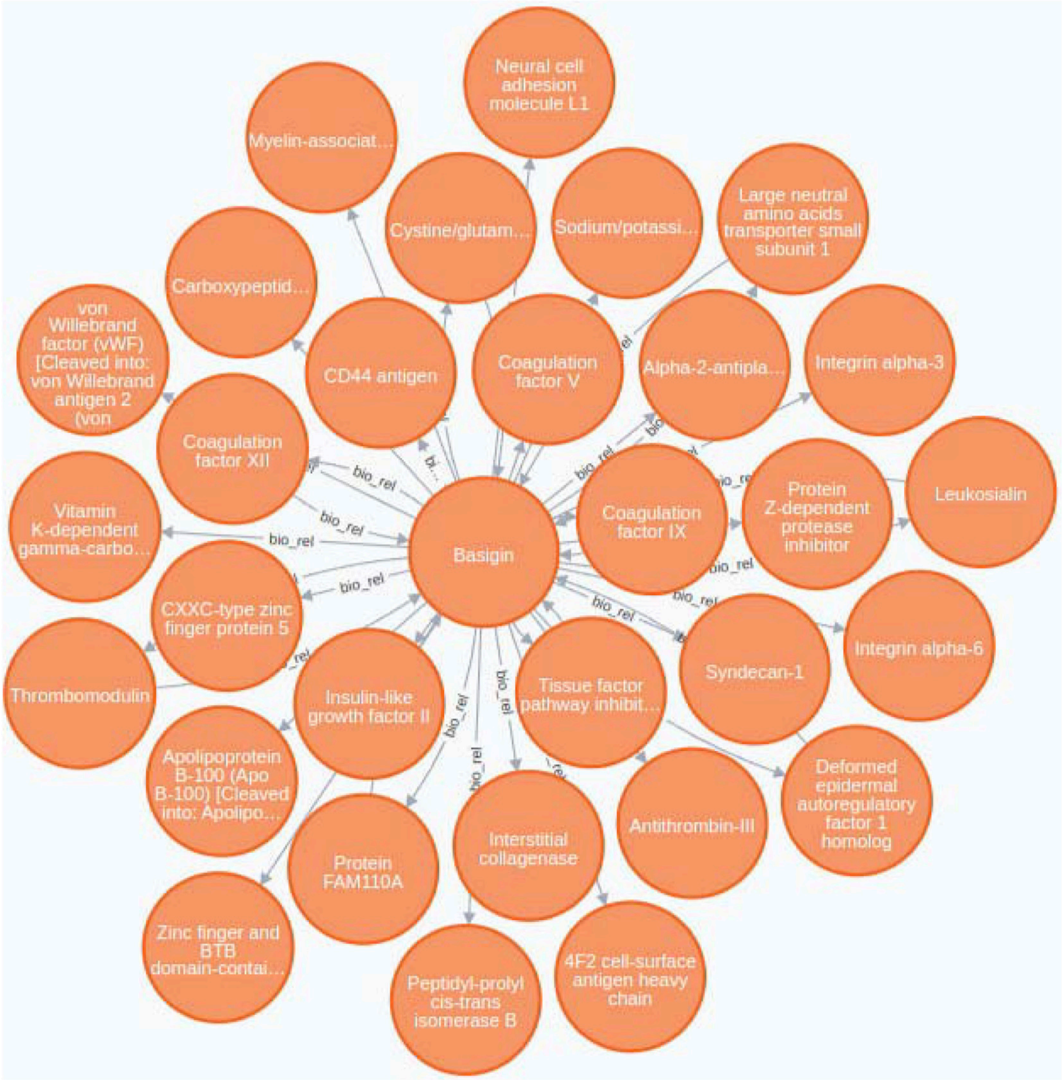

Neural cell adhesion molecule L1
Myelin-associat...
Cystine/glutam...
Sodium/potassi...
Large neutral amino acids transporter small subunit 1
Carboxypeptid...
von Willebrand factor (vWF) [Cleaved into: von Willebrand antigen 2 (von
CD44 antigen
Coagulation factor V
Alpha-2-antipla...
Integrin alpha-3
Coagulation factor XII
Vitamin K-dependent gamma-carbo...
Basigin
Coagulation factor IX
Protein Z-dependent protease inhibitor
Leukosialin
CXXC-type zinc finger protein 5
Thrombomodulin
Insulin-like growth factor II
Tissue factor pathway inhibit...
Syndecan-1
Integrin alpha-6
Apolipoprotein B-100 (Apo B-100) [Cleaved into: Apolipo...
Protein FAM110A
Interstitial collagenase
Antithrombin-III
Deformed epidermal autoregulatory factor 1 homolog
Zinc finger and BTB domain-contai...
Peptidyl-prolyl cis-trans isomerase B
4F2 cell-surface antigen heavy chain
bio_rel

FIGURE 8 | Basigin-Proteins interaction network. It has been created using the Neo4j user interface. In addition a limit of 30 nodes has been set. BioTagme and STRING edges have been merged in a single one.

Missing a crucial gene within a biological network can compromise scientists' efforts to understand certain molecular mechanisms. However, the most reliable approach to date remains the manual curation through careful and time-consuming literature analysis. On the other hand, a manually constructed network provides partial information due to the limited number of articles that a scientist could read.

Our case study tackle this issue by providing a practical example of how BioTagME can create valuable networks (**Figure 8**) by analyzing a large sets of PubMed abstracts. In addition, such a network has been compared with STRING to assess sensitivity and specificity.

Through BioTagMe, we inferred 426 true positive relations and 38 false negatives. Qualitatively, this network includes most of the interconnections mentioned in STRING, thus providing a reliable and comprehensive overview of the molecular function of *Basigin*. Quantitatively, BioTagME achieved a sensitivity of 91.8%, and a specificity of 94.8%.

3.2 Case Study 2

The second case study aims to build a general functional network related to the "blood coagulation pathway" and other biological entities (i.e. diseases, genes).

Blood coagulation is a complex chain process involving a series of stimulus responses in conjunction with coagulation factors and enzymes, whose intent is to stop blood fluxes when a vascular tissue injury occurs (Ngo et al. (2012)).

To evaluate the quality of BioTagME, our network (**Figure 6**) is compared with a "literature network" (generated by data and relationships into the external databases, **Table 1**) in terms of sensitivity and specificity.

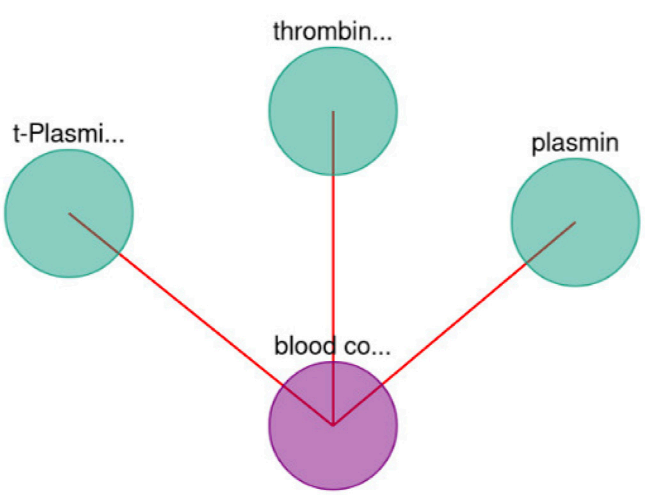


FIGURE 9 | Blood coagulation and enzymes interaction.

BioTagMe was able to infer 54 true positive and 23 false negative. Quantitatively, We achieved a sensitivity of 70.12%, and a specificity of 96.43%. Indeed, we could predict the relation between blood coagulation and PROS1 (**Figure 6**). Such gene plays a crucial role on the mechanism of PtdSer exposure during immunity and blood coagulation (Wang et al. (2022)).

Moreover, BioTagME could predict the relations among blood coagulation and the thrombin and plasmin enzymes (**Figure 9**). The role of Thrombin enzyme is to catalyze the initiation and propagation phases of blood coagulation. In addition, it converts soluble fibrinogen to insoluble fibrin (Becker et al. (2013)).

4 CONCLUSION

In this paper, we have implemented the BioTAGME framework for building offline biological knowledge graphs from all documents' titles and abstracts in PubMed. First, the graph's nodes (biological entities) have been extracted by TAGME. The edges, instead, have been predicted through the combination of the DT-Hybrid algorithm score and the TAGME relatedness computation. Such predicted edges have also been enriched with literature evidence resulting from laboratory experiments, biological, and biophysical processes (extracted from the connections among external databases), and protein-protein relationships in STRING. Moreover, an uploader module has been implemented to download and annotate new documents in PubMed to keep the graph up-to-date. Finally, the main pipeline (pipeline one) has been implemented using the Spark Framework to distribute the computation among several machines. Future works will include: 1) construction of knowledge-graphs based on open-access documents' title, abstract and full-text in PubMed and PubMed Central; 2) implementation and integration of new prediction algorithms to improve and increase the prediction of the relationship among biological entities; 3) implement a TAGME version based on a biological Wikipedia corpus (no biological pages will be pruned); 4) development of a new search panel to enable advanced queries in the knowledge-graph. Such a panel will provide: algorithms for community detection (clustering); matching, shortest path, and k-shortest path based on BioTagME score, nodes and edges types, publication date, etc; centrality measures; cypher free text for writing custom queries. Moreover, we will add a list of sentences (where possible) to describe predicted relationships.

4.1 Permission to Reuse and Copyright

Figures, tables, and images will be published under a Creative Commons CC-BY license, and permission must be obtained for the use of copyrighted material from other sources (including re-published/adapted/modified/partial figures and images from the internet). It is the responsibility of the authors to acquire the licenses, follow any citation instructions requested by third-party rights holders, and cover any supplementary charges.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

AP, PF, and AF conceived the work and coordinated the research. AD and SA designed and developed the system. LB and FB tested the system. AD and AP wrote the first draft of the paper. All the authors read and approved the final version of the paper.

FUNDING

AP, SA, and AF, have been partially supported by the following research projects: PO-FESR Sicilia 2014–2020 “DiOncoGen: Innovative diagnostics” (CUP G89J18000700007). AP, has been also partially supported by the following research project: “PROMOTE: Identificazione di nuovi biomarcatori per la diagnosi precoce di mesotelioma maligno pleurico in soggetti ex esposti a fibre asbestiformi”, University of Catania—Piano di incentivi per la ricerca 2020–2022. PF and LB have been supported by the EU H2020 programmes “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” (INFRAIA-01–2018-2019, grant # 871042), and by “Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us” (grant # 820437).

REFERENCES

Alaimo, S., Pulvirenti, A., Giugno, R., and Ferro, A. (2013). Drug-target Interaction Prediction through Domain-Tuned Network-Based Inference. *Bioinformatics* 29, 2004–2008. doi:10.1093/bioinformatics/btt307

Alaimo, S., Rapicavoli, R. V., Marceca, G. P., La Ferlita, A., Serebrennikova, O. B., Tsichlis, P. N., et al. (2021). PHENSIM: Phenotype Simulator. *PLoS Comput. Biol.* 17 (6), e1009069. doi:10.1371/journal.pcbi.1009069

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2010). Network Medicine: a Network-Based Approach to Human Disease. *Nat. Rev. Genet.* 12, 56–68. doi:10.1038/nrg2918

Beck, J. (2010). Report from the Field: PubMed central, an XML-Based Archive of Life Sciences Journal Articles. *Proceedings* 6. doi:10.4242/balisagevol6.beck01

Becker, R. C., Voora, D., and Shah, S. H. (2013). "Hemostasis and Thrombosis," in *Genomic and Personalized Medicine* (Amsterdam, Netherlands: Elsevier), 602–611. doi:10.1016/b978-0-12-382227-7.00052-5

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., et al. (2004). An Overview of Ensembl. *Genome Res.* 14, 925–928. doi:10.1101/gr.1860604

Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., et al. (2020). BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates. *Nucleic Acids Res.* 49, D498–D508. doi:10.1093/nar/gkaa1025

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2010). Reactome: a Database of Reactions, Pathways and Biological Processes. *Nucleic Acids Res.* 39, D691–D697. doi:10.1093/nar/gkq1018

[Dataset] Apache-Spark (2016). *Spark Sql, Dataframes and Datasets Guide.*

[Dataset] Databricks (2021). *Databricks/spark-xml: Xml Data Source for Spark Sql and Dataframes.*

[Dataset] Pion (2021). *Pion Network Library (Boost Licensed Open Source).*

Daugherty, L. C., Seal, R. L., Wright, M. W., and Bruford, E. A. (2012). Gene Family Matters: Expanding the HGNC Resource. *Hum. Genomics* 6. doi:10.1186/1479-7364-6-4

Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., et al. (2017). Reactome Pathway Analysis: a High-Performance In-Memory Approach. *BMC Bioinformatics* 18. doi:10.1186/s12859-017-1559-2

Ferragina, P., and Scaiella, U. (2010). *TAGME*. New York City: ACM Press. doi:10.1145/1871437.1871689

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2015). Cytoscape.js: a Graph Theory Library for Visualisation and Analysis. *Bioinformatics* 2015, btv557. doi:10.1093/bioinformatics/btv557

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., et al. (2017). Systematic Integration of Biomedical Knowledge Prioritizes Drugs for Repurposing. *Elife* 6, e26726. doi:10.7554/elife.26726

Hoyt, C. T., Domingo-Fernández, D., and Hofmann-Apitius, M. (2018). BEL Commons: an Environment for Exploration and Analysis of Networks Encoded in Biological Expression Language. *Database (Oxford)* 2018, bay126. doi:10.1093/database/bay126

Huang, H.-Y., Lin, Y.-C. -D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2019). miRTarBase 2020: Updates to the Experimentally Validated microRNA-Target Interaction Database. *Nucleic Acids Res.* 48, D148–D154. doi:10.1093/nar/gkz896

Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: a Map of Putative microRNA Target Sites in the Long Non-coding Transcriptome. *Bioinformatics* 28, 2062–2063. doi:10.1093/bioinformatics/bts344

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T. C. (2012). SemMedDB: a PubMed-Scale Repository of Biomedical Semantic Predications. *Bioinformatics* 28, 3158–3160. doi:10.1093/bioinformatics/bts591

Kim, J.-D., Wang, Y., Fujiwara, T., Okuda, S., Callahan, T. J., and Cohen, K. B. (2019). Open Agile Text Mining for Bioinformatics: the PubAnnotation Ecosystem. *the PubAnnotation ecosystem* 35, 4372–4380. doi:10.1093/bioinformatics/btz227

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2018). miRBase: from microRNA Sequences to Function. *Nucleic Acids Res.* 47, D155–D162. doi:10.1093/nar/gky1141

Lambrix, P., Tan, H., Jakoniene, V., and Strömbäck, L. (2007). *Biological Ontologies.* Berlin, Germany: Springer US, 85–99. doi:10.1007/978-0-387-48438-9_5

Lever, J., and Jones, S. (2017). Painless Relation Extraction with Kindred. *BioNLP* 2017, 176–183. doi:10.18653/v1/w17-2322

Ma, X., and Gao, L. (2012). Biological Network Analysis: Insights into Structure and Functions. *Brief. Funct. Genomics* 11, 434–442. doi:10.1093/bfgp/els045

McBride, B. (2004). *The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS*. Berlin, Germany: Springer Berlin Heidelberg, 51–65. doi:10.1007/978-3-540-24750-0_3

Muscolino, A., Di Maria, A., Rapicavoli, R. V., Alaimo, S., Bellomo, L., Billeci, F., et al. (2022). NETME: On-The-Fly Knowledge Network Construction from Biomedical Literature. *Appl. Netw. Sci.* 7. doi:10.1007/s41109-021-00435-x

Ngo, D.-H., Vo, T.-S., Ngo, D.-N., Wijesekara, I., and Kim, S.-K. (2012). Biological Activities and Potential Health Benefits of Bioactive Peptides Derived from marine Organisms. *Int. J. Biol. Macromolecules* 51, 378–383. doi:10.1016/j.ijbiomac.2012.06.001

Ossom Williamson, P., and Minter, C. I. J. (2019). Exploring PubMed as a Reliable Resource for Scholarly Communications Services. *jmla* 107. doi:10.5195/jmla.2019.433

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2019). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Rindflesch, T. C., and Fiszman, M. (2003). The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. *J. Biomed. Inform.* 36, 462–477. doi:10.1016/j.jbi.2003.11.003

Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Strauss, M., Geyer, P. E., et al. (2022). A Knowledge Graph to Interpret Clinical Proteomics Data. *Nat. Biotechnol.*. doi:10.1038/s41587-021-01145-6

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., et al. (2010). Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications. *J. Am. Med. Inform. Assoc.* 17, 507–513. doi:10.1136/jamia.2009.001560

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., et al. (2018). Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion. *Nucleic Acids Res.* 47, D955–D962. doi:10.1093/nar/gky1032

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2018). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47, D607–D613. doi:10.1093/nar/gky1131

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible. *Nucleic Acids Res.* 45, D362–D368. doi:10.1093/nar/gkw937

The UniProt Consortium (2016). UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi:10.1093/nar/gkw1099

Thorn, C. F., Klein, T. E., and Altman, R. B. (2013). "PharmGKB: The Pharmacogenomics Knowledge Base," in *Methods in Molecular Biology* (Totowa, NJ, USA: Humana Press), 311–320. doi:10.1007/978-1-62703-435-7_20

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2012). LNCipedia: a Database for Annotated Human lncRNA Transcript Sequences and Structures. *Nucleic Acids Res.* 41, D246–D251. doi:10.1093/nar/gks915

Wang, J., Yu, C., Zhuang, J., Qi, W., Jiang, J., Liu, X., et al. (2022). The Role of Phosphatidylserine on the Membrane in Immunity and Blood Coagulation. *Biomark Res.* 10. doi:10.1186/s40364-021-00346-0

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: Automated Concept Annotation for Biomedical Full Text Articles. *Nucleic Acids Res.* 47, W587–W593. doi:10.1093/nar/gkz389

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The Fair Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3. doi:10.1038/sdata.2016.18

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2007). DrugBank: a Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958

Wishart, D. S., Li, C., Marcu, A., Badran, H., Pon, A., Budinski, Z., et al. (2019). PathBank: a Comprehensive Pathway Database for Model Organisms. *Nucleic Acids Res.* 48, D470–D478. doi:10.1093/nar/gkz861

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-Cancer Association Database Constructed by Text Mining on Literature. *Bioinformatics* 29, 638–644. doi:10.1093/bioinformatics/btt014

Xiong, L., Edwards, C., and Zhou, L. (2014). The Biological Function and Clinical Utilization of CD147 in Human Diseases: A Review of the Current Scientific Literature. *Ijms* 15, 17411–17441. doi:10.3390/ijms151017411

Zhang, G., Shi, J., Zhu, S., Lan, Y., Xu, L., Yuan, H., et al. (2017). DiseaseEnhancer: a Resource of Human Disease-Associated Enhancer Catalog. *Nucleic Acids Res.* 46, D78–D84. doi:10.1093/nar/gkx920

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fgene.2022.855766

Benefits and Challenges of Pre-clustered Network-Based Pathway Analysis

*Miguel Castresana-Aguirre, Dimitri Guala and Erik L. L. Sonnhammer**

Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

Functional analysis of gene sets derived from experiments is typically done by pathway annotation. Although many algorithms exist for analyzing the association between a gene set and a pathway, an issue which is generally ignored is that gene sets often represent multiple pathways. In such cases an association to a pathway is weakened by the presence of genes associated with other pathways. A way to counteract this is to cluster the gene set into more homogenous parts before performing pathway analysis on each module. We explored whether network-based pre-clustering of a query gene set can improve pathway analysis. The methods MCL, Infomap, and MGclus were used to cluster the gene set projected onto the FunCoup network. We characterized how well these methods are able to detect individual pathways in multi-pathway gene sets, and applied each of the clustering methods in combination with four pathway analysis methods: Gene Enrichment Analysis, BinoX, NEAT, and ANUBIX. Using benchmarks constructed from the KEGG pathway database we found that clustering can be beneficial by increasing the sensitivity of pathway analysis methods and by providing deeper insights of biological mechanisms related to the phenotype under study. However, keeping a high specificity is a challenge. For ANUBIX, clustering caused a minor loss of specificity, while for BinoX and NEAT it caused an unacceptable loss of specificity. GEA had very low sensitivity both before and after clustering. The choice of clustering method only had a minor effect on the results. We show examples of this approach and conclude that clustering can improve overall pathway annotation performance, but should only be used if the used enrichment method has a low false positive rate.

Keywords: functional association networks, network clustering, biological mechanisms, pathway enrichment analysis, sensitivity increase

OPEN ACCESS

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Tao Huang,
Shanghai Institute of Nutrition and Health (CAS), China
Pier Luigi Martelli,
University of Bologna, Italy

****Correspondence:***
Erik L. L. Sonnhammer
erik.sonnhammer@dbb.su.se

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 January 2022*
Accepted: *25 April 2022*
Published: *10 May 2022*

Citation:
Castresana-Aguirre M, Guala D and Sonnhammer ELL (2022) Benefits and Challenges of Pre-clustered Network-Based Pathway Analysis. Front. Genet. 13:855766. doi: 10.3389/fgene.2022.855766

INTRODUCTION

The advance in high throughput experiments has led to a huge increase in the data available for understanding biological function. However, extracting function from high-throughput experiments is often not straightforward since genes and proteins are involved in many different biological mechanisms and pathways. The quest for biological insight from high-throughput experiments has therefore prompted the invention of a large number of pathway enrichment analysis tools.

The most recent family of pathway analysis methods are the network-based tools, such as EnrichNet (Glaab et al., 2012), NEAT (Signorelli et al., 2016), NEArender (Jeggari and Alexeyenko, 2017), BinoX (Ogris et al., 2017), and ANUBIX (Castresana-Aguirre and Sonnhammer, 2020). These

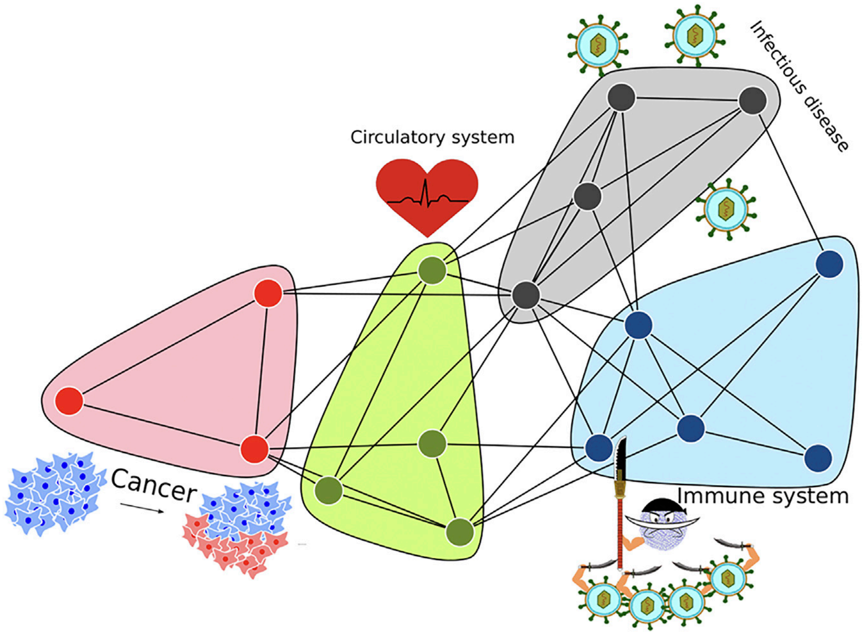


FIGURE 1 | Gene sets derived from experiments are often complex with multiple affected pathways. This illustration shows genes that belong to 4 pathways that are functionally distinct. The mixture of pathways may complicate the pathway enrichment analysis, especially for smaller pathways. By separating gene clusters prior to pathway analysis, a clearer picture of the pathway enrichment can be obtained.

methods require a functional association network, such as FunCoup (Persson et al., 2021) or STRING (Szklarczyk et al., 2021), where different types of data describing relationships between genes and/or proteins, are integrated to infer functional associations between genes. Using enrichment of network links, instead of overlap between gene sets, substantially improves the chances of detecting a relationship, as networks provide much more information (Ogris et al., 2017). Statistical significance of network-based pathway analysis methods is assessed based on the network crosstalk, i.e., links connecting the studied gene set and the pathway of interest. Methods such as BinoX rely on network randomization to obtain a null distribution, which is fit to a binomial distribution to compute the expected crosstalk. NEAT and NEArender compute the expected crosstalk based on the node degree of the query, the pathway and the network, with the difference that NEAT fits a hypergeometric distribution and NEArender a chi-square distribution, but their results are very similar. ANUBIX randomly samples gene sets of the same size as the original query set and fits the expected crosstalk to a beta-binomial distribution. While all these methods except ANUBIX have been shown to suffer from high false positive rates when testing random gene sets for enrichment (Castresana-Aguirre and Sonnhammer, 2020), we here included BinoX and NEAT, together with ANUBIX to study how clustering affects different methods.

Network-based methods provide the highest sensitivity of all the pathway enrichment families (Ogris et al., 2017; Castresana-Aguirre and Sonnhammer, 2020). However, experimental gene sets are often complex with multiple affected pathways, which increases noise and leads to decreased sensitivity. An example of this would be a gene set consisting of four functional modules where each one is enriched for a specific pathway (**Figure 1**). A pathway analysis method would struggle to detect each module's pathway association if the genes belonging to each module is only a small fraction of all genes in the gene set. Additionally, the studied gene set could contain noise in the form of other genes not related to the main phenotypes of the gene set, which could cause false negatives, impacting the sensitivity of pathway analysis.

Due to the ubiquitous use of pathway analysis methods and reliance on their output to interpret results from diverse and important fields of research such as drug development (Jhamb et al., 2019), biomarker discovery (Chen et al., 2017) and patient diagnosis (Lu et al., 2019), it is important to ensure that these methods can cope well with complex gene sets.

One way to achieve this is to reduce the mentioned complexity by separating the mix of affected pathways. Clustering is a technique that has been used to lower complexity of data by grouping similar entities in various fields, such as pattern recognition (Baraldi and Blonda, 1999; Chen and Huang, 2003), image analysis (Chen et al., 2015; Dhanachandra et al., 2015), and analysis of biological interaction networks (Ideker et al., 2002; Opresko et al., 2004; Mitra et al., 2013). In the field of pathway analysis, clustering is used in PathFindR (Ulgen et al., 2019) and GScluster (Yoon et al., 2019) to find subnetworks or modules in a gene set mapped to a protein-protein interaction (PPI) network, followed by gene overlap based pathway analysis. However, neither of these tools have evaluated the combination of clustering with state-of-the-art pathway analysis methods, nor have they compared the performance of used methods with and without clustering.

The approach we take here is applying clustering to decrease complexity of the gene sets, and then apply state-of-the-art network-based pathway enrichment methods. We first investigated whether top-performing clustering methods such as MCL, Infomap, and MGclus are able to extract single pathways from pathway mixtures. The performance of clustering in combination with the network-based pathway analysis methods BinoX, NEAT, and ANUBIX, as well as classical overlap-based Gene Enrichment Analysis (GEA), was evaluated using a benchmark constructed based on the KEGG pathway database.

MATERIALS AND METHODS

Clustering is a way to group objects into different communities, where the objects within each community are more similar to each other than to objects in the other communities (Malliaros and Vazirgiannis, 2013). When clustering is used in the context of a network it involves grouping nodes with high intra-module density, i.e., that are highly connected within a network neighborhood and less connected to the nodes outside said community. There are different types of clustering, e.g., connectivity clustering, centroid clustering, density clustering, distribution clustering, network-based clustering, etc. (Emmons et al., 2016). In our study we focus on network-based clustering, since we are mapping a query gene set onto a network. Since the purpose of this study is not to benchmark the clustering methods themselves, we decided to pick three methods. These methods are MGclus, which has been shown to work well with the FunCoup network (Frings et al., 2013), Infomap (Rosvall and Bergstrom, 2008), and MCL (Van Dongen, 2008), due to their superior performances compared to other methods (Lancichinetti and Fortunato, 2009; Shemirani et al., 2021).

Clustering Methods

MGclus defines modules based on the intra- versus inter-connectivity in a module and considers shared neighbors of nodes as evidence that they belong to the same module.

Both Infomap and MCL extract modules using random walks on the underlying network. MCL performs an iterative random walk along the edges of the network to discover where the flow tends to gather. These iterative random walks are calculated using Markov chains, where the transition probability matrix changes in each run. Infomap finds the optimal set of modules that minimizes the information required to describe a random walk through a network. The description is in two levels, coding for nodes and modules (Rosvall et al., 2009). All clustering algorithms were used with their standard configurations.

Pathway Analysis Tools

GEA is an overlap-based method that tests if the overlap between two sets of genes is higher than would be expected by chance. Statistical significance is assessed using a modified Fisher's exact test where random overlap is modeled from random samples of pairs of gene sets. This test is a conservative variation of Fisher's exact test, where 1 is subtracted from the observed overlap, as in DAVID's (Huang et al., 2009) EASE score. This means that GEA cannot determine statistical significance of overlaps smaller than 2 nodes.

BinoX assumes that the random crosstalk between two gene sets in the network is distributed according to the binomial distribution. It therefore randomizes the network and computes a distribution of pairs of randomly drawn gene sets to estimate the parameters of a binomially distributed random crosstalk. These parameters are used to determine the expected crosstalk. BinoX can assess whether a pathway is enriched or depleted for the studied gene set. A depleted pathway means that the gene set has fewer links to the pathway than expected by chance.

NEAT and NEArender use slightly different assumptions about the distribution of random crosstalk in the network. NEAT assumes a hypergeometric distribution of crosstalk while NEArender assumes a chi-square distribution. Therefore, instead of testing the observed crosstalk between the studied gene set and a pathway of interest using a sampled random distribution, they rely on the hypergeometric and chi-square test respectively to assess statistical significance. However, both methods compute the expected crosstalk in the same way, taking into account the degree of the gene set, the pathway and the network. Both methods can compute enrichment and depletion. Since NEAT and NEArender show very similar results, we only selected one of them (NEAT) for our benchmark.

ANUBIX is a novel network-based method that computes the enrichment of a gene set for a pathway of interest based on the network crosstalk. The observed crosstalk is assessed for statistical significance using a model of the null distribution of the random crosstalk in the network. This null distribution is modeled by drawing random samples of gene sets, of the same size as the studied gene set, from the genome, calculating their crosstalk with the pathway of interest and fitting the parameters of a beta-binomial distribution for the distribution of the random crosstalk. The procedure can be applied to one or multiple pathways of interest. The statistical significance of the observed crosstalk is only assessed for enrichment, where the observed crosstalk is larger than would be expected by chance.

Null Model Modification of ANUBIX

To generate a null distribution of random crosstalk, ANUBIX samples gene sets from the genome, at random. The assumptions behind this null distribution may be weak when the gene sets under study contain genes not present in the used functional association network or have node degrees that deviate from the expected degrees when drawing random genes. To make the underlying null model more accurate we used degree-aware node sampling (McCormack et al., 2013) to construct the underlying distribution. We achieved this by first grouping all network nodes into bins, one per degree if more than 100 nodes exist for a given degree, or bins representing a range of degrees if this was needed to obtain at least 100 nodes in the bin. Sampling to produce random gene sets was done by randomly selecting nodes from bins with the same degree as the nodes in the query set.

To assess the improvement of this modification, we generated 100 random gene sets by sampling from the whole genome and another 100 random gene sets by sampling from the subset of genes present in all Chemical and Genetic interaction (CGP) gene sets in the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011). Sampling was done such that the gene frequencies in the MSigDB gene sets were preserved. The size of the gene sets was fixed to 50 genes, which was the median size of all the gene sets in MSigDB.

Functional Association Network

Network-based pathway enrichment methods require a protein interaction network. In our study we used FunCoup, which is one of the most comprehensive functional association networks of genes/proteins available. FunCoup infers functional associations between genes by integrating different types of evidence using a redundancy-weighted naïve Bayesian approach, combined with orthology transfer. FunCoup's high coverage comes from the number and variety of different evidence types used, such as: mRNA and protein co-expression, co-evolution based on phylogenetic profile similarity, Protein-Protein and domain-domain interactions, sub-cellular co-localization, co-regulation via miRNA and transcription factors, as well as genetic interaction.

For this study, we used the *Homo sapiens* FunCoup 5 network. To avoid noise, we used the default link confidence cutoff of 0.8 resulting in a network of 612,276 links and 12,890 genes.

Pathway Database

For this study we use the 313 *H. sapiens* pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (v.96.0) (Kanehisa et al., 2016).

BENCHMARKS

Pathway Recovery for Each Clustering Method

Performance of clustering algorithms may vary depending on the properties of the network they are applied to, so we constructed a simple benchmark to assess this. We generated 100 gene sets by merging different KEGG pathways that had shared links, three pathways at a time. Then we applied the different clustering methods to these gene sets to produce modules. Each module was assigned to the pathway with the highest overlap, and the Jaccard index between the sets of assigned and true pathways was computed for each method. The Jaccard index distributions of the clustering methods were compared using Kruskal-Wallis and Wilcoxon tests.

True Positive Benchmark

KEGG pathways were bisected into two parts with similar number of nodes and total node degree. The overlap between the bisected parts was emulated based on the median overlap between gene sets in the MSigDB database and KEGG pathways. KEGG pathways were ordered by size and grouped into seven bins with an equal (or as equal as possible) number of pathways in each bin. We then sampled one pathway from each bin at random and merged them into a unique gene set. To decide how many pathways to join, we performed a pathway analysis study of Chemical and Genetic interaction (CGP) MSigDB gene sets against KEGG pathways using the null model modified ANUBIX. To keep a reasonable gene set size, and to avoid merging too many pathways, we used Bonferroni correction (Abdi, 2007) and a family-wise error rate (FWER) of 5% as a cutoff. This resulted in a median number of significantly enriched pathways of seven per gene set. We therefore chose to join seven pathways for the construction of the multi-pathway gene sets. Since our sampling was constrained by the binning procedure, to avoid having too much overlap between the constructed gene sets, but still retain a statically usable number of gene sets we generated 100 gene sets and ran pathway enrichment against the other parts of the bisected pathways. Since each gene set was constructed from seven different pathways and we were aiming to recover the other half of each of those pathways, we could at most have 700 true gene set-pathway associations or True Positives (TPs).

False Positive Benchmark

For the false positive (FP) benchmark we generated 100 random gene sets of the average size of the true positive gene sets, 280 genes. The generated gene sets were tested for enrichment against the true KEGG pathways. Considering their randomness, we did not expect to find any enriched pathways.

Performance Measures

Both the true positive and false positive benchmarks were applied with and without clustering of gene sets prior to pathway analysis. When clustering was applied, pathway enrichment was tested individually for each identified module. The pathways with the lowest *p*-value for each module were merged into a single list. The performance of each method was assessed by Receiver Operator Characteristics (ROC) curves (Bradley, 1997). For our analysis, we select only the pairs that were statistically significantly (FDR < 0.05) enriched after adjusting *p*-values using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The pipeline of the clustering implementation in pathway enrichment analysis is shown in **Figure 2**.

Adaptive Module Size Filtering

Applying clustering to the query gene sets increases the sensitivity of the underlying analysis. However, this often comes with an increase in false positives, mainly stemming from small modules. To control for this, we devised a filtering approach for small modules prior to the pathway enrichment analysis. To calibrate it, we generated 100 random gene sets for a range of sizes between 50 and 600 genes, increasing the size by 50 genes, and ran the clustered pathway enrichment pipeline against KEGG pathways. At FDR < 0.05, we studied which minimum module size cutoff was necessary to keep the FPR below 5%. With the selected range of gene set sizes, we observed that the required module size cutoff increased linearly with the query gene set size (**Supplementary Figure S1**), suggesting that the cutoff should be adapted to

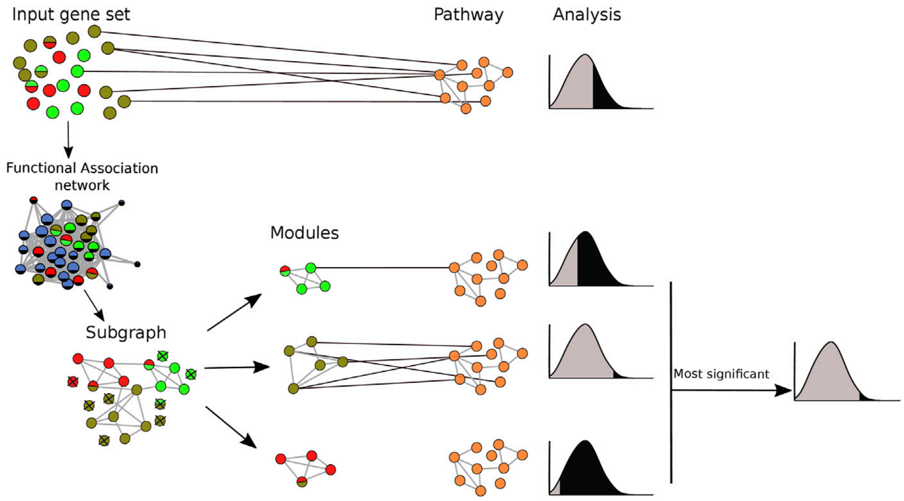


FIGURE 2 | Integrating clustering methods into pathway enrichment analysis. The input gene set was mapped onto an association network and clustering algorithms were applied. Based on this, the gene set was divided into several modules, and pathway enrichment analysis was run for each of the modules separately, keeping only the most significant result for each module-pathway pair. Pathway enrichment was also run in the original gene set for comparison purposes. Each node represents a gene. Different colors refer to membership in different modules or pathways. A multicolored circle indicates more than one membership.

different gene set sizes. This approach only works well for methods that already control the FPR well prior to clustering, here yielding good results only for ANUBIX. The adaptive module size filtering ensures an FPR level matching the set FDR level in ANUBIX when filtering out modules whose size is below 2% of the query gene set size, hence this filter was applied to ANUBIX here. For BinoX and NEAT this was however not possible to achieve without a massive loss of sensitivity, hence the filter could not be applied to them.

Clustered vs. Non-Clustered MSigDB Gene Sets Analysis

We ran pathway enrichment analysis against KEGG pathways for all the CGP MSigDB gene sets in two different scenarios, with and without pre-clustering the gene sets. To showcase that different gene sets are a mixture of different pathway or pathway families, for each MSigDB gene set, we studied how often a certain pathway subclass, as defined by KEGG, was targeted by the same gene set module. The KEGG database classifies pathways into 6 classes and 42 subclasses. The overlap in significantly enriched pathways between (A) with pre-clustering and (B) without pre-clustering was computed using the Jaccard Index as described in **Eq. 1**:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

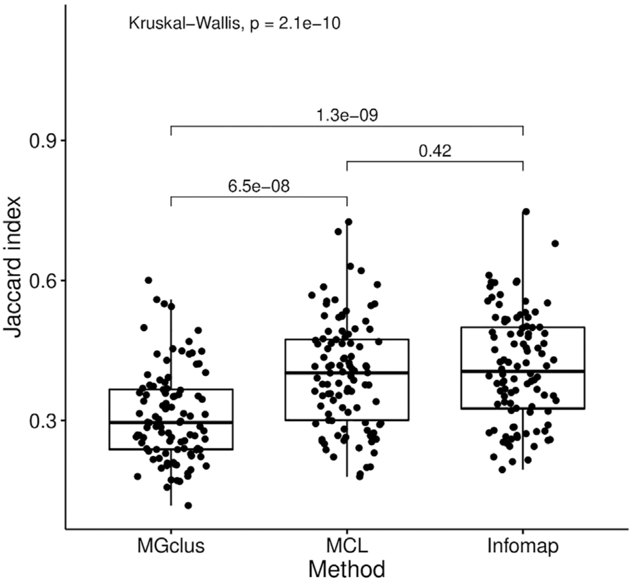


FIGURE 3 | Ability of the clustering methods Infomap, MCL, and MGclus to recover original KEGG pathways in multi-pathway gene sets. Three pathways were grouped into a single gene set prior to clustering. Each module was assigned to the pathway with the greatest overlap, and the Jaccard index of the overlap between true and assigned pathways was computed. The Jaccard index distributions of the clustering methods were compared using Kruskal-Wallis and Wilcoxon tests.

RESULTS

Gene sets derived from experiments typically represent multiple affected pathways. Therefore, mapping these gene sets onto a network such as FunCoup and applying network-based clustering algorithms to divide gene sets into more homogeneous subsets was expected to reduce noise and lead to more accurate pathway analysis. We investigated the effect of clustering on pathway

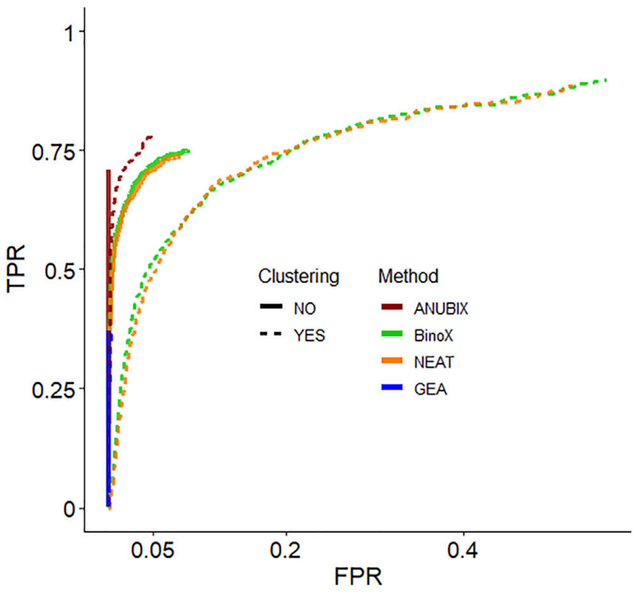


FIGURE 4 | Receiver Operating Characteristic (ROC) curves that measure the performance of each pathway analysis tool, with clustering (dotted lines) and without (solid lines). The used clustering algorithm was Infomap. Only the significantly enriched tests are shown (FDR < 0.05).

analysis using MGclus, MCL, and Infomap. To assess the clustering performance of these methods on data used in pathway analysis we applied them to gene sets constructed by joining multiple KEGG pathways. Infomap and MCL demonstrated the greatest ability to recover the original pathways with a mean Jaccard index of 41.2% for Infomap and 39.9% for MCL, followed by MGclus at 31% (**Figure 3**). The difference between Infomap and MCL was not significant (p = 0.42), however both Infomap and MCL were significantly different from MGclus, with $p = 1.3 \times 10^{-9}$ and $p = 6.5 \times 10^{-8}$, respectively.

The original null model of ANUBIX is suitable to capture non-randomness in pathways. However, it may not optimally handle biases present in the query gene set such as genes that are not in the network or genes with very high node degrees. To account for these biases and make the null model more strict we improved the random sampling step to take into account the degree distribution of the query genes. To assess the modified null model generation procedure we created two datasets of random gene sets: one by sampling from the whole genome, and another by sampling from the pool of genes present in the MSigDB CGP gene sets. For the first dataset, both the original and the null model modified ANUBIX had 0% FPR. However, for the second dataset the original ANUBIX had an FPR of 6.6%, while the FPR of the null model modified ANUBIX was only 0.2%.

We then devised a benchmark to show the effect of pre-clustering of query gene sets. The first part of the benchmark was intended to assess the ability to recover True Positive gene set-pathway pairs. Construction of the benchmark involved bisecting KEGG pathways, merging the first half of several pathways into a heterogeneous gene set and trying to detect enrichment between this gene set and the other bisected halves. In the second part of the benchmark we simulated False Positive gene set-pathway associations by generating random gene sets of the average size of the true positive gene sets. We then assessed the performance of pathway analysis methods: ANUBIX, BinoX, NEAT, and GEA, with, and without pre-clustering on this benchmark. **Figure 4** shows the results as a Receiver Operating Characteristic (ROC) curve for MCL and all pathway analysis algorithms. ROC curves when clustering by Infomap and MGclus are in **Supplementary Figure S2**. The ROC curves only show the statistically significant results at FDR < 0.05, and only for enrichment (i.e. not depletion).

Detailed True Positive Rate (TPR) and False Positive Rate (FPR) results are shown in **Table 1**. The best balanced performance prior to the application of clustering was demonstrated by ANUBIX, with a TPR of 71% and a FPR of 0%. BinoX and NEAT showed higher TPRs, of 75% and 74% respectively, but had a much higher FPR of 9% and 8%, respectively. As expected, GEA had a low TPR of only 37% due to the low coverage that overlap-based methods tend to have. However, it had a flawless specificity. A significant difference was observed between the results of ANUBIX and the other methods (McNemar´s test, p < 0.001).

When applying clustering of the gene sets prior to pathway analysis, we observed a statistically significant (McNemar´s test, p < 0.001) increase in TPR for all the network-based pathway enrichment methods ANUBIX, BinoX, and NEAT, but not for GEA, which decreased. The TPR for ANUBIX increased by at most 7 percentage points, when using Infomap, still maintaining an FPR not exceeding the requested FDR level of 5%. BinoX and NEAT exhibited higher increases in TPR of up to 14–15 percentage points. However, this increase came with a very high increase in FPR from 9% to 56–61% for BinoX and from 8% to 52–56% for NEAT. There is a significant difference between the results of the other methods and ANUBIX for all the clustering algorithms (p < 0.001).

We observed that almost all of the enrichments found without clustering were also found using pre-clustering of the query sets (**Figure 5**). For BinoX and NEAT the fraction of unique enrichments found without clustering were the lowest, below 2%, while for GEA they were the highest at 12–15%. Looking at enrichments only found by pre-clustering, these fractions were generally higher, 8–17%. We further noted that most of the associations, 99.6%, identified by GEA were also found by the network-based methods.

TABLE 1 | True positive rate (TPR) and false positive rate (FPR) for combinations of the clustering and pathway enrichment methods run at FDR = 0.05.

	ANUBIX		BinoX		NEAT		GEA	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
No clustering	0.71	0.00	0.75	0.09	0.74	0.08	0.37	0.00
MCL	0.73	0.03	0.90	0.57	0.88	0.53	0.35	0.00
MGclus	0.75	0.03	0.88	0.61	0.88	0.56	0.35	0.00
Infomap	0.78	0.05	0.90	0.56	0.88	0.52	0.36	0.00

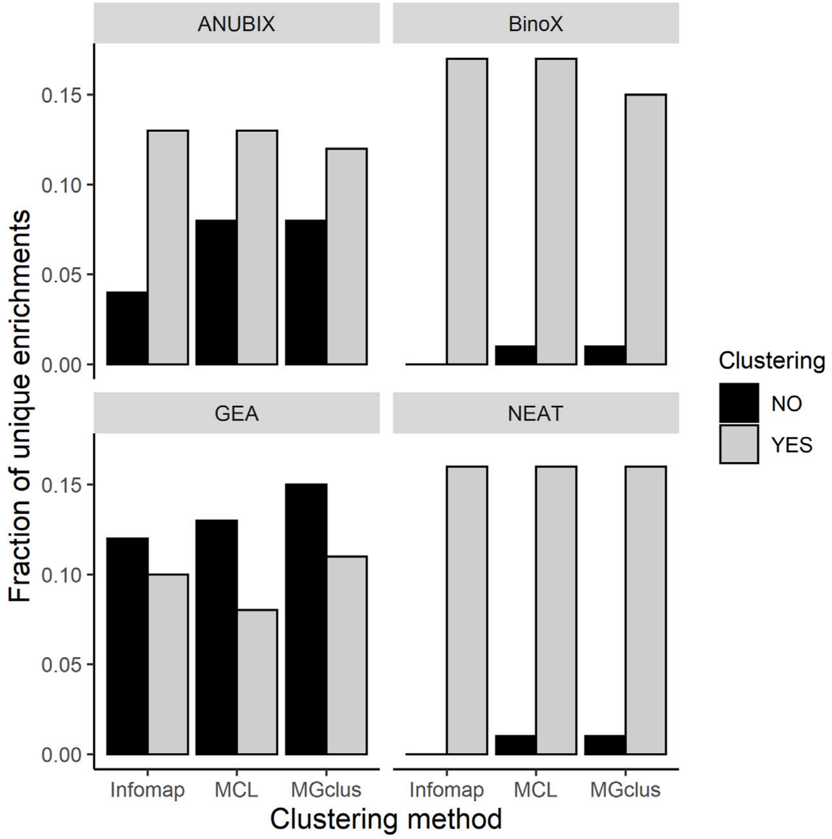


FIGURE 5 | Fractions of unique pathway enrichments found with pre-clustering relative to without pre-clustering, and vice versa, run at FDR = 0.05 for all the combinations of clustering methods and pathway enrichment tools.

Clustered Versus Non-Clustered Gene Sets Analysis

A large-scale analysis was carried out for 3302 gene sets from MSigDB/CGP against the 313 human pathways in KEGG, to observe possible benefits of applying clustering to experimental gene sets. Clustering was applied using Infomap and ANUBIX was used for the pathway enrichment analysis. Pathway enrichment analysis web server tools, such as PathBIX (Castresana-Aguirre et al., 2021) or PathwAX (Ogris et al., 2016), are implemented in a way that allows only single gene set queries. By analogy, we studied MSigDB gene sets by assuming independence between gene sets, i.e., multiple testing correction was only performed for the number of pathways each query is compared to.

Clustering of MSigDB gene sets occurred in 2703 of the 3302 gene sets. Pathway analysis without pre-clustering resulted in 129,044 significant (FDR < 0.05) crosstalks across 2,222 gene sets. Clustered analysis produced 122,819 significant crosstalks for 2,178 gene sets, of which 1,890 were shared with the non-clustering approach. The Jaccard index overlap (see Materials and Methods) of significant crosstalks between clustering and non-clustering was 52.5%, and 67.2% of the non-clustering crosstalks were found by the clustering approach while 70.6% of the clustering crosstalks were found by non-clustering.

To show that clustering helps to isolate different mechanisms within a gene set, we used the pathway subclasses as defined in the KEGG database and mapped them to the significant pathway crosstalks from the MSigDB large-scale analysis. Each pathway belongs to a KEGG subclass, and on average 95% of the significant pathways of a certain subclass had crosstalk to just one module in a gene set.

An Application of Clustered Pathway Enrichment Analysis

To illustrate the usefulness of clustering we provide an example with an MSigDB gene set, HAHTOLA_SEZARY_SYNDROM_UP (Hahtola et al., 2006). More examples can be found in **Supplementary File S1** where we provide all significant pathway enrichments found by pre-clustering using ANUBIX and Infomap but not without clustering. The selected example query set contains 99 up-regulated genes (**Supplementary Table S1**) from peripheral blood samples of Sezary syndrome patients compared to samples from healthy donors. Sezary syndrome is an aggressive form of cutaneous T-cell lymphoma (http://ghr.nlm.nih.gov/condition/sezary-syndrome) and is a rare disease driven by cancerous T-cells with one or several chromosomal

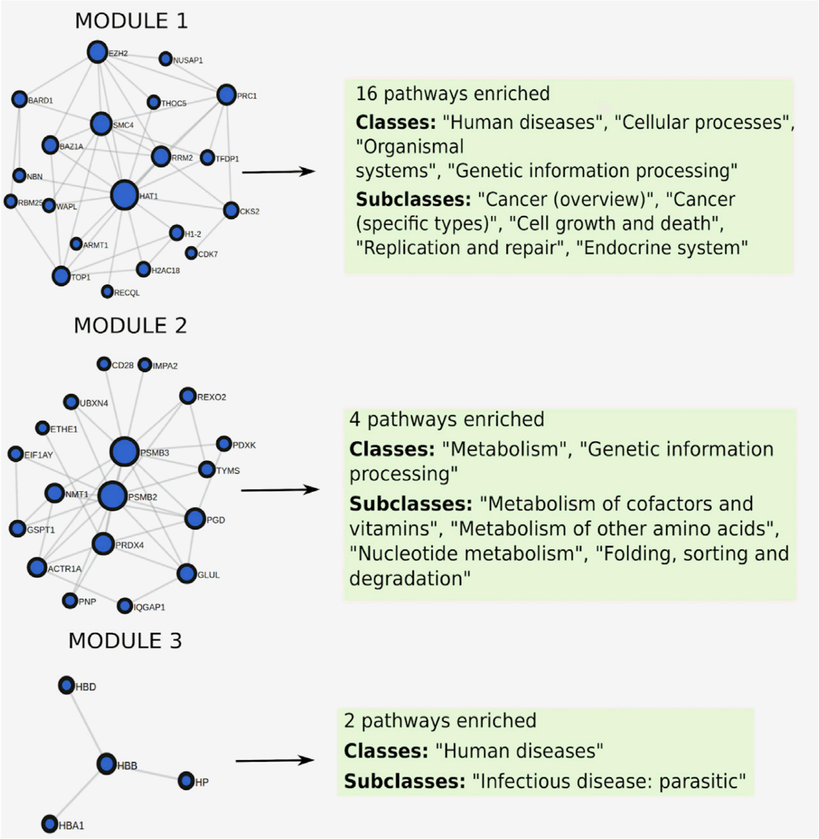

MODULE 1
EZH2
NUSAP1
BARD1
THOC5
PRC1
SMC4
BAZ1A
RRM2
TFDP1
NBN
HAT1
RBM25
WAPL
CKS2
H1-2
ARMT1
CDK7
H2AC18
TOP1
RECQL
16 pathways enriched
Classes: "Human diseases", "Cellular processes", "Organismal systems", "Genetic information processing"
Subclasses: "Cancer (overview)", "Cancer (specific types)", "Cell growth and death", "Replication and repair", "Endocrine system"
MODULE 2
CD28
IMPA2
REXO2
UBXN4
ETHE1
PDXK
EIF1AY
PSMB3
TYMS
NMT1
PSMB2
PGD
GSPT1
PRDX4
ACTR1A
GLUL
PNP
IQGAP1
4 pathways enriched
Classes: "Metabolism", "Genetic information processing"
Subclasses: "Metabolism of cofactors and vitamins", "Metabolism of other amino acids", "Nucleotide metabolism", "Folding, sorting and degradation"
MODULE 3
HBD
HBB
HP
HBA1
2 pathways enriched
Classes: "Human diseases"
Subclasses: "Infectious disease: parasitic"

FIGURE 6 | Clustered pathway enrichment analysis of the MSigDB gene set HAHTOLA_SEZARY_SYNDROM_UP. The gene set is divided into 3 modules by applying the network clustering algorithm Infomap. Each module finds different classes of pathways.

abnormalities. We used the web-server PathBIX, which provides both regular ANUBIX and clustered ANUBIX. We ran this gene set against the KEGG pathway database with a FunCoup cutoff of 0.8 and compared the results obtained from non-clustering and clustering. At FDR < 0.05, non-clustering finds 8 significantly enriched pathways, full results in **Supplementary Table S2**. The top seven pathways belonged to the KEGG classes of "Replication and repair" and "Cell growth and death", which are pathway classes affected by cancer. The eighth was the "Human T-cell leukemia virus 1 infection" pathway at FDR = 0.01. As opposed to the other seven unspecific cancer related pathways, the last one has been associated with Sezary syndrome (Pancake et al., 1995).

When clustering was applied to this gene set, it was split into three modules of size 20, 18, and 4, where each module was enriched for 16, 4, and 2 pathways respectively (**Figure 6**), full results in **Supplementary Table S3**. The first module retrieved all the enriched pathways found by the non-clustering approach, while finding additional enriched pathways belonging to the same pathway classes as the pathways found by non-clustering. Pathways relevant to cancer included "Fanconi anemia" (**Figure 7A**) at FDR = 2.8e−3, a bone marrow failure syndrome whose complications can result in leukemia (Cheung and Taniguchi, 2017), due to a failure in the repair of DNA interstrand crosslinks in the genome (Ceccaldi et al., 2016). The first module was further enriched in other cancer related pathways, such as "Transcriptional misregulation in cancer" at FDR = 1.77e−3. Furthermore, it was enriched in the "Viral carcinogenesis" pathway (FDR = 0.01). This pathway includes genes targeted by the Human T-cell leukemia virus 1 (HTL1 virus), which is thought to be the potential trigger for Sezary syndrome. This is as relevant as the HTL1 infection pathway identified by the non-clustering approach.

The second module finds pathways belonging to the metabolism class, such as "Glutathione metabolism" (**Figure 7B**) at FDR = 0.02, which is reasonable as glutathione has been proven to effectively block cell death in primary T cells from Sezary patients (Kiessling et al., 2009). Other metabolism pathways like "Purine metabolism" at FDR = 0.03, and "One carbon pool by folate" at FDR = 0.03, are reasonable as purine and folate are potential therapeutic drugs for Sezary syndrome (Oka and Miyagaki, 2019).

The third module finds pathways belonging to the class of parasitic infectious diseases, with "Malaria" at FDR = 3.79e−3 (**Figure 7C**) and "African trypanosomiasis" at FDR = 8.72e−4. Biomarkers such as miRNA are used for detecting infectious diseases. In malaria, some of the most expressed miRNAs are miR451 and miR92 (Babatunde et al., 2018), where the former is significantly correlated with diagnosis and prognosis of Sezary

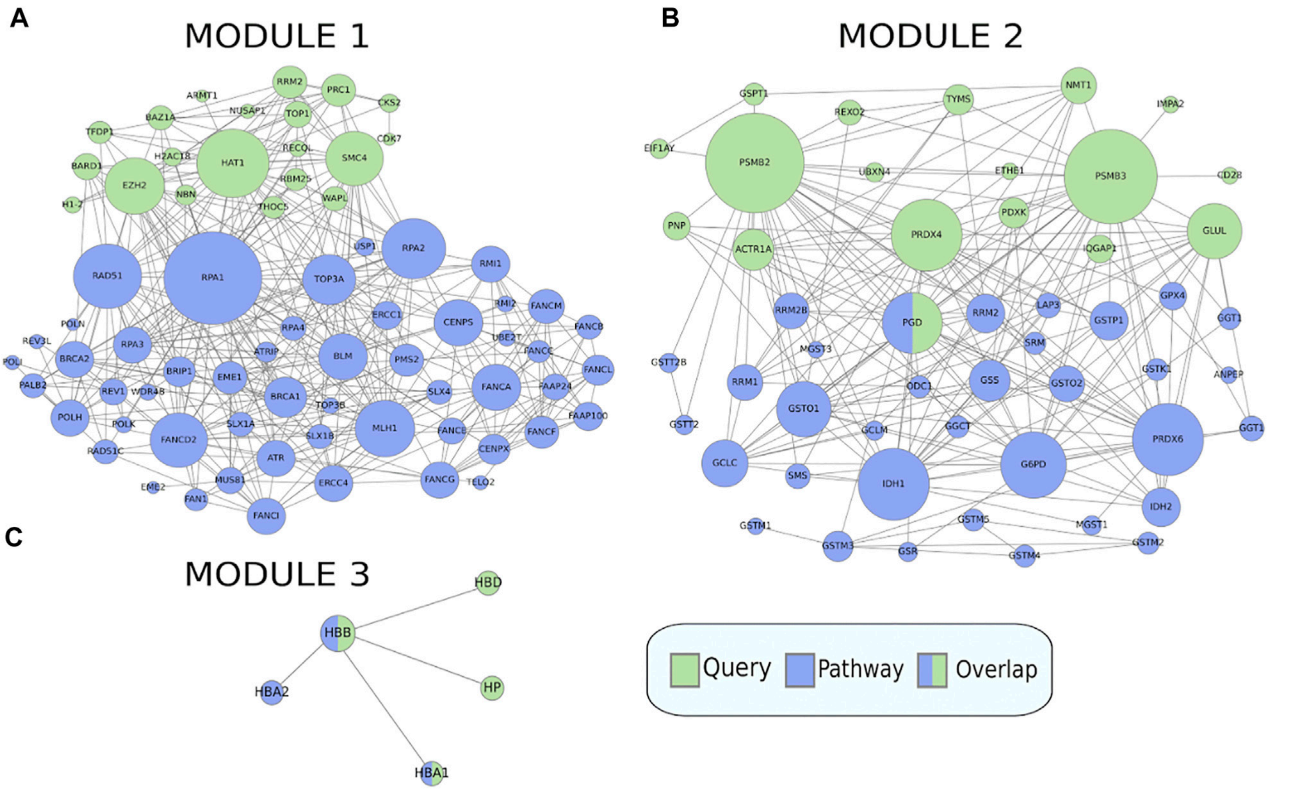


FIGURE 7 | Crosstalk between the three modules in the MSigDB gene set HAHTOLA_SEZARY_SYNDROM_UP and selected KEGG pathways (only genes linked in the network are shown). **(A)** The "Fanconi anemia" pathway which is significantly enriched for crosstalk to module 1. **(B)** The "Glutathione metabolism" pathway which is significantly enriched for crosstalk to module 2. **(C)** The "Malaria" pathway which is significantly enriched for crosstalk to module 3. The query gene set module genes are marked in green and the pathway genes are marked in blue.

syndrome, and the latter is downregulated in it (Narducci et al., 2011).

DISCUSSION

This study aimed at assessing the added benefit of pre-clustering gene sets prior to conducting pathway enrichment analysis. In order to achieve this we evaluated combinations of three network clustering methods in conjunction with one overlap-based and three network-based pathway analysis algorithms. Our findings indicate that pre-clustering increases sensitivity of pathway analysis with network-based methods but observed that it comes with the challenge of risking a high false positive rate. For two of these methods, the improvement in sensitivity came with an unacceptable loss of specificity. However, ANUBIX was able to substantially increase the sensitivity while keeping a high specificity.

The large-scale application of ANUBIX with clustering to the MSigDB gene sets against all KEGG pathways resulted in a similar number of significant enrichments as when no clustering was applied, but about a third of the enrichments were unique to each approach. We further observed that each network module within a gene set tended to be enriched by a different subclass of pathways. This supports the hypothesis that experimentally derived gene sets often represent mixtures of genes with different mechanisms, and isolating these provides a more informative analysis of the different mechanisms that are related to the condition under study. In this analysis we used Infomap for clustering as it was the best method in the benchmarks, and for the pathway enrichment analysis we used ANUBIX since it outperformed the other methods.

Before the pre-clustering analysis, we introduced a modification to the null model of ANUBIX. The new null model of ANUBIX evaluated in the study uses degree-aware sampling of genes in the network instead of randomly sampling genes from the whole genome. This null model modification resulted in a lower FPR compared to the original implementation, hence the modified version of ANUBIX was used in the rest of this study.

A previous benchmark showed that BinoX and NEAT suffer from a relatively high false positive rate (Castresana-Aguirre and Sonnhammer, 2020). To compute the crosstalk between a query gene set and a pathway, BinoX randomizes the network leading to a loss of the internal pathway structure. NEAT does not randomize the network to assess statistical significance but relies on the degrees of the query gene set, pathway, and the whole network, regardless of how that degree is distributed across the pathway. It has been demonstrated that there is a correlation between the FPs of these network-based methods and the fraction of intralinks of the pathways (Castresana-Aguirre and Sonnhammer, 2020), meaning that the less random the pathway topology is, the more prone it is to produce FPs. The distribution of crosstalk between a random gene set and a pathway often suffers from overdispersion, i.e., when the variance is larger than the mean. When this happens, the null distributions of crosstalk assumed by the different methods, binomial (BinoX) or hypergeometric (NEAT), are not appropriate. Both the overdispersion and the high false positive rate are resolved by ANUBIX. Instead of randomizing the whole network which distorts the pathway structure, ANUBIX assesses statistical significance by sampling random gene sets of the same size as the query gene set and computing an expected crosstalk distribution for each pathway. The resulting null distribution is fitted to a beta-binomial distribution, which has been demonstrated to accurately capture overdispersion (Young-Xu and Chan, 2008), and this is used to assess the significance of an observed crosstalk. Even though ANUBIX is the best performing method in that benchmark, we wanted to include other network-based methods to study if clustering could decrease their FPR. However, this issue became even more apparent when clustering was applied. We further observed that the average degree in the unclustered ANUBIX FP gene sets was 82 while the average degree of the genes in FP modules generated from those gene sets increased significantly ($p < 0.001$) to 150, 161, and 193 for Infomap, MCL and MGclus respectively. Statistical significance was assessed using a permutation test by computing the average degree for 2,000 data sets with 100 gene sets in each.

For this benchmark, we did not include quantitative pathway analysis tools, such as GSEA (Subramanian et al., 2005), CAMERA (Wu and Smyth, 2012) or SPIA (Tarca et al., 2009). In order to work, these methods require as input the differential expression of all genes. Several limitations were described previously (Subramanian et al., 2005) when selecting subsets of genes from such a list. Thus, clustering the whole set of genes into independent subsets is unlikely to be beneficial for these methods.

We have demonstrated that the application of clustering of query gene sets prior to pathway analysis improves the sensitivity of all studied pathway enrichment methods, and helps to elucidate complex mechanisms within an experimental gene set. However, pre-clustering is recommended to be used primarily with methods that can control the false positive rate well. The approach finds almost all associations found without clustering, while adding many new ones, and thus represents a powerful new tool in the quest for more accurate pathway analysis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MC-A: Conceptualization, Software, Data curation, Writing—Original Draft DG: Writing—Original Draft ES: Conceptualization, Writing—Original Draft.

FUNDING

This work was supported by the Swedish Research Council Project (2015- 05342). Open access funding provided by Stockholm University.

ACKNOWLEDGMENTS

We thank the Swedish Research Council for financial support, and Christoph Ogris and Sam de Meyer for preliminary studies of this subject.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.855766/full#supplementary-material

REFERENCES

Abdi, H. (2007). Bonferroni and Šidák Corrections for Multiple Comparisons. *Encycl. Meas. statistics* 3, 103–107.

Babatunde, K. A., Mbagwu, S., Hernández-Castañeda, M. A., Adapa, S. R., Walch, M., Filgueira, L., et al. (2018). Malaria Infected Red Blood Cells Release Small Regulatory RNAs through Extracellular Vesicles. *Sci. Rep.* 8, 884. doi:10.1038/s41598-018-19149-9

Baraldi, A., and Blonda, P. (1999). A Survey of Fuzzy Clustering Algorithms for Pattern Recognition. I. *IEEE Trans. Syst. Man. Cybern. B* 29, 778–785. doi:10.1109/3477.809032

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bradley, A. P. (1997). The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* 30, 1145–1159. doi:10.1016/s0031-3203(96)00142-2

Castresana-Aguirre, M., Persson, E, and Sonnhammer, ELL, (2021). PathBIX—a Web Server for Network-Based Pathway Annotation with Adaptive Null Models. *Bioinforma. Adv.* 1, vbab010. doi:10.1093/bioadv/vbab010

Castresana-Aguirre, M., and Sonnhammer, E. L. L. (2020). Pathway-specific Model Estimation for Improved Pathway Annotation by Network Crosstalk. *Sci. Rep.* 10, 13585. doi:10.1038/s41598-020-70239-z

Ceccaldi, R., Sarangi, P., and D'Andrea, A. D. (2016). The Fanconi Anaemia Pathway: New Players and New Functions. *Nat. Rev. Mol. Cell Biol.* 17, 337–349. doi:10.1038/nrm.2016.48

Chen, T., He, P., Tan, Y., and Xu, D. (2017). Biomarker Identification and Pathway Analysis of Preeclampsia Based on Serum Metabolomics. *Biochem. Biophysical Res. Commun.* 485, 119–125. doi:10.1016/j.bbrc.2017.02.032

Chen, X.-W., and Huang, T. (2003). Facial Expression Recognition: A Clustering-Based Approach. *Pattern Recognit. Lett.* 24, 1295–1302. doi:10.1016/s0167-8655(02)00371-9

Chen, Z., Qi, Z., Meng, F., Cui, L., and Shi, Y. (2015). Image Segmentation via Improving Clustering Algorithms with Density and Distance. *Procedia Comput. Sci.* 55, 1015–1022. doi:10.1016/j.procs.2015.07.096

Cheung, R. S., and Taniguchi, T. (2017). Recent Insights into the Molecular Basis of Fanconi Anemia: Genes, Modifiers, and Drivers. *Int. J. Hematol.* 106, 335–344. doi:10.1007/s12185-017-2283-4

Dhanachandra, N., Manglem, K., and Chanu, Y. J. (2015). Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Comput. Sci.* 54, 764–771. doi:10.1016/j.procs.2015.06.090

Emmons, S., Kobourov, S., Gallant, M., and Börner, K. (2016). Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLoS One* 11, e0159161. doi:10.1371/journal.pone.0159161

Frings, O., Alexeyenko, A., and Sonnhammer, E. L. L. (2013). MGclus: Network Clustering Employing Shared Neighbors. *Mol. Biosyst.* 9, 1670–1675. doi:10.1039/c3mb25473a

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: Network-Based Gene Set Enrichment Analysis. *Bioinformatics* 28, i451–i457. doi:10.1093/bioinformatics/bts389

Hahtola, S., Tuomela, S., Elo, L., Häkkinen, T., Karenko, L., Nedoszytko, B., et al. (2006). Th1 Response and Cytotoxicity Genes Are Down-Regulated in Cutaneous T-Cell Lymphoma. *Clin. Cancer Res.* 12, 4812–4821. doi:10.1158/1078-0432.ccr-06-0532

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks. *Bioinformatics* 18 (Suppl. 1), S233–S240. doi:10.1093/bioinformatics/18.suppl_1.s233

Jeggari, A., and Alexeyenko, A. (2017). NEArender: an R Package for Functional Interpretation of 'omics' Data via Network Enrichment Analysis. *BMC Bioinforma.* 18, 118. doi:10.1186/s12859-017-1534-y

Jhamb, D., Magid-Slav, M., Hurle, M. R., and Agarwal, P. (2019). Pathway Analysis of GWAS Loci Identifies Novel Drug Targets and Repurposing Opportunities. *Drug Discov. Today* 24, 1232–1236. doi:10.1016/j.drudis.2019.03.024

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070

Kiessling, M. K., Klemke, C. D., Kamiński, M. M., Galani, I. E., Krammer, P. H., and Gülow, K. (2009). Inhibition of Constitutively Activated Nuclear Factor-Kb Induces Reactive Oxygen Species- and Iron-dependent Cell Death in Cutaneous T-Cell Lymphoma. *Cancer Res.* 69, 2365–2374. doi:10.1158/0008-5472.can-08-3221

Lancichinetti, A., and Fortunato, S. (2009). Community Detection Algorithms: A Comparative Analysis. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 80, 056117. doi:10.1103/PhysRevE.80.056117

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi:10.1093/bioinformatics/btr260

Lu, Y., Wang, X., Dong, H., Wang, X., Yang, P., Han, L., et al. (2019). Bioinformatics Analysis of microRNA Expression between Patients with and without Latent Tuberculosis Infections. *Exp. Ther. Med.* 17, 3977–3988. doi:10.3892/etm.2019.7424

Malliaros, F. D., and Vazirgiannis, M. (2013). Clustering and Community Detection in Directed Networks: A Survey. *Phys. Rep.* 533, 95–142. doi:10.1016/j.physrep.2013.08.002

McCormack, T., Frings, O., Alexeyenko, A., and Sonnhammer, E. L. L. (2013). Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks. *PLoS One* 8, e54945. doi:10.1371/journal.pone.0054945

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative Approaches for Finding Modular Structure in Biological Networks. *Nat. Rev. Genet.* 14, 719–732. doi:10.1038/nrg3552

Narducci, M. G., Arcelli, D., Picchio, M. C., Lazzeri, C., Pagani, E., Sampogna, F., et al. (2011). MicroRNA Profiling Reveals that miR-21, miR486 and miR-214 Are Upregulated and Involved in Cell Survival in Sézary Syndrome. *Cell Death Dis.* 2, e151. doi:10.1038/cddis.2011.32

Ogris, C., Guala, D., Helleday, T., and Sonnhammer, E. L. L. (2017). A Novel Method for Crosstalk Analysis of Biological Networks: Improving Accuracy of Pathway Annotation. *Nucleic Acids Res.* 45, e8. doi:10.1093/nar/gkw849

Ogris, C., Helleday, T., and Sonnhammer, E. L. L. (2016). PathwAX: a Web Server for Network Crosstalk Based Pathway Annotation. *Nucleic Acids Res.* 44, W105–W109. doi:10.1093/nar/gkw356

Oka, T., and Miyagaki, T. (2019). Novel and Future Therapeutic Drugs for Advanced Mycosis Fungoides and Sézary Syndrome. *Front. Med.* 6, 116. doi:10.3389/fmed.2019.00116

Opresko, L. K., Gephart, J. M., and Mann, M.B. Editors (2004). *Advances in Systems Biology*. Boston, MA: Springer Science & Business Media, 547.

Pancake, B. A., Zucker-Franklin, D., and Coutavas, E. E. (1995). The Cutaneous T Cell Lymphoma, Mycosis Fungoides, Is a Human T Cell Lymphotropic Virus-Associated Disease. A Study of 50 Patients. *J. Clin. Investig.* 95, 547–554. doi:10.1172/jci117697

Persson, E., Castresana-Aguirre, M., Buzzao, D., Guala, D., and Sonnhammer, E. L. L. (2021). FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity. *J. Mol. Biol.* 433, 166835. doi:10.1016/j.jmb.2021.166835

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The Map Equation. *Eur. Phys. J. Spec. Top.* 178, 13–23. doi:10.1140/epjst/e2010-01179-1

Rosvall, M., and Bergstrom, C. T. (2008). Maps of Random Walks on Complex Networks Reveal Community Structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123. doi:10.1073/pnas.0706851105

Shemirani, R., Gillian, M B, Keith, B, Kristina, L, Christy, L A, Eimear, E K, et al. (2021)Selecting Clustering Algorithms for IBD Mapping, *bioRxiv*, 29, doi:10.1101/2021.08.11.456036

Signorelli, M., Vinciotti, V., and Wit, E. C. (2016). NEAT: an Efficient Network Enrichment Analysis Test. *BMC Bioinforma.* 17, 352. doi:10.1186/s12859-016-1203-6

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). Correction to 'The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets'. *Nucleic Acids Res.* 49, 10800. doi:10.1093/nar/gkab835

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics* 25, 75–82. doi:10.1093/bioinformatics/btn577

Ulgen, E., Ozisik, O., and Sezerman, O. U. (2019). pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data through Active Subnetworks. *Front. Genet.* 10, 858. doi:10.3389/fgene.2019.00858

Van Dongen, S. (2008). Graph Clustering via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* 30, 121–141. doi:10.1137/040608635

Wu, D., and Smyth, G. K. (2012). Camera: a Competitive Gene Set Test Accounting for Inter-gene Correlation. *Nucleic Acids Res.* 40, e133. doi:10.1093/nar/gks461

Yoon, S., Kim, J., Kim, S.-K., Baik, B., Chi, S.-M., Kim, S.-Y., et al. (2019). GScluster: Network-Weighted Gene-Set Clustering Analysis. *BMC Genomics* 20, 352. doi:10.1186/s12864-019-5738-6

Young-Xu, Y., and Chan, K. A. (2008). Pooling Overdispersed Binomial Data to Estimate Event Rate. *BMC Med. Res. Methodol.* 8, 58. doi:10.1186/1471-2288-8-58

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 16 June 2022
doi: 10.3389/fgene.2022.836856

Inference of a Boolean Network From Causal Logic Implications

Parul Maheshwari[1]*, *Sarah M. Assmann*[2] *and Reka Albert*[1,2]*

[1]Department of Physics, Penn State University, University Park, PA, United States, [2]Biology Department, Penn State University, University Park, PA, United States

Biological systems contain a large number of molecules that have diverse interactions. A fruitful path to understanding these systems is to represent them with interaction networks, and then describe flow processes in the network with a dynamic model. Boolean modeling, the simplest discrete dynamic modeling framework for biological networks, has proven its value in recapitulating experimental results and making predictions. A first step and major roadblock to the widespread use of Boolean networks in biology is the laborious network inference and construction process. Here we present a streamlined network inference method that combines the discovery of a parsimonious network structure and the identification of Boolean functions that determine the dynamics of the system. This inference method is based on a causal logic analysis method that associates a logic type (sufficient or necessary) to node-pair relationships (whether promoting or inhibitory). We use the causal logic framework to assimilate indirect information obtained from perturbation experiments and infer relationships that have not yet been documented experimentally. We apply this inference method to a well-studied process of hormone signaling in plants, the signaling underlying abscisic acid (ABA)—induced stomatal closure. Applying the causal logic inference method significantly reduces the manual work typically required for network and Boolean model construction. The inferred model agrees with the manually curated model. We also test this method by re-inferring a network representing epithelial to mesenchymal transition based on a subset of the information that was initially used to construct the model. We find that the inference method performs well for various likely scenarios of inference input information. We conclude that our method is an effective approach toward inference of biological networks and can become an efficient step in the iterative process between experiments and computations.

Keywords: Boolean network inference, Boolean model, network inference, network construction, stomatal closure, guard cell

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Franck Delaplace,
Université Paris-Saclay, France
Eugenio Martín Azpeitia,
University of Zurich, Switzerland

Correspondence:
Parul Maheshwari
parul.dkm@gmail.com
Reka Albert
rza1@psu.edu

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *16 December 2021*
Accepted: *23 May 2022*
Published: *16 June 2022*

Citation:
Maheshwari P, Assmann SM and Albert R (2022) Inference of a Boolean Network From Causal Logic Implications. Front. Genet. 13:836856. doi: 10.3389/fgene.2022.836856

1 INTRODUCTION

Network inference from expression information is an information extraction process where the inputs are knowledge of the identity of the components that make up a network and their states in a variety of contexts, and the output is a proposed regulatory network with edges and functions that define the dynamics between the biomolecules. For inference of a gene regulatory network, the input information comes from gene expression data, e.g., RNA-seq assays. Signal transduction networks can be inferred from data on protein expression and post-translational modifications, combined with information on small molecule mediators. Metabolic networks may be inferred from the knowledge of metabolite and enzyme concentrations. Various methods have been developed for network inference; each of these methods have their strengths and weaknesses.

Correlation measures (e.g., Pearson correlation coefficient) of the expression of gene or protein pairs can be used to construct a weighted gene or protein co-expression network (Huang et al., 2005). The correlation measures can also be combined with clustering methods such as hierarchical clustering or K-means clustering to obtain groups of co-expressed genes/proteins (Horvath et al., 2006). These networks show the extent of co-expression between genes/ proteins and may not be indicative of whether the gene products/proteins regulate each other or have any causal influence. Probabilistic graphical models like Bayesian networks use Bayesian inference to obtain conditional regulatory functions that indicate the probability that a target node has a certain state given the state of its regulators. This inference method often necessitates extensive data to calculate the conditional probability of the state of the target node given the state of the regulators (Sachs et al., 2005).

Network construction using edge inference from causal information (such as information from perturbation experiments) is a general method, applicable to any system, that represents an efficient alternative to network inference from state information (Albert et al., 2007; Kachalo et al., 2008, S.; Li et al., 2006). The input information is the identity of the components that make up a network and causal relationships between them, and the output is a proposed regulatory network. The causal effects used as input information include the positive or negative causal effect of one node on another (A → B), or information of the positive or negative effect of a node on the regulation of another node by a regulator (A → (B → C)). We will refer to the latter as a three-node causal effect. The inferred network incorporates each two-node causal effect as an edge or path of the corresponding sign. Experimentally documented direct interactions are always represented by edges. The inferred network incorporates each three-node causal effect as the intersection of two paths of the corresponding sign. Specifically, (A → (B → C)) will yield a positive path from B to C and a positive path from A to C, which intersect at an unknown mediator (a pseudo-vertex). Two reduction algorithms have been developed to simplify the resultant network while preserving each of the initially encoded causal relations: binary transitive reduction with critical edges, and pseudo-vertex collapse (Albert et al., 2007; Kachalo et al., 2008). The resulting network is the most parsimonious incorporation of the input information. This network synthesis method has been applied to various biological systems and resulted in equivalent networks compared to manual curation (Kachalo et al., 2008).

The Boolean modeling framework has been used successfully to model the dynamics of various types of biological networks (Wynn et al., 2012; Saadatpour and Albert, 2013; Abou-Jaoudé et al., 2016) as well as for model inference from state (e.g., gene/ protein expression or post-translational modification) data. Boolean models assume two possible states of each node, 1 (which can be interpreted as ON, active or above-threshold level) and 0 (interpreted as OFF, inactive, or below-threshold level). When a Boolean framework is used for network inference, a key pre-processing step is to discretize the data to either 0 or 1. Several methods are used for discretization of the relevant data for inference (Berestovsky and Nakhleh, 2013). One example is iterative k-means clustering where the data are iteratively clustered into fewer clusters until there are only two clusters that correspond to ON and OFF. The discretized data are then interpreted as Boolean states (e.g., activity). The inference process (described below) is performed in the same way independent of the entity whose state is described by the input data.

A traditional method to infer a regulatory network and Boolean functions from state information is to observe the time-course of the states of each node and perform an exhaustive search through all possible Boolean functions (with all subsets of nodes as possible regulators) to find the one that best fits the given data (Pandey et al., 2010; Berestovsky and Nakhleh, 2013; Dinh et al., 2017). This method is implemented in the software BoolNet (Müssel et al., 2010). This exhaustive search can be very time-consuming. Another difficulty is that it is often the case that not all of the combinations of the putative regulators' states are observed experimentally; thus, the inference is under-constrained and can be satisfied by multiple alternate set of regulators and multiple functions for each particular node.

A more effective method is to combine prior network information with state data (for example, known attractors or trajectories of the system) to infer the Boolean functions. Several methods preserve the prior knowledge network during the process of inferring the Boolean functions (La Rota et al., 2011; Ghaffarizadeh et al., 2017; Chevalier et al., 2020; Aghamiri & Delaplace, 2021). Such methods are implemented in the applications Griffin and SMBionet (Khalis et al., 2009; Munoz et al., 2018). Other methods refine the starting network by deleting or adding edges (Terfve et al., 2012; Azpeitia et al., 2013; Abou-Jaoudé et al., 2016; Dorier et al., 2016). Iterative experimental and computational analysis can then be used to further refine the Boolean network.

Here, we present a combined network and Boolean function inference method based on causal logic relationships between different network components (inferred from perturbation experiments), extending the work in Albert et al. (2007), Kachalo et al. (2008). We utilize the abundance of genetic or pharmacological perturbation (knockout and overexpression) experiments in the biological literature to infer causal logic relationships. We then infer a parsimonious network and a set of Boolean functions that recapitulates these causal relationships. Our method differs from other Boolean network inference methods in that it does not require snapshots or time courses of all the nodes' states, nor does it require a prior knowledge network. Our method covers the middle ground between curated (manual) network and model construction and automated network inference. It is closer to the former in that it aims to

TABLE 1 | Summary of the six different types of causal logic implication and their correspondence with the direct effect of the state of the regulator node (R) on the state of the target node (T). The first column lists the causal logic implication, the second column lists what that implication indicates about the definite knowledge of the state of the target node if the state of the regulator node is known, and the third column lists the corresponding Boolean rules. The "…" in the Boolean rule is a placeholder for any number of other regulators of the target node. The asterisk (*) denotes a future state of a node, i.e., T* refers to the future (or next timestep) state of the target node.

Causal logic implication	What does it mean for the state of T (independent of the state of the rest of the network)?	Equivalent Boolean rule
Sufficient	R = ON => T = ON	T* = R or …
Sufficient inhibitory	R = ON => T = OFF	T* = not R and …
Necessary	R = OFF => T = OFF	T* = R and …
Necessary inhibitory	R = OFF => T = ON	T* = not R or …

Sufficient and necessary	R = OFF => T = OFF
	R = ON => T = ON
		T* = R

Sufficient and necessary inhibitory	R = OFF => T = ON
	R = ON => T = OFF
		T* = not R

find the most parsimonious model and does not explicitly identify all the alternative models. Because of this reason, the resulting model should be verified by follow-up experiments, as all models should. This streamlined network and model inference method is aimed at making the model construction process less laborious and hence making it more accessible to the larger biological community.

2 MATERIALS AND METHODS

2.1 Background: Causal Logic Implications Between a Pair of Nodes in a Boolean Network

Causal logic, introduced in (Maheshwari and Albert, 2017), identifies causal relationships between pairs of nodes in a Boolean network as sufficient or necessary. This logic implication tells whether the sustained activity of the regulator node is sufficient or necessary to activate the target node (for a promoting edge) or deactivate the target node (for an inhibiting edge) regardless of the state of other regulators.

There are four categories of logic relationships between a regulator and its direct target: sufficient activator, sufficient inhibitor, necessary activator, and necessary inhibitor. All of these relationships are independent of the state of any other regulators. In other words, these are canalizing relationships (Kauffman, 1993). The logic relationships are summarized in **Table 1**. In the following we give two examples. If the sustained ON state of a regulator node leads to the sustained ON state of the target node, we say that the regulator is sufficient for the target. A regulator node being necessary for a target node means that the sustained OFF state of the regulator node leads to the sustained OFF state of the target node. Such necessary relationships are abundant in biology; for example, in an enzyme-catalyzed reaction both the presence of the reactant(s) and the activity of the enzyme are necessary for the production of the reaction's product.

An indirect regulator to target relationship can also have a logic implication; this relationship is mediated by a path or subgraph between the regulator and target node. For example, an indirect sufficient relationship between R and T can be mediated by a group of mediators M_i such that each M_i is necessary for the target node, the union of M_i is collectively sufficient for T, and R is sufficient for each Mi; see (Maheshwari and Albert, 2017) for a description of all the paths and subgraphs that mediate a logic implication. In these latter cases, the logic implication is independent of all other nodes in the network except for the nodes that make up the path/subgraph of the indirect regulation. An especially salient relationship is the combination of sufficient and necessary logic implication, i.e., when the state of a target node is completely determined by the state of a distant regulator node. A sufficient and necessary promoting relationship means that the state of the target node will be the same as the state of the regulator node while a sufficient and necessary inhibitory relationship means that the state of the target node will be the opposite of the state of the regulator node. More details on each of these causal logic relationships can be found in Maheshwari and Albert (2017).

2.2 Combining Causal Implications Incident on the Same Target Node

In a large and complex network, nodes can have multiple direct regulators, each of which may have a different causal logic implication on the target node. These logic implications must correspond to a single Boolean function that preserves each logic implication. Consequently, the resulting Boolean function is in the family of biologically meaningful functions (Raeymaekers, 2002) (i.e., no regulator is redundant or has an ambiguous effect), and also in the family of nested canalizing functions (Y. Li et al., 2013). Only certain combinations of logical regulators are able to preserve each logic implication. To see why this is the case, consider a hypothetical situation in which a target node (T) has a direct regulator (R1) that is sufficient. According to the definition of a sufficient regulator, the ON state of R1 always implies the ON state of T independent of the state of other regulators. In terms of Boolean functions, the existence of a sufficient direct regulator among multiple regulators implies a logic OR gate. This means that the effect of R1 is compatible with another direct regulator R2 that is also sufficient, making the update function *T* = R1 or R2.* Here T* indicates the next state of the target node T. The other

TABLE 2 | Compatibility of the causal logic implications of regulator nodes. The matrix lists the compatibility of different regulators with varying causal logic implications with ✓'s and ✗'s. The first row and the first column denote the logic implications of different regulators. A check (✓) entry denotes that the logic implications in the corresponding row title and column title are compatible while a cross (✗) entry denotes that they are incompatible.

Logic	Sufficient	Necessary	Sufficient inhibitory	Necessary inhibitory
Sufficient	✓	✗	✗	✓
Necessary	✗	✓	✓	✗
Sufficient inhibitory	✗	✓	✓	✗
Necessary inhibitory	✓	✗	✗	✓

case of compatibility is when R2 is a necessary inhibitor; in this case the function of the target is *T* = R1 or not R2.* Node T cannot have another direct regulator (R2) that is necessary, because the "necessary" classification of R2 (i.e., the OFF state of R2 implies the OFF state of T) contradicts the sufficiency of R1. In summary, sufficient regulators are incompatible with necessary regulators. Please note that this incompatibility does not mean that every regulator's effect on the target must always combine with a logic OR relation. For example, the Boolean rule for a target node can be *T* = R1 or (R2 and R3).* Here, neither of the regulators R2 or R3 are independently sufficient or independently necessary for T but they are still compatible with regulator R1.

We summarize the compatible logic implications in **Table 2** and describe them in words in the following. When a direct regulator is sufficient and necessary, it must be the only regulator of the target node. Similarly, when a direct regulator is sufficient and necessary inhibitory, it must be the only regulator of the target node. Necessary regulators are compatible with other necessary regulators and any other sufficient inhibitory regulators. Sufficient inhibitory regulators are compatible with other sufficient inhibitory regulators and any other necessary regulators. Sufficient regulators are compatible with other sufficient regulators and any necessary inhibitory regulators. Necessary inhibitory regulators are compatible with other necessary inhibitory regulators and any sufficient regulators.

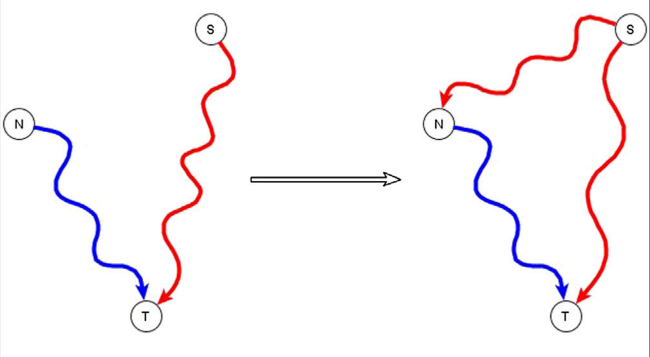


FIGURE 1 | Illustration of the co-pointing subgraph theorem for inferring logic implication between two regulators. The source node S is sufficient indirectly (*via* a path or a subgraph) for the target node T and the non-source node N is necessary indirectly for the target node T. The two subgraphs from S to T and from N to T are co-pointing subgraphs. This leads to inference of a direct or indirect causal logic implication that the signal node S is sufficient for the non-source node N.

2.3 Resolving Apparently Incompatible Implications by Inferring New Relationships

A subset of the incompatible relationships described in the previous subsection can be resolved if one or both of the apparently incompatible regulators is in reality an indirect regulator of the target node and if the two regulators are not independent of each other, but rather one of them has a logic implication on the other. This is expressed and proven in the co-pointing subgraph theorem of (Maheshwari and Albert, 2017). If a source node (S), i.e., a node with no regulators, is indirectly sufficient for a target node (T) and another node (N), which is not a source node, is directly or indirectly necessary for this target node, we say that there are two co-pointing subgraphs, one from S to T and one from N to T (Maheshwari and Albert, 2017)—see **Figure 1**. The co-pointing subgraph theorem from Maheshwari and Albert (2017) says that when there are two co-pointing subgraphs as in **Figure 1**, where source node S is sufficient and N is necessary to the target node, S must be sufficient for N. The simplest subgraph that satisfies this theorem is if the function of N is *N* = S*, and the function of the target is *T*= S and N.* Here we extend the applicability of this theorem to the situation in which S is not a source node and there is no path from N to S.

The co-pointing theorem can be used to resolve certain kinds of apparently incompatible logic implications of indirect regulators by inferring new edges. Situations like this happen often in genetic or pharmacological knockout experiments that aim to identify putative signal transduction mediators. If the experiment finds that the knockout of N disrupts the signal transduction process that initiates from signal S, we conclude that N is necessary for the target node. This might seem incompatible with the knowledge that the signal S is sufficient for the target but in fact it is consistent if N is a mediator of the pathway that establishes a sufficient relationship from S to the target node. Therefore, we infer that S is sufficient for N *via* an edge, a path, or a subgraph.

3 RESULTS

3.1 Our Proposed Method of Boolean Model Inference From Causal Logic Implications of Edges

We first give a high-level description of our inference process, then describe the details of each step in separate subsections. Boolean network inference using the causal logic method starts with a compilation of information regarding the interactions and inferred causal influences between different components of the network. When a target node has multiple regulators, we classify their effect on the target into three categories: direct relationships, indirect relationships that likely do not share mediators with any other relationships, and indirect relationships that likely share mediators. The first two categories are represented as edges in the Boolean network while the third may be implemented by paths or subgraphs. We compile the edges incident on each node into Boolean functions that best preserve their logic implications, resolving any incompatibilities. Finally, we evaluate the implementation of the mediator-sharing indirect causal logic relationships by paths and subgraphs, and if necessary add edges to reflect them, again resolving any incompatibilities in the Boolean functions. A Python implementation of this method is available in the GitHub repository https://github.com/parulm/suff_necc.

3.1.1 Distilling Biological Knowledge and the Results of Perturbation Experiments Into Logic Implications

The first step is to extract a library of information from experiments regarding the behavior of the system in normal and perturbed settings. This information is then organized as a list of causal influences and interactions, with more details indicated about each of the relationships whenever known. Each entry must include whether the interaction is promoting or inhibiting the target node, whether these relationships are direct (i.e., due to a single reaction or physical interaction) or indirect (mediated by other components) and the causal logic implication of the regulator on the target node (if known).

Certain types of biological information naturally lend themselves to causal representation. The causal effect associated with a biochemical reaction can readily be determined from the information that the presence of the reactant(s), together with the enzyme that catalyzes the reaction, leads to the production of the biomolecule that is the product of the reaction. Thus, in an enzyme-catalyzed reaction both the reactant(s) and the enzyme are necessary for the product. More generally, if an experimenter observes that the knockout of a node (regulator) leads to no (or below threshold) levels or activity of another node (target), one can conclude that the regulator node is necessary for the target. The "necessary" designation incorporates the assumption that the knockout of the source node would lead to the inactivity of the target in a different context as well. This assumption is widely made in the biological literature, as reflected by terms such as "necessary" and "required". If an experimenter observes that the sustained presence or constitutive activity of a regulator leads to high activity of another node (target), one can conclude that the regulator is sufficient for the target node. Given the fact that *in vivo* biological experiments involve multiple components in addition to the pair whose relationship is studied, the noted sufficient or necessary implication are provisional, conditioned on the presence or absence of other components (known or unknown) that define the biological context. Additional evidence that characterizes these possibly hidden components may necessitate the revision of the initial characterization.

Other types of biological information are better represented as multi-node relationships. Specifically, many biological experiments involve perturbing putative mediators and comparing an input-output relationship in the perturbed and normal systems. In these experiments, there are three essential entities, the input, the output and the mediator. This usually results in statements (three-node causal implications) of the form "A promotes (B induces C)" [see (Albert et al., 2017; S.; Li et al., 2006) for examples of such statements and how they were used during model construction]. In general, each such statement immediately leads to two derived statements. The first is that "B induces C", which usually implies that B is sufficient for C. The second statement is that "A promotes C". The causal logic implication of this statement is obtained by looking at the experiment regarding node A. The most frequently observed case is that knockout of A leads to a drastically reduced activity of C (below-threshold); in this case we conclude that A is necessary for C. The role of node A could also be inhibitory, leading to a statement of the type "A inhibits (B induces C)". The most frequently observed case is that constitutive activation of such inhibitory A leads to a below-threshold activity of C; in this case we conclude that A is a sufficient inhibitor of C.

A third regulatory relationship that can sometimes be inferred from a promoting or inhibiting three-node relationship depends on the use of the result on co-pointing subgraphs. If B affects C indirectly and there is no path from A to B, for certain types of causal logic of A on C we can infer that B regulates A according to a specific causal logic. We do this using the co-pointing subgraph theorem (Maheshwari and Albert, 2017). Given that B is sufficient for C, the co-pointing subgraph theorem applies for two causal logic implications of A on C. The first case is when A is necessary for C—we can conclude in this case that B is sufficient for A. The second case is when A is a sufficient inhibitor of C—then we can conclude that B is a sufficient inhibitor of A. However, if A is sufficient for C, no inference of any relationship between B and A can be made.

3.1.2 Assigning a Boolean Function for Each Node and Resolving Incompatibilities

We break up three-node causal implications into pairwise implications and assign a logic implication to each pairwise relationship as described above. We then consider each node of the network along with its regulators and the corresponding causal logic implications and use that information to obtain the Boolean function for the node. In the following we describe the method of determining Boolean functions in detail.

If a target node (T) has a single regulator (R), there are two general cases: the regulator is either promoting or inhibiting. If the regulator is promoting, its Boolean function would be $T^* = R$; and if the regulator is inhibiting, the Boolean function would be $T^* = not\ R$. When a target node has multiple regulators, we classify their effect on the target into three categories: direct relationships, indirect relationships that likely do not share mediators with any other relationships, and indirect relationships that likely share mediators. The third category consists of regulators of the target node that are likely connected to the target node by a path or subgraph, and this path may involve other, more direct regulators

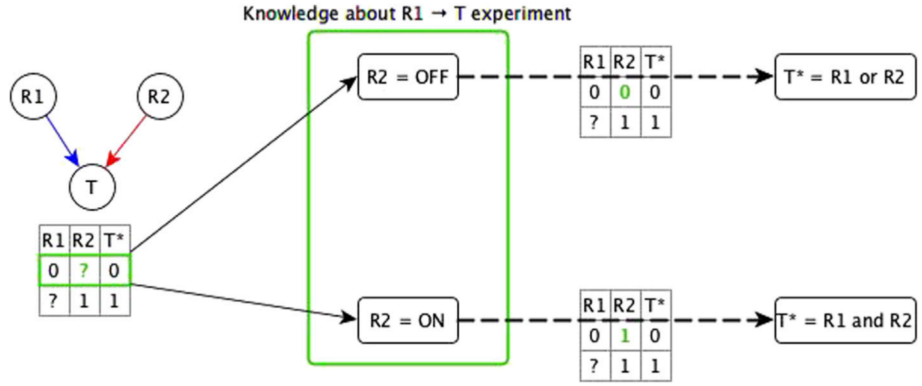


FIGURE 2 | Illustration of the use of knowledge about the experimental setting to resolve incompatibility. A target node T has two regulators; one is concluded to be necessary (R1) from experiment/literature and the other is concluded to be sufficient (R2) from experiment/literature. The truth table below the network diagram shows these relationships: R1 → T necessary relationship means that when R1 is constitutively OFF, T stabilizes to OFF; R2 → T sufficient relationship means that when R2 is constitutively ON, T stabilizes to ON. Since this is an incompatibility, we consider two assumptions about the state of R2 (denoted by green "?") in the experiment that concluded R1 to be necessary—the row marked in green in the truth table below the network. The first assumption is when R2 is OFF. In this case, there are two possible functions for T*: 1) T* = R1 and R2, and 2) T* = R1 or R2. If the rule were T* = R1 and R2, T would need to be OFF in the first row of the truth table, independent of the state of R1, hence, this leads us to conclude that the rule is T* = R1 or R2—with the corresponding truth table shown on the top dashed edge. When R2 = OFF, this rule becomes T* = R1, where R1 is indeed sufficient and necessary for T, making the "necessary" classification of R1 → T valid. The second assumption is when R2 is ON but T is OFF when R1 is OFF, which leads us to conclude that the rule is T* = R1 and R2 (see bottom dashed edge). When R2 = ON, this rule is reduced to T* = R1, where R1 is sufficient and necessary for T, making the "sufficient" conclusion about R1 → T valid. These two assumptions fall under the two templates of the dominant regulators resolution method. Due to commutativity of the Boolean operators, the resulting Boolean function would be the same whether the knowledge is about the R2 → T experiment or about the R1 → T experiment.

of the target node. For each of these regulators we need to evaluate, on a case-by-case basis, whether or not an edge from the regulator to the target is needed.

We start by considering the logic implications of direct relationships and indirect relationships that likely do not share mediators with any other relationships. In this case, there are two possibilities: all of the incoming logic implications are compatible, or the incoming edges have incompatible logic implications. In case of incompatible logic implications, we cannot directly define the Boolean function. These cases arise because the "sufficient" or "necessary" implication was premature and the target node's activity in fact depends on the specific combination of regulators. The ideal resolution for these incompatible logic implications would be to do biological experiments that test both knockout and constitutive activation of each regulator (see **Figure 2**). This solution is often impossible to execute due to technical challenges and/or the intertwined nature of biological systems. Hence, we make use of two theoretical resolution methods. One is automated while the other requires manual curation.

The automated resolution method for incompatible logic implication is the dominant regulators method; this has two templates described as follows. One of these two templates considers sufficient regulators as dominant (i.e., if any of the sufficient regulators is active then the target node activates); the other considers necessary regulators as dominant (i.e., if any of the necessary regulators is inactive then the target node inactivates). This resolution method assumes that during the experimental result that concluded the logic implication that is incompatible with the dominant logic implication the dominant regulators were in their non-canalizing state. In the following we describe each template.

The first automated template for resolving incompatibilities is to assume sufficient regulators are dominant. We impose this template by collecting all necessary regulators and marking them sufficient together, i.e., when all the necessary regulators are active, the target node will activate. Consider that target node T has regulators A, B, C, and D, where the edges A → T and B → T have sufficient logic implication while the edges C →T and D → T have necessary logic implication—see **Figure 3**. According to the first template, we group C and D together, resulting in the Boolean function *T* = A or B or (C and D)*. When we group the regulators C and D together and mark them as sufficient together, we are implicitly assuming information about the states of the other regulators, i.e., A and B, during the experiments concerning C and D. Specifically, we are assuming that A and B are OFF during the experiment involving knockout of C (or knockout of D); in this context the experiment shows that C (or D) is necessary for T, in agreement with the Boolean rule obtained by the first template. Since necessary regulators are compatible with sufficient inhibitory regulators, they can also be grouped together with sufficient inhibitory regulators. In the above example, if the edge from C to T were instead sufficient inhibitory, the resulting Boolean function would be *T* = A or B or (not C and D)*. Our code on the GitHub repository (https://github.com/parulm/suff_necc) lists the possible Boolean rules obtained by this automated method.

The second template in the automated resolution method is to give preference to the necessary logic implication and group the sufficient regulators. Going back to the example of **Figure 3**, the second template in this example (bottom right) leads to the Boolean rule *T* = (A or B) and C and D*. Since sufficient logic implication is compatible with necessary inhibitory logic

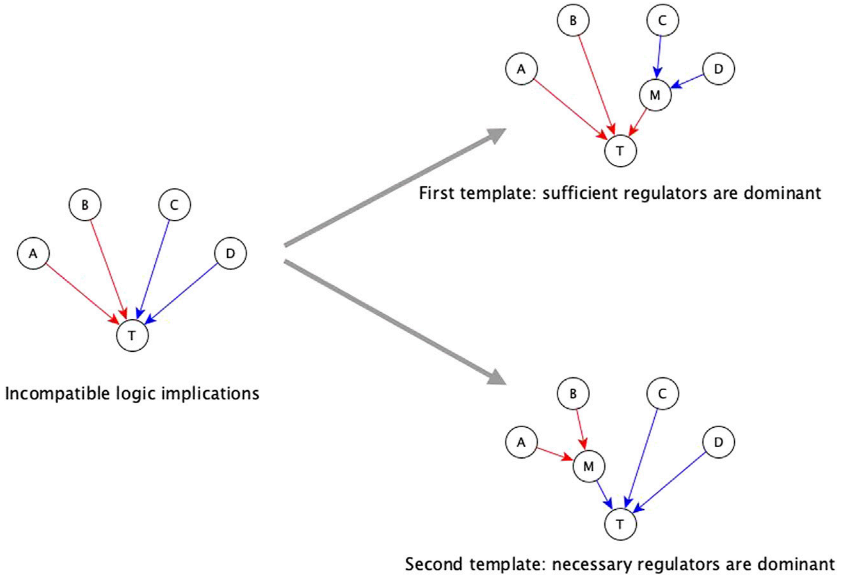

A
B
C
D
T
Incompatible logic implications
A
B
C
D
M
T
First template: sufficient regulators are dominant
A
B
C
D
M
T
Second template: necessary regulators are dominant

FIGURE 3 | Dominant regulators method for the resolution of incompatible logic implications. Red edges indicate sufficient causal logic implication while blue edges indicate necessary causal logic implication. In this example, regulators A and B are sufficient while regulators C and D are necessary for the target node T—this is an incompatible combination of logic implications. The two templates for resolving this by the dominant regulators method are shown on the right. The top-right case shows the first template where sufficient regulators are considered to be dominant, hence, the necessary regulators are grouped together, and this group is marked as sufficient. In this case, node M is a mediator node that is sufficient for T. The Boolean rule for the target node is: T* = A or B or (C and D), which is equivalent to T* = A or B or M; where M* = C and D. The bottom-right case shows the second template where necessary regulators are considered to be dominant, hence, the group of sufficient regulators is marked as necessary. In this case, M is a mediator node that is necessary for T. The Boolean rule for the target node is: T* = (A or B) and C and D which is equivalent to T* = M and C and D; where M* = A or B.

implication, sufficient regulators can be grouped with necessary inhibitory regulators as well. In the previous example, if the edge from B to T was necessary inhibitory, the Boolean rule would be *T* = (A or not B) and C and D*. These resolution templates are meant for a quick construction of the Boolean function from incomplete information.

In the manual curation resolution method, whenever we come across incompatible logic implications, we further browse the literature to find information about the states (ON/OFF) of other nodes of the network during the experiment that was used to infer the logic implication. We then use this knowledge to pick the more likely of the two possibilities detailed in the previous paragraph. Often, there are more complex possibilities for the Boolean function, which we handle on a case-by-case basis. In many of these scenarios, we construct an incomplete truth table from the literature knowledge and combine it with common biological knowledge to obtain a Boolean function.

The automated and manual resolution methods can also be applied simultaneously—we can obtain the two templates from the automated method and pick one if it satisfies the existing knowledge and the findings from the literature. The manual curation method or the two methods used simultaneously will be more thorough than just using the automated method. However, the automated method can be more useful to identify cases that need manual attention, particularly when there are many incompatibilities. Also, in scenarios where there is no additional literature information available, the automated method is something to rely on. Regardless of the method, the resulting function is one of multiple possibilities compatible with the incomplete input information. The function needs to be subjected to experimental verification followed by improvement as necessary.

3.1.3 Incorporating the Mediator-Sharing Indirect Relationships

After a draft network is constructed from the direct and indirect but independent relationships between different nodes, we look at the evidence for the remaining indirect relationships. Specifically, we look at whether such relationships are reflected by paths or subgraphs with logic implications in the network. If no relevant path or subgraph is present, we add an edge to reflect the relationship—see panels A, B, and C of **Figure 4**. In some cases, an edge directed to one of the regulators of the target would complete a path or subgraph and thus would be more appropriate, as illustrated in **Figure 4A** for node S—most of these instances are handled on a case-by-case basis. In some other cases, this edge is pointing directly to the target node—this would mean that the process behind the relationship of S and T is independent of the other edges after all—this is illustrated in **Figures 4B,C**. There are two such cases, one where the addition of such an edge is logically compatible with other regulators so we just connect the newly added regulator with the dominant Boolean operator—illustrated in **Figure 4B**. The second case is where the edge is incompatible with the other regulators—illustrated in **Figure 4C**. In this case, we make the newly added regulator the dominant regulator and update the Boolean rule accordingly. In the case where a path/subgraph

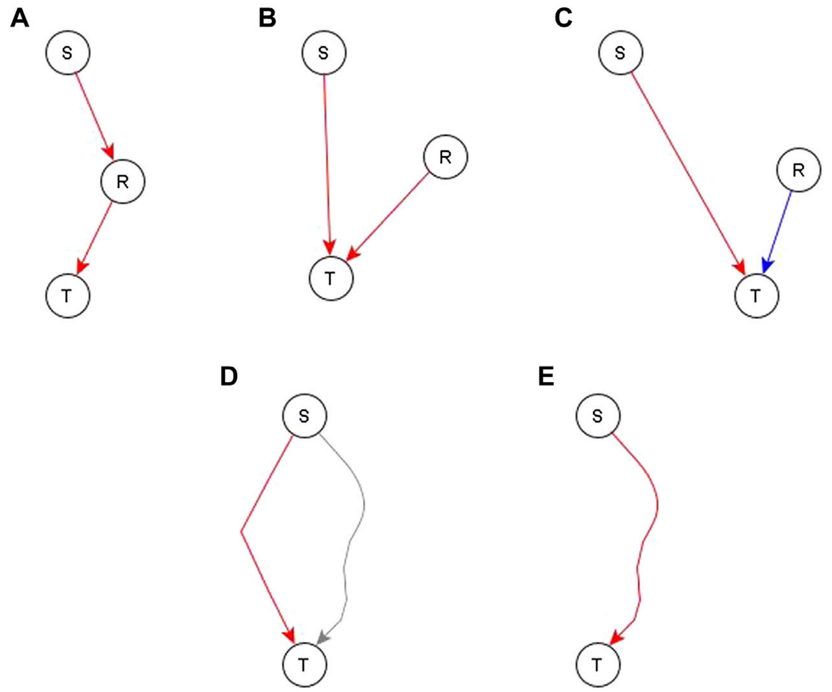


FIGURE 4 | Different ways to incorporate an indirect relationship from a regulator S to a target node T when S is sufficient for T. Panels **(A–C)** describe the case when there is no existing path or subgraph from the regulator (S) to the target node (T) and panels D and E describe the case when there is an existing path/subgraph from S to T. **(A)**. An existing regulator of T, namely R, is sufficient for T. If there is also biological support for a pathway or causal relationship from S to R, we complete a sufficient path from S to T by adding a sufficient edge from S to R. **(B)**. An existing regulator R of T is sufficient for T but there is no evidence to support a causal relationship from S to R. In this case, we construct an independent sufficient edge from S to T. **(C)**. An existing regulator R is necessary for T. We cannot be confident that R does not influence S, thus the co-pointing theorem cannot be applied. Since the causal logic relationship between S and T is "sufficient", we construct an independent sufficient edge from S to T. **(D)**. A path/subgraph exists from S to T, but its logic implication is not the same as the desired sufficient causal logic—this path is marked in gray. In this case, we add an independent sufficient edge from S to T to satisfy the "sufficient" logic. **(E)** A sufficient path/subgraph exists from S to T. In this case, the expected causal logic relationship already exists and hence we do not add any edges.

exists from the regulator to the target node, we have two cases. In the first case, the causal logic implication of the path/subgraph is not the same as the inferred causal logic implication of the new regulator. In this case, we add an edge from the newly added regulator to the target node—see **Figure 4D**. In the second case, the causal logic implication of the path/subgraph is the same as the inferred causal logic implication. In this case, we do not add a new edge—see **Figure 4E**.

Once all the experimental evidence is incorporated in the network, we use previously proposed logic reduction methods such as logic binary transitive reduction (l-BTR) to reduce this network and eliminate redundant edges (Albert et al., 2007). Logic binary transitive reduction consists of eliminating an edge from A to B if 1) it does not correspond to direct interactions and 2) there exists a path of the same sign and causal logic from A to B.

3.2 Application of the Network Inference Method to Biological Systems

3.2.1 The ABA Induced Stomatal Closure Network

We illustrate the inference process on a signal transduction network that is known to be complex and for which a significant (but still incomplete) amount of causal evidence exists. The ABA induced stomatal closure network is a plant signaling network that illustrates the process of closing of the stomatal pores on the surface of leaves induced by the plant hormone abscisic acid (ABA). ABA is produced in the plant in response to drought or other desiccating stress. This stomatal closure process involves the interconnected and interdependent activities of many ion transport proteins, enzymes and other biomolecules. This process is important to study since the stomatal pores are responsible for intake of CO_2 for photosynthesis and water loss in transpiration. In this case study, we build upon multiple previous studies on understanding this complex process by the means of a Boolean network model (Li, S. et al., 2006; Sun et al., 2014; Albert et al., 2017).

We do a careful analysis of the literature relevant to ABA induced stomatal closure. We derive pairwise relationships from three-node observations as described earlier. We find the associated causal logic corresponding to each pairwise relationship (see **Supplementary Table S1**). If the experiment reports strong qualitative results, we directly conclude the causal logic effect from there. There are two categories of such strong qualitative results. If the knockout of a gene (gene A) leads to a

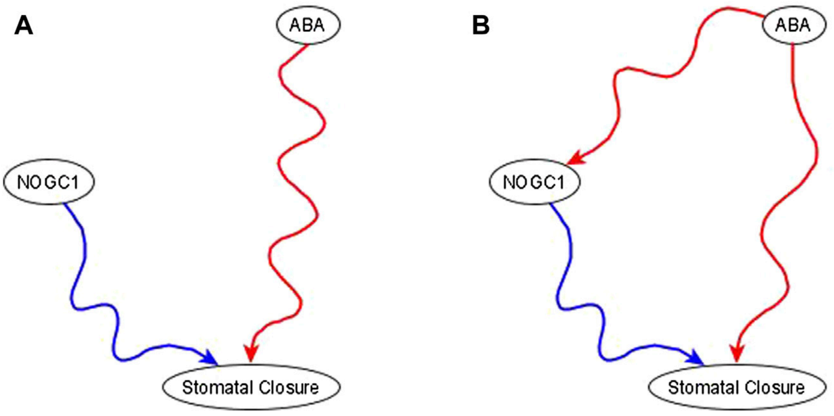


FIGURE 5 | Use of co-pointing subgraph inference method in the ABA signaling network. The signal, and source node of the network, ABA, is well-known to be sufficient for the target node stomatal closure and Nitric Oxide dependent Guanylate Cyclase (NOGC1) is a putative mediator of the signaling process. **(A)**. Experimental results show that NOGC1 knockout leads to a higher stomatal aperture, i.e., NOGC1 KO prevents the closing of the stomata, implying that NOGC1 is necessary for closure. **(B)** As per the co-pointing subgraph theorem, ABA must be sufficient for NOGC1. In a previously reported Boolean model of ABA induced closure (Albert et al., 2017), there is indeed a sufficient relationship from ABA to NOGC1.

drastic decrease in the activity of the protein product of another gene (gene B), we conclude that gene A is necessary for gene B. If there is evidence of a reaction or physical interaction between the products of gene A and B, we mark the edge as direct; otherwise it is marked "not direct". The second category is the observation that the supply of a molecule (X) leads to a drastic increase in the activity of a protein (Y), in such cases we conclude that X is sufficient for Y (directly or indirectly). In some cases, the causal logic implication has a lesser confidence (due to quantitative nuances or to expected combinatorial effects of multiple regulators), these cases are marked with an asterisk (*) in **Supplementary Table S1**. We take extra care in finding the Boolean functions in these cases as these relationships may actually be neither sufficient nor necessary. The relationship between the regulator D and target node T in the first template (top-right) of **Figure 3** is an example of such a complex relationship. In the cases where it is applicable, we also use the result on co-pointing subgraphs to infer edges, see **Supplementary Table S2**. We use the causal logic implication we find to infer the Boolean network by our method. Here, we present selected cases that exemplify the inference method.

Example of sufficient and necessary relationship. ABA activates RCARs (Park et al., 2009). RCARs is a collective node representing the PYR/PYL family of proteins, which are soluble ABA receptors that directly bind to ABA. Their strict dependence on ABA leads us to conclude a sufficient and necessary relationship from ABA to RCARs. This is further reinforced by the necessary nature of RCARs in the stomatal closure process reported in Gonzalez-Guzman et al. (2012).

Example of sufficient relationship. SPHK1 and SPHK2 are sphingosine kinases denoted together by one node as SPHK1/2. Phosphatidic acid (PA) interacts with both SPHK1 and SPHK2 and upon binding, it increases the activity of SPHK1/2. An increase in concentration of PA leads to increase in activity of SPHKs 1 and 2 as reported in Figures 4, 5 of (Guo et al., 2011). Hence, we conclude the logic implication of the edge from PA to SPHK1/2 to be sufficient.

Example of necessary relationship. Ca_c^{2+} promotes PLDα1 activity (Qin et al., 1997). Ca_c^{2+} is required for the activation of the enzyme PLDα1. The analysis in Qin et al. (1997) shows that a reduction in the Ca^{2+} concentration leads to a reduction in the PLDα1 activity—see Figures 3, 4 of (Qin et al., 1997). We conclude that Ca_c^{2+} is necessary for PLDα1 activity.

Example of the use of co-pointing subgraphs to characterize the indirect effect of ABA on nitric oxide-dependent guanylate cyclase (NOGC1). It is well-known that ABA is sufficient for stomatal closure (Joudoi et al., 2013; Albert et al., 2017). The results in Joudoi et al. (2013) show that knockout of *NOGC1* prevents stomatal closure (see Figure 2A of Joudoi et al. (2013). Hence, NOGC1 is necessary for closure. As its name indicates, NOGC1 is regulated by nitric oxide, thus it cannot be a source node. As per the co-pointing subgraph theorem (Maheshwari and Albert, 2017), this implies that ABA must be sufficient for NOGC1, which must be reflected by a sufficient path or subgraph in the resulting network (see **Figure 5**).

Example of adding an edge from an indirect regulator. When we observe an indirect regulator that already has a path to the target, we add an edge only if the path does not have the same logic implication—the case shown in **Figure 4D**. For example, the inference process is provided with the information that OST1 is sufficient for CaIM. There is already a path from OST1 to CaIM: OST1 → RBOH → ROS → GHR1 → CaIM, but the logic implication of this path is not "sufficient" and hence we add a sufficient edge from OST1 to CaIM. The resulting feed-forward loop is illustrated in **Figure 6**. As biological knowledge increases, this edge will likely be refined and populated by mediators or, refinement/correction of the existing path may render this edge unnecessary.

Example of an indirect relationship reflected by a path/subgraph. If an indirect regulator with a certain causal logic

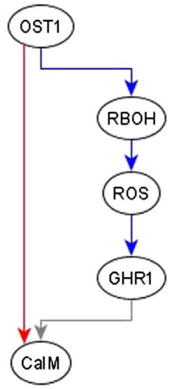


FIGURE 6 | Addition of sufficient edge from OST1 to CalM to reflect the causal logic inferred from the literature. The sufficient regulatory relationship of OST1 and CalM is not reflected in the path given by OST1 → RBOH → ROS → GHR1 → CalM. The path from OST1 to GHR1 is necessary but the edge from GHR1 to CalM is neither sufficient nor necessary since the rule for CalM is CalM* = Actin Reorganization or (NtSyp121 and GHR1 and MRP5) or not ABH1 or not ERA1 or OST1. Hence, the total path from OST1 to CalM does not have any logic implication. The sufficient edge from OST1 to CalM is hence added. Edges in red color are sufficient, in blue color are necessary, and in gray color are neither sufficient nor necessary.

implication already has a path or subgraph to the target node with the same logic implication, we do not add an edge—case shown in **Figure 4E**. This happens frequently in the ABA network. Here, we produce two examples that illustrate this. In the first case, we infer a necessary regulation of stomatal closure by SLAC1 [an ion channel that mediates anion efflux (AnionEM)] from the experimental observation that *SLAC1* knockout disrupts 8-nitro-cGMP -induced stomatal closure as shown in Figure 10C of (Joudoi et al., 2013). As shown in **Figure 7A**, there already exists a path, SLAC1 → AnionEM → H_2O Efflux → Closure, comprised of necessary edges. Hence, the logic implication is expressed by the path; we do not add the "necessary" edge from SLAC1 to Closure—shown as a dashed edge. In the second case, the input to our inference method indicates a necessary regulation of 8-nitro-cGMP by NOGC1—see **Figure 7B**. The path NOGC1 → cGMP → 8-nitro-cGMP is also necessary and hence we do not add the NOGC1 → 8-nitro-cGMP edge shown as a dashed edge.

The causal logic inference method is applied to 206 regulatory relationships collected in Supplementary Table S1 of (Albert et al., 2017), among which 107 relationships were known to be direct and 99 not known to be direct. An example application is available on the GitHub repository (see Methods). Among all these regulatory relationships, we could assign a logic implication to 196, of which 47 have a lower confidence (marked with an asterisk in **Supplementary Table S1**). We used the result on co-pointing subgraphs to infer 17 regulatory relationships, of which 8 resulted in the inference of a new edge. The remaining 9 cases corresponded to existing paths and subgraphs of the same logic implication. We then used the causal logic algorithm (Maheshwari and Albert, 2017) to look at the logic implications of the regulators for each node and construct the Boolean rules. In this process, 13 of the 62 nodes had incompatibility in the logic implications of the regulators. We used the dominant regulators method to resolve 7 of these cases. For the remaining 6 cases, we re-evaluated the causal logic implications and constructed the complete or incomplete truth table from data in the literature.

Our method, which only rarely needs manual interpretation and knowledge of the biology beyond the causal logic implication of an edge, re-discovered the Boolean rules of the ABA network correctly in 48 of 62 cases of inferring Boolean rules (see **Supplementary Text S1**). Following the second resolution method, we did an in-depth literature study for the remaining 14 cases. The methodology of this in-depth study involved constructing the incomplete truth table to find the exact rules. This methodology led to updated rules in 3 of the 14 cases, namely, PA (see **Supplementary Table S3**), AnionEM (see **Supplementary Table S4**), and OST1 (see **Supplementary Table S5**). Even after this update, only the rule for OST1 matched the previously reported rule (Albert et al., 2017). The remaining 13 discrepancies are marked in bold in **Supplementary Text S1** and are explained in detail in **Supplementary Text S2**. We believe they can be best resolved with new experimental results, leading to higher confidence in one of the possible Boolean functions.

3.2.2 The Network Corresponding to Epithelial to Mesenchymal Transition

As a second case study we consider another process whose underlying network is known to be complex: the epithelial to mesenchymal transition (EMT). Steinway et al. (2014) constructed a signal transduction network, and a Boolean model, whose outcome is the transcriptional downregulation of E-cadherin, which is a hallmark of EMT. As an additional test of our method we re-infer the Boolean model from a subset of the information that was used to construct the original model. We derive logical observations for every regulator—direct target pair from the Boolean functions of the EMT model, for a total of 127 edges (Steinway et al., 2014). We then modify this information to be more representative of characteristic use cases of the inference method. Specifically, we add indirect edges, or replace two-node paths by indirect edges, for a total of 18 changes to the input information. Some of these indirect edges correspond to a path of the same causal logic implication. Other indirect edges replace paths of the same causal implication. The logical observations used as inputs to the inference process are detailed in **Supplementary Table S6**. We use our inference method and find that it correctly resolves each modification.

1. Edges that correspond to a path of the same causal logic implication are reduced during the inference process.
   Example: The path TCF/LEF → GLI → SNAI1 is a sufficient path. So, the added TCF/LEF → SNAI1 sufficient indirect edge is redundant; it is reduced in the inference process – see **Figure 8**.

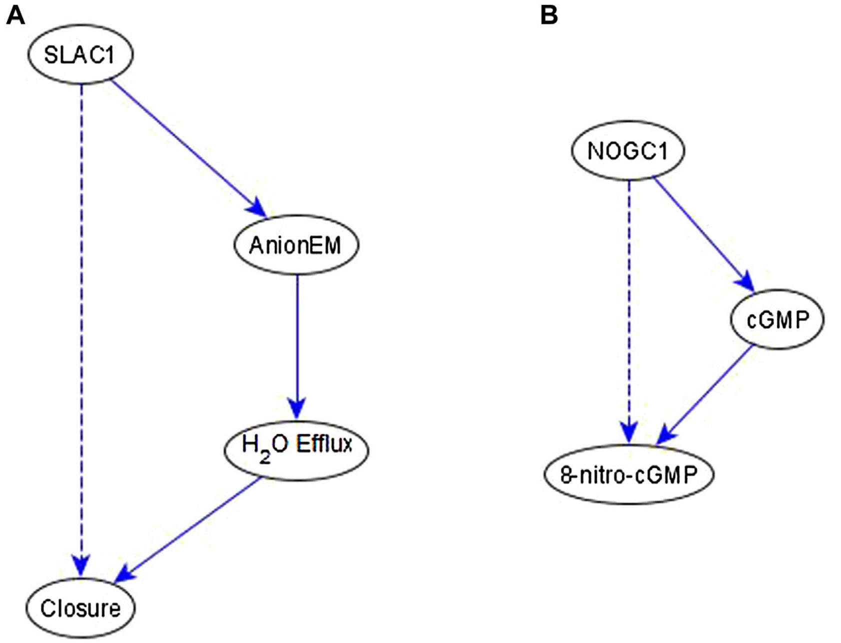


FIGURE 7 | Example of indirect regulators with the inferred causal logic implication reflected in a path. **(A)**. SLAC1 is inferred to be a necessary regulator of Closure, which is reflected by the necessary path formed by SLAC1, AnionEM, H_2O Efflux, and Closure. Hence the dashed edge from SLAC1 to Closure is not added to the network. **(B)** NOGC1 is inferred to be a necessary regulator of 8-nitro-cGMP, which is reflected by the necessary path formed by NOGC1, cGMP, and 8-nitro-cGMP. Hence the dashed edge from NOGC1 to 8-nitro-cGMP is not added to the network.

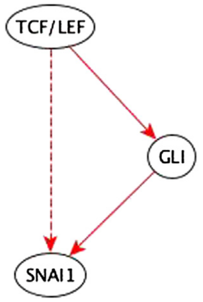


FIGURE 8 | Example of an input perturbation where an indirect regulator has the same logic implication reflected by a path formed by direct regulators. TCF/LEF is a sufficient indirect regulator of SNAI1 and a sufficient direct regulator of GLI which is a sufficient direct regulator of SNAI1. Hence the sufficient indirect edge from TCF/LEF to SNAI1 is reflected in the sufficient path TCF/LEF → GLI → SNAI1.

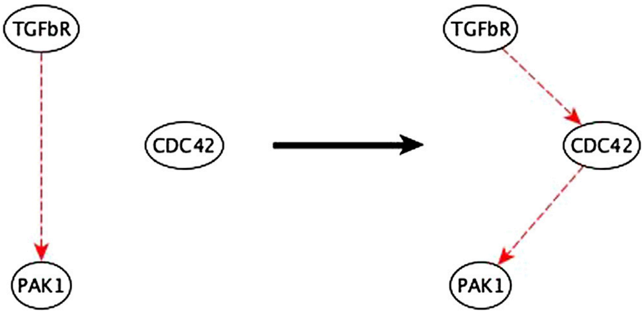


FIGURE 9 | Example of mediator node inference. The sufficient indirect regulation of PAK1 by TGFβR can be mediated by CDC42 as a sufficient path.

2. In cases where a two-node path is replaced by an indirect edge of the same logic implication, the inference method indicates potential mediator nodes, thus aiding the biologist in inferring the mediator – see **Figure 9**. We verified that all the suggested mediators were in fact actual mediators in the published EMT model.
3. A variant of the previous case is when a two-edge path between a regulator and a target is disrupted by deleting either the incoming or outgoing edge of the mediator node and is replaced by an indirect edge of the same logic implication. The inference method completes the path and infers a specific edge and logic implication for the mediator. This gives an even stronger aide for the biologist to infer the mediator compared to the previous case. An example of this case is listed in **Figure 10**.
4. The inference method used the co-pointing theorem to resolve discrepancies between incompatible indirect logic implications. An example is illustrated in **Figure 11**.

The Boolean functions resulting from the inference process are given in **Supplementary Text S3**. They are identical to the functions of the original EMT model.

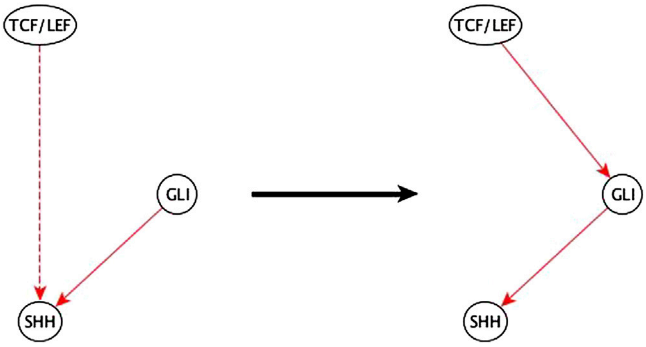


FIGURE 10 | Example of half-known mediator node inference. The inference input information reveals that TCF/LEF is indirectly sufficient for SHH and that GLI is directly sufficient for SHH. This helps the biologist infer that the indirect sufficient regulation of SHH by TCF/LEF could be *via* GLI and potentially achieved by a TCF/LEF → GLI sufficient edge.

To further test the accuracy and sensitivity of our method, we reduced the input information being provided to the code and assessed the accuracy of the resultant output. We provided ~80% of the initial input information (see **Supplementary Table S7**) and found that the method correctly infers the Boolean function of 41 of the 59 nodes in the network (see **Supplementary Text S8**), i.e., ~70% of the functions are correctly inferred.

4 DISCUSSION

In this work, we present a combined Boolean network inference method that infers the network topology and the Boolean function for each node by assigning a causal logic implication to pairs of network components based on parsimonious interpretation of the results of perturbation experiments. This method significantly reduces the manual work needed to construct a Boolean network and infer its update rules. Our method not only eases the model construction process but also points to conflicting elements of the network which can thereafter be used to guide follow-up experiments and hence improve biological understanding. In certain stages of the model construction process we have more than one option for Boolean functions, which can lead to an in-depth re-examination of the interpretation of experimental results. This often provides specific relationships to search for in the literature that might have been missed in the initial scan.

In addition to indicating the knowledge gaps that need filling, this inference method can also give hints about the direct or indirect nature of relationships. For example, if a regulator is not known to be direct and the underlying causal logic relationship is found to be fulfilled by a path or a subgraph, we have reason to believe that this relationship is indirect, and we have a list of putative mediators to consider. This was shown in the specific case of Ca_c^{2+} inhibition of PP2Cs in Maheshwari et al. (2019); the causal logic inference method makes it generally applicable.

Our application of this causal logic inference method to the well-studied ABA signaling process served as an excellent testbed for the method. A well-supported Boolean model of ABA-induced stomatal closure was reported in Albert et al. (2017), which we use to test the results of this inference method. The inferred Boolean functions of the ABA network (see **Supplementary Text S1**) particularly highlight the fact that this inference method gives a logical justification for choosing one of multiple possibilities in the face of insufficient knowledge. For example, the published and the inferred function for SLAH3 represent two different ways of resolving an incompatibility in the existing evidence; the inferred function is different from the published rule in Albert et al. (2017) on the basis of the stronger evidence of the sufficient inhibitory relationships between one of its regulators, ABI1, and SLAH3. Causal logic methodology working alongside other Boolean network analysis methods has helped us understand and improve the ABA network model (Maheshwari et al., 2019; Maheshwari et al., 2020). Despite the complexity of the network, we obtain promising results on this network using the causal logic inference method.

This method has limitations that should be addressed in future research. Any gaps or errors in the biological information used for inferring the causal logic could contribute to incorrect inferences. Furthermore, the assumption that a certain state of the regulator implies a state of the target node irrespective of the state of the other regulators does not always hold and instead the state of the target node is determined by a combination of the states of all regulators. Incompatibility between the regulators' designations is an indicator of the inappropriateness of the causal implication. We proposed methods to resolve incompatibility by weakening the assumption and replacing

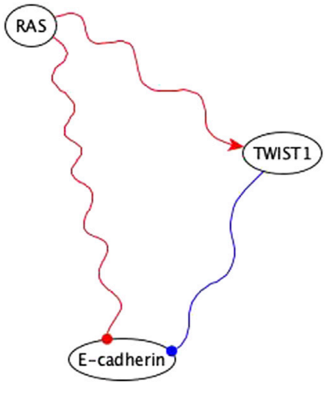


FIGURE 11 | Example of using co-pointing theorem for inference in the EMT network. RAS is sufficient inhibitory for E-cadherin *via* a subgraph and TWIST is a necessary inhibitory for E-cadherin. Sufficient inhibitory and necessary inhibitory logic implications are not compatible but since they share the same regulator, we can use the co-pointing theorem to conclude that RAS must be sufficient for TWIST1, which will result in the elimination of the edge from RAS to E-cadherin in the final version of the network.

it with multiple regulators being collectively sufficient or necessary. An undocumented regulator can also introduce incompatibility between the known regulators' designations. Developing systematic methodologies to consider undocumented regulators will be the topic of future work. In cases of observed failure of the causal logic implication the fallback is to use manual inference from the collective experimental evidence, see for example the rule for OST1 in the ABA network (see **Supplementary Table S5**).

We envision the use of this method as one step in the cycle between experiment and modeling: its use speeds up the construction of an initial parsimonious Boolean model and allows more effort to be dedicated to experimental verification of the model and to the resulting model improvement. Future applications of this method for the inference of other signal transduction or gene regulatory networks will help us further refine this theory to further decrease the manual interpretation required to obtain the Boolean functions. We also believe that one can expand the causal logic inference method to multi-level discrete networks, as have been constructed for stomatal response (Sun et al., 2014; Gan and Albert, 2016). In these networks, each biomolecule has multiple levels, for example, 0, 1, and 2, and each level is represented by an individual node that has corresponding Boolean functions for different levels of the regulator nodes. "Necessary" can be extended to mean that the lowest level of the regulator, i.e., when the regulator is inactive, implies the lowest level of the target, i.e., the inactivity of the target, and "sufficient" can be extended to mean that the highest level of the regulator implies the highest level of the target. Criteria for identification of the group of nodes that together are sufficient can be derived in various modeling frameworks, e.g., in threshold models a node may be activated if two out of its three possible activators are present.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

PM, SA, and RA designed the research and methodology and wrote the manuscript. PM and RA performed the analyses. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by NSF grant MCB-1715826 to SA and RA and NSF grant IIS-1814405 to RA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Jorge Gómez Tejeda Zañudo, Xiao Gan, and Gang Yang for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.836856/full#supplementary-material

REFERENCES

Abou-Jaoudé, W., Traynard, P., Monteiro, P. T., Saez-Rodriguez, J., Helikar, T., Thieffry, D., et al. (2016). Logical Modeling and Dynamical Analysis of Cellular Networks. *Front. Genet.* 7, 94. doi:10.3389/fgene.2016.00094

Aghamiri, S. S., and Delaplace, F. (2021). "TaBooN Boolean Network Synthesis Based on Tabu Search," in IEEE/ACM Transactions on Computational Biology and Bioinformatics (IEEE/ACM), 1. doi:10.1109/tcbb.2021.3063817

Albert, R., Acharya, B. R., Jeon, B. W., Zañudo, J. G. T., Zhu, M., Osman, K., et al. (2017). A New Discrete Dynamic Model of ABA-Induced Stomatal Closure Predicts Key Feedback Loops. *PLoS Biol.* 15 (9), e2003451. doi:10.1371/journal.pbio.2003451

Albert, R., DasGupta, B., Dondi, R., Kachalo, S., Sontag, E., Zelikovsky, A., et al. (2007). A Novel Method for Signal Transduction Network Inference from Indirect Experimental Evidence. *J. Comput. Biol.* 14 (7), 927–949. doi:10.1089/cmb.2007.0015

Azpeitia, E., Weinstein, N., Benítez, M., Mendoza, L., and Alvarez-Buylla, E. R. (2013). Finding Missing Interactions of the *Arabidopsis thaliana* Root Stem Cell Niche Gene Regulatory Network. *Front. Plant Sci.* 4, 110. doi:10.3389/fpls.2013.00110

Berestovsky, N., and Nakhleh, L. (2013). An Evaluation of Methods for Inferring Boolean Networks from Time-Series Data. *PloS One* 8 (6), e66031. doi:10.1371/journal.pone.0066031

Chevalier, S., Noël, V., Calzone, L., Zinovyev, A., and Paulevé, L. (2020). "Synthesis and Simulation of Ensembles of Boolean Networks for Cell Fate Decision," in International Conference on Computational Methods in Systems Biology (Springer), 193–209. doi:10.1007/978-3-030-60327-4_11

Dinh, J.-L., Farcot, E., and Hodgman, C. (2017). The Logic of the Floral Transition: Reverse-Engineering the Switch Controlling the Identity of Lateral Organs. *PLoS Comput. Biol.* 13 (9), e1005744. doi:10.1371/journal.pcbi.1005744

Dorier, J., Crespo, I., Niknejad, A., Liechti, R., Ebeling, M., and Xenarios, I. (2016). Boolean Regulatory Network Reconstruction Using Literature Based Knowledge with a Genetic Algorithm Optimization Method. *BMC Bioinforma.* 17 (1), 410–419. doi:10.1186/s12859-016-1287-z

Gan, X., and Albert, R. (2016). Analysis of a Dynamic Model of Guard Cell Signaling Reveals the Stability of Signal Propagation. *BMC Syst. Biol.* 10 (1), 78. doi:10.1186/s12918-016-0327-7

Ghaffarizadeh, A., Podgorski, G. J., and Flann, N. S. (2017). Applying Attractor Dynamics to Infer Gene Regulatory Interactions Involved in Cellular Differentiation. *Biosystems* 155, 29–41. doi:10.1016/j.biosystems.2016.12.004

Gonzalez-Guzman, M., Pizzio, G. A., Antoni, R., Vera-Sirera, F., Merilo, E., Bassel, G. W., et al. (2012). Arabidopsis PYR/PYL/RCAR Receptors Play a Major Role in Quantitative Regulation of Stomatal Aperture and Transcriptional Response to Abscisic Acid. *Plant Cell* 24 (6), 2483–2496. doi:10.1105/tpc.112.098574

Guo, L., Mishra, G., Taylor, K., and Wang, X. (2011). Phosphatidic Acid Binds and Stimulates Arabidopsis Sphingosine Kinases. *J. Biol. Chem.* 286 (15), 13336–13345. doi:10.1074/jbc.M110.190892

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., et al. (2006). Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Molecular Target. *Proc. Natl. Acad. Sci. U.S.A.* 103 (46), 17402–17407. doi:10.1073/pnas.0608396103

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Phys. Rev. Lett.* 94 (12), 128701. doi:10.1103/physrevlett.94.128701

Joudoi, T., Shichiri, Y., Kamizono, N., Akaike, T., Sawa, T., Yoshitake, J., et al. (2013). Nitrated Cyclic GMP Modulates Guard Cell Signaling inArabidopsis. *Plant Cell* 25 (2), 558–571. doi:10.1105/tpc.112.105049

Kachalo, S., Zhang, R., Sontag, E., Albert, R., and DasGupta, B. (2008). NET-SYNTHESIS: A Software for Synthesis, Inference and Simplification of Signal Transduction Networks. *Bioinformatics* 24 (2), 293–295. doi:10.1093/bioinformatics/btm571

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. USA: Oxford University Press.

Khalis, Z., Comet, J.-P., Richard, A., and Bernot, G. (2009). The SMBioNet Method for Discovering Models of Gene Regulatory Networks. *Genes, Genomes Genomics* 3 (1), 15–22.

La Rota, C., Chopard, J., Das, P., Paindavoine, S., Rozier, F., Farcot, E., et al. (2011). A Data-Driven Integrative Model of Sepal Primordium Polarity in Arabidopsis. *Plant Cell* 23 (12), 4318–4333. doi:10.1105/tpc.111.092619

Li, S., Assmann, S. M., and Albert, R. (2006). Predicting Essential Components of Signal Transduction Networks: A Dynamic Model of Guard Cell Abscisic Acid Signaling. *PLoS Biol.* 4 (10), e312. doi:10.1371/journal.pbio.0040312

Li, Y., Adeyeye, J. O., Murrugarra, D., Aguilar, B., and Laubenbacher, R. (2013). Boolean Nested Canalizing Functions: A Comprehensive Analysis. *Theor. Comput. Sci.* 481, 24–36. doi:10.1016/j.tcs.2013.02.020

Maheshwari, P., Du, H., Sheen, J., Assmann, S. M., and Albert, R. (2019). Model-driven Discovery of Calcium-Related Protein-Phosphatase Inhibition in Plant Guard Cell Signaling. *PLoS Comput. Biol.* 15 (10), e1007429. doi:10.1371/journal.pcbi.1007429

Maheshwari, P., and Albert, R. (2017). A Framework to Find the Logic Backbone of a Biological Network. *BMC Syst. Biol.* 11 (1), 122. doi:10.1186/s12918-017-0482-5

Maheshwari, P., Assmann, S. M., and Albert, R. (2020). A Guard Cell Abscisic Acid (ABA) Network Model that Captures the Stomatal Resting State. *Front. Physiol.* 11, 927. doi:10.3389/fphys.2020.00927

Muñoz, S., Carrillo, M., Azpeitia, E., and Rosenblueth, D. A. (2018). Griffin: A Tool for Symbolic Inference of Synchronous Boolean Molecular Networks. *Front. Genet.* 9, 39. doi:10.3389/fgene.2018.00039

Müssel, C., Hopfensitz, M., and Kestler, H. A. (2010). BoolNet—An R Package for Generation, Reconstruction and Analysis of Boolean Networks. *Bioinformatics* 26 (10), 1378–1380.

Pandey, S., Wang, R. S., Wilson, L., Li, S., Zhao, Z., Gookin, T. E., et al. (2010). Boolean Modeling of Transcriptome Data Reveals Novel Modes of Heterotrimeric G-Protein Action. *Mol. Syst. Biol.* 6 (1), 372. doi:10.1038/msb.2010.28

Park, S.-Y., Fung, P., Nishimura, N., Jensen, D. R., Fujii, H., Zhao, Y., et al. (2009). Abscisic Acid Inhibits Type 2C Protein Phosphatases via the PYR/PYL Family of START Proteins. *Science* 324 (5930), 1068–1071. doi:10.1126/science.1173041

Qin, W., Pappan, K., and Wang, X. (1997). Molecular Heterogeneity of Phospholipase D (PLD). *J. Biol. Chem.* 272 (45), 28267–28273. doi:10.1074/jbc.272.45.28267

Raeymaekers, L. (2002). Dynamics of Boolean Networks Controlled by Biologically Meaningful Functions. *J. Theor. Biol.* 218 (3), 331–341. doi:10.1006/jtbi.2002.3081

Saadatpour, A., and Albert, R. (2013). Boolean Modeling of Biological Regulatory Networks: A Methodology Tutorial. *Methods* 62 (1), 3–12. doi:10.1016/j.ymeth.2012.10.012

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science* 308 (5721), 523–529. doi:10.1126/science.1105809

Steinway, S. N., Zañudo, J. G. T., Ding, W., Rountree, C. B., Feith, D. J., Loughran, T. P., et al. (2014). Network Modeling of TGFβ Signaling in Hepatocellular Carcinoma Epithelial-To-Mesenchymal Transition Reveals Joint Sonic Hedgehog and Wnt Pathway Activation. *Cancer Res.* 74 (21), 5963–5977. doi:10.1158/0008-5472.CAN-14-0225

Sun, Z., Jin, X., Albert, R., and Assmann, S. M. (2014). Multi-Level Modeling of Light-Induced Stomatal Opening Offers New Insights into its Regulation by Drought. *PLoS Comput. Biol.* 10 (11), e1003930. doi:10.1371/journal.pcbi.1003930

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M. K., et al. (2012). CellNOptR: A Flexible Toolkit to Train Protein Signaling Networks to Data Using Multiple Logic Formalisms. *BMC Syst. Biol.* 6 (1), 133. doi:10.1186/1752-0509-6-133

Wynn, M. L., Consul, N., Merajver, S. D., and Schnell, S. (2012). Logic-Based Models in Systems Biology: A Predictive and Parameter-Free Network Analysis Method. *Integr. Biol.* 4 (11), 1323–1337. doi:10.1039/c2ib20193c

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

METHODS
published: 13 July 2022
doi: 10.3389/fgene.2022.855770

Optimal Sparsity Selection Based on an Information Criterion for Accurate Gene Regulatory Network Inference

*Deniz Seçilmiş[1], Sven Nelander[2] and Erik L. L. Sonnhammer[1]**

[1]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Solna, Sweden, [2]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

Accurate inference of gene regulatory networks (GRNs) is important to unravel unknown regulatory mechanisms and processes, which can lead to the identification of treatment targets for genetic diseases. A variety of GRN inference methods have been proposed that, under suitable data conditions, perform well in benchmarks that consider the entire spectrum of false-positives and -negatives. However, it is very challenging to predict which single network sparsity gives the most accurate GRN. Lacking criteria for sparsity selection, a simplistic solution is to pick the GRN that has a certain number of links per gene, which is guessed to be reasonable. However, this does not guarantee finding the GRN that has the correct sparsity or is the most accurate one. In this study, we provide a general approach for identifying the most accurate and sparsity-wise relevant GRN within the entire space of possible GRNs. The algorithm, called SPA, applies a "GRN information criterion" (GRNIC) that is inspired by two commonly used model selection criteria, Akaike and Bayesian Information Criterion (AIC and BIC) but adapted to GRN inference. The results show that the approach can, in most cases, find the GRN whose sparsity is close to the true sparsity and close to as accurate as possible with the given GRN inference method and data. The datasets and source code can be found at https://bitbucket.org/sonnhammergrni/spa/.

Keywords: sparsity selection, information criteria, gene regulatory network inference, gene expression data, noise in gene expression

Edited by:
Marco Pellegrini,
Italian National Research Council, Italy

Reviewed by:
Eugen Pircalabelui,
Catholic University of Louvain,
Belgium
Maria Rodriguez Martinez,
IBM Research—Zurich, Switzerland

****Correspondence:***
Erik L. L. Sonnhammer
erik.sonnhammer@dbb.su.se

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *15 January 2022*
Accepted: *30 May 2022*
Published: *13 July 2022*

Citation:
Seçilmiş D, Nelander S and Sonnhammer ELL (2022) Optimal Sparsity Selection Based on an Information Criterion for Accurate Gene Regulatory Network Inference. Front. Genet. 13:855770. doi: 10.3389/fgene.2022.855770

INTRODUCTION

Genes are responsible for orchestrating the biochemical processes in a living organism, which is only possible through a well-organized system of gene regulatory interactions called a gene regulatory network (GRN). An alteration of the system may result in complex genetic diseases, and potential treatment targets for these diseases can be identified by inferring reliable GRNs as they can reveal important mechanisms in the underlying system.

Despite the importance of an accurate GRN inference, it has been difficult to achieve due to several data-related issues such as biases and noise (Tjärnberg et al., 2015; Tjärnberg et al., 2017). The application of preprocessing approaches to noisy datasets followed by GRN inference through an accurate method has been shown on *in silico* data to overcome the noise-related obstacles in the inference to a degree (Seçilmiş et al., 2020; Seçilmiş et al., 2021), when the accuracy of the GRN inference can be measured with a known true network which is available for synthetically generated data. The accuracy is most commonly measured by the area under the precision-recall and receiver-operating characteristic curves (AUPR and AUROC, respectively) (Huynh-Thu et al., 2010; Madar et al., 2010; Marbach et al., 2012; Bellot et al., 2015), that consider the entire range of sparsities, from an empty to a full network.

For practical purposes, however, it is important to be able to infer the single best GRN, which should be as close to the true GRN as possible. In a benchmark with simulated data from a known true network, this can be assessed by accuracy measures such as the F1-score. However, in the absence of a known true network when using real biological datasets where underlying novel genetic interactions are yet to be identified as potential treatment targets, none of these measurements can be used to evaluate the accuracy of the inferred GRNs. In such situations, the selection of the best GRN is of critical importance and most often made by an arbitrary cut-off on the sparsity, which is usually ~1–3 three links per gene on average for biological reasons (Martínez-Antonio et al., 2008; Seçilmiş et al., 2020). However, this approach does not guarantee the selected GRN to represent the best and most optimal model within the space of all possibilities in terms of both accuracy and the information content. Previous attempts have been published, for instance, Tjärnberg et al. (2013) proposed a method to reconstruct the gene expression from a set of inferred GRNs whose sparsity ranges from full to empty, and showed that this approach works well for informative data but not when the noise level is high. Morgan et al. (2020) proposed a method for assessing GRN quality based on cross-validated fitting of the GRN's topology to expression data which was applied to select an optimal GRN for a biological dataset.

Methods such as LASSO (Tibshirani, 1996; Friedman et al., 2010) use a regularization approach through an internal penalty term (called the L1-norm) to obtain a sparse GRN. However, they do not offer any guidance on what value to set the L1 penalty to find the optimal sparsity. To this end, one could potentially use the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), since these approaches would, in principle, minimize the information loss with the minimum required number of independent variables across all given models. These approaches have previously been used in combination with penalty-based GRN inference approaches (Menéndez et al., 2010), such as the graphical LASSO (Friedman et al., 2008). However, it failed for AIC and is also not applicable to non-penalty-based GRN inference methods such as GENIE3 (Huynh-Thu et al., 2010).

Here, we present SPA, a sparsity selection algorithm that is inspired by the AIC and BIC in terms of introducing a penalty term to the goodness of fit, but is developed particularly for GRN inference to identify the most mathematically optimal and accurate GRN within a set of GRNs from varying sparsities inferred by any inference method. The main idea behind the algorithm is to determine the optimal model in which regulator genes are alone capable of explaining target genes with minimum information loss, given the gene expression data and its perturbation design.

METHODS

SPA: The Sparsity Selection Algorithm

SPA is a model selection pipeline that takes as input *S* GRNs with different sparsities inferred by any inference method, gene expression measured, and the perturbation design. It then assesses the quality of each input GRN *i (1, . . . , S)* based on its information content as detailed in Algorithm 1, and identifies the model minimizing GRNIC as the best GRN (**Figure 1**).

Algorithm 1.

```
-
$A_i$: the inferred GRN $i$ (1, …, S) matrix, NxN (genes x genes)
$P$: the perturbation design matrix, NxM (genes x experiments)
$Y$: the true gene expression matrix, NxM (genes x experiments)
$Y_{est}$\{i\}: the gene expression matrix predicted from $A_i$, NxM (genes x experiments)
-   For each sparsity $i$, given GRN model $A_i$
        -   Estimate predicted gene expression $Y_{est}\{i\} = -A_i^{\dagger} \times P$, where $A_i^{\dagger}$ is the Moore-
            Penrose inverse of $A_i$. To stabilize the inversion of $A_i$, and to obtain gene
            expression values in a suitable range, its diagonal elements are first set to -
            1, and via singular value decomposition, singular values below 1/max($Y$) are
            set to zero.
        -   Standard normalize $Y_{est}$\{i\} and $Y$, the measured gene expression, into $Y_{estn}$\{i\}
            and $Y_n$
        -   Calculate prediction error, i.e. badness of fit, $\varepsilon_i = \| Y_{estn}\{i\} - Y_n \|_{Fro}$
        -   Determine the penalty term, $k_i$, as the number of regulators (genes with
            outgoing links) in $A_i$
-   $E$: vector of errors across all sparsities, $\varepsilon_i \in E$
-   $K_{temp}$: vector of penalty terms across all sparsities, $k_i \in K_{temp}$
-   Calculate the normalized badness of fit vector $L$ in the interval [0,1] by
    $L_{temp}$ = e^[($E$-min($E$))/(max($E$)-min($E$))]
    $L = (L_{temp} - \min(L_{temp})) / (\max(L_{temp}) - \min(L_{temp}))$
-   Calculate the normalized penalty term vector in the interval [0,1] by
    $K = (K_{temp} - \min(K_{temp}))/(\max(K_{temp})-\min(K_{temp}))$
-   Calculate the model selection criterion (GRNIC) as a vector for all sparsities by Eq.
    1
-   Determine the best GRN $A_i$ as the one minimizing GRNIC.
```

The assessment of a GRN model's information content is made by an information criterion inspired by AIC and BIC, called GRNIC from GRN Information Criterion, and calculated according to **Eq. 1**. This criterion aims to balance the error in predicting the underlying gene expression (badness of fit) and the number of variables (regulators) in the model. Therefore, the GRN that minimizes GRNIC is expected to include a set of variables which alone are sufficient enough to reconstruct the measured gene expression, without needing more variables.

$$GRNIC = K + L. \tag{1}$$

In **Eq. 1**, *K* refers to the normalized penalty term, here set to the number of genes that regulate at least one other gene in the GRN, and *L* denotes the normalized badness of fit, here based on the prediction errors of the estimated gene expression from the GRN, calculated as described in Algorithm 1. The model for predicting the gene expression from a GRN i as $-\boldsymbol{A_i}^{\dagger} \times \boldsymbol{P}$ is derived by Tjärnberg et al. (2015). It is preceded with a conditional step of removing singular values below 1/max($\boldsymbol{Y}$) that is almost never used but is included as a safeguard against unstable inversions.

The normalization step to the terms of GRNIC was added as their units do not naturally relate to each other, making them incomparable without it. Finding the minimum GRNIC rewards low badness of fit (high goodness of fit) considering a penalty that increases with the number of variables. To implement GRNIC as close to AIC as possible, we here took *e* to the power of the badness of fit as an opposite equivalent of taking the natural logarithm of the goodness of fit.

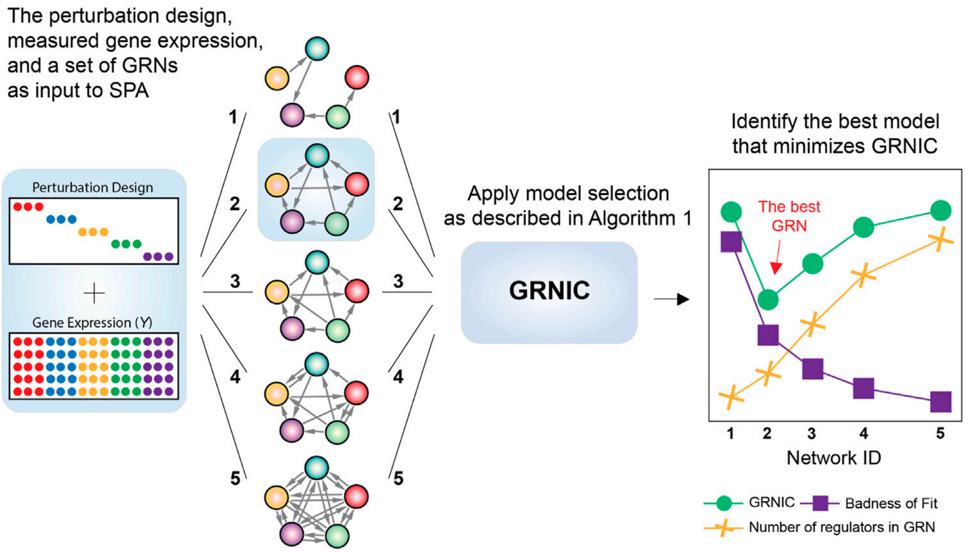

The perturbation design, measured gene expression, and a set of GRNs as input to SPA
Perturbation Design
+
Gene Expression (Y)
1
2
3
4
5
Apply model selection as described in Algorithm 1
GRNIC
Identify the best model that minimizes GRNIC
The best GRN
1 2 3 4 5
Network ID
GRNIC
Badness of Fit
Number of regulators in GRN

FIGURE 1 | Workflow of SPA. SPA takes a set of inferred GRNs with varying sparsities, the measured gene expression in fold changes, and the perturbation design as input. It then uses the GRN Information Criterion (GRNIC) as described in Algorithm 1 and identifies the GRN that minimizes GRNIC as the best GRN.

GRN Inference

We applied two different types of approaches: non-penalty-based and penalty-based methods. As the non-penalty-based method we chose GENIE3, and for the penalty-based methods, we chose LASSO and Ridge regression. The code for the used methods is available at https://bitbucket.org/sonnhammergrni/genespider.

GENIE3 was run with the following parameter settings: number of regulators: all genes; tree method: random forests; and number of trees: 1,000. Note that reverse edge direction is used because it is considerably more accurate. This resulted in a fully connected GRN with directed interactions that all have positive weights. All GRNs were then extracted whose sparsity ranged from 1–5 interactions per gene on average. For a 100-gene dataset, this corresponds to 401 different sparsities. The rationale behind this is that a biologically relevant GRN would contain ~1–3 interactions per gene on average (Martínez-Antonio et al., 2008; Marbach et al., 2012).

LASSO was run as described by Tjärnberg et al. (2015) using the glmnet Matlab package with alpha = 1, and GRNs were inferred with as many sparsities as can be obtained given the data. This was followed by extracting the GRNs whose sparsity ranges from 1–5 interactions per gene on average, following the same aforementioned biological motivation.

Ridge regression was also run using the glmnet Matlab package with alpha = 0. Different sparsities were obtained by applying cutoffs to the full GRN that Ridge regression outputs.

Data

Five 100-gene subnetworks were extracted from the complete *E. coli* GRN available in the GeneNetWeaver network and data generation tool (Schaffter et al., 2011) to be used as the "true" GRNs (in other words, the underlying regulatory system, where the topological properties of the *E. coli* network are preserved), for gene expression data generation. Autoregulatory interactions (self-loops) exist in these true GRNs for the system's stability, but none of them was used later on when measuring inference accuracy, to make it solely determined by non-self-loops. To maximize the regulatory effect in subsets, all genes were requested to be a transcription factor, yet only a fraction of all genes had regulatory effects: 0.53, 0.52, 0.53, 0.50, and 0.53 for the five true GRNs. The vertices (genes) were drawn randomly with the "greedy" edge selection (a GeneNetWeaver network extraction setting). The sparsity of the extracted true GRNs is defined as the number of interactions per gene on average, and ranges between 1.5 and 1.95 excluding self-loops. For each true subnetwork, noise-free steady-state single knockdown gene expression data were generated from ordinary differential equations (a data generation setting in GeneNetWeaver). Fold changes in gene expression following the system's perturbations were calculated as log base two of the ratio between experiment and wild-type expression. The gene expression data created by GeneNetWeaver are an NxN matrix $\boldsymbol{Y}$ (N = 100) of single replicate experiments, which places the perturbation indications as −1 on the diagonal in the experiment design matrix $\boldsymbol{P}$. Then, concatenating these matrices three times with themselves yields a three-replicate matrix of size Nx3N. We separately generated the corresponding Gaussian noise matrices with two different signal-to-noise ratios (SNRs) corresponding to "high" and "low" noise levels from **Supplementary Equation S1** and added these to the $\boldsymbol{Y}$ matrix.

RESULTS

We performed GRN inference using GENIE3, LASSO, and Ridge regression on synthetic datasets and measured the GRN inference accuracy in terms of the F1-score. Given a set of GRNs of different sparsities for each method, we applied GRNIC model selection

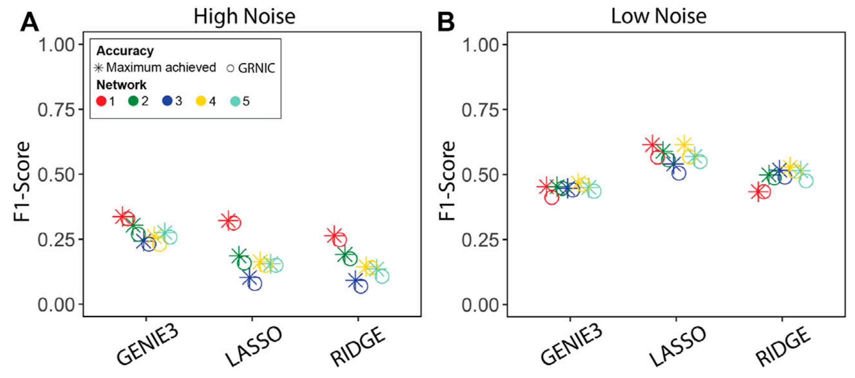


FIGURE 2 | Performance evaluation of the sparsity selection pipeline in terms of the F1-score. F1-scores of the inferred GRNs from datasets generated by GeneNetWeaver with **(A)** high and **(B)** low noise levels. Each panel contains F1-scores from five datasets for two categories: GRNIC (circle) and maximum achieved in inference (star).

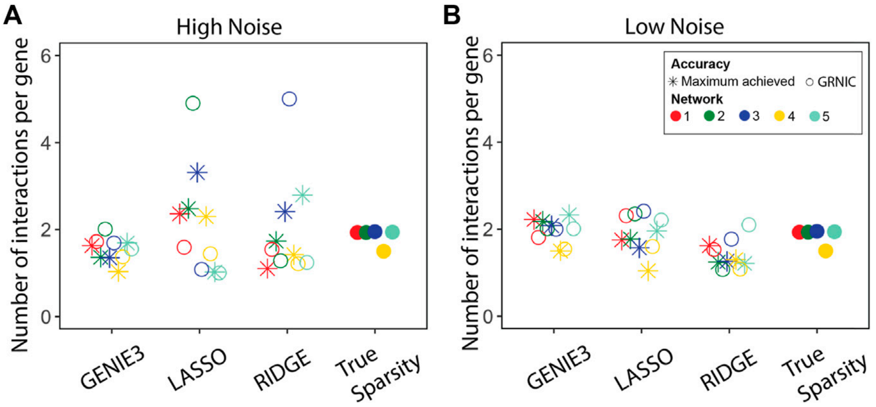


FIGURE 3 | Performance evaluation of the sparsity selection pipeline in terms of sparsity. The sparsity of the inferred GRNs (the GRN having the maximum F1-score, and the one selected by GRNIC) is shown in terms of the average number of links for **(A)** high and **(B)** low noise levels for the five networks. The sparsities of the five true GRNs are shown in an extra column to the right.

criterion as described in the Methods section (Algorithm 1; **Eq. 1**). F1-scores of the selected GRNs were then compared with the maximum F1-score obtained from the investigated range (**Figure 2**). In addition to this, we evaluated the sparsities of the selected GRNs (**Figure 3**).

The results show that the maximum accuracy of the GRN inference method can be nearly achieved by SPA using GRNIC, for all applied inference methods. In particular, we observed a noise-related trend in the accuracy in terms of the F1-score, where the GRN inference accuracy increased relative to the decreasing noise, most notably for LASSO and Ridge regression, from “high” to “low” noise. SPA was able to identify GRNs very close to the maximum accuracy GRNs for all methods in most datasets. There was a slight deviation from the maximum achieved F1-score for network 1 for GENIE3, networks 1 and 4 for LASSO, and network 5 for Ridge regression, at the low noise level. At the high noise level, SPA was able to identify GRNs whose F1-scores are almost identical to the maximum achieved for all methods.

We analyzed the two terms of the GRN information criterion (GRNIC) from **Eq. 1**, that is, the penalty term (K) and badness of fit term (L), separately, and assessed their effect on GRNIC (**Supplementary Figures S1–S5**). We observed that in most cases, the two terms of GRNIC behave as expected, where the badness of fit decreases relative to the increasing number of regulators in the model (see e.g. **Supplementary Figure S1A**). However, there are a few cases that do not behave as expected, which we investigate as follows.

Issue of GRNIC Curve Not Finding a Minimum at the True Sparsity

We observed a few cases where, even though both the penalty term (K) and the badness of fit term (L) behave as expected, the resulting GRNIC values are not minimized around the true sparsity, and instead peak here (see e.g. **Supplementary Figure S2E**). This situation can occur if the increase in the number of regulators goes faster than the decrease in the badness of fit, causing the penalty term to dominate the badness of fit when calculating GRNIC. There is no obvious solution to this issue, but either an improved function that better captures the badness of fit or an adjustment to the penalty term K could potentially resolve it.

Issue of Aberrant Badness of Fit

The badness of fit is expected to decrease with the addition of regulators, that is, going from sparser GRNs to denser ones, because of the increasing number of variables. However, a very clear parabolic curve is formed by the badness of fit values from the GENIE3 GRNs at the high noise level dataset generated from network 4 (**Supplementary Figure S4A**) and less clearly from network 3 (**Supplementary Figure S3A**). This type of behavior, however, does not prevent SPA from finding the optimal sparsity. Another concerning behavior of the badness of fit was observed for Ridge regression GRNs at the low noise level, especially for networks 3–5, where the badness of fit increases relative to the increasing number of regulators (**Supplementary Figures S2F, S4F**), which is the opposite of what is expected. We have not found any clear explanation as to why these two situations occur. In both types of aberrant behavior, the increased badness of fit with increased GRN density indicates that the larger models incorporate links that are less predictive, for instance, false positives that have a strong negative impact on the goodness of fit.

To explore how the badness of fit curves compare to what is expected by chance, we applied an experiment-wise random shuffling to the normalized gene expression matrices reconstructed by the GRN inference methods. Both badness of fit curves, the actual and shuffled, are visualized together in **Supplementary Figures. S6–S10**. The trend observed in most of the actual badness of fit curves, that is, gradually decreasing badness of fit with decreasing sparsity, was lost for the shuffled curves, which also had a very stochastic behavior. This adds further support to the validity of the applied badness of fit for the purpose of assessing the ability of a GRN to reconstruct the underlying gene expression.

Despite a few aberrant cases, the GRNs identified by SPA are almost as accurate as of the maximum achieved accuracy. To further compare the sparsities of the GRNs identified by SPA with those that achieved the maximum accuracy levels among others, we calculated the number of interactions per gene on average (**Figure 3**).

For GENIE3, SPA selected GRNs closer to the true sparsity than the other methods, and for most networks, it also came closer than the most accurate sparsity. This suggests that while GENIE3 predictions are not optimal, the criteria applied by SPA find the set of regulators that are optimal to reconstruct the underlying gene expression at both noise levels. The situation for LASSO and Ridge regression is not as promising as observed for GENIE3 at the high noise level. Some sparsities were overestimated while some others were underestimated compared with the true sparsity levels. However, the GRNs that achieved the maximum F1-scores also deviated from the true sparsity levels, suggesting that at this noise level, it is a difficult task to accurately reconstruct GRNs from the underlying data. This hypothesis is supported by the sparsity comparison at the low noise level, where both SPA GRNs and those which achieved the maximum F1-scores have similar sparsities to the true levels. In some cases, for example, for networks 3 and 5 at low noise levels, sparsities of the GRNs identified by SPA are closer to the true sparsity levels than those which achieved the maximum F1-scores. This means, for some cases, SPA is able to eliminate malefic interactions/regulators without sacrificing a significant portion of accuracy.

DISCUSSION

The ability of SPA in identifying a GRN that approaches the maximum achieved accuracy with a biologically relevant sparsity solves an old and vexing problem in the field. It can provide guidance for selecting the most optimal and accurate GRN from a set of inferred ones, which is important, for instance, when the ultimate goal is to determine novel treatment targets for underlying genetic diseases from biological data in the absence of a true GRN.

There are a few obstacles to doing this, of which the most important one is noise in gene expression. This study shows that, even though SPA identifies the most accurate GRN, its accuracy may still not be good enough for a biological discovery if the noise levels are high, referring to possibly unreliable predictions. Therefore, when using SPA, one should always note that the highest possible accuracy that SPA can achieve is only limited to the applied GRN inference method's ability to compensate for the noise.

SPA relies on the prediction accuracy of a set of input GRNs in reconstructing the underlying gene expression, given the perturbation design. Therefore, its usage is, to some extent, limited to those methods inferring signed GRNs where not only the direction but also the sign of the interaction, that is, whether activation or inhibition, is known, if one wants to ensure mathematical suitability. However, our application to GRNs inferred by GENIE3 showed that SPA can overcome this limitation and still identify an accurate GRN at a reasonable sparsity. It would be possible to further extend its application to undirected GRNs (Faith et al., 2007; de Matos Simoes and Emmert-Streib, 2012) to see to what degree SPA can find the most optimal GRN in such cases. However, we consider this problem out of the scope for this particular study since the main motivation behind identifying the most optimal GRN is to be able to understand the causality in gene regulation, and the most straightforward way of achieving this goal is to apply methods which are suitable for this purpose.

The replacement of the goodness of fit term by the prediction error in calculating the information-theoretical criteria required a few adjustments in their formulation, including a normalization step for the estimated gene expression data, and a scaling step to the badness of fit and number of variables to allow for a fair comparison between the two terms of the information criterion. This was necessary to allow for a comparison between predicted and measured gene expression since, depending on the magnitude of the GRN content and measured gene expression, the predicted gene expression can vary significantly, potentially confounding biases. A series of normalization steps on the predicted gene expression allowed both measured and predicted gene expression to be in the same range, therefore providing a more balanced comparison to the problem. The results support the methodology behind both the applied badness of fit calculation and the general formula of SPA.

The GRNIC algorithm includes a step in which the exponential of the badness of fit is calculated. This was carried out to implement GRNIC as close to AIC as possible, as an opposite equivalent of taking the natural logarithm of the goodness of fit, and also, we noticed that it improved performance. We also evaluated a number of alternative transformation functions such as square, cube, and logarithm of 1 minus the scaled badness of fit, to see which one performed the best. We concluded that even if some cases were improved with other functions, on the whole, there was no improvement and we, therefore, prefer to stay with the formulation closest to the original AIC. A potential future improvement could be to adapt the function to certain properties in the predicted and/or measured expression data. Because applying a transformation function can radically change the scale of the badness of fit, we apply normalization to ensure that it is in the same range as the penalty term.

In addition to the overall accuracy of SPA in identifying the most accurate GRNs near the true sparsity levels, we also focused on a few aberrant cases, some of which were possible to explain in terms of the negative effect of high noise levels, while some other questions raised by SPA remained unanswered. These may be answered by other researchers in the field together with what is present in this study, possibly inspiring an even better solution to the model selection problem in a larger context.

In conclusion, the implemented sparsity selection approach introduces a great advance to the field since achieving the highest possible accuracy is now made possible with the combination of a GRN inference method and SPA. We foresee that more novel gene regulatory interactions will be identified from the best possible GRNs using our algorithm, and potential treatment targets will be proposed.

DATA AVAILABILITY STATEMENT

The datasets and source code can be found at https://bitbucket.org/sonnhammergrni/spa/.

AUTHOR CONTRIBUTIONS

DS implemented the algorithm, performed the analyses, and wrote the manuscript; SN provided funding and support; and ES supervised the study and revised the manuscript.

FUNDING

This work was supported by the Swedish Foundation for Strategic Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.855770/full#supplementary-material

REFERENCES

Bellot, P., Olsen, C., Salembier, P., Oliveras-Vergés, A., and Meyer, P. E. (2015). NetBenchmark: a Bioconductor Package for Reproducible Benchmarks of Gene Regulatory Network Inference. *BMC Bioinforma.* 16, 312. doi:10.1186/s12859-015-0728-4

de Matos Simoes, R., and Emmert-Streib, F. (2012). Bagging Statistical Network Inference from Large-Scale Gene Expression Data. *PLoS One* 7, e33624. doi:10.1371/journal.pone.0033624

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol.* 5, e8. doi:10.1371/journal.pbio.0050008

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics* 9, 432–441. doi:10.1093/biostatistics/kxm045

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* 5, e12776. doi:10.1371/journal.pone.0012776

Madar, A., Greenfield, A., Vanden-Eijnden, E., and Bonneau, R. (2010). DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator. *PLoS One* 5, e9803. doi:10.1371/journal.pone.0009803

Marbach, D., Costello, J. C., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., et al. (2012). Wisdom of Crowds for Robust Gene Network Inference. *Nat. Methods* 9, 796–804. doi:10.1038/nmeth.2016

Martínez-Antonio, A., Janga, S. C., and Thieffry, D. (2008). Functional Organisation of *Escherichia coli* Transcriptional Regulatory Network. *J. Mol. Biol.* 381, 238–247. doi:10.1016/j.jmb.2008.05.054

Menéndez, P., Kourmpetis, Y. A., ter Braak, C. J., and van Eeuwijk, F. A. (2010). Gene Regulatory Networks from Multifactorial Perturbations Using Graphical Lasso: Application to the DREAM4 Challenge. *PLoS One* 5, e14147. doi:10.1371/journal.pone.0014147

Morgan, D., Studham, M., Tjärnberg, A., Weishaupt, H., Swartling, F. J., Nordling, T. E. M., et al. (2020). Perturbation-based Gene Regulatory Network Inference to Unravel Oncogenic Mechanisms. *Sci. Rep.* 10, 14149. doi:10.1038/s41598-020-70941-y

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: In Silico Benchmark Generation and Performance Profiling of Network Inference Methods. *Bioinformatics* 27, 2263–2270. doi:10.1093/bioinformatics/btr373

Seçilmiş, D., Hillerton, T., Morgan, D., Tjärnberg, A., Nelander, S., Nordling, T. E. M., et al. (2020). Uncovering Cancer Gene Regulation by Accurate Regulatory Network Inference from Uninformative Data. *NPJ Syst. Biol. Appl.* 6, 37. doi:10.1038/s41540-020-00154-6

Seçilmiş, D., Hillerton, T., Nelander, S., and Sonnhammer, E. L. L. (2021). Inferring the Experimental Design for Accurate Gene Regulatory Network Inference. *Bioinformatics* 37, 3553–3559. doi:10.1093/bioinformatics/btab367

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tjärnberg, A., Morgan, D. C., Studham, M., Nordling, T. E. M., and Sonnhammer, E. L. L. (2017). GeneSPIDER - Gene Regulatory Network Inference

Benchmarking with Controlled Network and Data Properties. *Mol. Biosyst.* 13, 1304–1312. doi:10.1039/c7mb00058h

Tjärnberg, A., Nordling, T. E., Studham, M., Nelander, S., and Sonnhammer, E. L. (2015). Avoiding Pitfalls in L1-Regularised Inference of Gene Networks. *Mol. Biosyst.* 11, 287–296. doi:10.1039/c4mb00419a

Tjärnberg, A., Nordling, T. E., Studham, M., and Sonnhammer, E. L. (2013). Optimal Sparsity Criteria for Network Inference. *J. Comput. Biol.* 20, 398–408. doi:10.1089/cmb.2012.0268

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 11 August 2022
DOI 10.3389/fgene.2022.886649

OPEN ACCESS

EDITED BY
Marco Pellegrini,
Italian National Research Council, Italy

REVIEWED BY
Kedar Sharma,
National Institute of Environmental Health Sciences (NIH), United States
Lorenzo Di Rienzo,
Italian Institute of Technology (IIT), Italy

*CORRESPONDENCE
Roded Sharan,
roded@tauex.tau.ac.il

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 28 February 2022
ACCEPTED 04 July 2022
PUBLISHED 11 August 2022

CITATION
Gross B and Sharan R (2022), Multi-task learning for predicting SARS-CoV-2 antibody escape.
Front. Genet. 13:886649.
doi: 10.3389/fgene.2022.886649

Multi-task learning for predicting SARS-CoV-2 antibody escape

Barak Gross and Roded Sharan*

School of Computer Science, Tel Aviv University, Tel Aviv, Israel

The coronavirus pandemic has revolutionized our world, with vaccination proving to be a key tool in fighting the disease. However, a major threat to this line of attack are variants that can evade the vaccine. Thus, a fundamental problem of growing importance is the identification of mutations of concern with high escape probability. In this paper we develop a computational framework that harnesses systematic mutation screens in the receptor binding domain of the viral Spike protein for escape prediction. The framework analyzes data on escape from multiple antibodies simultaneously, creating a latent representation of mutations that is shown to be effective in predicting escape and binding properties of the virus. We use this representation to validate the escape potential of current SARS-CoV-2 variants.

KEYWORDS

multi-task learning, neural network, escape prediction, coronavirus, receptor binding domain

1 Introduction

Since 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), accounted for more than 500 million infections and more than six million deaths worldwide according to World Health Organization (WHO, 2022). Though the virus mutates rapidly, only a small minority of mutations are expected to impact the virus phenotype and increase its fitness advantage. Such mutations might alter properties of the virus such as: pathogenicity, infectivity, transmissibility and/or antigenicity. Due to the virus' high infectivity and rapid mutability, in early stages of the pandemic such mutations of concern started to appear. For example, D614G was noted to be increasing in frequency in April 2020 and to have emerged independently several times in the global SARS-CoV-2 population. Subsequent studies indicated that D614G confers a moderate advantage for infectivity (Hou et al., 2020; Yurkovetskiy et al., 2020) and transmissibility (Volz et al., 2022).

Although several antibodies and vaccines showed good clinical results, recognizing mutations that impact the escape from antibodies and vaccines is still a major question in the battle against SARS-CoV2. The receptor binding domain (RBD) region is a sub-region of the SARS-CoV-2 spike glycoprotein that mediates viral attachment to ACE2 receptors. The RBD is a major determinant of host range and a dominant target of neutralizing antibodies, promoting systematic studies of mutations to the RBD region and their impact on a variety of attributes including binding (Starr et al., 2020), antibody escape (Starr et al., 2021a; Greaney et al., 2021b; Starr et al., 2021b) and more. There are fewer studies that consider multiple mutations (Li et al., 2020; Barton et al., 2021), covering only a handful of them due to the infeasible number of experiments needed to iterate over all possibilities.

The on-going emergence of variants with dozens of mutations motivates computational approaches to study the effect of multiple mutations. These include structural-based approaches (Rienzo et al., 2021; Bozdaganyan et al., 2022) or approaches that directly use deep mutational scans. An example is the escape calculator (Greaney et al., 2021a) which aggregates data about antibody escape using an interpolation based approach, thus allowing for a quantitative scoring of the antigenic effects of arbitrary combinations of mutations. Deep learning methods (Goodfellow et al., 2016) have become the method of choice for diverse data science applications including the analysis of coronavirus data. Specifically, Hie et al. (2021) applied a masked language model approach to a data set of more than 1 million SARS-CoV-2 sequences. Using the language model they ranked mutations based on semantic change (distance between wildtype and mutated sequence) and grammaticality (probability for mutation under the resulting model), thus aiding in identification of mutations that evade the immune system. But they did not address any antibodies or vaccines in their work.

In this paper, we try to combine the best of both worlds–aggregating escape data based on experimental data a la (Greaney et al., 2021a), while using deep learning methods, like in (Hie et al., 2021)—to tackle the challenge of predicting antibody escape potential. Our approach uses the paradigm of multi-task learning, where multiple learning tasks are solved at the same time in order to exploit commonalities and differences across tasks. We show that using a multi-task approach to learn escape data endows us with a representation that can be useful in multiple prediction scenarios. We further apply our approach to analyze the common variants of concern.

2 Results

We developed a framework to assess the effect of mutations in the RBD on viral escape, both with a single-task approach and a multi-task approach. We tested our framework using experimental antibody escape data and compared the multi-task and single-task approaches. We demonstrate that multi-task learning helps reduce variance and improve performance. Moreover, we show that using multi-task learning yields an informative representation of the RBD sequence that can be subsequently used to predict multiple properties.

2.1 Multi-task learning improves antibody escape recognition

Our main training data set is taken from Greaney et al. (2021b) and contains systematic single amino-acid substitutions in the RBD region and their effects on escape probability with respect to each one of several antibodies. From the aforementioned data two tasks were derived: classifying a mutation as significant for escape and predicting (regressing) its escape probability. To this end, we developed neural network models that either consider one antibody at a time (single-task) or multiple antibodies simultaneously (multi-task). The architectures and training process of these models are detailed in the Materials and Methods.

Figure 1 depicts the (distribution of) Pearson correlation between predicted and measured escape probabilities across 9 antibodies, comparing between the single-task and multi-task approaches. Similarly, Figure 2 depicts the area under the ROC curve for the corresponding classification task. It is evident that the multi-task approach reduces variance and increases mean performance for both regression and classification tasks, respectively.

2.2 Analysis of the induced embedding

After establishing the utility of our predictive model, We aim to further use it to find an informative representation of mutations that is more compact than the sequence of amino acids, while also preserving antibody escape information. Such a representation will allow us to test the predictive power of our model with respect to yet unseen properties. As a first test, we calculate viral escape of single amino-acid substitution from new, yet unseen antibodies: LY-CoV016, REGN10987 and REGN10933 Starr et al. (2021b). Figure 3 shows that the embedding-based predictions outperform the original neural network. This result indicates the power of the latent representation compared to the original amino-acid sequence.

As a second test, we checked the utility of the representation in predicting the effect of a single amino-acid substitution on the binding of the spike protein to ACE2. Specifically, binding affinity is given as the difference between the log of the dissociation constant of the mutation with respect to wildtype. As the binding is vital for viral entry, we assumed the learned representation could encompass useful information about it. Figure 3 confirms this assumption and shows that using the learned representation leads to improved predictions. In conclusion, the embedding was able to encode useful data regarding sites and mutations and apply them to new tasks successfully.

3 Materials and methods

3.1 Data representation

Greaney et al. compiled a data set containing the escape information of about 2,000 single amino-acid substitutions in the RBD with respect to nine monoclonal antibodies Greaney et al.

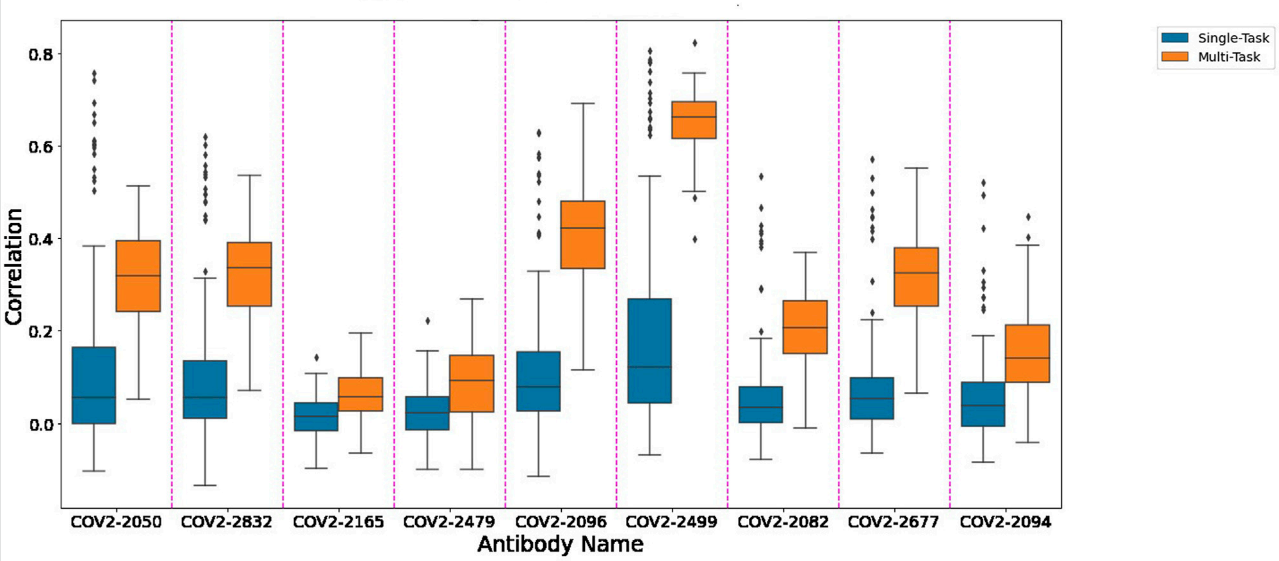


FIGURE 1
A comparison of single-task and multi-task performance in predicting escape probability.

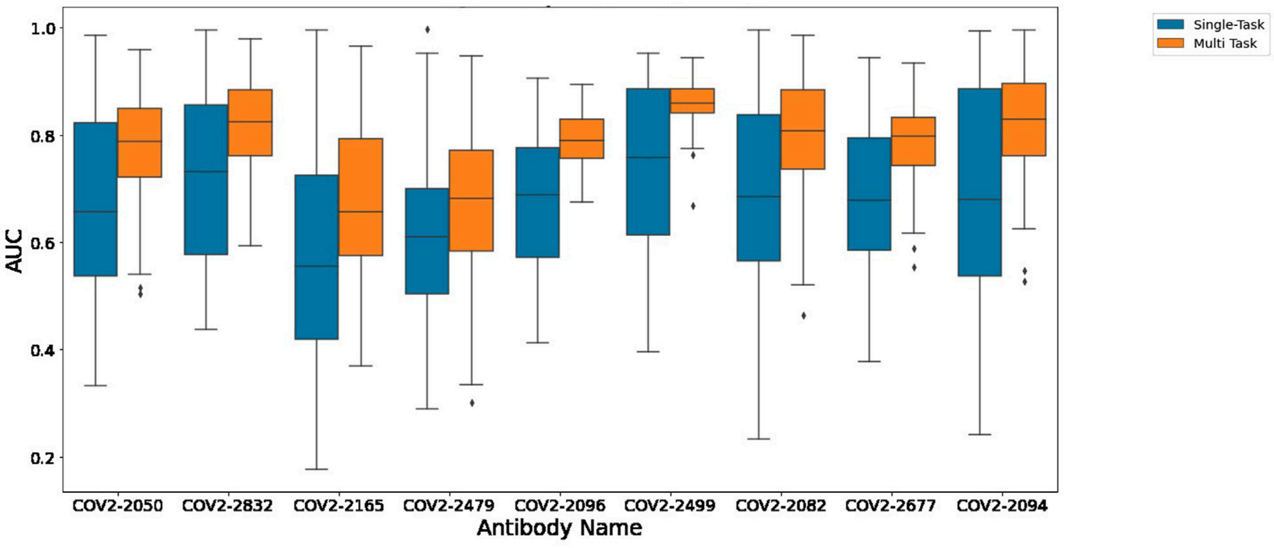


FIGURE 2
A comparison of single-task and multi-task performance in binary escape prediction.

(2021b). Since the amino-acid changes are all in the RBD region, we can treat our input as a subsequence of the original Spike protein, reducing the representation to a 201-long character string.

The original escape information is given as probabilities. In order to create the viral-escape classification task from the continuous data, we followed Starr et al. Starr et al. (2021b) and chose 10 times the median escape across all sites as

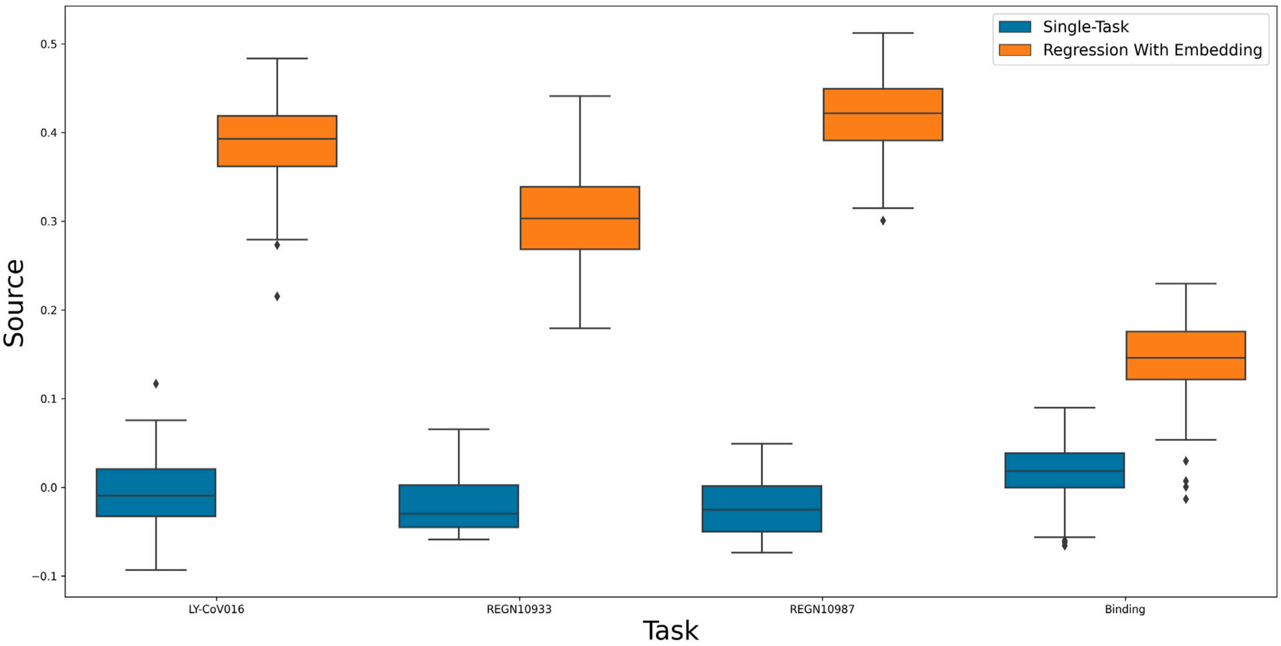


FIGURE 3
A comparison of a single-task neural network to linear regression of multi-task induced embedding.

threshold for significant escape for each antibody and gave a label of 1 to samples that exceeded the threshold.

3.2 Neural network architecture and training

We use a neural network that receives as an input a string of fixed length $n = 201$ over an alphabet of size $m = 20$ (number of amino acids). The model applies one-hot encoding on every character, resulting in a binary vector of size m. It then applies a linear transformation to each vector to create a "character-level" embedding. These embeddings are concatenated and fed to a fully-connected layer, creating a "sequence-level" embedding. The final output layer is a linear layer with size equal to the number of prediction tasks k, followed by k task-dependent activation functions. In our case, $k = 9$, each task corresponds to an antibody in our train data, while our output activation functions are all sigmoids. This means our output will be 9 probabilities each corresponding to an escape probability of a different antibody.

When we refer to a model as a "single-task model" it means that $k = 1$, when $k > 1$ we refer to the model as a "multi-task model". This means, that when comparing between multi-task and single task, we will have k single-task model, each corresponding to one antibody, while having a single multi-task model with k outputs. For training and evaluation we randomly split the data into 30% test and 70% train, run the model 100 times and report the performance distribution obtained using box-plots. Performance is measured in the binary case using the area under the ROC curve and in the continuous case using Pearson's correlation between predictions and true value in test set. We use the Adam optimizer Kingma and Ba (2015) with a learning rate of 1e-4 and a maximum of 100 epochs. Our model loss function is the sum of all the tasks' loss functions, where for each task we use the cross entropy loss function.

3.3 Training using a fixed embedding

Utilizing our multi-task model's last hidden layer as a latent representation of mutations, we can predict other RBD properties such as binding. To this end, we add a linear layer after the embedding layer whose weights are trained using linear regression. When calculating escape probabilities, we use the sigmoid activation function in our output layer, so the training with fixed embedding is done *via* logistic regression (more precisely, linear regression on the inverse sigmoid of the escape data), meaning the task is identical to binding regression.

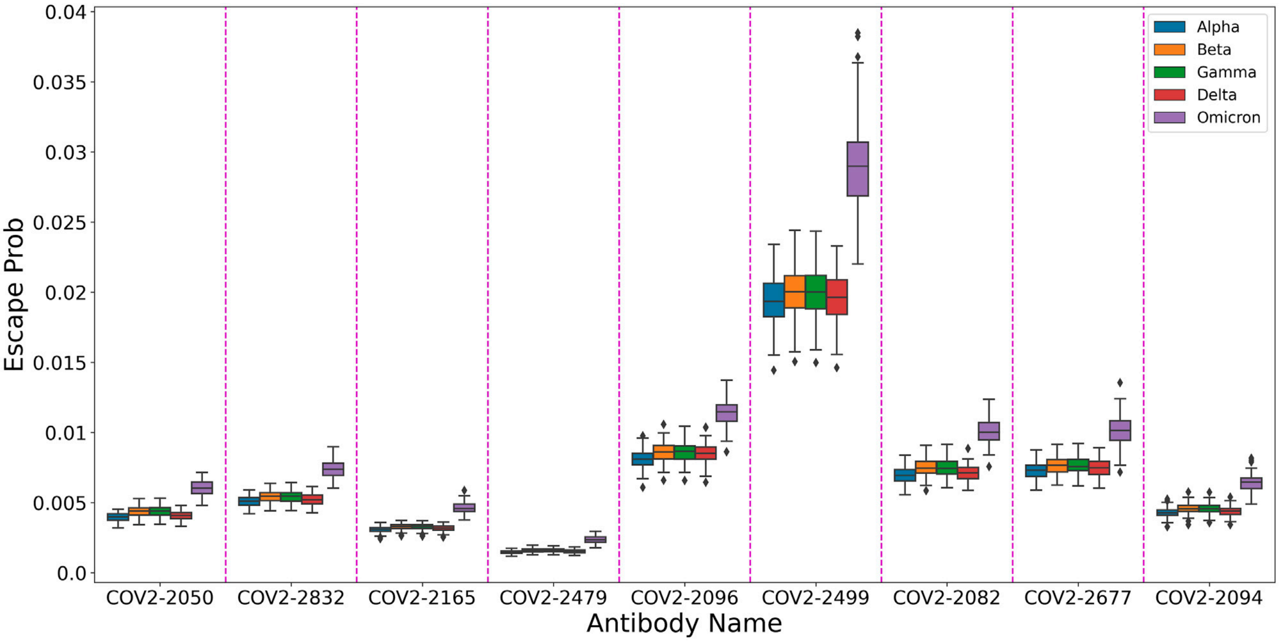


FIGURE 4
Predicted escape from antibodies for different SARS-CoV-2 variants.

3.4 Data and materials availability

Code and data are available at https://github.com/bgmoshe/multi_tasking_antibodies.

4 Discussion

In this paper we develop a computational framework that harnesses systematic mutation screens in the receptor binding domain of the viral Spike protein for escape prediction. Unlike (Bozdaganyan et al., 2022) and (Greaney et al., 2021a), who demonstrate an approach to quantify binding to antibodies, we do not assume a predefined relation between the effect of different mutations, allowing us to have a more general model that is learned automatically from data. Furthermore, in contrast to (Hie et al., 2021) we can quantify mutation escape potential with respect to each antibody. Our framework allows us to infer a latent representation of mutations that preserves escape information. This is particularly useful for predictions regarding yet unseen antibodies or variants.

In order to showcase this attribute, We used our trained model to predict the escape probabilities of variants of concern (as defined by the World Health Organization) as shown in Figure 4. The figure highlights the result of Planas et al. (2021) that Omicron has higher probability of evading antibodies than previous variants. We suggest that using our multi-task model one can provide information on the effect of multiple mutations at different sites, thus allowing researchers to focus on more likely variants of concern.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Both authors conceived the idea for the paper and designed the algorithm. BG implemented the algorithm and RS supervised.

Funding

RS was supported by a QBI/UCSF-TAU joint grant in Computational Biology and Drug Discovery.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Barton, M. I., MacGowan, S. A., Kutuzov, M. A., Dushek, O., Barton, G. J., and van der Merwe, P. A. (2021). Effects of common mutations in the SARS-CoV-2 spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *eLife* 10, e70658. doi:10.7554/eLife.70658

Bozdaganyan, M. E., Shaitan, K. V., Kirpichnikov, M. P., Sokolova, O. S., and Orekhov, P. S. (2022). Computational analysis of mutations in the receptor-binding domain of sars-cov-2 spike and their effects on antibody binding. *Viruses* 14, 295. doi:10.3390/v14020295

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep learning*. Cambridge, MA, USA: MIT Press. Availableat: http://www.deeplearningbook.org.

Greaney, A. J., Starr, T. N., and Bloom, J. D. (2021a). An antibody-escape calculator for mutations to the SARS-CoV-2 receptor-binding domain. bioRxiv., 2021.12.04.471236. doi:10.1101/2021.12.04.471236

Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., et al. (2021b). Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* 29, 44–57.e9. doi:10.1016/j.chom.2020.11.007

Hie, B., Zhong, E. D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. *Science* 371, 284–288. doi:10.1126/science.abd7331

Hou, Y. J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K. H., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication *ex vivo* and transmission *in vivo*. *Science* 370, 1464–1468. doi:10.1126/science.abe8499

Kingma, D. P., and Ba, J. (2015). "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), San Diego, May 7-9, 2015.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., et al. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* 182, 1284–1294.e9. doi:10.1016/j.cell.2020.07.012

Planas, D., Saunders, N., Maes, P., Guivel-Benhassine, F., Planchais, C., Buchrieser, J., et al. (2021). Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature* 602, 671–675. doi:10.1038/s41586-021-04389-z

Rienzo, L. D., Monti, M., Milanetti, E., Miotto, M., Boffi, A., Tartaglia, G. G., et al. (2021). Computational optimization of angiotensin-converting enzyme 2 for sars-cov-2 spike molecular recognition. *Comput. Struct. Biotechnol. J.* 19, 3006–3014. doi:10.1016/j.csbj.2021.05.016

Starr, T. N., Greaney, A. J., Addetia, A., Hannon, W. W., Choudhary, M. C., Dingens, A. S., et al. (2021b). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* 371, 850–854. doi:10.1126/science.abf9302

Starr, T. N., Greaney, A. J., Dingens, A. S., and Bloom, J. D. (2021a). Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep. Med.* 2, 100255. doi:10.1016/j.xcrm.2021.100255

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182, 1295–1310.e20. doi:10.1016/j.cell.2020.08.012

Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, A., et al. (2022). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* 184, 64–75.e11. doi:10.1016/j.cell.2020.11.020

World Health Organization WHO (2022). *WHO coronavirus (COVID-19) dashboard. 2022.* Availableat: https://covid19.who.int (Accessed 02, 21).

Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., et al. (2020). Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* 183, 739–751. doi:10.1016/j.cell.2020.09.032

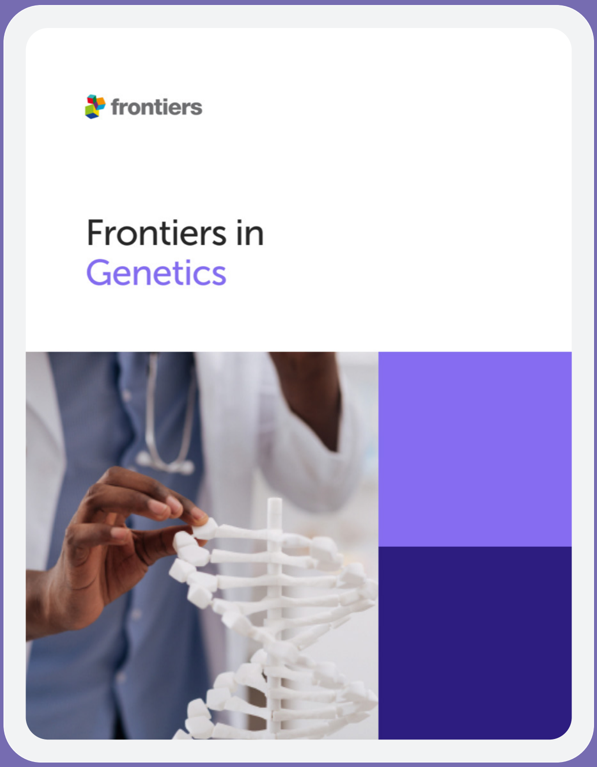
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact