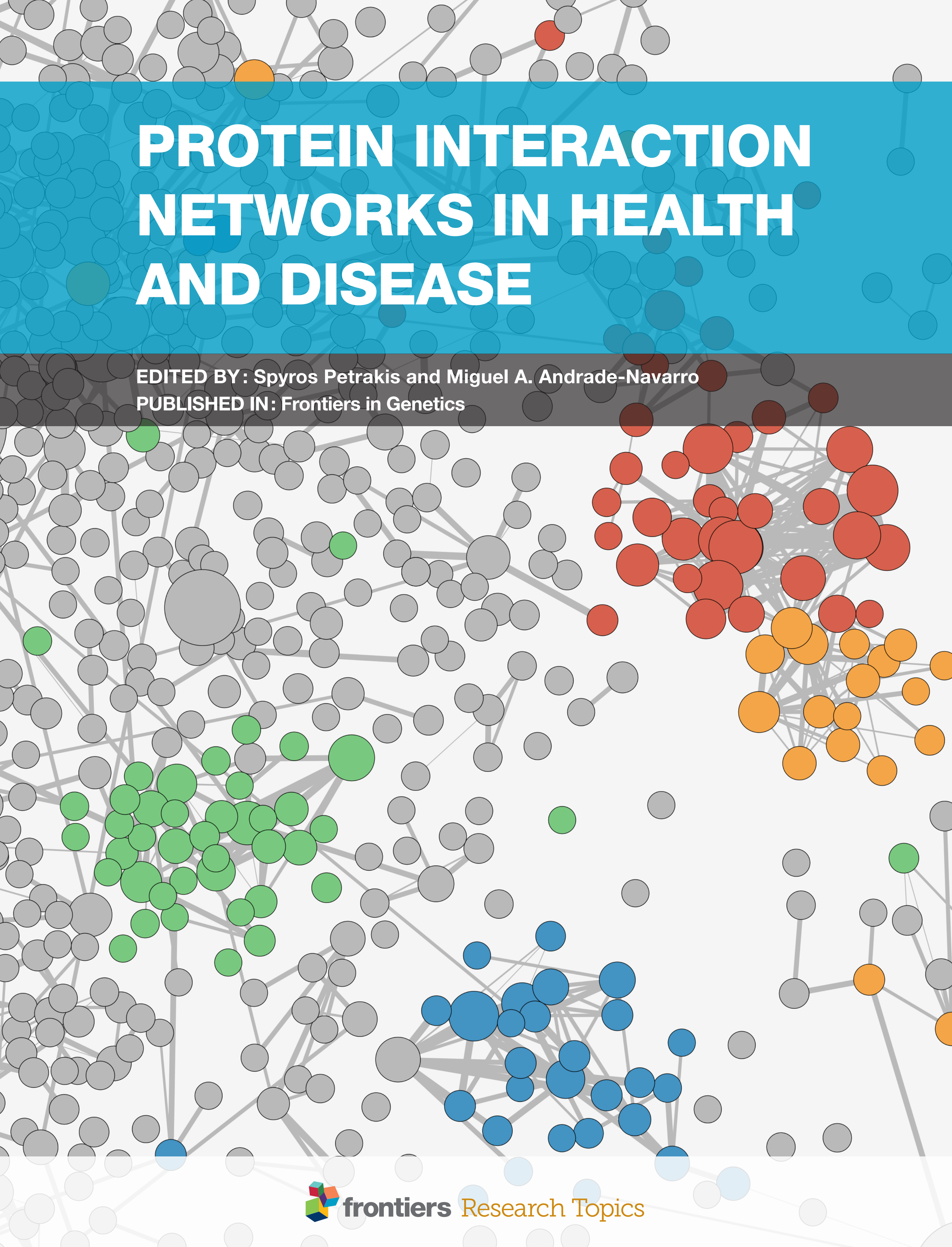


PROTEIN INTERACTION NETWORKS IN HEALTH AND DISEASE

EDITED BY: Spyros Petrakis and Miguel A. Andrade-Navarro

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88919-982-2
DOI 10.3389/978-2-88919-982-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

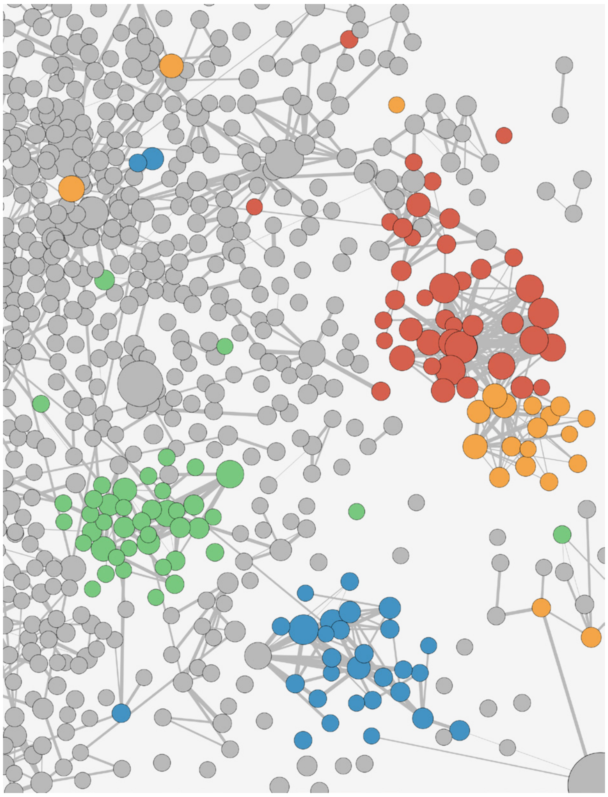
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

PROTEIN INTERACTION NETWORKS IN HEALTH AND DISEASE

Topic Editors:
Spyros Petrakis, Centre for Research and Technology Hellas, Greece
Miguel A. Andrade-Navarro, Johannes-Gutenberg University & Institute of Molecular Biology, Germany



Network biologists have found that proteins associated to human disorders form disease modules in the human protein interaction network (depicted with different color in the figure). Moreover, modules that are topologically close to each other correspond to diseases with similar phenotypes or symptoms, or to conditions that co-occur. As a result, the development of computational approaches to predict protein interactions and detect spurious ones is crucial in the emerging field of network medicine.
Image by Gregorio Alanis-Lobato

The identification and mapping of protein-protein interactions (PPIs) is a major goal in systems biology. Experimental data are currently produced in large scale using a variety of high-throughput assays in yeast or mammalian systems. Analysis of these data using computational tools leads to the construction of large protein interaction networks, which help researchers identify novel protein functions.

However, our current view of protein interaction networks is still limited and there is an active field of research trying to further develop this concept to include important processes: the topology of interactions and their changes in real time, the effects of competition for binding to the same protein region, PPI variation due to alternative splicing or post-translational modifications, etc.

In particular, a clinically relevant topic for development of the concept of protein interaction networks is the consideration of mutant isoforms, which may be responsible for a pathological condition. Mutations in proteins may result in loss of normal interactions and appearance of novel abnormal interactions that may affect a protein's function and biological cycle.

This Research Topic presents novel findings and recent achievements in the field of protein interaction networks with a focus on disease. Authors describe methods for the identification and quantification of PPIs, the annotation and analysis of networks, considering PPIs and protein complexes formed by mutant proteins associated with pathological conditions or genetic diseases.

Citation: Petrakis, S., Andrade-Navarro, M. A., eds. (2016). Protein Interaction Networks in Health and Disease. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-982-2

Table of Contents

Section 1. Identification and Quantification of Protein-Protein Interactions

Section 2. Annotation of Protein Interaction Networks

frontiers in Genetics

EDITORIAL
published: 14 June 2016
doi: 10.3389/fgene.2016.00111

Editorial: Protein Interaction Networks in Health and Disease

Spyros Petrakis[1]* and Miguel A. Andrade-Navarro[2,3]*

[1] Centre for Research and Technology Hellas, Institute of Applied Biosciences, Thessaloniki, Greece, [2] Faculty of Biology, Johannes-Gutenberg University, Mainz, Germany, [3] Institute of Molecular Biology, Mainz, Germany

Keywords: protein interactions, protein function, protein network, disease, systems biology

OPEN ACCESS

Edited and reviewed by:
Raina Robeva,
Randolph-Macon College, USA

****Correspondence:***
Spyros Petrakis
spetrak@certh.gr;
Miguel A. Andrade-Navarro
andrade@uni-mainz.de

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

Received: *20 May 2016*
Accepted: *01 June 2016*
Published: *14 June 2016*

Citation:
Petrakis S and Andrade-Navarro MA (2016) Editorial: Protein Interaction Networks in Health and Disease. Front. Genet. 7:111. doi: 10.3389/fgene.2016.00111

The Editorial on the Research Topic

Protein Interaction Networks in Health and Disease

The identification and annotation of protein-protein interactions (PPIs) is of great importance in systems biology. Big data produced from experimental or computational approaches allow not only the construction of large protein interaction maps but also expand our knowledge on how proteins build up molecular complexes to perform sophisticated tasks inside a cell. However, if we want to accurately understand the functionality of these complexes, we need to go beyond the simple identification of PPIs. We need to know when and where an interaction happens in the cell and also understand the flow of information through a protein interaction network.

Another perspective of the research on PPI networks is the study of their relation to disease. In disease conditions, mutations that alter the secondary structure of one protein might perturb its PPIs, as well. Thereafter many things can go wrong via cascading effects, caused by the inter-relatedness of the mutated protein to other proteins through the PPI network. Such perturbations could block the formation of a protein complex or lead to the formation of new protein complexes and the activation of abnormal signaling pathways. These events could alter the cellular transcriptome profile and further contribute to disease pathogenesis. That is why the maintenance of the proper structure and functionality of a PPI network is crucial for cellular homeostasis. Its disruption can cause complex effects and understanding them requires advanced methods for analysis.

The aim of this Research Topic is to present novel findings and recent achievements in the field of PPI networks. Thematically, it is divided into two parts. First, we present methods for the identification and quantification of PPIs; second, we describe computational approaches to annotate interactomes and extract information related to disease prediction or disease progression.

The first four articles deal with the identification and quantification of PPIs. In the first work, Suter et al. describe the application of next generation sequencing (NGS) for the characterization of binary PPIs. Authors present an accurate method to analyze yeast two-hybrid data by NGS and also interpret interaction data via quantitative statistics. They also discuss how this methodology can be used to discover differential PPIs allowing the identification of disease mechanisms (Suter et al.).

The next two review articles describe mass spectrometry (MS) based approaches. Yang et al. present methods that can determine the relative abundance of purified proteins in a sample enabling the identification of transient PPIs in different conditions. Additionally, when combined with proximity tagging methods, MS may illuminate spatial or temporal PPIs, especially those of signaling pathways whose perturbation may underlie human diseases (Yang et al.). Meyer and Selbach indicate how MS can be used to identify dynamic changes in the interactome. Stable isotope labeling in aminoacids and affinity purification-MS can shed light on the dynamic behavior of proteins even at different stages of an experiment following perturbation. Authors also describe how MS may identify the stoichiometry of proteins in complexes. These methodologies can be employed to study the dynamic changes of PPIs under normal and disease conditions (Meyer and Selbach).

In the next article, Buntru et al. review novel cell-based assays for the detection of PPIs and discuss their strengths and weaknesses. Compared to traditional genetic or biochemical methods, these techniques provide quantitative information of PPIs even in the context of living cells. This information can be used to prioritize a large number of PPIs, allowing researchers to better describe the biological systems and improve our understanding of disease processes (Buntru et al.).

The second part of the Research Topic is comprised of seven papers dealing with the annotation of protein interaction networks. Alanis-Lobato describes computational mining tools to improve the reliability of protein networks and predict new interactions based on the topological characteristics of their components. He also provides examples on how the integration of clinical data can highlight disease modules in these networks or indicate similarities between diseases (Alanis-Lobato).

Pelassa and Fiumara study the functional role of homopolymeric amino acid repeats (AARs) in proteins and their PPIs. AARs are considered to mediate PPIs and in some cases correlate with human diseases, such as polyglutamine expansions involved in Huntington's disease. The authors describe a computational screening of the human interactome and show that AAR-containing network components have a high degree of connectivity. They also indicate an overlap between AARs and interaction domains suggesting that AARs play an important role in shaping protein interaction networks (Pelassa and Fiumara).

Lecca and Re present WG-Cluster, a novel algorithm for the detection of modular structures in protein networks. This tool combines network node and edge weight information of connected proteins improving the biological interpretability of a PPI. The authors also apply their technique in biological datasets from patients with colorectal cancer and indicate differentially active cellular processes in normal vs. tumor conditions (Lecca and Re).

In the next article, Chen and colleagues use the dynamical network biomarkers method to detect early disease signals in a breast cancer cell model. The authors pinpoint critical network changes and highlight a number of pathways associated with the pre-transition from the normal state to a cancer cell progression stage. They also suggest the use of these signals as targets for disease intervention (Chen et al.).

Databases collecting data on experimentally verified PPIs are a valuable resource for the research community. In particular, there are studies that extract biological knowledge from analysis of these global data. However, the different intensity with which different proteins have been studied, influences the amount of data that is available for certain proteins leading to wrong statements. Schaefer et al. study the biases that affect the human PPI data due to research heterogeneity, propose measures to correct this and show an application to proteins involved in cancer.

Yeger-Lotem and Sharan present computational approaches to construct tissue or disease-specific interactomes. They indicate how the combination of transcriptome profiles with proteomics data could categorize PPIs from large networks according to their occurrence in specific tissues. In parallel, they present the effect of disease-causing mutations on protein stability and subsequently on the integrity and structure of protein networks in an affected tissue (Yeger-Lotem and Sharan).

In the last article of this topic, Theofilatos et al. argue about the challenges of computational analysis of PPI data and present future goals such as biomarker discovery or identification of pathogenic PPIs and their drug targeting. Authors also support that the integration of environmental or clinical data in protein networks will allow their in-depth study and the construction of personalized interactomes (Theofilatos et al.).

In the past decade, network biology focused on the representation of the binary interaction of proteins. Today, the field of PPI research capitalizes and hops above the establishment of such previous work and resources, identifies existing limitations, and proposes further avenues of investigation, as reflected in this Frontiers Research Topic. A tight connection between experimental and computational efforts is a hallmark of the articles that we present here, which set the tone that PPI research will follow in the next years. If anything remains unchanged, this is our awareness of the fact that diseases are often caused by the malfunction of large protein complexes. This holds as the main motivation of research in the field, which screams for more complete and reliable interactomes, ultimately crucial in order to identify relevant pathogenic mechanisms and design therapeutic intervention strategies.

AUTHOR CONTRIBUTIONS

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

Conflict of Interest Statement: 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.

PERSPECTIVE
published: 17 December 2015
doi: 10.3389/fgene.2015.00346

Next-Generation Sequencing for Binary Protein–Protein Interactions

Bernhard Suter[1], Xinmin Zhang[2], C. Gustavo Pesce[1], Andrew R. Mendelsohn[1,3], Savithramma P. Dinesh-Kumar[4] and Jian-Hua Mao[5]*

[1] Next Interactions, Inc., Richmond, CA, USA, [2] BioInfoRx, Inc., Madison, WI, USA, [3] Regenerative Sciences Institute, Sunnyvale, CA, USA, [4] Department of Plant Biology, University of California, Davis, Davis, CA, USA, [5] Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

The yeast two-hybrid (Y2H) system exploits host cell genetics in order to display binary protein–protein interactions (PPIs) via defined and selectable phenotypes. Numerous improvements have been made to this method, adapting the screening principle for diverse applications, including drug discovery and the scale-up for proteome wide interaction screens in human and other organisms. Here we discuss a systematic workflow and analysis scheme for screening data generated by Y2H and related assays that includes high-throughput selection procedures, readout of comprehensive results via next-generation sequencing (NGS), and the interpretation of interaction data via quantitative statistics. The novel assays and tools will serve the broader scientific community to harness the power of NGS technology to address PPI networks in health and disease. We discuss examples of how this next-generation platform can be applied to address specific questions in diverse fields of biology and medicine.

Keywords: protein–protein interactions, yeast two-hybrid, interactome mapping, next-generation sequencing, quantitative interaction profiles

OPEN ACCESS

Edited by:
Spyros Petrakis,
Aristotle University of Thessaloniki,
Greece

Reviewed by:
Rosalba Giugno,
University of Catania, Italy
David Georges Biron,
Centre National de la Recherche
Scientifique, France

****Correspondence:***
Bernhard Suter
suter@nextinteractions.com

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

***Received:** 20 September 2015*
***Accepted:** 26 November 2015*
***Published:** 17 December 2015*

Citation:
Suter B, Zhang X, Pesce CG, Mendelsohn AR, Dinesh-Kumar SP and Mao J-H (2015) Next-Generation Sequencing for Binary Protein–Protein Interactions. Front. Genet. 6:346. doi: 10.3389/fgene.2015.00346

INTRODUCTION

Networks of protein–protein interactions (PPIs) govern essentially all biological processes and mechanisms, such as receptor-ligand recognition, immune responses, intracellular and extracellular signaling, growth regulation, and development. Early on, PPI networks or “interactomes” were recognized as the next frontier in biomedicine after the completion of the human genome project (Mendelsohn and Brent, 1999). The role of interaction networks in complex diseases is now a central focus in network biology (Vidal et al., 2011; Sharma et al., 2015). Innovative concepts and technologies are therefore required to satisfy a broad and unmet need for highly reliable and efficient mapping of PPIs.

The charting of interactomes is in many ways more challenging than that of genomes. Proteins are encoded by multiple transcript isoforms and are localized in diverse cellular compartments with distinct milieus. Moreover, variations in amino acids and post-translational modifications affect and determine PPIs. Hence, from a practical perspective, working with proteins is more demanding than working with DNA. For these reasons, to this date, our technical capabilities for systematic approaches toward PPI networks remain limited, when compared with the routine deciphering of genomes, transcriptomes, and exomes at high efficiency and low cost by next-generation sequencing (NGS) technologies (Shendure, 2011; Shendure and Lieberman Aiden, 2012; Mardis, 2013).

Over the last 30 years, diverse technologies have been developed to detect PPIs that are based on different principles with individual strengths and weaknesses. Affinity purification followed by mass-spectrometry (AP-MS) is the standard method to identify protein complexes (Bensimon et al., 2012; Dunham et al., 2012). On the other hand, a variety of assays, such as yeast two-hybrid (Y2H), as well as protein fragment complementation (PCA) in yeast and various mammalian assays, are currently applied for the *in vivo* screening of binary interactions to identify direct binding partners (Stynen et al., 2012). These assays rely on the reconstitution of PPIs *in vivo* and the direct or indirect activation of reporters for selection and scoring of interactions.

Since its inception (Fields and Song, 1989; Gyuris et al., 1993), Y2H has emerged as a widely applied approach for the exploration of novel PPIs and interactome-wide screens (Vidal and Fields, 2014). The assay relies on the splitting of a transcription factor into its DNA binding and activation domains. In most implementations, the bait protein is fused to the DNA binding domain, whereas the prey or a library of prey cDNAs is fused to the activation domain. A physical interaction between bait and prey reconstitutes the transcription factor and activates one or several reporter genes, allowing selection of yeast cells expressing interacting bait-prey pairs. After selection for growth, only a small minority of cells with interacting proteins is enriched over a large background of cells containing non-interacting proteins. Y2H provides therefore a genetic selection system, in which interaction partners can be identified by sequencing the DNA encoding the prey proteins that interact with a defined bait protein.

A variety of other existing *in vivo* assays for screening binary PPIs can be considered alternative implementation of Y2H principles, such as split ubiquitin system for membrane proteins (Obrdlik et al., 2004; Jones et al., 2014), the reverse Y2H screening system and the two-bait interaction trap to explore the effect of allelic variants on PPIs (Vidal et al., 1996; Xu et al., 1997). The yeast one-hybrid technique is a variant for the identification of proteins that bind to DNA motifs and transcription factor binding sites (Fuxman Bass et al., 2015). With yeast three-hybrid (Y3H) the goal is identification of proteins binding to small molecule drugs (protein-drug interactions; PDIs; Moser and Johnsson, 2013).

In this article, we give an overview on existing methods that present different solutions to use NGS as readout for Y2H data. We also present our own experimental and bioinformatics platform that we developed for this purpose and discuss how NGS can overcome the existing limitations of Y2H and diverse other binary interaction assays.

YEAST TWO-HYBRID TECHNOLOGIES AND MAPPING OF INTERACTOMES

High-throughput Y2H assays have been instrumental in proteome-wide screens for the mapping of PPIs that were so far undertaken in human and various model organisms (Uetz et al., 2000; Rual et al., 2005; Stelzl et al., 2005; Yu et al., 2008; Simonis et al., 2009; Rolland et al., 2014). A recent focus for high-throughput Y2H is on differential PPIs of normal and disease-associated alleles occurring in the human population (Dittmer et al., 2014; Sahni et al., 2015). In matrix-based Y2H procedures, comprehensive collections of bait and prey strains are combined in high-throughput, using robotic infrastructure (Uetz et al., 2000, 2004; Stelzl et al., 2005). Yeast clones are arrayed on defined matrix positions, therefore PPIs are scored as visual readouts, eliminating the need to do DNA sequencing for identification. Moreover, the use of annotated full-length open reading frames (ORFs) also circumvents potential artifacts that are associated with cDNA libraries. On the other hand, the requirement for preassembled and defined libraries restricts this method to human and well-defined model organisms for which ORF collections have been made. Moreover, the automated setup that is required for this approach is expensive and not readily available for many researchers.

Despite the importance of Y2H as a discovery system, most Y2H results, also those generated in high-throughput experiments, are not based on truly quantitative measurements. This contrasts with gene expression and protein–DNA interactions which have been systematically explored with DNA microarrays and NGS. Notably, the use of DNA microarrays for parallel identification of Y2H screening results was recognized early on (Cho et al., 1998). More recently, a microarray-Y2H screening and scoring system was introduced and applied to identify interaction partners of huntingtin and ataxin-1, two important determinants for neurodegenerative diseases (Suter et al., 2013). Using the Qi-Sampler repeat sampling tool (Fontaine et al., 2011), microarray-Y2H results were benchmarked against sets of known positives (golden sets) and other gene sets for statistical enrichments. High-confidence microarray-Y2H interactions correlated with positives from the literature and PPIs that were confirmed with luminescence-based mammalian interactome mapping as an alternative assay. Moreover, the quantitative scoring of interaction data and comparison to background controls allowed the elimination of many non-specific binders or sticky prey proteins.

The first adaptation of NGS technology for Y2H came from the lab of Marc Vidal (Yu et al., 2011). In the Stitch-Seq method, the sequences of putatively interacting bait and prey proteins are concatenated so that they comprise a single amplicon for a massive and parallel NGS readout. The method was successfully used to generate high-throughput Y2H datasets (Rolland et al., 2014). The Y2H-Seq approach by the group of Ulrich Stelzl relies on the combination of NGS with matrix Y2H (Weimann et al., 2013). It demonstrated the advantages of the NGS readout for scalability by sequencing the results of hundreds of separate screens through barcode indexing in a single Illumina run. A higher interaction coverage in the screened interactome space was achieved by increasing the sensitivity for detection of PPIs. The Y2H-Seq screens resulted in a network of 523 interactions involving 22 methyltransferases or demethylases for previously undiscovered cellular roles in non-histone protein methylation. However, while Y2H-Seq and Stitch-Seq are powerful tools and pioneering implementations of NGS for Y2H, they are intended for interactome screenings with ORF libraries and aim primarily at increasing scale and sensitivity but do not fully exploit the quantitative potential of NGS.

A NEXT GENERATION SOLUTION FOR Y2H SCREENS

We believe that the perceived shortcomings of Y2H such as inconsistent or non-reproducible results, lack of quantitation, laborious procedures, and above all, high rates of false positive results can be traced to the lack of an adequate readout system. With next-generation interaction screening (NGIS), we developed an innovative concept and methodology to harness the power of NGS technologies for the exploration of PPIs. The application of NGS removes the main restrictions on Y2H imposed by the cost of DNA sequencing. Replacing conventional Sanger sequencing with NGS leads to a massively increased throughput while reducing the cost of sequencing per screen to a small fraction of the conventional readouts (1,000–10,000-fold or more). Currently we are providing screening services for clients that include experimental work, data analysis, and the use of a cloud-based platform (**Figure 1**). NGIS procedures can be applied to every available Y2H and Y2H variant setup for binary interaction screens.

The technical principle of NGIS is shown in **Figure 2**. Tissue- or organ derived cDNA libraries that were cloned into Y2H prey vectors are combined with individual Y2H bait strains via cDNA transformation and mating procedures, and grown on selective medium. Selected prey cDNA clones are then amplified and products are fragmented and sequenced at their entire lengths with Illumina MiSeq or HiSeq. Most important, entire pools are sequenced after unbiased selection without the need to isolate individual clones. Another benefit of the NGIS protocol is that multiple repeat screens can now be undertaken to screen at maximum sensitivity, such that a weak enrichment corresponding to a single clone can be detected in a larger overall population and maximum coverage of the interaction space is achieved. With bioinformatics tools and algorithms adapted from RNA-Seq analytical methods, NGIS data can be processed to assign fold change and false discovery rate for every cDNA clone being sequenced in the assay. By comparing replicated bait results with controls (unrelated baits), the maximum information can be extracted out of the assays, scoring false positives and also taking into account the occurrence of false negatives and the reproducibility of the screening results.

With substantial cost reduction for screening and sequencing, it is worthwhile to generate large repeat datasets only for the purpose of screening the background of non-specific interactions. Indeed, non-specific Y2H activation by a subset of prey cDNAs (sticky preys) often makes up a majority of all hits in a

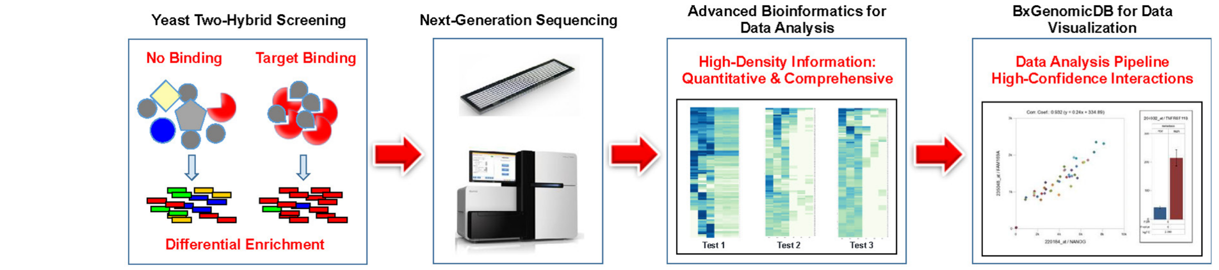

Yeast Two-Hybrid Screening
No Binding
Target Binding
Differential Enrichment
Next-Generation Sequencing
Advanced Bioinformatics for Data Analysis
High-Density Information: Quantitative & Comprehensive
Test 1
Test 2
Test 3
BxGenomicDB for Data Visualization
Data Analysis Pipeline High-Confidence Interactions

FIGURE 1 | Pipeline for Next-generation interaction sequencing (NGIS). Specific target binding in Y2H (or related assays) results in distinct populations of cDNAs that are identified and quantified via NGS. Interactions are scored and interpreted in a bioinformatics pipeline with quantitative statistics.

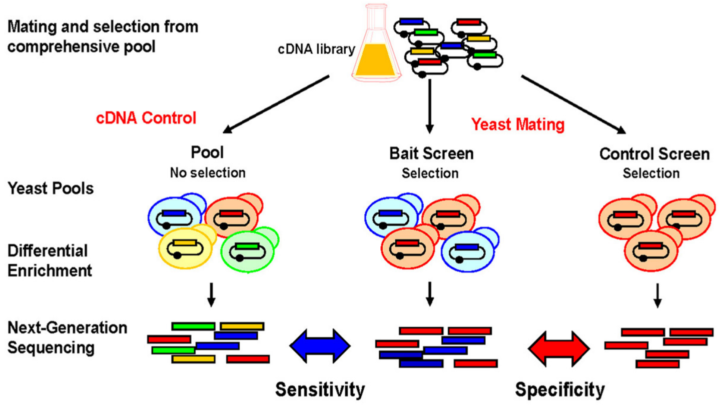


FIGURE 2 | Screening principle and applications for NGIS. Experimental scheme for NGS based interaction profiling. Bait-specific screening results/profiles are compared to original cDNA pools and control screens to uncover background and non-specific interactions. Prey cDNAs that interact with the bait are enriched and quantitated using next-generation sequencing (NGS) and bioinformatics analysis. Bait-specific enrichments (blue) can be quantitatively distinguished from non-specific enrichments (red) and non-selected preys (orange, green).

Y2H screen (Uetz, 2002). Hence, without prior knowledge, conventional Y2H requires specificity tests to confirm each identified PPI after the screening procedure is done, usually by isolation of cDNAs and retests with control strains (Vidalain et al., 2004). Using the NGIS screening scheme, bait specific DNA enrichments can be scored for specificity and non-specific interactions can be excluded *a priori*. This closes an existing gap to other technologies, such as AP-MS for which control datasets for background contaminants are routinely applied to distinguish bona fide interactors from non-specific contaminants (Lavallée-Adam et al., 2011; Mellacheruvu et al., 2013). Importantly, Y2H screening data can be viewed and interpreted as interaction profiles, comparable to transcription profiles in RNA sequencing (Trapnell et al., 2012; Law et al., 2014). Quantitative comparisons between different screen sets allow data mining and predictions for gene function that are impossible to do with the conventional Y2H readout by Sanger sequencing (Suter et al., 2013).

With NGIS interaction and interactome profiles, binary interaction screens can be adapted in several ways and toward different goals (**Table 1**). The primary goal in most Y2H screens is to define the function of proteins by identifying their molecular neighborhoods and to find specific targets that are relevant in diseases, e.g., proteins with functions in cancer or host receptors for pathogen effector proteins in microbial pathogenesis. NGIS interaction profiles and gene enrichment analysis help to understand the function of proteins of interest and the search for relevant interaction targets. An approach related to ours, Quantitative Interactor Sequencing (Qi-Seq), applied the split-ubiquitin system and Illumina NGS to screen for plant host targets for the HopZ2 effector protein that is secreted by the Gram-negative bacterial pathogen *Pseudomonas syringae*, and identified the *Arabidopsis thaliana* MLO2 protein as a target (Lewis et al., 2012).

Besides the discovery of novel PPIs, NGIS also provides a systematic approach to address changes in interaction profiles introduced by variants and polymorphisms in proteins that underlie phenotypes in complex and inheritable diseases. A number of studies have shown that Y2H assay is well-suited to detect changes in PPIs that are introduced by disease-specific alleles or random-generated amino acid mutations (Vidal et al., 1996; Xu et al., 1997; Dreze et al., 2009; Rolland et al., 2014). A recent study profiled the interactions of several thousand missense mutations across a spectrum of Mendelian disorders (Sahni et al., 2015). The analysis indicated that two-thirds of disease-associated alleles perturb PPIs, while common variants from healthy individuals rarely affect interactions. Our NGIS platform provides a rapid way to compare PPI patterns from wild-type and mutant versions of the same protein. Quantitative Y2H data will not only show presence or absence of individual PPIs, but also shift in overall interaction patterns, which may cause gain or loss of protein function.

TABLE 1 | Solutions provided by NGIS for diverse problems and applications.

Area	Problem	Solution
Biological pathways	Mechanism of diverse diseases	Comparative interaction profiling
Microbial pathogenesis	Host virulence determinants	Comparative interaction profiling
Complex and inheritable diseases	Variants of unknown significance	Parallel interaction fingerprints
Protein engineering	Determinants of protein and peptide binding	Complete interaction landscapes
Drug discovery	Search for drug targets	Three-hybrid target discovery

PERSPECTIVES AND FUTURE CHALLENGES

An immediate use of NGS based interaction screens with Y2H or Y2H variant techniques can be seen in the extraction of valuable and specific leads from quantitative and comprehensive interaction profiles. PPI profiles can be from wild-type and mutant proteins, as well as from isoforms of the same proteins, and also from full length proteins and their individual domains. Often, researchers are not interested in the complete set PPIs exhibited by a target of interest, but rather in a set of PPIs that are altered in disease. By providing an effective way to discover differential or regulated PPIs, NGIS could therefore constitute an important application to explore biological pathways and disease mechanisms.

Other areas in which NGIS could have an impact are protein engineering and target discovery for small molecule drugs (see **Table 1**). Considering that protein domains rather than full-length proteins are at the basic level of proteome organization, screening for protein fragments often reveals specific interaction sites and also PPIs that are masked in full length-proteins by steric hindrance. The value of fragment-based Y2H approaches was demonstrated previously (Boxem et al., 2008; Waaijers et al., 2013). NGS with complex cDNA libraries for high-resolution mapping of interaction sites could therefore be instrumental to achieve a full coverage of the protein interaction space. Reducing the lengths of interaction motifs further down to peptides, NGIS can also be applied for peptide aptamers for which Y2H has been instrumental (Bickle et al., 2006; Hamdi and Colas, 2012). We can also envision a role for NGIS procedures for the selection and optimization of scaffolds for aptamer displays. For example, libraries of novel aptamer scaffolds could be selected that can be targeted to diseased tissue and used both extra- and intracellularly. Scaffolds could then be optimized for functional interactions with proteins of interest.

Within the proper framework, NGIS could also, in principle, be applied for Y2H-based protein-drug interactions, such as Y3H to screen for novel protein targets that bind to known drugs (Moser and Johnsson, 2013), or to address the disruption of PPIs and protein complexes by small molecule binding (Flusin et al., 2012). The screening and selection in small volumes of liquid culture as opposed to large volumes of agar plates is a prerequisite for efficient screens in the presence of drugs. Quantitative analysis of NGIS data could be used to effectively distinguish drug-specific from non-specific interactions.

By providing quantitative measurements, reproducibility by repeat assays, background controls for false positives, streamlined scoring and statistical analysis, NGIS overcomes existing bottlenecks of Y2H, thus providing a valuable technology and service platform. In addition, reconstruction of the components for Y2H fusion expression and reporter selection could increase accuracy, speed, automation, and cost-effectiveness for Y2H screens. A wide repertoire of sequence elements and well-characterized parts is now available for this purpose, although less attention had been paid to PPI and interaction affinities than to transcription parameters (Galdzicki et al., 2011). Regulated promoters that could compensate for differential expression of individual bait proteins could allow a better comparison between different interaction profiles. Another area for improvements is the use of new reporter assays to score interactions. For example, fluorescence measurements by cytometry for Y2H were already recognized as an alternative to the existing reporter systems (Chen et al., 2008). We expect that improved Y2H and Y2H-like assays will unlock the full potential of interaction screening and therefore provide a great benefit for biological and biomedical sciences.

AUTHOR CONTRIBUTIONS

BS developed NGS for Y2H screens is responsible for content and wrote the article. J-HM and SD-K are scientific collaborators and advisors for the NGIS technology, XZ does the bioinformatics and codeveloped the concept. GP and AM, helped develop the Y2H procedures. All authors read and commented the manuscript.

FUNDING

Original research on the NGIS technology was supported by the National Science Foundation (NSF) with Small Business Innovation Research (SBIR) grant no 121608.

REFERENCES

Bensimon, A., Heck, A. J., and Aebersold, R. (2012). Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* 81, 379–405. doi: 10.1146/annurev-biochem-072909-100424

Bickle, M. B., Dusserre, E., Moncorgé, O., Bottin, H., and Colas, P. (2006). Selection and characterization of large collections of peptide aptamers through optimized yeast two-hybrid procedures. *Nat. Protoc.* 1, 1066–1091. doi: 10.1038/nprot.2006.32

Boxem, M., Maliga, Z., Klitgord, N., Li, N., Lemmens, I., Mana, M., et al. (2008). A protein domain-based interactome network for *C. elegans* early embryogenesis. *Cell* 134, 534–545. doi: 10.1016/j.cell.2008.07.009

Chen, J., Zhou, J., Bae, W., Sanders, C. K., Nolan, J. P., and Cai, H. (2008). A yEGFP-based reporter system for high-throughput yeast two-hybrid assay by flow cytometry. *Cytometry A* 73, 312–320. doi: 10.1002/cyto.a.20525

Cho, R. J., Fromont-Racine, M., Wodicka, L., Feierbach, B., Stearns, T., Legrain, P., et al. (1998). Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3752–3757. doi: 10.1073/pnas.95.7.3752

Dittmer, T. A., Sahni, N., Kubben, N., Hill, D. E., Vidal, M., Burgess, R. C., et al. (2014). Systematic identification of pathological lamin A interactors. *Mol. Biol. Cell* 25, 1493–1510. doi: 10.1091/mbc.E14-02-0733

Dreze, M., Charloteaux, B., Milstein, S., Vidalain, P. O., Yildirim, M. A., Zhong, Q., et al. (2009). 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog. *Nat. Methods* 6, 843–849. doi: 10.1038/nmeth.1394

Dunham, W. H., Mullin, M., and Gingras, A. C. (2012). Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* 12, 1576–1590. doi: 10.1002/pmic.201100523

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245–246. doi: 10.1038/340245a0

Flusin, O., Saccucci, L., Contesto-Richefeu, C., Hamdi, A., Bardou, C., Poyot, T., et al. (2012). A small molecule screen in yeast identifies inhibitors targeting protein-protein interactions within the vaccinia virus replication complex. *Antiviral Res.* 96, 187–195. doi: 10.1016/j.antiviral.2012.07.010

Fontaine, J. F., Suter, B., and Andrade-Navarro, M. A. (2011). QiSampler: evaluation of scoring schemes for high-throughput datasets using a repetitive sampling strategy on gold standards. *BMC Res. Notes* 4:57. doi: 10.1186/1756-0500-4-57

Fuxman Bass, J. I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., et al. (2015). Human gene-centered transcription factor networks for enhancers and disease variants. *Cell* 161, 661–673. doi: 10.1016/j.cell.2015.03.003

Galdzicki, M., Rodriguez, C., Chandran, D., Sauro, H. M., and Gennari, J. H. (2011). Standard biological parts knowledgebase. *PLoS ONE* 6:e17005. doi: 10.1371/journal.pone.0017005

Gyuris, J., Golemis, E., Chertkov, H., and Brent, R. (1993). Cdi1, a human G1 and S phase protein phosphatase that associates with Cdk2. *Cell* 75, 791–803. doi: 10.1016/0092-8674(93)90498-F

Hamdi, A., and Colas, P. (2012). Yeast two-hybrid methods and their applications in drug discovery. *Trends Pharmacol. Sci.* 33, 109–118. doi: 10.1016/j.tips.2011.10.008

Jones, A. M., Xuan, Y., Xu, M., Wang, R. S., Ho, C. H., Lalonde, S., et al. (2014). Border control–a membrane-linked interactome of *Arabidopsis*. *Science* 344, 711–716. doi: 10.1126/science.1251358

Lavallée-Adam, M., Cloutier, P., Coulombe, B., and Blanchette, M. (2011). Modeling contaminants in AP-MS/MS experiments. *J. Proteome Res.* 10, 886–895. doi: 10.1021/pr100795z

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15, R29. doi: 10.1186/gb-2014-15-2-r29

Lewis, J. D., Wan, J., Ford, R., Gong, Y., Fung, P., Nahal, H., et al. (2012). Quantitative Interactor Screening with next-generation Sequencing (QIS-Seq) identifies *Arabidopsis thaliana* MLO2 as a target of the *Pseudomonas syringae* type III effector HopZ2. *BMC Genomics* 13:8. doi: 10.1186/1471-2164-13-8

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* (*Palo Alto Calif.*) 6, 287–303. doi: 10.1146/annurev-anchem-062012-092628

Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J. P., St-Denis, N. A., Li, T., et al. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* 10, 730–736. doi: 10.1038/nmeth.2557

Mendelsohn, A. R., and Brent, R. (1999). Protein interaction methods–toward an endgame. *Science* 284, 1948–1950. doi: 10.1126/science.284.5422.1948

Moser, S., and Johnsson, K. (2013). Yeast three-hybrid screening for identifying anti-tuberculosis drug targets. *Chembiochem* 14, 2239–2242. doi: 10.1002/cbic.201300472

Obrdlik, P., El-Bakkoury, M., Hamacher, T., Cappellaro, C., Vilarino, C., Fleischer, C., et al. (2004). K+ channel interactions detected by a genetic system optimized for systematic studies of membrane protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12242–12247. doi: 10.1073/pnas.0404467101

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178. doi: 10.1038/nature04209

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647–660. doi: 10.1016/j.cell.2015.04.013

Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., et al. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* 24, 3005–3020. doi: 10.1093/hmg/ddv001

Shendure, J. (2011). Next-generation human genetics. *Genome Biol.* 12, 408. doi: 10.1186/gb-2011-12-9-408

Shendure, J., and Lieberman Aiden, E. (2012). The expanding scope of DNA sequencing. *Nat. Biotechnol.* 30, 1084–1094. doi: 10.1038/nbt.2421

Simonis, N., Rual, J. F., Carvunis, A. R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., et al. (2009). Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat. Methods* 6, 47–54. doi: 10.1038/nmeth.1279

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029

Stynen, B., Tournu, H., Tavernier, J., and Van Dijck, P. (2012). Diversity in genetic in vivo methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiol. Mol. Biol. Rev.* 76, 331–382. doi: 10.1128/MMBR.05021-11

Suter, B., Fontaine, J. F., Yildirimman, R., Raskó, T., Schaefer, M. H., Rasche, A., et al. (2013). Development and application of a DNA microarray-based yeast two-hybrid system. *Nucleic Acids Res.* 41, 1496–1507. doi: 10.1093/nar/gks1329

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Uetz, P. (2002). Two-hybrid arrays. *Curr. Opin. Chem. Biol.* 6, 57–62. doi: 10.1016/S1367-5931(01)00288-5

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627. doi: 10.1038/35001009

Uetz, P., Rajagopala, S. V., Dong, Y. A., and Haas, J. (2004). From ORFeomes to protein interaction maps in viruses. *Genome Res.* 14, 2029–2033. doi: 10.1101/gr.2583304

Vidal, M., Braun, P., Chen, E., Boeke, J. D., and Harlow, E. (1996). Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. *Proc. Natl. Acad. Sci. U.S.A.* 93, 10321–10326. doi: 10.1073/pnas.93.19.10321

Vidal, M., Cusick, M. E., and Barabasi, A. L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Vidal, M., and Fields, S. (2014). The yeast two-hybrid assay: still finding connections after 25 years. *Nat. Methods* 11, 1203–1206. doi: 10.1038/nmeth.3182

Vidalain, P. O., Boxem, M., Ge, H., Li, S., and Vidal, M. (2004). Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* 32, 363–370. doi: 10.1016/j.ymeth.2003.10.001

Waaijers, S., Koorman, T., Kerver, J., and Boxem, M. (2013). Identification of human protein interaction domains using an ORFeome-based yeast two-hybrid fragment library. *J. Proteome Res.* 12, 3181–3192. doi: 10.1021/pr400047p

Weimann, M., Grossmann, A., Woodsmith, J., Özkan, Z., Birth, P., Meierhofer, D., et al. (2013). A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat. Methods* 10, 339–342. doi: 10.1038/nmeth.2397

Xu, C. W., Mendelsohn, A. R., and Brent, R. (1997). Cells that register logical relationships among proteins. *Proc. Natl. Acad. Sci. U.S.A.* 94, 12473–12478. doi: 10.1073/pnas.94.23.12473

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 104–110. doi: 10.1126/science.1158684

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., et al. (2011). Next-generation sequencing to generate interactome datasets. *Nat. Methods* 8, 478–480. doi: 10.1038/nmeth.1597

Conflict of Interest Statement: 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.

frontiers
in Genetics

MINI REVIEW
published: 01 December 2015
doi: 10.3389/fgene.2015.00344

Illuminating Spatial and Temporal Organization of Protein Interaction Networks by Mass Spectrometry-Based Proteomics

Jiwen Yang[1], *Sebastian A. Wagner*[2] *and Petra Beli*[1]*

[1] *Institute of Molecular Biology, Mainz, Germany,* [2] *Department of Medicine, Hematology and Oncology, Goethe University, Frankfurt, Germany*

Protein–protein interactions are at the core of all cellular functions and dynamic alterations in protein interactions regulate cellular signaling. In the last decade, mass spectrometry (MS)-based proteomics has delivered unprecedented insights into human protein interaction networks. Affinity purification-MS (AP-MS) has been extensively employed for focused and high-throughput studies of steady state protein–protein interactions. Future challenges remain in mapping transient protein interactions after cellular perturbations as well as in resolving the spatial organization of protein interaction networks. AP-MS can be combined with quantitative proteomics approaches to determine the relative abundance of purified proteins in different conditions, thereby enabling the identification of transient protein interactions. In addition to affinity purification, methods based on protein co-fractionation have been combined with quantitative MS to map transient protein interactions during cellular signaling. More recently, approaches based on proximity tagging that preserve the spatial dimension of protein interaction networks have been introduced. Here, we provide an overview of MS-based methods for analyzing protein–protein interactions with a focus on approaches that aim to dissect the temporal and spatial aspects of protein interaction networks.

Keywords: mass spectrometry-based proteomics, protein–protein interactions, transient interactions, spatial interactions

OPEN ACCESS

Edited by:
Spyros Petrakis,
Aristotle University of Thessaloniki,
Greece

Reviewed by:
Andrew K. Ottens,
Virginia Commonwealth University,
USA
Andreas Zanzoni,
Institut National de la Santé et de la
Recherche Médicale – Aix Marseille
Université, France

Correspondence:
Petra Beli
p.beli@imb-mainz.de

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

Received: *15 September 2015*
Accepted: *20 November 2015*
Published: *01 December 2015*

Citation:
Yang J, Wagner SA and Beli P (2015)
Illuminating Spatial and Temporal
Organization of Protein Interaction
Networks by Mass
Spectrometry-Based Proteomics.
Front. Genet. 6:344.
doi: 10.3389/fgene.2015.00344

PROTEIN INTERACTIONS ARE DEFINED BY TEMPORAL AND SPATIAL CONSTRAINTS

Protein–protein interactions are at the core of all cellular functions and dynamic alterations in protein interactions regulate cellular signaling (Scott and Pawson, 2009). Accurate and comprehensive mapping of protein–protein interaction networks is essential for understanding the regulatory mechanisms of cellular processes and signaling pathways as well as for identifying perturbed cellular signaling underlying human diseases. Proteins can form stable interactions and function as part of permanent protein assemblies, however a large proportion of protein–protein interactions are defined by temporal and spatial constraints. Protein–protein interactions can be dynamically altered in response to the intrinsic and extrinsic stimuli (Perkins et al., 2010). Transient protein interactions are frequently induced by posttranslational modifications (PTMs) and, depending on their cellular function, have a range of affinities and lifetimes (Nooren and Thornton, 2003; Seet et al., 2006). Prominent examples include the recruitment of DNA repair factors to sites of DNA lesions, cell cycle-regulated interactions and the formation of receptor signaling complexes after growth factor stimulation. Furthermore, protein–protein interactions are restricted by cellular compartments and can be regulated by protein re-localization to specific cellular structures or organelles. The transient nature and spatial organization are therefore important features that need to be considered when analyzing protein–protein interaction networks (**Figure 1**).

MASS SPECTROMETRY-BASED PROTEOMICS FOR ANALYSIS OF PROTEIN–PROTEIN INTERACTIONS

Mass spectrometry (MS)-based proteomics has become an indispensable tool in modern molecular and cell biology research (Larance and Lamond, 2015). In shotgun or bottom up proteomics approaches, proteins are extracted from cells or tissues and digested into peptides using specific proteases (Aebersold and Mann, 2003). The resulting peptides are separated according to hydrophobicity using high-pressure liquid chromatography and identified by tandem MS (LC-MS/MS).

The most commonly employed approach to study protein–protein interactions *in vivo* is affinity purification-MS (AP-MS; Gingras et al., 2007; Vermeulen et al., 2008; Meyer and Selbach, 2015). In AP-MS workflows, a protein of interest (bait protein) is co-purified with its interaction partners and the purified proteins are subsequently identified by LC-MS/MS. Purification of the bait protein can be achieved using antibodies that specifically bind to the endogenous bait protein. Alternatively, epitope tags can be employed that enable robust and reproducible purification of the bait protein and its interaction partners using highly specific affinity matrices. The latter approach is especially beneficial when antibodies recognizing the bait protein are not available; however, the introduction of epitope tags usually involves overexpression of the bait protein and can lead to non-physiological interactions.

The power of AP-MS for high-throughput discovery of protein–protein interactions has been exemplified by recent landmark studies from the Mann and Gygi laboratories that demonstrated systematic analyses of human protein–protein interactions and mapped 28,500 and 23,744 unique interactions, respectively (Hein et al., 2015; Huttlin et al., 2015). These studies represent a milestone in the long-term effort to comprehensively map human protein–protein interactions.

In addition to AP-MS, co-fractionation strategies have been employed to study cellular organelles and protein complexes. The Mann laboratory has employed biochemical fractionation based on density gradient centrifugation to define the composition of cellular organelles (Andersen et al., 2003; Foster et al., 2006). More recently, Havugimana et al. (2012) and Wan et al. (2015) employed extensive biochemical fractionation and MS to determine the composition of soluble protein complexes in human cells and in cells from diverse metazoan model organisms.

RESOLVING TRANSIENT PROTEIN–PROTEIN INTERACTIONS

Most studies conducted have so far investigated steady state protein–protein interactions, leaving the temporal and spatial aspects of protein–protein interactions largely disregarded. Mapping transient protein–protein interactions during cellular signaling and in response to cellular perturbations remains a major future challenge. For instance, changes in protein interactions induced by growth factor stimulation or cellular stress, as well as interactions between PTM-catalyzing enzymes and substrates, can often not be captured using conventional methods for analyses of protein interactions. Accordingly, efforts are ongoing to design proteomics methods that permit analysis of transient and low affinity protein interactions.

AP-MS Combined with Quantitative Mass Spectrometry-Based Proteomics

Affinity purification combined with quantitative MS-based proteomics can be used to identify dynamic protein–protein interactions (**Figure 2**). In this approach, affinity purifications are performed under different conditions and the relative abundance of interaction partners is then determined by quantitative MS-based approaches, including metabolic and chemical labeling as well as label-free methods (Ong and Mann, 2005; Bantscheff et al., 2012). Affinity purification is often combined with stable isotope labeling with amino acids in cell culture (SILAC) to monitor protein interactomes after different types of cellular perturbations, including DNA damage (Mosbech et al., 2012; Brown et al., 2015) and ligand stimulation (Satpathy et al., 2015). In addition, this approach has been applied to study the temporal dynamics of protein interactions during cell cycle progression (Hubner et al., 2010; Pagliuca et al., 2011).

Recently, data-independent acquisition (DIA) methods have been employed to map changes in protein–protein interactions after cellular perturbations. Analysis of peptide samples from affinity purification experiments has typically been performed using data-dependent acquisition methods (DDA). Due to the semi-stochastic precursor ion selection in DDA methods, the complete set of peptides can often not be reproducibly identified across all samples. In DIA methods, fragment spectra for the entire mass range are acquired by co-isolating precursor ions in isolation windows of selected m/z ranges. Collins et al. (2013) have described a method for mapping dynamic changes in protein–protein interactions by combining affinity purification with DIA using MS-sequential window acquisition of all theoretical spectra (MS-SWATH). The authors have analyzed interaction partners of 14-3-3β in cells stimulated with insulin-like growth factor for different time periods and reproducibly quantified 1,967 proteins across all samples. A similar approach has been used by Lambert et al. (2013) to map the interaction partners of wild type and mutant forms of CDK4 as well as to probe the effects of Hsp90 inhibition on CDK4 interactions.

In Vivo Reversible Crosslinking

A complementary approach to affinity purification that aims to capture transient and low affinity protein–protein interactions is reversible chemical crosslinking (Hall and Struhl, 2002; Vasilescu et al., 2004; Klockenbusch and Kast, 2010; Smith et al., 2011). Chemicals that mediate protein crosslinks, such as formaldehyde, are applied to cells before lysis to “freeze” protein–protein

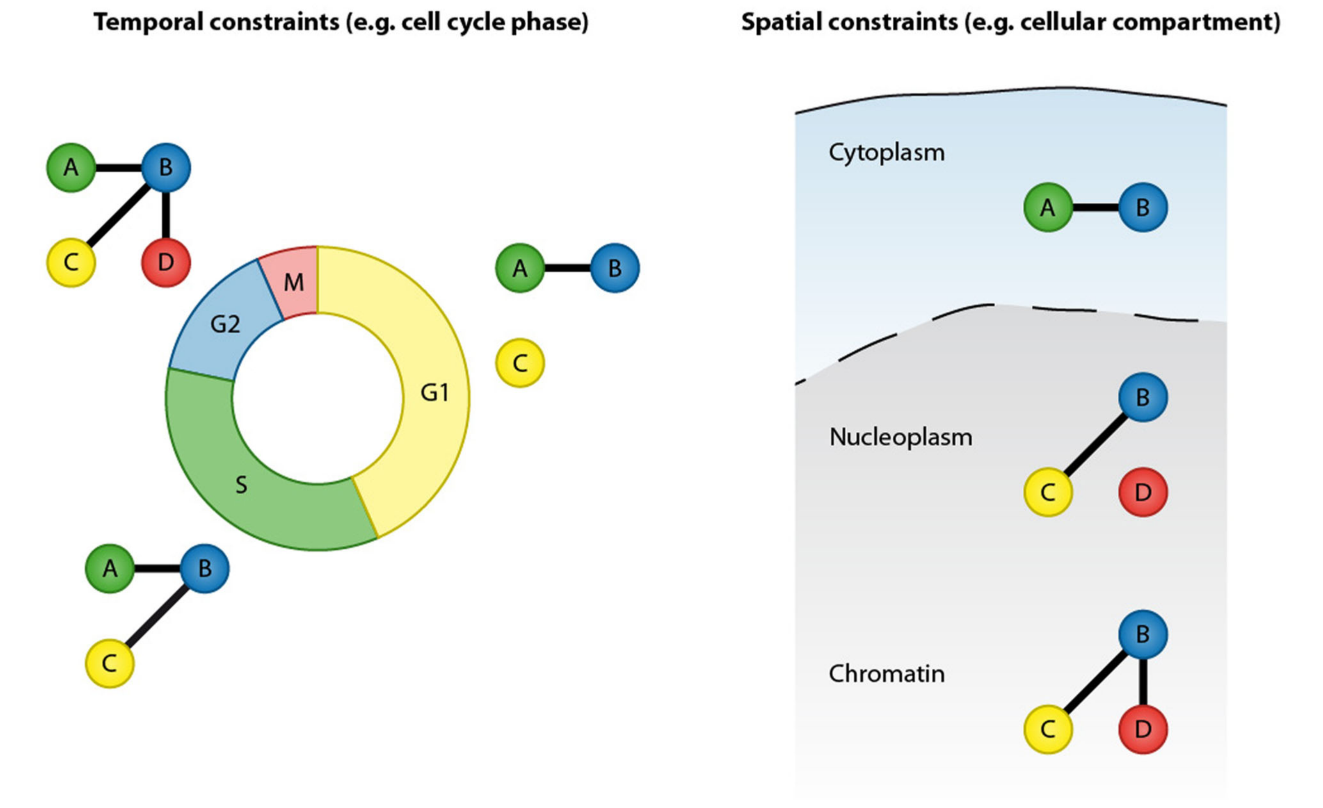

Temporal constraints (e.g. cell cycle phase)
Spatial constraints (e.g. cellular compartment)
A
B
C
D
M
G2
G1
S
Cytoplasm
Nucleoplasm
Chromatin

FIGURE 1 | Protein–protein interactions are defined by temporal and spatial constraints. Many protein interactions are transient and occur only at specific time points, for instance in a particular cell cycle stage. These transient interactions can be mediated by posttranslational modification or by dynamic changes in expression of the binding partners. In addition to temporal constrains, protein interactions are spatially restricted by cellular compartments.

interactions *in vivo* by forming reversible covalent bonds between adjacent amino acids, thereby providing a snapshot of the protein interactome (**Figure 2**). Following crosslinking, cells are lysed and proteins are subjected to conventional affinity purification protocols. Crosslinks are reversed after purification, often simply by boiling, and affinity-purified proteins are identified by LC-MS/MS. In addition to formaldehyde, other crosslinkers that are commonly used for protein–protein interaction studies are NHS-esters and imidates that react with primary amines in the proteins to yield stable amide bonds. If crosslinking is combined with epitope tagging of the bait protein and purification with affinity matrices such as GFP-Trap and Ni-NTA, cell lysis and washing can be performed under denaturing conditions, thus enabling the recovery of poorly soluble proteins and reducing contamination with non-physiological interactions that might occur during the purification (Tagwerker et al., 2006). Formaldehyde-based crosslinking and purification under denaturing conditions has been employed to identify interaction partners of Skp1, an essential component of the SCF ubiquitin ligase complex, and to map the dynamic interaction partners of the 26S proteasome across cell cycle phases (Tagwerker et al., 2006; Kaake et al., 2010). The fact that the crosslinking procedure requires optimization for different cell types and bait proteins might be the reason that this technique has not so far been frequently used for the investigation of transient protein–protein interactions.

Co-fractionation Combined with Quantitative Mass Spectrometry

Kristensen et al. (2012) have developed a method that employs quantitative MS based on SILAC and high-performance size-exclusion chromatography to monitor changes in the cellular interactome in response to growth factor stimulation (**Figure 2**). Using this approach, the authors have identified 350 proteins whose association with a complex increased or decreased after cells were stimulated with the epidermal growth factor. A particular feature of this method is that it allows mapping of dynamic changes in the cellular interactome without the need to overexpress bait proteins and perform affinity purifications. In addition, size-exclusion chromatography enables the heterogeneity of protein complexes within the cells to be determined, by monitoring the distribution of a protein among different complexes. Another advantage of this method is that it provides the possibility to analyze the interactome within a single subcellular compartment, thereby providing a spatial dimension and avoiding the risk of non-physiological interactions that can occur after cell lysis and loss of cellular compartmentalization.

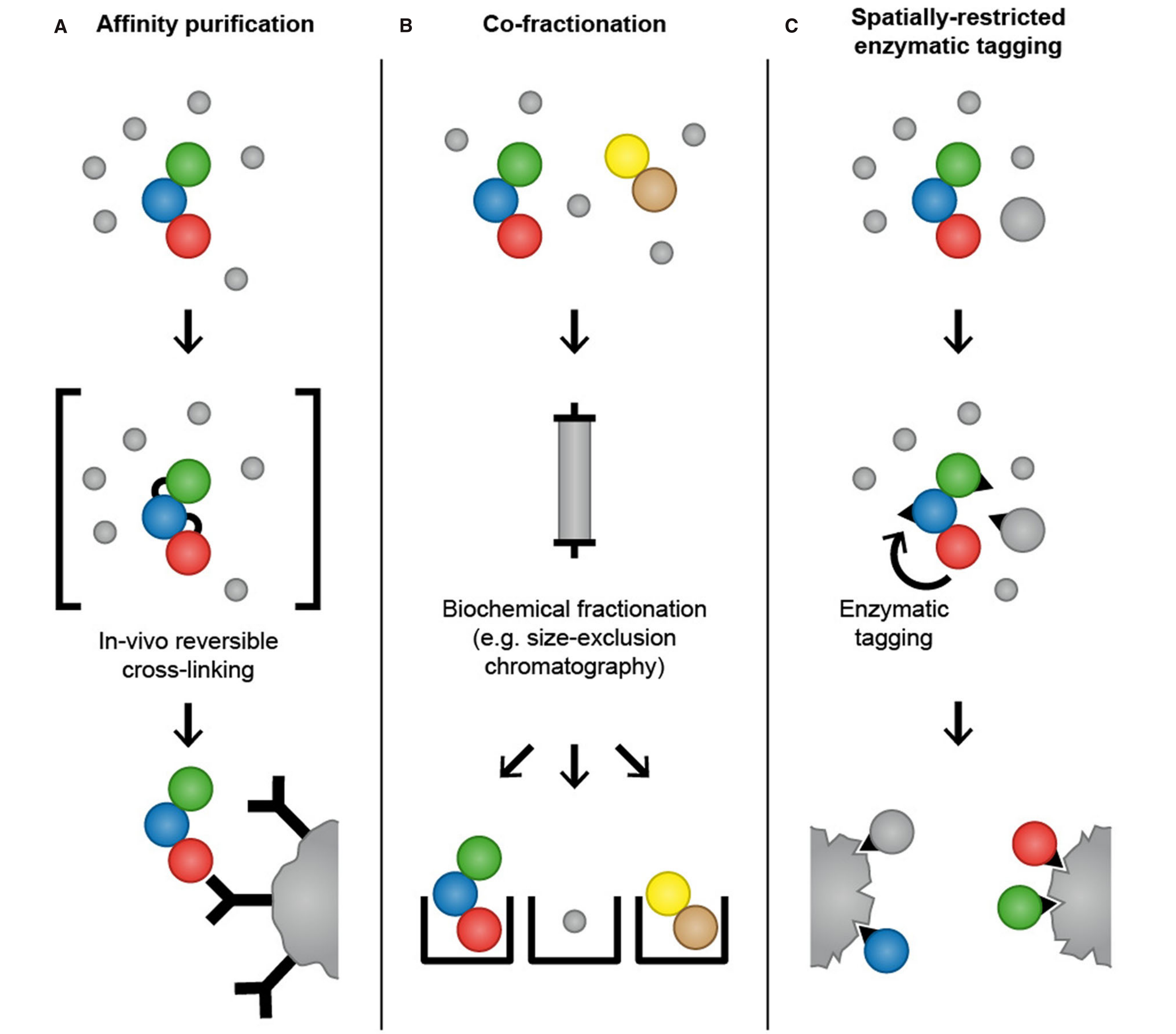

A Affinity purification
B Co-fractionation
C Spatially-restricted enzymatic tagging
In-vivo reversible cross-linking
Biochemical fractionation (e.g. size-exclusion chromatography)
Enzymatic tagging

FIGURE 2 | Mass spectrometry-based proteomics methods for analysis of temporal and spatial aspects of protein–protein interactions. In affinity purification approaches, an antibody that specifically binds to endogenously expressed bait protein is used to purify the protein of interest and its interaction partners. Alternatively, a bait protein fused to an epitope tag is ectopically expressed in cells and purified using affinity matrices or tag-specific antibodies. To increase the probability of capturing transient and weak interactions, chemicals that mediate protein–protein crosslinks can be applied to cells before lysis to "freeze" interactions by forming reversible covalent bonds between adjacent amino acids **(A)**. In co-fractionation-based methods, proteins are subjected to extensive fractionation, for instance by high-performance size-exclusion chromatography, and the precise co-elution of two proteins is used as evidence for their interaction **(B)**. In spatially restricted enzymatic tagging BirA* or APEX is fused to a protein of interest and ectopically expressed in cells. Biotinylation of proximal proteins is triggered by the addition of biotin for 24 h (BioID) or biotin-phenol for 1 min (APEX). Cells are lysed under denaturing conditions and biotinylated proteins are recovered using streptavidin followed by LC-MS/MS analysis **(C)**.

RESOLVING SPATIAL ORGANIZATION OF PROTEIN–PROTEIN INTERACTIONS BY PROXIMITY TAGGING

In addition to defining transient protein–protein interactions, another challenge lies in resolving the spatial organization of protein interaction networks. In affinity purification approaches, proteins localized to different cellular compartments are mixed during cell lysis and subjected to purification under native conditions, which might lead to the formation of non-physiological interactions. Recently developed methods for spatially restricted enzymatic tagging using the promiscuous biotin ligase BirA* (BioID) or the engineered ascorbate peroxidase (APEX) can be employed to overcome this problem and preserve the spatial dimension of interactions (Roux et al., 2012; Rhee et al., 2013).

Biotin Ligase-Based Proximity Tagging (BioID)

BirA is a biotin ligase from *E. coli* that activates biotin to biotinoyl 5-AMP (bioAMP) in an ATP-dependent reaction (Chapman-Smith and Cronan, 1999). Biotinoyl 5′-AMP is then transferred to substrate proteins containing a specific BirA recognition sequence (Beckett et al., 1999). An engineered mutant form of BirA (R118G) with abolished substrate specificity and reduced affinity for biotinoyl 5′-AMP promiscuously biotinylates proteins in its proximity (Choi-Rhee et al., 2004; Cronan, 2005). Roux et al. (2012) devised a method called BioID in which the promiscuous biotin ligase BirA* is fused to a protein of interest and expressed in mammalian cells. After incubation of the cells with biotin, the BirA*-fusion protein biotinylates proteins in its proximity (**Figure 2**). Subsequently, cells are lysed under denaturing conditions and biotinylated proteins are selectively isolated using streptavidin and identified by LC-MS/MS. The authors tested the utility of BioID by fusing BirA* to the nuclear envelope (NE) component lamin A (LaA) that is highly insoluble and therefore difficult to study with conventional methods for interactome analysis. Analysis of biotinylated proteins in cells expressing BirA*-LaA by LC-MS/MS identified known LaA interactors as well as the novel NE component SLAP75 (Roux et al., 2012). BioID possesses several advantages over conventional affinity purification. Firstly, BirA*-based biotinylation of proteins occurs in living cells and therefore non-physiological interactions that might occur after cell lysis and loss of cellular compartmentalization are avoided. Secondly, proximity-dependent biotinylation by the promiscuous biotin ligase BirA* can capture low affinity interactions that will frequently be lost in conventional affinity purification. Furthermore, BioID allows the use of denaturing lysis conditions, which helps to identify proteins that are insoluble under commonly used native lysis conditions and reduces contamination with non-specific binders. However, BioID also has limitations that should be considered during experimental design. Activated biotin targets primary amines (predominantly lysine residues) and the efficacy of the biotinylation depends on the number and availability of primary amines in proteins (Roux et al., 2013). As result, the abundance of the purified biotinylated proteins does not necessarily correlate with the strength or stoichiometry of the association. Moreover, biotinoyl 5′-AMP has a half-life of minutes, which might lead to a large labeling radius (Rhee et al., 2013). In the BioID-LaA experiment, the authors showed that histone proteins constitute only a small fraction of the identified proteins, although they are lysine rich and highly abundant in the nucleus, which provides evidence against the idea that BioID generates widespread biotinylation (Roux et al., 2012). Importantly, BioID does not distinguish interaction from proximity, which needs to be taken into account during data analysis. BioID has been successfully employed to identify interaction partners of proteins and to characterize the composition of subcellular organelles, such as the centrosomes and the nuclear pore, which are otherwise refractory to traditional approaches (Couzens et al., 2013; Firat-Karalar et al., 2014; Coyaud et al., 2015; Dingar et al., 2015; Rodriguez-Fraticelli et al., 2015; Zhou et al., 2015). A recent study employed BioID to identify over 50 putative substrates of the ubiquitin ligase $SCF^{\beta\text{-TrCP1/2}}$ indicating a potential application of BioID for the analysis of substrates of PTM-catalyzing enzymes (Coyaud et al., 2015). The Gingras laboratory has performed a side-by-side comparison of AP-MS and BioID for analyzing interaction partners of chromatin-associated proteins (Lambert et al., 2015). Interestingly, they concluded that BioID enables the identification of a larger number of interaction partners and that identified interaction partners are significantly less abundant than interaction partners identified by AP-MS. Another observation from this study is the relatively small overlap between the interaction partners identified by AP-MS and BioID, suggesting that both approaches have a bias for specific subsets of proteins and might have a complementary value for comprehensive identification of protein interaction partners.

Ascorbate Peroxidase-Based Proximity Tagging

Another enzymatic proximity tagging approach developed by the Ting laboratory uses an engineered ascorbate peroxidase (APEX) (Martell et al., 2012). APEX is a monomeric mutant derived from the plant APEX with increased enzymatic activity. Like wild type peroxidase, APEX catalyzes H_2O_2-dependent polymerization and local deposition of DAB (3,3′-diaminobenzidine), which subsequently recruits electron dense osmium, yielding electron microscopy (EM) contrast (Lam et al., 2015). Based on the observation that APEX is active in all cellular compartments and withstands strong EM fixation, Martell et al. (2012) demonstrated the utility of APEX for EM analysis of a variety of mammalian organelles and specific proteins.

In addition to DAB, APEX also oxidizes numerous phenol derivatives such as biotin-phenol to phenoxyl radicals that covalently react with electron-rich amino acids. In cells expressing APEX fused to a protein of interest, biotinylation of proximal proteins is initiated by incubating cells with biotin-phenol and H_2O_2 for 1 min. The proximal proteins can subsequently be purified using streptavidin under denaturing conditions and identified by LC-MS/MS analysis (**Figure 2**). Rhee et al. (2013) selected mitochondria as a model organelle for testing APEX-based identification of organelle proteins. To test the spatially restricted labeling capacity of APEX, mitochondrial matrix-targeted APEX was used to investigate the protein composition of the mitochondrial matrix and inner mitochondrial membrane. Using LC-MS/MS, the authors have identified 495 proteins, 94% of which had prior mitochondrial annotation. Thirty-one of those 495 proteins had never been correlated with mitochondria and are therefore potentially novel mitochondrial proteins. Of note, only subunits with exposure to matrix space were identified, indicating that phenoxyl radicals do not pass through the inner mitochondrial membrane, proving further the specificity of APEX-based proximity tagging (Lam et al., 2015).

APEX-based proximity tagging can provide spatially and temporally resolved proteomic maps and can be potentially employed to study weak and dynamic protein interactions as well as enzyme-substrate relations. APEX requires only 1 min to label proximal proteins rather than the 24 h required for the BioID method. It therefore, has a better temporal resolution and offers a better platform to study transient protein–protein interactions under different conditions and time points. Furthermore, phenoxyl radicals are short lived (<1 ms) and therefore have a small labeling radius (<20 nm). It is worth mentioning that APEX can also be used to confirm the subcellular localization of target proteins using EM or fluorescent microscopy. To date, the applicability of APEX beyond the mapping of proteins in membrane-bound cellular organelles has not been demonstrated, and it remains to be addressed if APEX-based proximity tagging is suitable for analysis of interaction partners of individual proteins or protein substrates of PTM-catalyzing enzymes.

CONCLUSION

Mass spectrometry-based proteomics has delivered unprecedented insights into human protein interaction networks. To date, most studies have focused on mapping steady-state protein–protein interactions. Future challenges remain in the identification of transient and low affinity interactions during cellular signaling, as well as in understanding the spatial organization of protein interaction networks. Although affinity purification combined with quantitative MS-based proteomics is a powerful approach for the identification of dynamic protein interactions, transient and low affinity interactions, such as those induced by growth factor stimulation or cellular stress, are frequently lost. *In vivo* chemical crosslinking, in which chemicals that form reversible covalent bonds are applied to cells before lysis to "freeze" protein–protein interactions can help to identify these interactions. The need to optimize the crosslinking procedure for different cell types and bait proteins hinders the routine use of this method for analyzing transient protein interactions. In addition to AP-MS, approaches based on protein co-fractionation combined with quantitative MS have been successfully employed to analyze transient protein interactions during cellular signaling. Spatially restricted enzymatic tagging approaches, such as BioID and APEX, preserve the spatial organization of protein interaction networks and enable analysis of protein interactions in insoluble structures, thereby complementing AP-MS. Importantly, these approaches do not enable a distinction to be made between interaction partners and non-interacting proximal proteins. Therefore, combining affinity purification and spatially restricted enzymatic tagging could help to produce a more accurate and comprehensive picture of protein–protein interaction networks of interest. This strategy has the potential to become a standard procedure for protein interaction studies, as has already been exemplified by a recent study that focused on chromatin-associated protein complexes (Lambert et al., 2015)

AUTHOR CONTRIBUTIONS

JY, SW, and PB prepared and wrote the manuscript.

ACKNOWLEDGMENTS

We thank the members of the PB and SW group for helpful discussions. The research in PB's group is supported by the Emmy Noether Program of the German Research Foundation (DFG, BE 5342/1-1) and the Marie Curie Career Integration Grant from the European Research Council (grant agreement number: 630763). This work was supported by the LOEWE program Ubiquitin Networks (Ub-Net) of the State of Hesse (Germany). The Institute of Molecular Biology (IMB) is funded by the Boehringer Ingelheim Foundation.

REFERENCES

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* 422, 198–207. doi: 10.1038/nature01511

Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426, 570–574. doi: 10.1038/nature02166

Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* 404, 939–965. doi: 10.1007/s00216-012-6203-4

Beckett, D., Kovaleva, E., and Schatz, P. J. (1999). A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.* 8, 921–929. doi: 10.1110/ps.8.4.921

Brown, J. S., Lukashchuk, N., Sczaniecka-Clift, M., Britton, S., le Sage, C., Calsou, P., et al. (2015). Neddylation promotes ubiquitylation and release of Ku from DNA-damage sites. *Cell Rep.* 11, 704–714. doi: 10.1016/j.celrep.2015.03.058

Chapman-Smith, A., and Cronan, J. E. J. (1999). Molecular biology of biotin attachment to proteins. *J. Nutr.* 129, 477S–484S.

Choi-Rhee, E., Schulman, H., and Cronan, J. E. (2004). Promiscuous protein biotinylation by *Escherichia coli* biotin protein ligase. *Protein Sci.* 13, 3043–3050. doi: 10.1110/ps.04911804

Collins, B. C., Gillet, L. C., Rosenberger, G., Röst, H. L., Vichalkovski, A., Gstaiger, M., et al. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods* 10, 1246–1253. doi: 10.1038/nmeth.2703

Couzens, A. L., Knight, J. D. R., Kean, M. J., Teo, G., Weiss, A., Dunham, W. H., et al. (2013). Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci. Signal.* 6, rs15. doi: 10.1126/scisignal.2004712

Coyaud, E., Mis, M., Laurent, E. M. N., Dunham, W. H., Couzens, A. L., Robitaille, M., et al. (2015). BioID-based identification of Skp Cullin F-box (SCF)$^{\beta\text{-TrCP1/2}}$ E3 ligase substrates. *Mol. Cell. Proteomics* 14, 1781–1795. doi: 10.1074/mcp.M114.045658

Cronan, J. E. (2005). Targeted and proximity-dependent promiscuous protein biotinylation by a mutant Escherichia coli biotin protein ligase. *J. Nutr. Biochem.* 16, 416–418. doi: 10.1016/j.jnutbio.2005.03.017

Dingar, D., Kalkat, M., Chan, P.-K., Srikumar, T., Bailey, S. D., Tu, W. B., et al. (2015). BioID identifies novel c-MYC interacting partners in cultured cells and xenograft tumors. *J. Proteomics* 118, 95–111. doi: 10.1016/j.jprot.2014.09.029

Firat-Karalar, E. N., Rauniyar, N., Yates, J. R. III, and Stearns, T. (2014). Proximity interactions among centrosome components identify regulators of centriole duplication. *Curr. Biol.* 24, 664–670. doi: 10.1016/j.cub.2014.01.067

Foster, L. J., de Hoog, C. L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V. K., et al. (2006). A mammalian organelle map by protein correlation profiling. *Cell* 125, 187–199. doi: 10.1016/j.cell.2006.03.022

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 8, 645–654. doi: 10.1038/nrm2208

Hall, D. B., and Struhl, K. (2002). The VP16 activation domain interacts with multiple transcriptional components as determined by protein–protein cross-linking *in vivo*. *J. Biol. Chem.* 277, 46043–46050. doi: 10.1074/jbc.M208911200

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., et al. (2012). A census of human soluble protein complexes. *Cell* 150, 1068–1081. doi: 10.1016/j.cell.2012.08.011

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163, 712–723. doi: 10.1016/j.cell.2015.09.053

Hubner, N. C., Bird, A. W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., et al. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals *in vivo* protein interactions. *J. Cell Biol.* 189, 739–754. doi: 10.1083/jcb.200911091

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., et al. (2015). The BioPlex network: a systematic exploration of the human interactome. *Cell* 162, 425–440. doi: 10.1016/j.cell.2015.06.043

Kaake, R. M., Milenković, T., Przulj, N., Kaiser, P., and Huang, L. (2010). Characterization of cell cycle specific protein interaction networks of the yeast 26S proteasome complex by the QTAX strategy. *J. Proteome Res.* 9, 2016–2029. doi: 10.1021/pr1000175

Klockenbusch, C., and Kast, J. (2010). Optimization of formaldehyde cross-linking for protein interaction analysis of non-tagged integrin beta1. *J. Biomed. Biotechnol.* 2010, 927585. doi: 10.1155/2010/927585

Kristensen, A. R., Gsponer, J., and Foster, L. J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* 9, 907–909. doi: 10.1038/nmeth.2131

Lam, S. S., Martell, J. D., Kamer, K. J., Deerinck, T. J., Ellisman, M. H., Mootha, V. K., et al. (2015). Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods* 12, 51–54. doi: 10.1038/nmeth.3179

Lambert, J.-P., Ivosev, G., Couzens, A. L., Larsen, B., Taipale, M., Lin, Z.-Y., et al. (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat. Methods* 10, 1239–1245. doi: 10.1038/nmeth.2702

Lambert, J.-P., Tucholska, M., Go, C., Knight, J. D. R., and Gingras, A.-C. (2015). Proximity biotinylation and affinity purification are complementary approaches for the interactome mapping of chromatin-associated protein complexes. *J. Proteomics* 118, 81–94. doi: 10.1016/j.jprot.2014.09.011

Larance, M., and Lamond, A. I. (2015). Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* 16, 269–280. doi: 10.1038/nrm3970

Martell, J. D., Deerinck, T. J., Sancak, Y., Poulos, T. L., Mootha, V. K., Sosinsky, G. E., et al. (2012). Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.* 30, 1143–1148. doi: 10.1038/nbt.2375

Meyer, K., and Selbach, M. (2015). Quantitative affinity purification mass spectrometry: a versatile technology to study protein–protein interactions. *Front. Genet.* 6:237. doi: 10.3389/fgene.2015.00237

Mosbech, A., Gibbs-Seymour, I., Kagias, K., Thorslund, T., Beli, P., Povlsen, L., et al. (2012). DVC1 (C1orf124) is a DNA damage–targeting p97 adaptor that promotes ubiquitin-dependent responses to replication blocks. *Nat. Struct. Mol. Biol.* 19, 1084–1092. doi: 10.1038/nsmb.2395

Nooren, I. M. A., and Thornton, J. M. (2003). Diversity of protein–protein interactions. *EMBO J.* 22, 3486–3492. doi: 10.1093/emboj/cdg359

Ong, S.-E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* 1, 252–262. doi: 10.1038/nchembio736

Pagliuca, F. W., Collins, M. O., Lichawska, A., Zegerman, P., Choudhary, J. S., and Pines, J. (2011). Quantitative proteomics reveals the basis for the biochemical specificity of the cell-cycle machinery. *Mol. Cell* 43, 406–417. doi: 10.1016/j.molcel.2011.05.031

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein–protein interactions: structural, functional, and network properties. *Structure* 18, 1233–1243. doi: 10.1016/j.str.2010.08.007

Rhee, H.-W., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A., et al. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* 339, 1328–1331. doi: 10.1126/science.1230593

Rodriguez-Fraticelli, A. E., Bagwell, J., Bosch-Fortea, M., Boncompain, G., Reglero-Real, N., Garcia-Leon, M. J., et al. (2015). Developmental regulation of apical endocytosis controls epithelial patterning in vertebrate tubular organs. *Nat. Cell Biol.* 17, 241–250. doi: 10.1038/ncb3106

Roux, K. J., Kim, D. I., and Burke, B. (2013). BioID: a screen for protein–protein interactions. *Curr. Protoc. Protein Sci.* 74, Unit 19.23. doi: 10.1002/0471140864.ps1923s74

Roux, K. J., Kim, D. I., Raida, M., and Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* 196, 801–810. doi: 10.1083/jcb.201112098

Satpathy, S., Wagner, S. A., Beli, P., Gupta, R., Kristiansen, T. A., Malinova, D., et al. (2015). Systems-wide analysis of BCR signalosomes and downstream phosphorylation and ubiquitylation. *Mol. Syst. Biol.* 11, 810. doi: 10.15252/msb.20145880

Scott, J. D., and Pawson, T. (2009). Cell signaling in space and time: where proteins come together and when they're apart. *Science* 326, 1220–1224. doi: 10.1126/science.1175668

Seet, B. T., Dikic, I., Zhou, M.-M., and Pawson, T. (2006). Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.* 7, 473–483. doi: 10.1038/nrm1960

Smith, A. L., Friedman, D. B., Yu, H., Carnahan, R. H., and Reynolds, A. B. (2011). ReCLIP (reversible cross-link immuno-precipitation): an efficient method for interrogation of labile protein complexes. *PLoS ONE* 6:e16206. doi: 10.1371/journal.pone.0016206

Tagwerker, C., Flick, K., Cui, M., Guerrero, C., Dou, Y., Auer, B., et al. (2006). A tandem affinity tag for two-step purification under fully denaturing conditions: application in ubiquitin profiling and protein complex identification combined with *in vivo* cross-linking. *Mol. Cell. Proteomics* 5, 737–748. doi: 10.1074/mcp.M500368-MCP200

Vasilescu, J., Guo, X., and Kast, J. (2004). Identification of protein–protein interactions using *in vivo* cross-linking and mass spectrometry. *Proteomics* 4, 3845–3854. doi: 10.1002/pmic.200400856

Vermeulen, M., Hubner, N. C., and Mann, M. (2008). High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. *Curr. Opin. Biotechnol.* 19, 331–337. doi: 10.1016/j.copbio.2008.06.001

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., et al. (2015). Panorama of ancient metazoan macromolecular complexes. *Nature* 525, 339–344. doi: 10.1038/nature14877

Zhou, Z., Rawnsley, D. R., Goddard, L. M., Pan, W., Cao, X.-J., Jakus, Z., et al. (2015). The cerebral cavernous malformation pathway controls cardiac development via regulation of endocardial MEKK3 signaling and KLF expression. *Dev. Cell* 32, 168–180. doi: 10.1016/j.devcel.2014.12.009

Conflict of Interest Statement: 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.

frontiers in Genetics

MINI REVIEW
published: 14 July 2015
doi: 10.3389/fgene.2015.00237

Quantitative affinity purification mass spectrometry: a versatile technology to study protein–protein interactions

*Katrina Meyer and Matthias Selbach**

Proteome Dynamics, Max Delbrück Center for Molecular Medicine, Berlin, Germany

While the genomic revolution has dramatically accelerated the discovery of disease-associated genes, the functional characterization of the corresponding proteins lags behind. Most proteins fulfill their tasks in complexes with other proteins, and analysis of protein–protein interactions (PPIs) can therefore provide insights into protein function. Several methods can be used to generate large-scale protein interaction networks. However, most of these approaches are not quantitative and therefore cannot reveal how perturbations affect the network. Here, we illustrate how a clever combination of quantitative mass spectrometry with different biochemical methods provides a rich toolkit to study different aspects of PPIs including topology, subunit stoichiometry, and dynamic behavior.

Keywords: mass spectrometry based proteomics, quantitative proteomics, protein–protein interaction, stoichiometry, cross-linking

OPEN ACCESS

Edited by:
Miguel Andrade,
Johannes-Gutenberg University of Mainz, Germany

Reviewed by:
Ina Hoeschele,
Virginia Polytechnic Institute and State University, USA
Scott H. Harrison,
North Carolina A&T State University, USA

Correspondence:
Matthias Selbach,
Proteome Dynamics, Max Delbrück Center for Molecular Medicine,
Robert-Rössle-Street 10,
D-13092 Berlin, Germany
matthias.selbach@mdc-berlin.de

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

Received: *27 March 2015*
Accepted: *25 June 2015*
Published: *14 July 2015*

Citation:
Meyer K and Selbach M (2015) Quantitative affinity purification mass spectrometry: a versatile technology to study protein–protein interactions. Front. Genet. 6:237. doi: 10.3389/fgene.2015.00237

Introduction

Proteins do not act in isolation but typically mediate their biological functions by interacting with other proteins (Charbonnier et al., 2008). Owing to the central importance of protein–protein interactions (PPIs) in biology, methods have been developed to study multiple aspects of PPIs (Meyerkord and Fu, 2015). For example, X-ray crystallography and NMR provide detailed spatial information about interaction interfaces. Surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), and förster resonance energy transfer (FRET) provide binding affinities and kinetics. However, all of those methods require *a priori* knowledge of the interaction partners and suffer from the drawback of a low throughput. Technologies like protein microarrays, phage display and the yeast two-hybrid system permit high-throughput screens for PPIs. However, these approaches rely on *in vitro* assays or heterologous biological systems. Therefore, it is not clear if PPIs detected by these methods occur in the relevant *in vivo* context.

Affinity purification combined with mass spectrometry (AP-MS) has emerged as a particularly attractive method for PPI mapping (Gingras et al., 2007). A major advantage is that this method allows unbiased detection of PPIs under physiological conditions. Importantly, AP-MS can assess PPIs in relevant biological contexts such as mammalian cell lines or even tissues. Moreover, AP-MS experiments have the advantage that they can provide quantitative information (q-AP-MS). This greatly increases the confidence in interaction partners that are identified and can also be used to study the impact of perturbations on PPIs.

We argue that q-AP-MS is one of the most powerful technologies to map PPIs in health and disease. The aim of this Mini Review is to briefly explain the general principle of q-AP-MS and to emphasize the versatility of AP-MS to investigate various aspects of PPIs including quantities, topology, subunit stoichiometry, and dynamic behavior. We will begin with a brief introduction to quantitative shotgun proteomics.

Quantitative Shotgun Proteomics

The principle idea of shotgun proteomics is that protein samples are first digested into peptides (Aebersold and Mann, 2003). These peptides are then separated by high performance liquid chromatography (HPLC) and directly ("online") transferred into a mass spectrometer. This instrument performs two important tasks. First, it measures the mass to charge ratios (m/z) and intensity of the peptides eluting from the HPLC column (MS^1). Second, in order to determine the amino acid sequence, the instruments selects individual peptides for fragmentation and records the resulting fragment spectra (MS^2). Data generated in this manner is then compared to protein databases for peptide and protein identification (Eng et al., 2011).

Until a decade ago, the field of proteomics has used mass spectrometry mainly to draw qualitative conclusions about the existence of a protein in a given sample. The reason for this is that the intensities of peaks in a mass spectrum are not directly proportional to the amounts of the corresponding peptides. Hence, mass spectrometry is intrinsically not a quantitative technology (Ong and Mann, 2005). However, over the past several years various technologies have been developed to enable proteome-wide quantification using mass spectrometry (Gstaiger and Aebersold, 2009; Cox and Mann, 2011; Bantscheff et al., 2012). One idea relies on the incorporation of stable heavy isotopes into proteins through metabolic (SILAC) or chemical labeling approaches. This permits different cell populations to be mixed and analyzed together, since the mass-shift introduced by the labeling makes them distinguishable. Relative changes in peptide intensities reflect differences in the abundance of the proteins under distinct experimental conditions. Alternatively, proteins can be quantified using computational methods ("label-free quantification"; **Figure 1B**). This may be based solely on how often peptides have been chosen for fragmentation (spectral counting) or on all intensities obtained from precursor peptide scans. Care should be taken when employing the first approach, since it provides only very rough abundance estimates (Rinner et al., 2007; Gingras and Raught, 2012). While the choice of a quantification approach depends on various factors, stable isotope-based methods are generally more precise than label free approaches since samples can be combined and analyzed together (Sury et al., 2010; Lau et al., 2014). For example, while stable isotope-based methods can detect even minor changes in protein abundance, label free approaches typically require a twofold change or more (Cox et al., 2014).

Specificity and Sensitivity

A major challenge in AP-MS is to distinguish true interaction partners from non-specific contaminants. An early idea to address this problem was tandem affinity purification (TAP; Puig et al., 2001). Here, the protein of interest is expressed as a fusion with two different biochemical tags. Two consecutive rounds of affinity purification are then employed in order to remove non-specific contaminants. Although this approach has been used successfully in many studies, it has two major disadvantages. First, only very stable complexes survive the procedure, which means that TAP cannot be used to study more dynamic interactions. Second, the sensitivity of modern mass spectrometers is so high that they still detect many non-specific binders after TAP. An alternative idea is to use a single purification step and to exclude non-specific contaminants based on prior knowledge. The "contaminant repository for affinity purification" (CRAPome) was built for this purpose and contains information about frequently observed unspecific binders (Mellacheruvu et al., 2013). While this is generally a good idea, one important limitation is that the non-specific background depends on specific experimental conditions. In other words, not all proteins in the CRAPome are necessarily contaminants in a specific experiment, nor are all contaminants in a specific experiment contained in the CRAPome.

Quantitative proteomics offers an attractive solution to address these challenges (**Figure 1A**). In quantitative AP-MS (q-AP-MS), the quantity of proteins that co-purify with the bait is compared to a negative control (Vermeulen et al., 2008; Paul et al., 2011). In this set-up, true interaction partners can be identified by their specific abundance ratio while non-specific contaminants bind equally well under both conditions, which results in a 1:1 ratio. Hence, q-AP-MS uses quantification to filter out non-specific contaminants. This greatly increases confidence in identified interaction partners, even under mild biochemical purification conditions.

Perturbations

One of the major advantages of q-AP-MS is that it can assess dynamic changes in PPIs upon perturbation (**Figure 1C**). To this end, the proteins which co-purify with a bait protein under normal and perturbed conditions are compared in a quantitative manner. An early example of this general principle employed the immobilized SH2-domain of the adapter protein Grb2 to study epidermal growth factor (EGF) receptor signaling (Blagoev et al., 2003). SH2 domains interact with specific tyrosine-phosphorylated motifs. Therefore, the immobilized domain was used in cells stimulated with EGF to pull down interacting proteins. Cells that had not been stimulated served as a negative control. Subsequently, a quantitative comparison of the two pull-down contexts revealed proteins recruited to Grb2 upon activation by EGF. After this pioneering work, the same idea was used to assess dynamic PPIs during cell signaling with different experimental designs. For example, immobilized peptides carrying specific posttranslational modifications and their unmodified counterparts were used to identify modification-dependent interactions (Selbach et al., 2009; Bartke et al., 2010; Francavilla et al., 2013). Immunoprecipitation of endogenous or epitope-tagged proteins before and after stimulation has also been frequently employed (Collins et al., 2013; Zheng et al., 2013; Sury et al., 2015). Finally, quantification can reveal differences in the interaction partners of wild-type

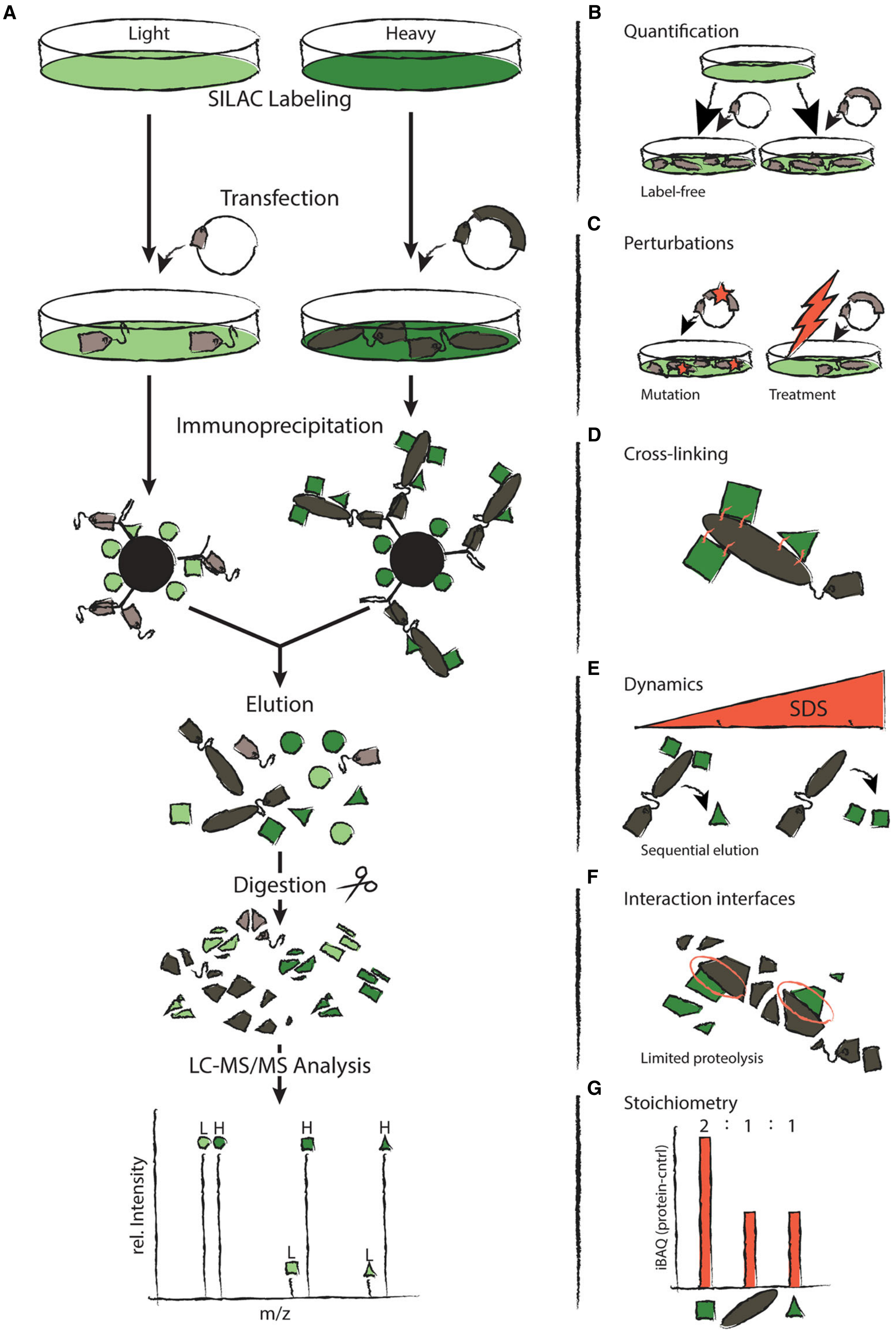


FIGURE 1 | Approaches to q-AP-MS experiments. (A) The left hand side depicts the typical workflow of a SILAC-based q-AP-MS experiment. Differentially SILAC-labeled cells are transfected with a tagged protein of interest or a control vector containing only the tag, respectively. Proteins are immunoprecipitated with antibodies directed against the tag. Samples are mixed prior to elution. Eluted proteins are cleaved into peptides and analyzed by Liquid-Chromatography Mass Spectrometry (LC-MS). **(B–G)** The right hand side depicts how q-AP-MS can be employed to study different aspects of PPIs. **(B)** Label-free quantification provides an alternative to SILAC. **(C)** Immunoprecipitation can compare changes in PPIs upon perturbation. **(D)** Transient interactions and complex structure can be studied by cross-linking. **(E)** Submodule composition and PPI dynamics can be revealed by sequential elution with increasing concentrations of SDS. **(F)** Limited proteolysis provides a means to detect interaction interfaces. **(G)** The stoichiometry of complexes can be revealed by comparing abundances of the different subunits.

proteins and disease-associated variants (Lambert et al., 2013; Hosp et al., 2015). If those mutations map to a protein with unknown function, an AP-MS experiment can provide valuable insights based on the known functions of identified interaction partners.

In vivo Interactions

Many AP-MS studies make use of overexpressed and/or tagged proteins as baits. However, this may interfere with the normal *in vivo* function of the protein and thus lead to false-positive or false-negative results. Overexpression artifacts can be limited when tagged proteins are expressed at near-endogenous levels, for example using bacterial artificial chromosomes (Hubner et al., 2010). However, it is still possible that the tag interferes with protein function. It has been shown recently that even cloning scars between the protein and the tag can lead to false-positive identifications (Banks et al., 2015). This problem can be addressed by targeting the endogenous protein with specific antibodies. While this has been employed successfully (Malovannaya et al., 2011; Lundby et al., 2014), an important caveat is that antibody cross-reactivity may lead to false-positive results. In case of tagged proteins the specificity can be assessed using untransfected cells as negative controls, but this is not possible when the endogenous protein is targeted. To address this issue, many published studies have used control antibodies. However, due to differences in the cross-reactivity of various antibodies, this strategy is questionable. A better control is to knock down the protein of interest in the control condition, which makes it possible to use the same antibody for comparison (Selbach and Mann, 2006). Nevertheless, the lack of good antibodies is an important limitation and one of the reasons why epitope-tagged proteins still dominate such studies.

Another important consideration is that the interaction partners identified in cell lines may not necessarily be relevant *in vivo*. More and more studies therefore purify proteins and their interaction partners directly from animal models (Cheeseman et al., 2004; Angrand et al., 2006; Bartoi et al., 2010; Rees et al., 2011; Hanack et al., 2015). With the advent of genome editing techniques such as CRISPR it is now possible to generate genomic tag knock-ins in an efficient manner (Sander and Joung, 2014). This makes it much easier to create tagged versions of endogenous proteins for *in vivo* interactome mapping and tissue culture experiments. Most of the methods discussed here are generally applicable to any organism. Even the SILAC approach, which was originally developed for metabolic labeling of tissue culture cells, has since been extended to a number of model organisms (Kirchner and Selbach, 2012). Thus, we expect that *in vivo* interaction proteomics will become more widespread.

Cross-linking

Upon cell lysis, proteins are brought into an artificial environment. This can result in the loss of weak or transient interactions or the formation of *in vitro* interactions in the lysate. One way to address this problem is *in vivo* cross-linking (Kaake et al., 2014; **Figure 1D**). Newly formed covalent bonds between interacting proteins permit stringent purification conditions which minimize *in vitro* interactions and preserve transient interactions (Tardiff et al., 2007; Fang et al., 2012). Moreover, the identification of cross-linked peptides can provide valuable information about the structure of proteins and complexes (Rappsilber, 2011; Walzthoeni et al., 2013). Despite these advantages, most AP-MS experiments performed today do not employ cross-linking. One reason is that cross-linked peptides are typically less abundant and are thus more difficult to identify than regular peptides. To address this problem, several strategies that enrich for cross-linked peptides have been developed (Rinner et al., 2008; Nessen et al., 2009).

Interaction Interfaces

Cross-linking requires that target sites be accessible, which makes it difficult to apply this approach to interfaces buried within a protein complex. This limitation is actually used as an advantage in several other methods to provide information about interaction interfaces. For example, protein painting employs small molecular dyes which adhere to the accessible surfaces of protein complexes, excluding binding interfaces (Luchini et al., 2014). During the subsequent digestion, only peptides within interaction interfaces are accessible to trypsin and can thus be identified. Limited proteolysis (Feng et al., 2014) is an approach that is complementary to protein painting, in that it reveals only peptides outside interaction interfaces that are accessible to trypsin (**Figure 1F**). Another possibility is to treat samples with heavy (i.e., deuterated) water: hydrogen-deuterium exchange (HDX; Mandell et al., 2005) relies on the fact that amides hidden within protein–protein interfaces are not in direct contact with the solvent and will exchange their hydrogen atoms at a lower rate than more accessible amides. The corresponding changes in the peptide mass can then be detected using mass spectrometry. These techniques are not only useful in the study of PPIs but can additionally provide information about protein structure (Chorev et al., 2015).

Stoichiometry

The approaches mentioned above typically rely on relative quantification. Thus, they can be used to distinguish specific interaction partners from contaminants and to quantify dynamic changes in PPIs upon perturbation. However, these methods can only compare the same protein under different conditions. They do not provide information about the stoichiometry of the distinct members of a complex. One way to compare different proteins in a complex is to measure their absolute abundances using synthetic isotope-labeled reference peptides as spike-in standards (Schmidt et al., 2010). For a large number of proteins, this is tedious and expensive. The SH-quant approach therefore incorporates an additional reference peptide into the affinity tag that is used for the pull-down (Wepf et al., 2009). This permits quantification of the bait and also of prey proteins, in the event they have been used as baits in another experiment. This "correlational quantification" allows the measurement of protein complex stoichiometry and absolute protein complex abundances. Alternatively, the stoichiometry of protein complexes can also be analyzed through a combination of affinity purification and intensity-based absolute quantification (iBAQ; **Figure 1G**; Schwanhausser et al., 2011; Smits et al., 2013). The latter approach has the advantage that it is easy to implement and does not require the tagging of multiple baits. It is also important to keep in mind that the same bait protein can be part of multiple protein complexes. Therefore, not all proteins that co-purify with a bait are necessarily members of the same complex. Distinguishing between these different complexes requires the individual pull-down of all components.

Dynamic Interactions

Not all of the specific interaction partners of a protein necessarily belong to a stable complex. Some interaction partners interact only transiently. The dynamic behavior of proteins can be investigated by mixing protein samples at different stages of an AP-MS experiment. Metabolic labeling approaches such as SILAC allow a mixing of samples directly after cells are harvested (Ong et al., 2002). While this minimizes experimental differences in sample handling, it also results in the loss of dynamic interactions with high on/off rates: During incubation with antibodies, these dynamic interaction partners will be exchanged between both conditions and reach equilibrium over time. Alternatively, samples may first be mixed after affinity purification. When both protocols are performed in parallel on the same samples, the data can be used to identify the dynamic components in protein complexes (Mousson et al., 2008; Wang and Huang, 2008). A related idea uses increasing concentrations of SDS to elute precipitated proteins sequentially (**Figure 1E**; Texier et al., 2014). These data can be used to dissect the submodular composition of complexes due to their different binding properties.

Binding affinity is a particularly relevant quantity with regard to characterizing the interaction between two proteins. Typically, binding affinities are measured using methods such as ITC or SPR assays which require considerable quantities of purified proteins. q-AP-MS experiments can also be designed in a way to provide information about binding affinities (Sharma et al., 2009): First, a known quantity of an immobilized bait is incubated with cell extracts to pull down interactors. Next, the supernatant from this experiment is used in a second pull-down with the same bait. The quantification of the proteins in both pull-downs can then be used to infer the dissociation constants of the interactions. While so far this technique has only been used to calculate equilibrium dissociation constants (K_ds) of proteins interacting with small molecules and peptides, it should be generally applicable to a range of ligands, including entire proteins, used as baits.

Conclusions

The examples described above show that a combination of quantitative shotgun proteomics with various biochemical methods can provide a rich toolkit to explore various aspects of PPIs. This can be employed to (i) identify binding partners with high specificity, (ii) assess the stoichiometry of complexes, (iii) provide information about interaction interfaces, (iv) analyze binding affinities, and (v) study dynamic changes of PPIs upon perturbation. Bearing in mind possible pitfalls (Duncan et al., 2010), mass spectrometers can thus be regarded as “Swiss army knives” for PPI research. Since instruments are becoming faster, more sensitive, easier to operate and cheaper, we expect these approaches to become available to more and more scientists.

References

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* 422, 198–207. doi: 10.1038/nature01511

Angrand, P. O., Segura, I., Volkel, P., Ghidelli, S., Terry, R., Brajenovic, M., et al. (2006). Transgenic mouse proteomics identifies new 14-3-3-associated proteins involved in cytoskeletal rearrangements and cell signaling. *Mol. Cell. Proteomics* 5, 2211–2227. doi: 10.1074/mcp.M600147-MCP200

Banks, C. A., Boanca, G., Lee, Z. T., Florens, L., and Washburn, M. P. (2015). Proteins interacting with cloning scars: a source of false positive protein–protein interactions. *Sci. Rep.* 5, 8530. doi: 10.1038/srep08530

Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* 404, 939–965. doi: 10.1007/s00216-012-6203-4

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* 143, 470–484. doi: 10.1016/j.cell.2010.10.012

Bartoi, T., Rigbolt, K. T., Du, D., Kohr, G., Blagoev, B., and Kornau, H. C. (2010). GABAB receptor constituents revealed by tandem affinity purification from transgenic mice. *J. Biol. Chem.* 285, 20625–20633. doi: 10.1074/jbc.M109.049700

Blagoev, B., Kratchmarova, I., Ong, S. E., Nielsen, M., Foster, L. J., and Mann, M. (2003). A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nat. Biotechnol.* 21, 315–318. doi: 10.1038/nbt790

Charbonnier, S., Gallego, O., and Gavin, A. C. (2008). The social network of a cell: recent advances in interactome mapping. *Biotechnol. Annu. Rev.* 14, 1–28. doi: 10.1016/S1387-2656(08)00001-X

Cheeseman, I. M., Niessen, S., Anderson, S., Hyndman, F., Yates, J. R. III, Oegema, K., et al. (2004). A conserved protein network controls assembly of the outer kinetochore and its ability to sustain tension. *Genes Dev.* 18, 2255–2268. doi: 10.1101/gad.1234104

Chorev, D. S., Ben-Nissan, G., and Sharon, M. (2015). Exposing the subunit diversity and modularity of protein complexes by structural mass spectrometry approaches. *Proteomics* doi: 10.1002/pmic.201400517 [Epub ahead of print].

Collins, B. C., Gillet, L. C., Rosenberger, G., Rost, H. L., Vichalkovski, A., Gstaiger, M., et al. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods* 10, 1246–1253. doi: 10.1038/nmeth.2703

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 13, 2513–2526. doi: 10.1074/mcp.M113.031591

Cox, J., and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* 80, 273–299. doi: 10.1146/annurev-biochem-061308-093216

Duncan, M. W., Aebersold, R., and Caprioli, R. M. (2010). The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* 28, 659–664. doi: 10.1038/nbt0710-659

Eng, J. K., Searle, B. C., Clauser, K. R., and Tabb, D. L. (2011). A face in the crowd: recognizing peptides through database search. *Mol. Cell. Proteomics* 10, R111 009522. doi: 10.1074/mcp.R111.009522

Fang, L., Kaake, R. M., Patel, V. R., Yang, Y., Baldi, P., and Huang, L. (2012). Mapping the protein interaction network of the human COP9 signalosome complex using a label-free QTAX strategy. *Mol. Cell. Proteomics* 11, 138–147. doi: 10.1074/mcp.M111.016352

Feng, Y., De Franceschi, G., Kahraman, A., Soste, M., Melnik, A., Boersema, P. J., et al. (2014). Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* 32, 1036–1044. doi: 10.1038/nbt.2999

Francavilla, C., Rigbolt, K. T., Emdal, K. B., Carraro, G., Vernet, E., Bekker-Jensen, D. B., et al. (2013). Functional proteomics defines the molecular switch underlying FGF receptor trafficking and cellular outputs. *Mol. Cell.* 51, 707–722. doi: 10.1016/j.molcel.2013.08.002

Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 8, 645–654. doi: 10.1038/nrm2208

Gingras, A. C., and Raught, B. (2012). Beyond hairballs: the use of quantitative mass spectrometry data to understand protein–protein interactions. *FEBS Lett.* 586, 2723–2731. doi: 10.1016/j.febslet.2012.03.065

Gstaiger, M., and Aebersold, R. (2009). Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* 10, 617–627. doi: 10.1038/nrg2633

Hanack, C., Moroni, M., Lima, W. C., Wende, H., Kirchner, M., Adelfinger, L., et al. (2015). GABA Blocks Pathological but Not Acute TRPV1 Pain Signals. *Cell* 160, 759–770. doi: 10.1016/j.cell.2015.01.022

Hosp, F., Vossfeldt, H., Heinig, M., Vasiljevic, D., Arumughan, A., Wyler, E., et al. (2015). Quantitative interaction proteomics of neurodegenerative disease proteins. *Cell Rep.* 11, 1134–1146. doi: 10.1016/j.celrep.2015.04.030

Hubner, N. C., Bird, A. W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., et al. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals *in vivo* protein interactions. *J. Cell Biol.* 189, 739–754. doi: 10.1083/jcb.200911091

Kaake, R. M., Wang, X., Burke, A., Yu, C., Kandur, W., Yang, Y., et al. (2014). A new *in vivo* cross-linking mass spectrometry platform to define protein–protein interactions in living cells. *Mol. Cell. Proteomics* 13, 3533–3543. doi: 10.1074/mcp.M114.042630

Kirchner, M., and Selbach, M. (2012). *In vivo* quantitative proteome profiling: planning and evaluation of SILAC experiments. *Methods Mol. Biol.* 893, 175–199. doi: 10.1007/978-1-61779-885-6_13

Lambert, J. P., Ivosev, G., Couzens, A. L., Larsen, B., Taipale, M., Lin, Z. Y., et al. (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat. Methods* 10, 1239–1245. doi: 10.1038/nmeth.2702

Lau, H. T., Suh, H. W., Golkowski, M., and Ong, S. E. (2014). Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics. *J. Proteome Res.* 13, 4164–4174. doi: 10.1021/pr500630a

Luchini, A., Espina, V., and Liotta, L. A. (2014). Protein painting reveals solvent-excluded drug targets hidden within native protein-protein interfaces. *Nat. Commun.* 5, 4413. doi: 10.1038/ncomms5413

Lundby, A., Rossin, E. J., Steffensen, A. B., Acha, M. R., Newton-Cheh, C., Pfeufer, A., et al. (2014). Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics. *Nat. Methods* 11, 868–874. doi: 10.1038/nmeth.2997

Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., et al. (2011). Analysis of the human endogenous coregulator complexome. *Cell* 145, 787–799. doi: 10.1016/j.cell.2011.05.006

Mandell, J. G., Baerga-Ortiz, A., Falick, A. M., and Komives, E. A. (2005). Measurement of solvent accessibility at protein-protein interfaces. *Methods Mol. Biol.* 305, 65–80. doi: 10.1385/1-59259-912-5:065

Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J. P., St-Denis, N. A., Li, T., et al. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* 10, 730–736. doi: 10.1038/nmeth.2557

Meyerkord, C. L., and Fu, H. (2015). "Methods in Molecular Biology," in *Protein-Protein Interactions: Methods and Applications,* 1st Edn, eds C. L. Meyerkord and H. Fu (New York: Humana Press), 620. doi: 10.1007/978-1-4939-2425-7

Mousson, F., Kolkman, A., Pijnappel, W. W., Timmers, H. T., and Heck, A. J. (2008). Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes. *Mol. Cell. Proteomics* 7, 845–852. doi: 10.1074/mcp.M700306-MCP200

Nessen, M. A., Kramer, G., Back, J., Baskin, J. M., Smeenk, L. E., De Koning, L. J., et al. (2009). Selective enrichment of azide-containing peptides from complex mixtures. *J. Proteome Res.* 8, 3702–3711. doi: 10.1021/pr900257z

Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., et al. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 1, 376–386. doi: 10.1074/mcp.M200025-MCP200

Ong, S. E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* 1, 252–262. doi: 10.1038/nchembio736

Paul, F. E., Hosp, F., and Selbach, M. (2011). Analyzing protein–protein interactions by quantitative mass spectrometry. *Methods* 54, 387–395. doi: 10.1016/j.ymeth.2011.03.001

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., et al. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* 24, 218–229. doi: 10.1006/meth.2001.1183

Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* 173, 530–540. doi: 10.1016/j.jsb.2010.10.014

Rees, J. S., Lowe, N., Armean, I. M., Roote, J., Johnson, G., Drummond, E., et al. (2011). *In vivo* analysis of proteomes and interactomes using Parallel Affinity Capture (iPAC) coupled to mass spectrometry. *Mol. Cell. Proteomics* 10, M110 002386. doi: 10.1074/mcp.M110.002386

Rinner, O., Mueller, L. N., Hubalek, M., Muller, M., Gstaiger, M., and Aebersold, R. (2007). An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat. Biotechnol.* 25, 345–352. doi: 10.1038/nbt1289

Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L. N., Beck, M., Schmidt, A., et al. (2008). Identification of cross-linked peptides from large sequence databases. *Nat. Methods* 5, 315–318. doi: 10.1038/nmeth.1192

Sander, J. D., and Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* 32, 347–355. doi: 10.1038/nbt.2842

Schmidt, C., Lenz, C., Grote, M., Luhrmann, R., and Urlaub, H. (2010). Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. *Anal. Chem.* 82, 2784–2796. doi: 10.1021/ac902710k

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global quantification of mammalian gene expression control. *Nature* 473, 337–342. doi: 10.1038/nature10098

Selbach, M., and Mann, M. (2006). Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat. Methods* 3, 981–983. doi: 10.1038/nmeth972

Selbach, M., Paul, F. E., Brandt, S., Guye, P., Daumke, O., Backert, S., et al. (2009). Host cell interactome of tyrosine-phosphorylated bacterial proteins. *Cell Host. Microbe.* 5, 397–403. doi: 10.1016/j.chom.2009.03.004

Sharma, K., Weber, C., Bairlein, M., Greff, Z., Keri, G., Cox, J., et al. (2009). Proteomics strategy for quantitative protein interaction profiling in cell extracts. *Nat. Methods* 6, 741–744. doi: 10.1038/nmeth.1373

Smits, A. H., Jansen, P. W., Poser, I., Hyman, A. A., and Vermeulen, M. (2013). Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res.* 41, e28. doi: 10.1093/nar/gks941

Sury, M. D., Chen, J. X., and Selbach, M. (2010). The SILAC fly allows for accurate protein quantification *in vivo*. *Mol. Cell. Proteomics* 9, 2173–2183. doi: 10.1074/mcp.M110.000323

Sury, M. D., Mcshane, E., Hernandez-Miranda, L. R., Birchmeier, C., and Selbach, M. (2015). Quantitative proteomics reveals dynamic interaction of c-Jun N-terminal kinase (JNK) with RNA transport granule proteins splicing factor proline- and glutamine-rich (Sfpq) and non-POU domain-containing octamer-binding protein (Nono) during neuronal differentiation. *Mol. Cell. Proteomics* 14, 50–65. doi: 10.1074/mcp.M114.039370

Tardiff, D. F., Abruzzi, K. C., and Rosbash, M. (2007). Protein characterization of *Saccharomyces cerevisiae* RNA polymerase II after *in vivo* cross-linking. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19948–19953. doi: 10.1073/pnas.0710179104

Texier, Y., Toedt, G., Gorza, M., Mans, D. A., Van Reeuwijk, J., Horn, N., et al. (2014). Elution profile analysis of SDS-induced subcomplexes by quantitative mass spectrometry. *Mol. Cell. Proteomics* 13, 1382–1391. doi: 10.1074/mcp.O113.033233

Vermeulen, M., Hubner, N. C., and Mann, M. (2008). High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. *Curr. Opin. Biotechnol.* 19, 331–337. doi: 10.1016/j.copbio.2008.06.001

Walzthoeni, T., Leitner, A., Stengel, F., and Aebersold, R. (2013). Mass spectrometry supported determination of protein complex structure. *Curr. Opin. Struct. Biol.* 23, 252–260. doi: 10.1016/j.sbi.2013.02.008

Wang, X., and Huang, L. (2008). Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. *Mol. Cell. Proteomics* 7, 46–57. doi: 10.1074/mcp.M700261-MCP200

Wepf, A., Glatter, T., Schmidt, A., Aebersold, R., and Gstaiger, M. (2009). Quantitative interaction proteomics using mass spectrometry. *Nat. Methods* 6, 203–205. doi: 10.1038/nmeth.1302

Zheng, Y., Zhang, C., Croucher, D. R., Soliman, M. A., St-Denis, N., Pasculescu, A., et al. (2013). Temporal regulation of EGF signalling networks by the scaffold protein Shc1. *Nature* 499, 166–171. doi: 10.1038/nature12308

Conflict of Interest Statement: 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.

MINI REVIEW
published: 04 May 2016
doi: 10.3389/fgene.2016.00074

Current Approaches Toward Quantitative Mapping of the Interactome

Alexander Buntru[†]*, Philipp Trepte*[†]*, Konrad Klockmeier, Sigrid Schnoegl and Erich E. Wanker**

Max Delbrueck Center for Molecular Medicine, Berlin, Germany

Protein–protein interactions (PPIs) play a key role in many, if not all, cellular processes. Disease is often caused by perturbation of PPIs, as recently indicated by studies of missense mutations. To understand the associations of proteins and to unravel the global picture of PPIs in the cell, different experimental detection techniques for PPIs have been established. Genetic and biochemical methods such as the yeast two-hybrid system or affinity purification-based approaches are well suited to high-throughput, proteome-wide screening and are mainly used to obtain qualitative results. However, they have been criticized for not reflecting the cellular situation or the dynamic nature of PPIs. In this review, we provide an overview of various genetic methods that go beyond qualitative detection and allow quantitative measuring of PPIs in mammalian cells, such as dual luminescence-based co-immunoprecipitation, Förster resonance energy transfer or luminescence-based mammalian interactome mapping with bait control. We discuss the strengths and weaknesses of different techniques and their potential applications in biomedical research.

Keywords: PPI analysis, FRET, DULIP, FCCS, PLA, Interactome Mapping, BiFC, LUMIER, BRET, Quantification of protein-protein interactions

OPEN ACCESS

Edited by:
Spyros Petrakis,
Centre for Research
and Technology – Hellas, Greece

Reviewed by:
Pablo Meyer,
International Business Machines
Corporation, USA
Tero Aittokallio,
Institute for Molecular Medicine
Finland, Finland

****Correspondence:***
Erich E. Wanker
ewanker@mdc-berlin.de

[†]*These authors have contributed equally to this work.*

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

Received: *23 December 2015*
Accepted: *18 April 2016*
Published: *04 May 2016*

Citation:
Buntru A, Trepte P, Klockmeier K, Schnoegl S and Wanker EE (2016) Current Approaches Toward Quantitative Mapping of the Interactome. Front. Genet. 7:74. doi: 10.3389/fgene.2016.00074

INTRODUCTION

Physical interactions between proteins are crucial to most biological processes. Hence, major efforts have been made to systematically identify protein–protein interactions (PPIs) using the yeast two-hybrid (Y2H) system and affinity purification–mass spectrometry (AP/MS) approaches (Stelzl et al., 2005; Yu et al., 2008; Guruharsha et al., 2011). However, these methods are mainly suited for providing qualitative data, especially at the large scale. For a more comprehensive functional description of interactions, additional information is required. Knowledge of interaction strength, e.g., is of particular importance. It informs us of binding affinities and lifetimes of protein complexes, which are critical for the dynamic regulation of cellular systems (Perkins et al., 2010; Hieb et al., 2012). In summary, a better understanding of complex cellular processes not only requires knowledge of which proteins interact but also of the characteristics of interactions. To obtain such insight, quantitative experimental techniques for the detection of PPIs in mammalian cells have moved into focus (Hieb et al., 2012; Chen et al., 2015). These include biochemical methods such as quantitative affinity-purification and mass spectrometry (qAP–MS; Hosp et al., 2015) or genetic methods such as using luminescence-based mammalian interactome mapping with bait control (LUMIER with BACON; Taipale et al., 2014). Using qAP–MS, e.g., the association of proteins with neurodegenerative disease proteins such as amyloid precursor protein (APP), presenilin-1 and ataxin-1 (ATXN-1) have been quantitatively analyzed and the effects of disease-causing mutations on PPIs have been systematically assessed in pull-down assays (Hosp et al., 2015). The quantitative investigation of PPIs using LUMIER with BACON revealed a comprehensive Hsp90–client interaction network, which provided insight into previously unknown organization principles of functional chaperone modules in mammalian cells (Taipale et al., 2014).

A recent study suggests that about 60% of disease-causing mutations in proteins influence their association with other proteins. It was estimated that half of those mutations leads to a complete loss of protein interactions while the other half only perturbs a particular subset of interactions (Sahni et al., 2015). A pathological poly-glutamine expansion in ATXN-1, causally related to spinocerebellar ataxia type 1 (SCA1), e.g., was found to induce binding of the protein to RBM17 rather than CiC, thereby promoting disease (Lim et al., 2008). To detect such changes in affinity and to map how interaction profiles of individual proteins are changed through mutations, methods that allow quantitative PPI analysis are urgently needed.

However, the available methodologies do not yet permit a full quantitative assessment of PPIs at the cellular level. Current methods to study binary PPIs in mammalian cells can broadly be classified in two groups. Assays like bimolecular fluorescence complementation (BiFC), bimolecular luminescence complementation (BiLC) and proximity ligation assay (PLA) yield a quantitative readout without allowing conclusions about interaction strengths, while assays like Förster resonance energy transfer (FRET), bioluminescence resonance energy transfer (BRET), fluorescence cross-correlation spectroscopy (FCCS), dual luminescence-based co-immunoprecipitation (DULIP) and LUMIER with BACON provide a quantitative readout that can be used to determine binding strengths. In this paper, we will review recent developments in quantitative PPI detection technologies and provide an overview of relevant applications of these methods in biomedical research. We focus on genetic approaches in mammalian cells, as mass spectrometry-based methods have been recently reviewed elsewhere (Meyer and Selbach, 2015). Protein microarrays also provide important insights on PPIs and can provide quantitative readouts (MacBeath and Schreiber, 2000; Jones et al., 2006). They also have been reviewed elsewhere and will not be discussed here (Wolf-Yadlin et al., 2009).

An overview of the discussed methods and their capabilities is provided in **Table 1**.

FLUORESCENCE CROSS-CORRELATION SPECTROSCOPY

Fluorescence correlation spectroscopy (FCS) was described for the first time over 40 years ago (Magde et al., 1974; Macháň, 2014). It was developed to measure chemical reaction rates and diffusion coefficients by analyzing the thermodynamic fluctuations in the fluorescence intensity of a system. FCS is now a well-established biophysical method, which in combination with confocal microscopy is routinely used to obtain quantitative information about the abundance of fluorescently tagged proteins

TABLE 1 | Overview of capabilities of binary PPI detection methods in mammalian cells.

Method	FCCS	BiFC/BiLC	PLA	FRET	BRET	LUMIER	DULIP
PPI assay principle	Co-migration of proteins	Protein fragment complementation	Proximity-based ligation of oligonucleotides	Förster resonance energy transfer	Bioluminescence resonance energy transfer	Co-immunoprecipitation	Co-immunoprecipitation
PPI read-out	Fluorescence	Fluorescence/ luminescence	Fluorescence/ luminescence	Fluorescence	Luminescence	Luminescence	Luminescence
Quantification of binding strengths	Yes	No	No	Yes	Yes	Yes	Yes
Detection of PPIs in intact cells	Yes	Yes	No	Yes	Yes	No	No
Detection of PPIs in lysed cells	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subcellular localization of PPIs	No	Yes	Yes	Yes	Yes	No	No
Detection of endogenous PPIs	No	No	Yes	No	No	No	No
Expression of tagged proteins required	Yes	Yes	No	Yes	Yes	Yes	Yes
Throughput of PPI detection assay	Medium	High	Low	High	High	High	High

Several methods can be used for high-throughput interaction mapping. Fluorescence and luminescence play a prominent role in current approaches.

in living cells (Macháň, 2014). Through the expansion of the method to dual-color FCCS it became possible to quantify interactions of labeled proteins *in vivo* under physiological conditions (Schwille et al., 1997).

Fluorescence cross-correlation spectroscopy allows the measurement of protein mobility, concentration and interactions by exploiting the temporal fluorescence fluctuations of two fluorescently labeled particles under a confocal laser scanning microscope diffusing through a minute focal volume (**Figure 1A**). As a distinct number of fluorescently labeled molecules diffuse through the focal volume (Haustein, 2014), the fluorescence signals fluctuate in a manner dependent on the mobility and concentration of the investigated proteins. An autocorrelation function of the fluctuating fluorescence signals provides the diffusion coefficients and concentrations of molecules. Importantly, FCCS utilizes two spectrally different fluorophores, e. g., monomerized green or red fluorescent proteins, to label a pair of proteins (Bacia et al., 2006). If the differently labeled proteins are associated with each other, they pass through the effective volume in a synchronized way. This causes simultaneous fluctuation of their fluorescence signals leading to an increase in the amplitude of the cross-correlation function, allowing the determination of *in vivo* interaction strengths for proteins of interest (Boeke et al., 2014). However, co-migration does not fully prove a direct binary interaction of two-labeled molecules. Thus, validation with other methods that are more stringent in this regard is necessary (Shi et al., 2009).

Quantitative *in vivo* FCCS analysis, e.g., revealed binding strengths for PPIs involved in the extracellular signal-regulated kinase/mitogen-activated protein kinase (ERK/MAPK) pathway (Sadaie et al., 2014). The generated quantitative data was utilized to perform computer-assisted simulations to model the ERK-/MAPK-signaling cascade, uncovering that Shc binding to EGFR is critical for the regulation of the pathway. Similarly, systematic FCCS studies of 41 PPIs revealed important information about the regulation of clathrin-mediated endocytosis in yeast (Boeke et al., 2014). Through the *in vivo* measurement of interaction strengths for selected interactions likely to be involved in endocytosis the protein Ede1 was discovered as a crucial scaffold for the organization of this process. These results highlight the application power of FCCS for quantitative detection of PPIs in cells and show that quantitative PPI information improves our current understanding of signal transduction pathways. Through the systematic application of FCCS it seems feasible that comprehensive, quantitative interactome maps can be generated in the future.

BIMOLECULAR COMPLEMENTATION METHODS: PROTEIN-FRAGMENT COMPLEMENTATION ASSAY (PCA), BiFC, AND BiLC

Protein-fragment complementation assays have been utilized for a long time to detect PPIs in yeast or mammalian cells (Johnsson and Varshavsky, 1994; Kerppola, 2006). PCAs are classical reporter assays, in which a fluorescent protein or enzyme, e.g., is split in two and the parts are then fused to the N- or C-terminal end of the potential interactors. If the proteins of interest interact, the fragments unite, emitting measurable fluorescence or displaying quantifiable enzymatic activity. Different PCA variants have been used successfully in small- or proteome-scale applications to detect PPIs (Tarassov et al., 2008; Sung et al., 2013; Petschnigg et al., 2014).

One of the most commonly used PCA variants is the BiFC assay, which is based on the reconstitution of a fluorescent protein such as the green fluorescent protein (GFP) (Kerppola, 2008; Kodama and Hu, 2012). In an application of BiFC, an N-terminal GFP fragment (NGFP) containing the first 157 amino acids and a C-terminal GFP fragment (CGFP) fragment containing 81 terminal amino acids were fused to peptides that are known to assemble into antiparallel leucine zippers (Ghosh et al., 2000). The interaction of the peptides led to the reconstitution of functional GFP molecules that exhibited a single excitation maximum at 475 nm and an emission maximum at 506 nm. Today, multiple BiFC assays with many different split fluorescent proteins (FPs) are available for application, including ECFP, EGFP, EYFP (**Figure 1B**), Venus, Citrine, Cerulean, or mCherry [reviewed in Miller et al., 2015]. However, all PPI detection assays with split-FP variants suffer from spontaneous self-assembly of the utilized fragments, which results in relatively high background fluorescence in cells. To overcome this limitation, variants of the Venus-based BiFC PPI detection system with an improved signal-to-noise ratio were developed (Kodama and Hu, 2010). Another disadvantage of most if not all currently available BiFC methods is that split–FP fusions form irreversible protein complexes *in vitro* and *in vivo*, which can lead to false-positive results. Also, they only allow measuring the association of protein complexes but not their dissociation (Ciruela et al., 2010).

A related PCA is the BiLC assay, which uses luciferases rather than fluorescent proteins (**Figure 1C**). As for BiFC, several variants of the method are available that utilize different luciferases such as firefly (Paulmurugan et al., 2002), *Renilla* (Paulmurugan and Gambhir, 2003), *Gaussia* (Remy and Michnick, 2006), or NanoLuc® (Dixon et al., 2015). Importantly, the reconstitution of the luciferase fragments is reversible in these assays, allowing the detection of both association and dissociation of fusion proteins in living cells in real-time (Remy and Michnick, 2006). Compared to BiFC, BiLC assays offer a higher signal-to-noise ratio, which is very advantageous for the large-scale detection of PPIs in cells. BiLC assays were also utilized to study the localization of PPIs in cells (Kaihara et al., 2003). However, due to the relatively low number of emitted photons this can be a very challenging task (Kato, 2012).

The power of BiFC-based PPI detection methods lies in their ability to identify weak or transient interactions in cells (Miller et al., 2015). This is because fusion proteins are stabilized in complementation assays and fluorescent signals are only observed when the tagged fusions interact directly. The relatively weak interaction between the SH3 domain of c-Abl and the

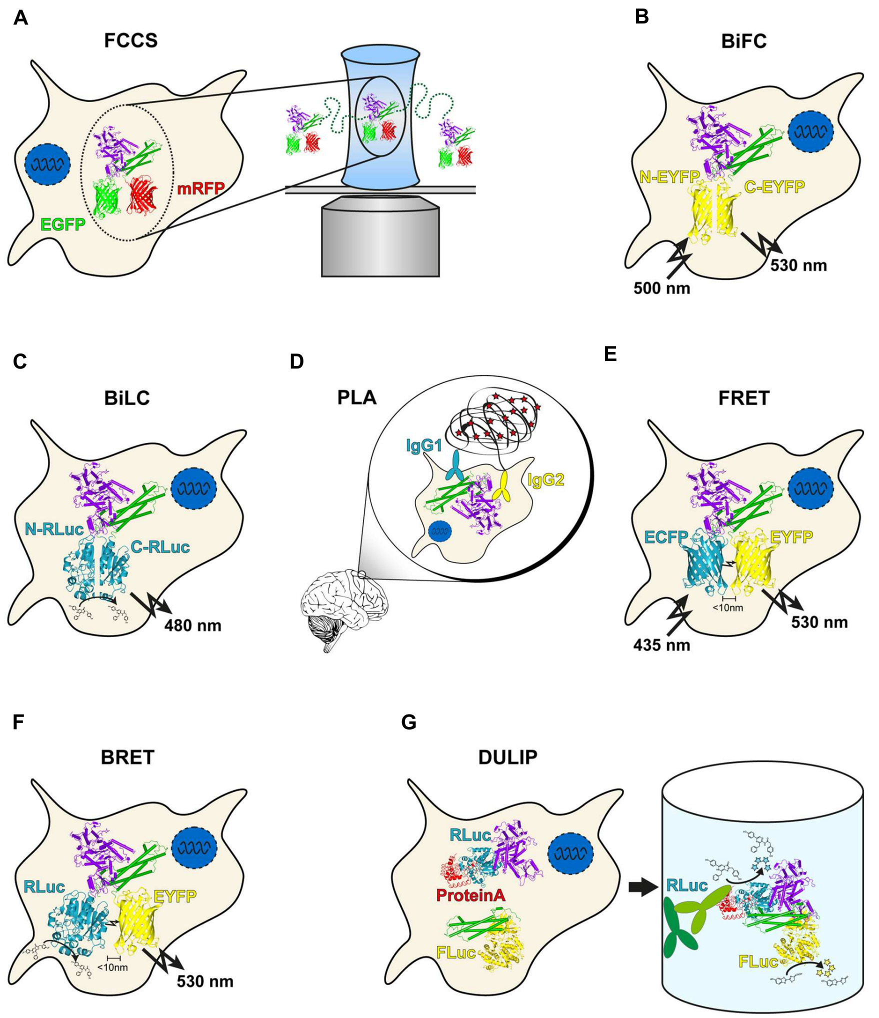


FIGURE 1 | Overview of genetic protein–protein interaction (PPI) methods. (A) In Fluorescence cross-correlation spectroscopy (FCCS) measurements, co-migration of two fluorescently labeled molecules through a focal volume is quantified. **(B)** bimolecular fluorescence complementation (BiFC) utilizes two non-fluorescent fragments of EGFP or a variant. Upon interaction of the two labeled proteins, the fragments can reassociate, resulting in fluorescence. **(C)** The principle of bimolecular luminescence complementation (BiLC) is similar to BiFC but is based on two fragments of a luciferase. In contrast to BiFC, the reassociation is reversible. **(D)** Close proximity of two DNA oligomer-labeled antibodies allows circularization of two additional oligomers after hybridization. The product is amplified in a rolling circle reaction and subsequently detected with fluorescently labeled probes. **(E)** During Förster resonance energy transfer (FRET), energy is transferred non-radiatively from an excited donor molecule to an acceptor molecule. In case the acceptor is also a fluorophore, the transmitted energy is emitted at a longer wavelength (the so called sensitized emission). **(F)** bioluminescence resonance energy transfer (BRET) is similar to FRET with the difference that a luciferase serves as a donor molecule. **(G)** In dual luminescence-based co-immunoprecipitation (DULIP) assays, two proteins of interest are fused to firefly or *Renilla* luciferase, respectively. An additional PA-tag allows precipitation of the bait protein from the lysate. If an interaction occurs, co-precipitation of the prey protein is indicated by luminescence arising from the firefly luciferase.

poly-proline peptide p41 ($K_d = 1.5\ \mu M$), e.g., could be readily detected in intact cells using a YFP-based BiFC assay (Morell et al., 2007). Recently, a recombinase enhanced bimolecular luciferase complementation (ReBiL) platform was established that allows the detection of low-affinity PPIs in living cells. It enabled the discovery of the interaction between the E3 ubiquitin ligase FANCL and the ubiquitin-conjugating enzyme UBE2T ($K_d = 0.454\ \mu M$), two key players in DNA repair processes (Li et al., 2014).

BiFC and BiLC allow rapid, sensitive investigation of PPIs in cells with a quantitative data output both in focused experiments as well as on the proteome scale (Sung and Huh, 2010). To assess binding affinities of interacting proteins in cells, however, both unbound and bound protein molecules would need to be quantified. This is not possible with BiFC or BiLC assays because only interacting fusion proteins show fluorescence or luminescence complementation (**Figures 1B,C**). Finally, it is important to note that the lack of information on unbound FPs in BiFC assays may lead to false positive as well as false negative results in systematic PPI screenings, simply because the expected bait and prey fusions may not be properly expressed in cells.

PROXIMITY LIGATION ASSAYS

The proximity ligation assay utilizes antibodies to which short single-stranded DNA oligonucleotides, often termed PLA probes, have been attached (Fredriksson et al., 2002; Söderberg et al., 2006; Weibrecht et al., 2010). When bound to two proteins that are in close proximity in biological systems (distance < 30 nm), these antibody–DNA probes facilitate the ligation of additional DNA molecules by ligases and subsequent amplification by polymerase chain reaction or a rolling circle mechanism. The amplified DNA molecules function as templates for the binding of fluorescently labeled oligonucleotide probes that act as surrogate markers for interacting proteins (**Figure 1D**). The dual recognition by PLA probes required for the formation of DNA reporter molecules decreases non-specific signals because only ligated reporters are amplified (Weibrecht et al., 2010).

Proximity ligation assays have the advantage over methods like BiFC or FCCS that associations between proteins can be identified and quantified without additional tags. The only requirement is the availability of specific, high-affinity antibodies against the proteins of interest that can be modified with DNA oligonucleotides. In the last 10 years, multiple variants of PLAs have been developed, which can be applied to the detection of protein–protein, protein–DNA, and protein–RNA interactions (Swartzman et al., 2010; Hansen et al., 2014). Furthermore, the method was adapted for the identification of interactions dependent on post translational modifications. Recently, e.g., an SH2-PLA was established, which allows the quantification of interactions between an SH2 domain and phosphotyrosines in the EGFR using a microtiter plate format (Thompson et al., 2015). This method, which is highly sensitive and has a large dynamic range, has a wide array of applications both in basic and translational cancer research. Similarly, an *in situ* PLA variant was successfully applied to detect the Erα/Src/PI3K protein complex in breast cancer cells and patient samples (Poulard et al., 2014), suggesting that the method has the potential to be utilized as diagnostic tool.

Although several studies have generated quantitative information about PPIs using PLAs, e.g., through secondary methods like color segmentation image analysis (Gajadhar and Guha, 2010; Leuchowius et al., 2010; Pacchiana et al., 2014), the currently available variants cannot be utilized to define binding strengths of interactions. To obtain such information, knowledge about the abundance of both bound and unbound protein molecules would be required. However, PLAs remain powerful tools to validate interactions initially identified in high-throughput screens under physiological conditions.

FRET-BASED METHODS

The fundamental theory of FRET was established in the first half of the 20th century (Cario and Franck, 1922). Its great potential for biological research, however, has only been realized in the past 20 years, after different techniques had been developed that allowed the application of FRET in biological systems (Mills et al., 2003; Wallrabe and Periasamy, 2005; Ma et al., 2014). This, in particular, includes the combination of FRET with microscopy techniques, which allow the investigation of PPIs with temporal and spatial resolution *in vivo* (Sun et al., 2013). FRET is a distance-dependent process in which, through dipole–dipole interactions, an exited fluorophore molecule (the donor) transfers energy non-radiatively to another fluorophore molecule (the acceptor), resulting in acceptor emission (Lakowicz, 2013). Alternatively, dark quenchers can be used as acceptors for studying, e.g., membrane–protein interactions (Cho et al., 2016). There are three main conditions that need to be met for efficient FRET: (i) there must be "spectral overlap" of the donor's emission and the acceptor's excitation spectra, (ii) the donor and acceptor fluorophores (termed FRET pair) must be in close proximity and (iii) the dipoles of the donor and acceptor must be aligned (Lakowicz, 2013). Due to the fact that FRET efficiency is proportional to the inverse of the sixth power of the distance between the donor and the acceptor, only fluorophores that are in very close proximity (<10 nm) show FRET (Clegg, 1995). Thus, FRET allows the detection of direct interactions between proteins, whereas methods such as FCCS, PLA, DULIP, or LUMIER with BACON cannot distinguish between proteins that directly interact or are only present in the same complex (Li et al., 2015).

To measure FRET with microscopic techniques several basic approaches have been developed. This includes acceptor photobleaching (Szabà et al., 1992), fluorescence life-time imaging microscopy (Wallrabe and Periasamy, 2005), spectral imaging (Chen, 2011), and sensitized emission, which still is the most commonly applied FRET method. Sensitized emission measurements can be performed using standard confocal and wide-field microscopes with appropriate filters or fluorescence microplate readers. Three channels are normally required for the imaging of donor, acceptor and FRET signals. The sensitized emission method, also called three-cube FRET, is based on the detection of acceptor fluorescence after donor excitation (Gordon et al., 1998; Mattheyses and Marcus, 2015). However, it is important to note that usually it is not possible to visualize sensitized emission directly due to contamination of the FRET signal by both donor and direct acceptor fluorescence. Thus, the measurement has to be corrected for donor bleed-through and acceptor cross-excitation, which can be performed through the calculation of calibration factors obtained from measurements with reference samples containing either donor or acceptor molecules alone (Mattheyses and Marcus, 2015). Currently, various algorithms are available to correct for these fluorescence contaminations, which all give comparable results (Zal and Gascoigne, 2004; Chen et al., 2006). Subsequent normalization to the donor or acceptor protein level (or a combination of both) provides a quantitative FRET signal (Hoppe et al., 2002; Zal and Gascoigne, 2004; Chen et al., 2006; Elder et al., 2009).

To study PPIs with FRET, the proteins of interest need to be tagged with appropriate donor and acceptor fluorophores. This is possible through the production of genetically encoded fusions with fluorescent protein tags in cells using multiple expression plasmids (Hochreiter et al., 2015). This includes FRET pairs such as ECFP/EYFP (**Figure 1E**), mTurquoise/mCitrine or EGFP/mCherry that are commonly applied for the investigation of PPIs in cells (Day and Davidson, 2012; Mattheyses and Marcus, 2015). A major strength of FRET-based interaction studies in living cells is that quantitative information about PPIs can be obtained. This is achieved through saturation experiments in which FRET is monitored in cells coexpressing a constant amount of donor-tagged protein with increasing amounts of acceptor-tagged protein or *vice versa* (Carriba et al., 2008; Martínez-Muñoz et al., 2014). Through such an approach, $FRET_{50}$ values can be calculated, which provide an indication about the binding strength of tagged interacting proteins. However, it needs to be noted that FRET measurements in living cells can provide information about binding affinities only when the absolute concentrations of investigated proteins are known. Such information, however, is generally not available without additional measurements in standard FRET-based PPI studies (Sun et al., 2013). Nevertheless, a recent study demonstrated that reliable *in vivo* binding affinities between the proteins glutathione (GSH) and glutathione-*S*-transferase (GST) can be obtained from FRET measurements in intact cells (Chen et al., 2015). Thus, FRET microscopy and spectroscopy are powerful techniques that can provide highly reliable information about the binding strengths of PPIs, even at subcellular resolution.

BRET-BASED METHODS

Bioluminescence resonance energy transfer is a biophysical technique that, similar to FRET, can be readily applied for quantifying PPI strengths in living cells (Pfleger and Eidne, 2006). One distinction between the two methods is that FRET involves energy transfer between two fluorophores, one of which requires extrinsic excitation by a suitable light source, whereas BRET occurs after oxidation of a substrate (e.g., coelenterazine) through a luciferase enzyme (**Figure 1F**). Previous studies indicate that different luciferase enzymes such as *Renilla* luciferase (Rluc) or NanoLuc in combination with various fluorophores (e.g., EYFP) are suitable for in-cell BRET experiments and for the quantification of PPIs using $BRET_{50}$ values (Hamdan et al., 2006; Szalai et al., 2014; Brown et al., 2015). The assembly of G protein-coupled receptors, e.g., was successfully studied in mammalian cells with the help of BRET (Stoddart et al., 2015). Furthermore, it was shown that a sequential BRET–FRET technique (termed SRET) is able to detect the interactions between three proteins *in vivo* (Carriba et al., 2008). Combined BRET and FRET methods are powerful tools to analyze the assembly of higher-order protein complexes and the effects of posttranslational modifications on PPIs. Recently, a BRET–FRET approach was applied to study the oligomerization of the proteins CCR5, CD4 and CXCR4, which are of critical importance for the infection of cells by HIV-1 (Martínez-Muñoz et al., 2014). Thus, novel fluorescence and luminescence-based methods allow the systematic quantitative analysis of protein complexes in cell models. They might be advanced for routine validation of PPIs identified in high-throughput screens with qualitative assays (Rolland et al., 2014).

LUCIFERASE-BASED CO-IMMUNOPRECIPITATION METHODS

Co-immunoprecipitation (Co-IP) is commonly used to detect PPIs in protein extracts (Phizicky and Fields, 1995). However, identifying interactions with Co-IPs is laborious and time consuming, making the method unsuitable for systematic screening. To overcome these limitations, a luminescence-based Co-IP assay – termed LUMIER – was developed, which provides at least semi-quantitative PPI information and can be performed in microtiter plates (Barrios-Rodiles et al., 2005). Here, bait and prey proteins are co-produced as FLAG and *Renilla* fusions in mammalian cells and interactions are detected by luciferase enzymatic assays in co-immunoprecipitates. LUMIER has the advantage that large numbers of bait/prey pairs can be systematically tested for putative interactions under relatively well-defined assay conditions. The method was successfully applied for the generation of a dynamic PPI network for the TGF beta pathway (Barrios-Rodiles et al., 2005) as well as for the identification of inhibitors of the Wnt pathway (Miller et al., 2009), indicating that it is suitable for the elucidation of novel signaling pathway components with high confidence.

The original LUMIER assay has the disadvantage that the FLAG-tagged bait proteins cannot be quantified in co-immunoprecipitates, which may lead to false negative results in large-scale PPI screenings. To overcome this limitation, an improved version of the LUMIER assay was recently established (Taipale et al., 2012, 2014), which was termed LUMIER with bait control (LUMIER with BACON). Here, the immunoprecipitated FLAG-tagged bait proteins are systematically quantified by ELISA. LUMIER with BACON, which can also be performed in microtiter plates, facilitates the calculation of quantitative interaction scores that can be used for hierarchical clustering of PPIs and the prediction of potential functional modules. Applying LUMIER with BACON, a quantitative chaperone interaction network was generated that enabled the identification of regulators of cellular proteostasis (Taipale et al., 2014).

A dual luciferase reporter pull-down (DLR-PD) assay for the detection of PPIs in mammalian cells was also reported (Jia et al., 2011). In this assay, bait and prey proteins are co-produced in cells as firefly and *Renilla* luciferase fusions, respectively. In addition, the expressed bait protein harbors a HAVI-tag that is recognized and biotinylated by the co-produced biotin-protein ligase BirA. The DLR-PD assay was shown to successfully detect nuclear and cytoplasmic PPIs in HEK293 cell lysates, suggesting that the method can be applied for PPI screening. However, pull-down assays with beads are not easy to scale up for high-throughput applications. To overcome this limitation, most recently a DULIP assay was developed for interactome mapping in mammalian cells (Trepte et al., 2015). This method can be performed in 384-well microtiter plates and can be automated for large-scale interaction screens (**Figure 1G**).

In DULIP assays the bait and prey proteins are co-produced as *Renilla* and firefly luciferase fusions in mammalian cells, respectively. In addition, the expressed bait protein harbors a protein A (PA) tag (Li, 2010) that allows the co-precipitation of bait/prey complexes in microtiter plates. The successful expression of bait and prey fusion proteins as well as the success of bait/prey co-precipitation can be quantified using DULIP. This enables the calculation of quantitative, normalized interaction ratios for all tested protein pairs, which can be utilized to create quantitative PPI interaction maps. The method, e.g., was capable of detecting the effects of point mutations on the interaction strength of synaptic proteins (Trepte et al., 2015), suggesting that it might be suitable for more comprehensive investigations of the effects of disease-causing mutations on PPIs. Taken together, luminescence-based assays are powerful PPI detection methods that, in the future, might allow us to obtain quantitative information about interactions in large-scale systematic studies.

CONCLUSIONS AND OUTLOOK

Resulting from multiple high-throughput PPI screening efforts with genetic and biochemical methods (Stelzl et al., 2005; Yu et al., 2008; Rolland et al., 2014), we currently possess large databases with unexplored interactions. Their further characterization requires quantitative experimental strategies that are easy to implement in laboratories and allow the identification of interactions at medium to high throughput in mammalian cells. Recent developments indicate that quantitative PPI information can be generated *in vivo* with methods such as FCCS, BRET, DULIP, or LUMIER with BACON (**Table 1**). This opens new avenues for interactomics researchers because the dynamics and strengths of PPIs can be assessed for the first time with these techniques. Also computational approaches to predict or filter PPIs relevant to a given question will profit enormously from direct prioritization of PPIs based on quantitative interaction data. It seems now possible to capture a broad range of high-, medium- and low-affinity interactions and to link this information to specific cellular processes. In the long run, this will enable us to describe the molecular principles of biological systems in more detail and to improve our understanding of disease processes. We suggest that truly quantitative interactome research is now within reach and efforts need to be intensified to obtain comprehensive quantitative PPI data sets in living cells.

AUTHOR CONTRIBUTIONS

AB, PT, KK, and EW wrote the initial manuscript. AB, PT, KK, SS, and EW revised the manuscript and approved the final version.

FUNDING

This work was supported by grants from the German Research Foundation (SFB740), the Federal Ministry of Education and Research (NGFN-Plus, NeuroNet, MooDS, Integrament, Mutanom), the European Union (EuroSpin and SynSys), and the Helmholtz Association (MSBN and HelMA) to EW. The relevant grant numbers are SFB740: 740/2-11, NeuroNet: 01GS08169-73, MooDS: 01GS08150, Integrament: 01ZX1314C, Mutanom: 01GS08108, EuroSpin: Health-F2-2009-241498, SynSys: HEALTH-F2- 2009-242167, and HelMA: HA-215. The funders had no role in the preparation of the manuscript.

REFERENCES

Bacia, K., Kim, S. A., and Schwille, P. (2006). Fluorescence cross-correlation spectroscopy in living cells. *Nat. Methods* 3, 83–89. doi: 10.1038/nmeth822

Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., et al. (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* 307, 1621–1625. doi: 10.1126/science.1105776

Boeke, D., Trautmann, S., Meurer, M., Wachsmuth, M., Godlee, C., Knop, M., et al. (2014). Quantification of cytosolic interactions identifies Ede1 oligomers as key organizers of endocytosis. *Mol. Syst. Biol.* 10, 756–756. doi: 10.15252/msb.20145422

Brown, N. E., Blumer, J. B., and Hepler, J. R. (2015). Bioluminescence resonance energy transfer to detect protein-protein interactions in live cells. *Methods Mol. Biol.* 1278, 457–465. doi: 10.1007/978-1-4939-2425-7_30

Cario, G., and Franck, J. (1922). Über zerlegung von wasserstoffmolekülen durch angeregtel quecksilberatome. *Z. Phys. A At. Nucl.* 11, 161–166. doi: 10.1007/BF01328410

Carriba, P., Navarro, G., Ciruela, F., Ferré, S., Casadó, V., Agnati, L., et al. (2008). Detection of heteromerization of more than two proteins by sequential BRET-FRET. *Nat. Methods* 5, 727–733. doi: 10.1038/nmeth.1229

Chen, H., Puhl, H. L., and Ikeda, S. R. (2006). Measurement of FRET efficiency and ratio of donor to acceptor concentration in living cells. *Biophys. J.* 91, L39–L41. doi: 10.1529/biophysj.106.088773

Chen, W., Avezov, E., Schlachter, S. C., Gielen, F., Laine, R. F., Harding, H. P., et al. (2015). A method to quantify FRET stoichiometry with phasor plot analysis and acceptor lifetime ingrowth. *Biophys. J.* 108, 999–1002. doi: 10.1016/j.bpj.2015.01.012

Chen, Y.-C. (2011). Spectral resolution in conjunction with polar plots improves the accuracy and reliability of FLIM measurements and estimates of FRET efficiency. *J. Microsci.* 244, 21–37. doi: 10.1111/j.1365-2818.2011.03488.x

Cho, W., Kim, H., and Hu, Y. (2016). High-throughput fluorometric assay for membrane-protein interaction. *Methods Mol. Biol.* 1376, 163–174. doi: 10.1007/978-1-4939-3170-5_14

Ciruela, F., Vilardaga, J.-P., and Fernández-Dueñas, V. (2010). Lighting up multiprotein complexes: lessons from GPCR oligomerization. *Trends Biotechnol.* 28, 407–415. doi: 10.1016/j.tibtech.2010.05.002

Clegg, R. M. (1995). Fluorescence resonance energy transfer. *Curr. Opin. Biotechnol.* 6, 103–110. doi: 10.1016/0958-1669(95)80016-6

Day, R. N., and Davidson, M. W. (2012). Fluorescent proteins for FRET microscopy: monitoring protein interactions in living cells. *Bioessays* 34, 341–350. doi: 10.1002/bies.201100098

Dixon, A. S., Schwinn, M. K., Hall, M. P., Zimmerman, K., Otto, P., Lubben, T. H., et al. (2015). NanoLuc complementation reporter optimized for accurate measurement of protein interactions in cells. *ACS Chem. Biol.* 11, 400–408. doi: 10.1021/acschembio.5b00753

Elder, A. D., Domin, A., Schierle, G. S. K., Lindon, C., Pines, J., Esposito, A., et al. (2009). A quantitative protocol for dynamic measurements of protein interactions by Förster resonance energy transfer-sensitized fluorescence emission. *J. R. Soc. Interface* 6, S59–S81. doi: 10.1098/rsif.2008.0381.focus

Fredriksson, S., Gullberg, M., Jarvius, J., Olsson, C., Pietras, K., Gústafsdóttir, S. M., et al. (2002). Protein detection using proximity-dependent DNA ligation assays. *Nat. Biotechnol.* 20, 473–477. doi: 10.1038/nbt0502-473

Gajadhar, A., and Guha, A. (2010). A proximity ligation assay using transiently transfected, epitope-tagged proteins: application for in situ detection of dimerized receptor tyrosine kinases. *Biotechniques* 48, 145–152. doi: 10.2144/000113354

Gordon, G. W., Berry, G., Liang, X. H., Levine, B., and Herman, B. (1998). Quantitative fluorescence resonance energy transfer measurements using fluorescence microscopy. *Biophys. J.* 74, 2702–2713.

Ghosh, I., Hamilton, A. D., and Regan, L. (2000). Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *J. Am. Chem. Soc.* 122, 5658–5659. doi: 10.1021/ja994421w

Guruharsha, K. G., Rual, J.-F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., et al. (2011). A protein complex network of *Drosophila melanogaster*. *Cell* 147, 690–703. doi: 10.1016/j.cell.2011.08.047

Hamdan, F. F., Percherancier, Y., Breton, B., and Bouvier, M. (2006). Monitoring protein-protein interactions in living cells by bioluminescence resonance energy transfer (BRET). *Curr. Protoc. Neurosci.* Chap. 5, Unit 5.23 doi: 10.1002/0471142301.ns0523s34

Hansen, M. C., Nederby, L., Henriksen, M. O.-B., Hansen, M., and Nyvold, C. G. (2014). Sensitive ligand-based protein quantification using immuno-PCR: a critical review of single-probe and proximity ligation assays. *Biotechniques* 56, 217–228.

Haustein, E. (2014). Fluorescence correlation spectroscopy: principles and applications. *Cold Spring Harb. Protoc.* 2014, 709–725. doi: 10.1101/pdb.top081802

Hieb, A. R., D'Arcy, S., Kramer, M. A., White, A. E., and Luger, K. (2012). Fluorescence strategies for high-throughput quantification of protein interactions. *Nucleic Acids Res.* 40:e33. doi: 10.1093/nar/gkr1045

Hochreiter, B., Garcia, A. P., and Schmid, J. A. (2015). Fluorescent proteins as genetically encoded FRET biosensors in life sciences. *Sensors* 15, 26281–26314. doi: 10.3390/s151026281

Hoppe, A., Christensen, K., and Swanson, J. A. (2002). Fluorescence resonance energy transfer-based stoichiometry in living cells. *Biophys. J.* 83, 3652–3664. doi: 10.1016/S0006-3495(02)75365-4

Hosp, F., Vossfeldt, H., Heinig, M., Vasiljevic, D., Arumughan, A., Wyler, E., et al. (2015). Quantitative interaction proteomics of neurodegenerative disease proteins. *Cell Rep.* 11, 1134–1146. doi: 10.1016/j.celrep.2015.04.030

Jia, S., Peng, J., Gao, B., Chen, Z., Zhou, Y., Fu, Q., et al. (2011). Relative quantification of protein-protein interactions using a dual luciferase reporter pull-down assay system. *PLoS ONE* 6:e26414. doi: 10.1371/journal.pone.0026414

Johnsson, N., and Varshavsky, A. (1994). Split ubiquitin as a sensor of protein interactions in vivo. *Proc. Natl. Acad. Sci. U.S.A.* 91, 10340–10344. doi: 10.1073/pnas.91.22.10340

Jones, R. B., Gordus, A., Krall, J. A., and Macbeath, G. (2006). A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* 439, 168–174. doi: 10.1038/nature04177

Kaihara, A., Kawai, Y., Sato, M., Ozawa, T., and Umezawa, Y. (2003). Locating a protein-protein interaction in living cells via split *Renilla* luciferase complementation. *Anal. Chem.* 75, 4176–4181. doi: 10.1021/ac0300800

Kato, N. (2012). Luciferase and bioluminescence microscopy for analyses of membrane dynamics in living cells. *J. Membr. Sci. Technol.* 2:e109. doi: 10.4172/2155-9589.1000e109

Kerppola, T. K. (2006). Visualization of molecular interactions by fluorescence complementation. *Nat. Rev. Mol. Cell Biol.* 7, 449–456. doi: 10.1038/nrm1929

Kerppola, T. K. (2008). Bimolecular fluorescence complementation (BiFC) analysis as a probe of protein interactions in living cells. *Annu. Rev. Biophys.* 37, 465–487. doi: 10.1146/annurev.biophys.37.032807.125842

Kodama, Y., and Hu, C.-D. (2010). An improved bimolecular fluorescence complementation assay with a high signal-to-noise ratio. *Biotechniques* 49, 793–805. doi: 10.2144/000113519

Kodama, Y., and Hu, C.-D. (2012). Bimolecular fluorescence complementation (BiFC): a 5-year update and future perspectives. *Biotechniques* 53, 285–298. doi: 10.2144/000113943

Lakowicz, J. R. (2013). *Principles of Fluorescence Spectroscopy*, 3rd Edn. Berlin: Springer. doi: 10.1007/978-0-387-46312-4

Leuchowius, K.-J., Jarvius, M., Wickström, M., Rickardson, L., Landegren, U., Larsson, R., et al. (2010). High content screening for inhibitors of protein interactions and post-translational modifications in primary cells by proximity ligation. *Mol. Cell. Proteomics* 9, 178–183. doi: 10.1074/mcp.M900331-MCP200

Li, X., Wang, W., and Chen, J. (2015). From pathways to networks: connecting dots by establishing protein-protein interaction networks in signaling pathways using affinity purification and mass spectrometry. *Proteomics* 15, 188–202. doi: 10.1002/pmic.201400147

Li, Y. (2010). Commonly used tag combinations for tandem affinity purification. *Biotechnol. Appl. Biochem.* 55, 73–83. doi: 10.1042/BA20090273

Li, Y.-C., Rodewald, L. W., Hoppmann, C., Wong, E. T., Lebreton, S., Safar, P., et al. (2014). A versatile platform to analyze low-affinity and transient protein-protein interactions in living cells in real time. *Cell Rep.* 9, 1946–1958. doi: 10.1016/j.celrep.2014.10.058

Lim, J., Crespo-Barreto, J., Jafar-Nejad, P., Bowman, A. B., Richman, R., Hill, D. E., et al. (2008). Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1. *Nature* 452, 713–718. doi: 10.1038/nature06731

Ma, L., Yang, F., and Zheng, J. (2014). Application of fluorescence resonance energy transfer in protein studies. *J. Mol. Struct.* 1077, 87–100. doi: 10.1016/j.molstruc.2013.12.071

MacBeath, G., and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763.

Macháň, R. (2014). Recent applications of fluorescence correlation spectroscopy in live systems. *FEBS Lett.* 588, 3571–3584. doi: 10.1016/j.febslet.2014.03.056

Magde, D., Elson, E. L., and Webb, W. W. (1974). Fluorescence correlation spectroscopy. II. An experimental realization. *Biopolymers* 13, 29–61. doi: 10.1002/bip.1974.360130103

Martínez-Muñoz, L., Barroso, R., Dyrhaug, S. Y., Navarro, G., Lucas, P., Soriano, S. F., et al. (2014). CCR5/CD4/CXCR4 oligomerization prevents HIV-1 gp120IIIB binding to the cell surface. *Proc. Natl. Acad. Sci. U.S.A.* 111, E1960–E1969. doi: 10.1073/pnas.1322887111

Mattheyses, A. L., and Marcus, A. I. (2015). Förster resonance energy transfer (FRET) microscopy for monitoring biomolecular interactions. *Methods Mol. Biol.* 1278, 329–339. doi: 10.1007/978-1-4939-2425-7_20

Meyer, K., and Selbach, M. (2015). Quantitative affinity purification mass spectrometry: a versatile technology to study protein-protein interactions. *Front. Genet.* 6:237. doi: 10.3389/fgene.2015.00237

Miller, B. W., Lau, G., Grouios, C., Mollica, E., Barrios-Rodiles, M., Liu, Y., et al. (2009). Application of an integrated physical and functional screening approach to identify inhibitors of the Wnt pathway. *Mol. Syst. Biol.* 5:315. doi: 10.1038/msb.2009.72

Miller, K. E., Kim, Y., Huh, W.-K., and Park, H.-O. (2015). Bimolecular fluorescence complementation (bifc) analysis: advances and recent applications for genome-wide interaction studies. *J. Mol. Biol.* 427, 2039–2055. doi: 10.1016/j.jmb.2015.03.005

Mills, J. D., Stone, J. R., Rubin, D. G., Melon, D. E., Okonkwo, D. O., Periasamy, A., et al. (2003). Illuminating protein interactions in tissue using confocal and two-photon excitation fluorescent resonance energy transfer microscopy. *J. Biomed. Opt.* 8, 347–356. doi: 10.1117/1.1584443

Morell, M., Espargaro, A., Avilés, F. X., and Ventura, S. (2007). Detection of transient protein-protein interactions by bimolecular fluorescence complementation: the Abl-SH3 case. *Proteomics* 7, 1023–1036. doi: 10.1002/pmic.200600966

Pacchiana, R., Abbate, M., Armato, U., Dal Prà, I., and Chiarini, A. (2014). Combining immunofluorescence with in situ proximity ligation assay: a novel imaging approach to monitor protein-protein interactions in relation to subcellular localization. *Histochem. Cell Biol.* 142, 593–600. doi: 10.1007/s00418-014-1244-8

Paulmurugan, R., and Gambhir, S. S. (2003). Monitoring protein-protein interactions using split synthetic *Renilla* luciferase protein-fragment-assisted complementation. *Anal. Chem.* 75, 1584–1589. doi: 10.1021/aco20731c

Paulmurugan, R., Umezawa, Y., and Gambhir, S. S. (2002). Noninvasive imaging of protein-protein interactions in living subjects by using reporter protein complementation and reconstitution strategies. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15608–15613. doi: 10.1073/pnas.242594299

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure* 18, 1233–1243. doi: 10.1016/j.str.2010.08.007

Petschnigg, J., Groisman, B., Kotlyar, M., Taipale, M., Zheng, Y., Kurat, C. F., et al. (2014). The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nat. Methods* 11, 585–592. doi: 10.1038/nmeth.2895

Pfleger, K. D. G., and Eidne, K. A. (2006). Illuminating insights into protein-protein interactions using bioluminescence resonance energy transfer (BRET). *Nat. Methods* 3, 165–174. doi: 10.1038/nmeth841

Phizicky, E. M., and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* 59, 94–123.

Poulard, C., Rambaud, J., Le Romancer, M., and Corbo, L. (2014). Proximity ligation assay to detect and localize the interactions of ERα with PI3-K and Src in breast cancer cells and tumor samples. *Methods Mol. Biol.* 1204, 135–143. doi: 10.1007/978-1-4939-1346-6_12

Remy, I., and Michnick, S. W. (2006). A highly sensitive protein-protein interaction assay based on Gaussia luciferase. *Nat. Methods* 3, 977–979. doi: 10.1038/nmeth979

Rolland, T., Ta An, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Sadaie, W., Harada, Y., Matsuda, M., and Aoki, K. (2014). Quantitative in vivo fluorescence cross-correlation analyses highlight the importance of competitive effects in the regulation of protein-protein interactions. *Mol. Cell. Biol.* 34, 3272–3290. doi: 10.1128/MCB.00087-14

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647–660. doi: 10.1016/j.cell.2015.04.013

Schwille, P., Meyer-Almes, F. J., and Rigler, R. (1997). Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution. *Biophys. J.* 72, 1878–1886. doi: 10.1016/S0006-3495(97)78833-7

Shi, X., Foo, Y. H., Sudhaharan, T., Chong, S.-W., Korzh, V., and Ahmed, S. (2009). Determination of dissociation constants in living zebrafish embryos with single wavelength fluorescence cross-correlation spectroscopy. *Biophys. J.* 97, 678–686. doi: 10.1016/j.bpj.2009.05.006

Söderberg, O., Gullberg, M., Jarvius, M., Ridderstråle, K., Leuchowius, K.-J., Jarvius, J., et al. (2006). Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nat. Methods* 3, 995–1000. doi: 10.1038/nmeth947

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029

Stoddart, L. A., White, C. W., Nguyen, K., Hill, S. J., and Pfleger, K. D. G. (2015). Fluorescence- and bioluminescence-based approaches to study GPCR ligand binding. *Br. J. Pharmacol.* doi: 10.1111/bph.13316 [Epub ahead of print].

Sun, Y., Rombola, C., Jyothikumar, V., and Periasamy, A. (2013). Förster resonance energy transfer microscopy and spectroscopy for localizing protein-protein interactions in living cells. *Cytometry A* 83, 780–793. doi: 10.1002/cyto.a.22321

Sung, M.-K., and Huh, W.-K. (2010). *In vivo* quantification of protein-protein interactions in *Saccharomyces cerevisiae* using bimolecular fluorescence complementation assay. *J. Microbiol. Methods* 83, 194–201. doi: 10.1016/j.mimet.2010.08.021

Sung, M.-K., Lim, G., Yi, D.-G., Chang, Y. J., Yang, E. B., Lee, K., et al. (2013). Genome-wide bimolecular fluorescence complementation analysis of SUMO interactome in yeast. *Genome Res.* 23, 736–746. doi: 10.1101/gr.148346.112

Swartzman, E., Shannon, M., Lieu, P., Chen, S.-M., Mooney, C., Wei, E., et al. (2010). Expanding applications of protein analysis using proximity ligation and qPCR. *Methods* 50, S23–S26. doi: 10.1016/j.ymeth.2010.01.024

Szabà, G., Pine, P. S., Weaver, J. L., Kasari, M., and Aszalos, A. (1992). Epitope mapping by photobleaching fluorescence resonance energy transfer measurements using a laser scanning microscope system. *Biophys. J.* 61, 661–670. doi: 10.1016/S0006-3495(92)81871-4

Szalai, B., Hoffmann, P., Prokop, S., Erdélyi, L., Várnai, P., and Hunyady, L. (2014). Improved methodical approach for quantitative BRET analysis of G Protein coupled receptor dimerization. *PLoS ONE* 9:e109503. doi: 10.1371/journal.pone.0109503

Taipale, M., Krykbaeva, I., Koeva, M., Kayatekin, C., Westover, K. D., Karras, G. I., et al. (2012). Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell* 150, 987–1001. doi: 10.1016/j.cell.2012.06.047

Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., Lin, Z.-Y., Larsen, B., et al. (2014). A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. *Cell* 158, 434–448. doi: 10.1016/j.cell.2014.05.039

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Serna Molina, M. M., Shames, I., et al. (2008). An in vivo map of the yeast protein interactome. *Science* 320, 1465–1470. doi: 10.1126/science.1153878

Thompson, C. M., Bloom, L. R., Ogiue-Ikeda, M., and Machida, K. (2015). SH2-PLA: a sensitive in-solution approach for quantification of modular domain binding by proximity ligation and real-time PCR. *BMC Biotechnol.* 15:60. doi: 10.1186/s12896-015-0169-1

Trepte, P., Buntru, A., Klockmeier, K., Willmore, L., Arumughan, A., Secker, C., et al. (2015). DULIP: a dual luminescence-based co-immunoprecipitation assay for interactome mapping in mammalian cells. *J. Mol. Biol* 427, 3355–3388. doi: 10.1016/j.jmb.2015.08.003

Wallrabe, H., and Periasamy, A. (2005). Imaging protein molecules using FRET and FLIM microscopy. *Curr. Opin. Biotechnol.* 16, 19–27. doi: 10.1016/j.copbio.2004.12.002

Weibrecht, I., Leuchowius, K.-J., Clausson, C.-M., Conze, T., Jarvius, M., Howell, W. M., et al. (2010). Proximity ligation assays: a recent addition to the proteomics toolbox. *Expert Rev. Proteomics* 7, 401–409. doi: 10.1586/epr.10.10

Wolf-Yadlin, A., Sevecka, M., and Macbeath, G. (2009). Dissecting protein function and signaling using protein microarrays. *Curr. Opin. Chem. Biol.* 13, 398–405. doi: 10.1016/j.cbpa.2009.06.027

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 104–110. doi: 10.1126/science.1158684

Zal, T., and Gascoigne, N. R. J. (2004). Photobleaching-corrected FRET efficiency imaging of live cells. *Biophys. J.* 86, 3923–3939. doi: 10.1529/biophysj.103.022087

Conflict of Interest Statement: 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.

frontiers in Genetics

MINI REVIEW
published: 23 September 2015
doi: 10.3389/fgene.2015.00296

Mining protein interactomes to improve their reliability and support the advancement of network medicine

Gregorio Alanis-Lobato [1, 2]*

[1] Faculty of Biology, Institute of Molecular Biology, Johannes Gutenberg University of Mainz, Mainz, Germany, [2] Integrative Systems Biology Lab, Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

OPEN ACCESS

Edited by:
Spyros Petrakis,
Aristotle University of Thessaloniki,
Greece

Reviewed by:
Anshu Bhardwaj,
Council of Scientific and Industrial Research, India
Maria I. Klapa,
Foundation for Research & Technology Hellas, Greece

***Correspondence:**
Gregorio Alanis-Lobato,
Institute of Molecular Biology,
Ackermannweg 4, 55128 Mainz,
Germany
g.alanis-lobato@imb-mainz.de

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

Received: *29 May 2015*
Accepted: *07 September 2015*
Published: *23 September 2015*

Citation:
Alanis-Lobato G (2015) Mining protein interactomes to improve their reliability and support the advancement of network medicine.
Front. Genet. 6:296.
doi: 10.3389/fgene.2015.00296

High-throughput detection of protein interactions has had a major impact in our understanding of the intricate molecular machinery underlying the living cell, and has permitted the construction of very large protein interactomes. The protein networks that are currently available are incomplete and a significant percentage of their interactions are false positives. Fortunately, the structural properties observed in good quality social or technological networks are also present in biological systems. This has encouraged the development of tools, to improve the reliability of protein networks and predict new interactions based merely on the topological characteristics of their components. Since diseases are rarely caused by the malfunction of a single protein, having a more complete and reliable interactome is crucial in order to identify groups of inter-related proteins involved in disease etiology. These system components can then be targeted with minimal collateral damage. In this article, an important number of network mining tools is reviewed, together with resources from which reliable protein interactomes can be constructed. In addition to the review, a few representative examples of how molecular and clinical data can be integrated to deepen our understanding of pathogenesis are discussed.

Keywords: interactome, proteome, network, reliability, prediction, medicine, disease, pathogenesis

1. Introduction

The existence of living cells is not possible without organized and coordinated communication between proteins. Failure of the control mechanisms that underlie these delicate relationships can lead to disease or even death (Lesk, 2007). This highlights that the study of the complex network of interactions between proteins is crucial to improve our understanding of the intricate mechanisms that make life possible (Lesk, 2007; Loscalzo and Barabasi, 2011). To ease the analysis of this involved biological machine, it is commonly represented as a network of nodes, linked to each other if there is evidence of their physical or functional relationship.

Today we have access to vast Protein Interaction Networks (PINs) from different organisms, due to high-throughput experimental techniques that are often an improved variation of yeast-two-hybrid screenings, or of co-immunoprecipitation followed by mass spectrometry (Vidal et al., 2011). Nevertheless, these networks are incomplete and contain a significant number of false positive interactions (Kuchaiev et al., 2009). However, it is fortunate that their structural properties are not different from those observed in good quality social or technological networks (Albert and Barabási, 2002; Liu et al., 2011; Cannistraci et al., 2013a) (**Figure 1A**). These topological similarities have prompted the development of tools, based on node-connectivity properties, aimed at improving the reliability and completeness of complex networks (Cannistraci et al., 2013a).

The reliability indices and predictions resulting from the application of these methods can be integrated with other sources of high-quality protein interactions (PIs). With these, one can construct reliable PINs that, together with clinical and genetic data, represent the fundamental pieces of information used in the emerging field of network medicine (Barabási et al., 2011; Loscalzo and Barabasi, 2011).

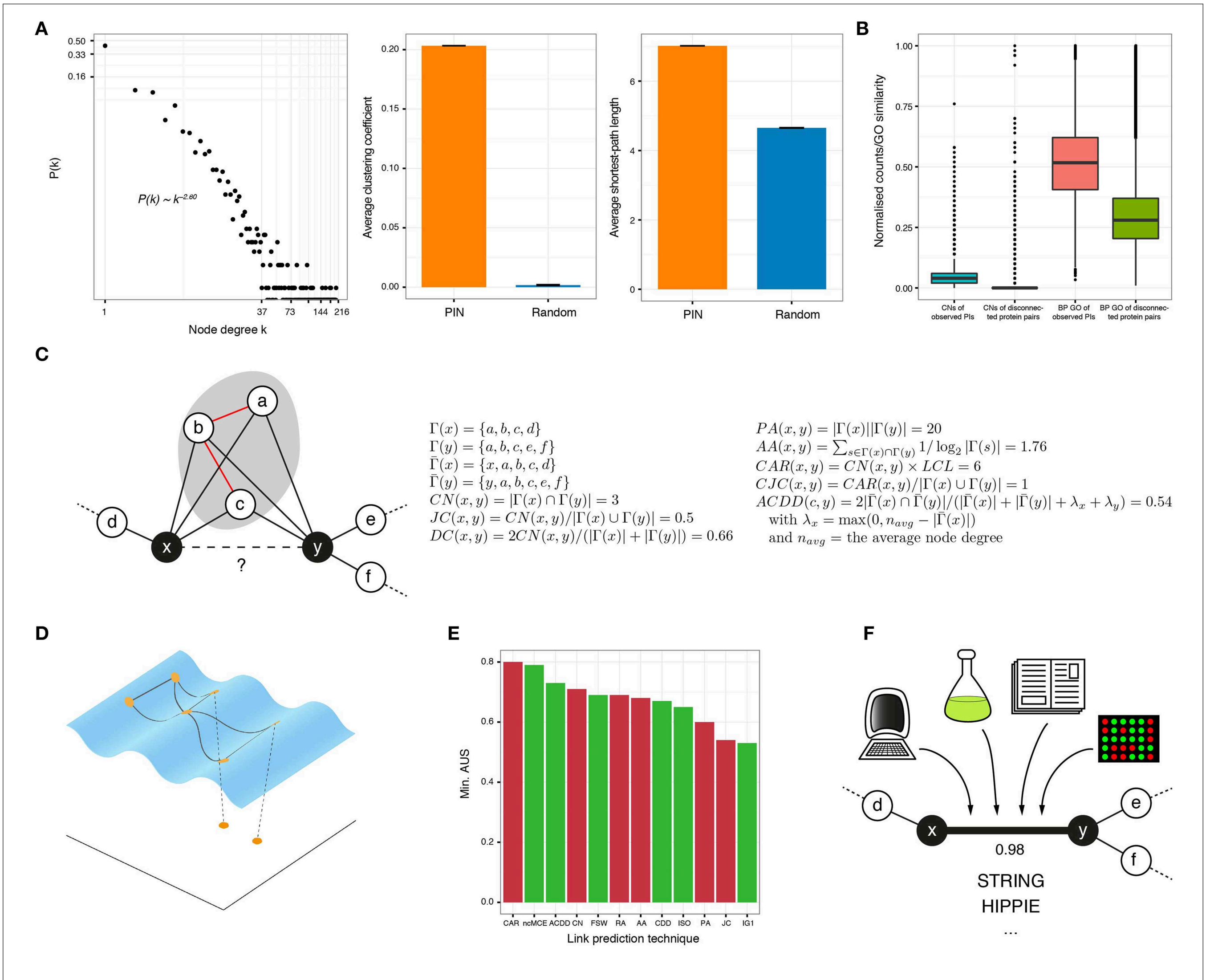


FIGURE 1 | (A) In protein interactomes, only a few proteins, known as hubs, have a high number of interactors (node degree) and the rest interact only with a small number of proteins (left). In addition, PINs are highly clustered (middle) and every protein is easily reachable from anywhere in the network (right), compared to graphs with the same number of randomly linked nodes. **(B)** The number of common neighbors (normalized by the maximum) and the gene ontology (GO) similarity (biological process or BP shown) of protein interactions is higher than for disconnected protein pairs in the observed network. **(C)** The goal of neighborhood-based link prediction and reliability assessment is to assign a likelihood score to the observed or potential interaction between two proteins x and y. The formulae for representative link predictors are listed and applied to the toy network on the left. $\Gamma(x)$ is the set of neighbors of node x, $\bar{\Gamma}(x)$ is the same set but including x and the local community links (LCL) are highlighted in red. **(D)** There is compelling evidence that complex networks, like PINs, lie on low-dimensional manifolds embedded in high-dimensional space. When protein networks are mapped to low dimensions, good candidate interactions lie in close proximity. **(E)** The link prediction performance of several of the topological techniques discussed in this review, measured by the minimum Area Under the Sparsification curve (AUS) amongst four networks (for details of these datasets, see Cannistraci et al., 2013b). Red bars correspond to methods proposed for networks in general and green bars to methods proposed for bio-networks. **(F)** High-quality PI resources, like STRING or HIPPIE, assign a confidence score to each of their reported interactions, based on the different evidence sources supporting them.

2. Topological Reliability Assessment and Prediction of Protein Interactions

The observable network topologies of biological systems are not complete and contain spurious interactions. In addition, the mechanisms that lead two proteins to interact are not fully understood yet. As a consequence, traditional machine learning algorithms cannot be easily applied to PINs. Not only is the definition of features to discriminate between interacting and non-interacting proteins a challenging task, but also the construction of positive and negative sets of interactions to train these algorithms. For example, two unlinked proteins in the observable network cannot be considered as part of the negative set: it could very well be that they are disconnected due to experimental constraints that prevented scientists from observing their interaction. Alternatively it could be that, two linked proteins represent a false positive that is part of the dataset because one of the interactors is, for example, a sticky protein (Saito et al., 2002).

In this context, the assignment of likelihood scores to connected and disconnected pairs of proteins, on the mere basis of the observable network topology, is a convenient means to improve the degree of confidence and completeness of PINs (Cannistraci et al., 2013a). Although reliability assessment of PIs deals with connected proteins pairs and PI prediction with disconnected pairs, the methods used for one or the other are the same. The following subsections account for the most important techniques to perform these functions. A more in-depth description of these approaches can be found in, for example, Lü and Zhou (2011).

2.1. Neighborhood-based Techniques

In 2001, Newman found that the relative probability of collaboration between scientists increases with their number of common acquaintances (Newman, 2001). **Figure 1B** shows that this is also applicable to PINs: the number of common neighbors (CNs) is higher for connected protein pairs than for disconnected ones, in a high quality human interactome. This inspired the creation of the CN index, which assigns high likelihood scores to protein pairs with many CNs.

Newman's findings triggered the development of a myriad of neighborhood-based approaches (Lü and Zhou, 2011). Some of them are only normalizations of CN, like Jaccard's index (Jaccard, 1912) or the Dice Similarity (Dice, 1945), but others really depart from it. For example, Preferential Attachment (PA) (Newman, 2001) is the product between the number of neighbors of the two nodes being analyzed, and Adamic and Adar (2003) and Resource Allocation (Zhou et al., 2009) assign higher likelihood scores to node pairs whose CNs do not interact with other components. Other indices, like Local Path (Lü et al., 2009) or Katz (Katz, 1953), not only take the number of CNs into account but also the neighbors of these CNs and so on, up to a user-specified depth.

In 2013, Cannistraci and colleagues introduced a paradigm shift in topological link prediction, by noting that the presence of a tightly connected set of CNs increases the probability of interaction between non-adjacent nodes (Cannistraci et al., 2013a). Thus, they introduced a family of neighborhood-based approaches by changing the formulation of popular techniques with the inclusion of the number of links between CNs. The simplest example is the so-called Cannistraci-Alanis-Ravasi index (CAR) that multiples this number by CN.

Although the above mentioned techniques can be applied to PINs, they were formulated for networks in general and do not consider any particular biological assumption. The pioneers of PI reliability assessment and prediction are Saito and colleagues. In 2002, after observing that the partners of sticky proteins and self-activators do not interact with anything else in PINs, they proposed the Interaction Generality index (IG1), which assigns low reliability scores to protein pairs whose neighbors have very few partners (Saito et al., 2002). They later introduced the IG2, which postulates that closed-loop motifs are indicative of PIs (Saito et al., 2003).

Another two indices put forward in the context of protein interactomes are the Interaction Reliability by Alternative Paths index (IRAP) and its successor IRAP* (Chen et al., 2006b). According to these indices, the likelihood that two proteins interact increases if there is a large number of alternative network paths through which they can communicate. Unfortunately, these techniques, together with IG2, are computationally demanding, which prompted the development of more efficient and accurate methods (Chen et al., 2006a) such as the Functional Similarity Weight (FSW) and the Adjusted Czekanowski-Dice Dissimilarity (Chua et al., 2006; Liu et al., 2009; Alanis-Lobato et al., 2013). These approaches are interesting because they bet for a lenient integration of the CN and PA indices: protein pairs with lots of common interactors are good candidate PIs, but if one of the two proteins has very few partners, the confidence score is penalized.

All the afore-mentioned techniques represent, in general, an efficient and accurate way to identify protein pairs that are good candidates for interaction (see the formulation of some of them and their application to a toy example in **Figure 1C**). However, they all strongly depend on topological information to work properly. As a consequence, they perform poorly when applied to very sparse networks, like the PINs of non-model or poorly annotated organisms (You et al., 2010).

2.2. Maximum Likelihood Techniques

Maximum likelihood approaches, introduced mainly for link prediction, rely on the underlying community structure of complex networks. In the Hierarchical Random Graph (Clauset et al., 2008), the space of all possible dendrograms of a network is searched to get the ones that best fit its hierarchical structure. Non-adjacent pairs of nodes that have high average probability of being connected within these dendrograms represent good candidates for interaction. In the Stochastic Block Model (Guimerà and Sales-Pardo, 2009), in which a network is partitioned into groups, the probability that two nodes are connected depends on the groups to which they belong. An important issue with these approaches is that they are computationally expensive and not parameter-free (Lü and Zhou, 2011).

2.3. Network Embedding Techniques

Data analysts are regularly faced with the problem of finding meaningful low-dimensional representations of high-dimensional data. Algorithms such as Multidimensional Scaling or Principal Component Analysis embed data to low dimensions by preserving inter-sample distances or covariances but, if the dataset under study contains non-linear structure, they fail to provide useful mappings (Tenenbaum et al., 2000). To solve this issue, non-linear dimensionality reduction algorithms, such as Isometric Feature Mapping (ISOMAP), are commonly employed. Under the hypothesis that the biological features that lead to a PI are complex and non-linear, one could assume that PINs are shaped over a manifold embedded in a high-dimensional space, where interacting proteins are geometrically close to each other and disconnected pairs are far apart (Kuchaiev et al., 2009; You et al., 2010; Cannistraci et al., 2013a). This highlights that if a reasonable measure of dissimilarity between proteins is established, a pairwise dissimilarity matrix can be constructed and used to reveal the low-dimensional geometry of the analyzed network. Good candidates for interaction are finally determined via closeness relationships in the reduced space (**Figure 1D**).

Nataša Pržulj and her colleagues are pioneers in the modeling of PINs with geometric graphs. Their computational experiments show close matches between important topological properties of PINs and geometric random graphs (Przulj et al., 2004). Their results support the hypothesis that PINs do have an underlying geometric structure. These conclusions resulted from the embedding of networks to low dimensions, using the shortest-paths between nodes as dissimilarity and investigating whether proteins pairs that map close to each other are indeed more likely to interact (Higham et al., 2008; Kuchaiev et al., 2009). In 2010, You and co-workers extended this idea with the application of FSW to the PIN after embedding, with the aim to refine the identification of candidate PIs (You et al., 2010).

Around the same time period, a group of physicists and network scientists were independently developing a framework to model complex networks, resting on the assumption that a hidden metric space underlies them and shapes their topology (Boguñá et al., 2009). Contrary to Pržulj and You, who map PINs to a Euclidean space, this group's hypothesis is that complex networks respect the rules of hyperbolic spaces (Krioukov et al., 2010, 2012). This choice is reasonable: trees (subgraphs touching all network nodes without cycles), which abstract the skeleton or hierarchy of complex networks, need an exponential amount of space to branch [the total number of nodes at depth d in a b-ary tree is $(b^{d+1} - 1)/(b - 1)$] and only hyperbolic spaces expand exponentially, providing enough space for a complex network to grow (Krioukov et al., 2010). This premise evolved into a model able to produce scale-free and strongly clustered networks, by simply distributing nodes at random in hyperbolic space and connecting those that are hyperbolically close to each other (Papadopoulos et al., 2012). In addition, the fact that two nodes are connected in a real network correlates strikingly well with short hyperbolic distances between them (Krioukov et al., 2010; Papadopoulos et al., 2012). These results confirm that complex networks, like PINs, do possess an intrinsic organization shaped by geometric principles that agree well with hyperbolic ones. However, current algorithms to map networks to hyperbolic space depend on a Metropolis-Hastings algorithm that requires some manual intervention to converge in a reasonable amount of time (Papadopoulos et al., 2012). More computationally efficient methods are currently under development.

Finally, in the non-centered Minimum Curvilinear Embedding (ncMCE), a technique that has been successfully applied in different fields (Alanis-Lobato et al., 2015), the Minimum Spanning Tree (MST) is extracted from the network under scrutiny to construct a matrix of pairwise distances between nodes over the MST. The network is then projected to low-dimensions by singular value decomposition of this matrix and, in contrast to previous approaches, that assign likelihood scores by directly measuring Euclidean distances between node pairs (Kuchaiev et al., 2009; You et al., 2010), in ncMCE the network is reconstructed in the reduced space so that its edges are weighted by the distances between connected nodes. Likelihood scores are then the shortest-paths between nodes in this low-dimensionally projected, weighted network (Cannistraci et al., 2013b). It is not surprising that this technique achieves a remarkable performance in the prediction of PIs: measuring distances between proteins over the MST, corresponds to navigating one of the discrete representations of the hyperbolic geometry underlying the network under study. As previously mentioned, hyperbolic spaces are smooth versions of the trees abstracting the hierarchy of PINs (Krioukov et al., 2010).

2.4. General Framework for Measuring the Effectiveness of These Techniques

In order to benchmark the accuracy of a link prediction technique, the following framework is commonly employed:

1. Remove L randomly selected PIs from the observable network topology.
2. Assign confidence scores to disconnected protein pairs in the pruned network with a topological technique and sort them decreasingly (best candidate interactions positioned at the top of this list).
3. Take L protein pairs from the top of the sorted list and compute the proportion present in the set of interactions removed in 1. This is a measure of the technique's *precision*.
4. Repeat steps 1–3 t times, removing different sets of randomly selected PIs.
5. Repeat steps 1–4 removing $2L$, $3L$, etc. interactions, up to the point where the network loses connectivity. This allows for the construction of a sparsification curve (SC), whose points are the mean precisions of the technique applied at each sparsification level.

This evaluation depicts the ability of a topological approach to predict accurately under the presence of less and less network information. Nonetheless, it has an intrinsic problem because, as discussed above, some of the candidate interactions with high confidence scores may not be part of the randomly removed set of PIs. However, they may represent good candidates that current technologies cannot measure. Moreover, members of the removed set of links may be false positives that good link predictors are correctly discarding by giving them low scores.

Subsequently, researchers have opted for using Gene Ontology (GO) similarities (Yu et al., 2010) to discriminate between good and bad candidate PIs. This is based on the *guilt-by-association* principle (Oliver, 2000), which states that if two proteins are involved in similar bio-processes, they are more likely to interact (see **Figure 1B**). Although Resnik's index (Resnik, 1999) is the prevailing GO similarity, Wang's index is worth mentioning because it was formulated specifically for the GO (Wang et al., 2007). Another interesting method improves GO similarities by considering the inherent uncertainty originating from the GO incompleteness (Yang et al., 2012).

Figure 1E presents the minimum area under the SC for most of the topological techniques described in this section, when they are applied to four yeast networks for the link prediction task (Cannistraci et al., 2013a,b). This figure depicts the robustness of each technique, as their worst performance is exposed. Despite the good results of some of these methods, there is still room for improvement, and development of approaches that consider the scale-free structure and geometry of PINs remain active subjects of research (Papadopoulos et al., 2012; Zhu et al., 2013).

3. Resources for High Confidence Protein Interactions

Proteins with a high likelihood to interact can considerably reduce the universe of possible pairs to test in the lab and guide wet-lab validations. These interactions can then be integrated with available repositories of high-quality PIs that attach confidence scores to each reported interaction (see **Figure 1F**). One of such resources is the Search Tool for the Retrieval of Interacting Genes (STRING), which provides a combined score that indicates higher confidence when more than one source of evidence supports an interaction (Szklarczyk et al., 2011). STRING evidence sources include computational associations (neighborhood-based, co-occurrence, co-expression, text mining), high-throughput experiments, other databases, and interactions identified in other organisms. The current version of STRING (available at http://string-db.org) provides an interactive network viewer and access to interactions between almost 10 million proteins, from more than 2000 organisms (Szklarczyk et al., 2015).

The Human Integrated Protein-Protein Interaction rEference (HIPPIE) retrieves interactions from major expert-curated databases and calculates a score for each PI, reflecting its combined experimental evidence. This score is a function of the number of studies supporting the interaction, the quality of the experimental techniques used to measure it and the number of organisms in which it is present (Schaefer et al., 2012). In HIPPIE (http://cbdm-01.zdv.uni-mainz.de/∼mschaefer/hippie/), one can query the interactors of a protein or a set of proteins and explore the resulting network in an interactive viewer. Furthermore, the results can be filtered by PI type, tissue, functions, directionality and inhibitory/activating effect (Schaefer et al., 2013).

Another worth-mentioning resource is INstruct (http://instruct.yulab.org/). It collects interactions from eight major expert-curated databases and filters out low-quality PIs, to keep only those supported by domain-domain interactions obtained from co-crystal structures (Wang et al., 2012; Meyer et al., 2013). INstruct provides a web-based interface to query its extremely high-quality PINs for 7 different species. The network properties depicted in **Figures 1A–C** correspond to the INstruct PIN for human.

It is important to stress that when querying interactions from these resources, high-confidence should be preferred over size. In a recent study, Rolland and colleagues assembled PIs from 7 public databases and found that interactions supported by multiple sources can be validated at rates that are significantly higher than those of PIs supported by a single method (Rolland et al., 2014). This is critical, because meaningful results about human health and disease can only be achieved when using high-confidence PINs.

4. Protein Interaction Networks in Health and Disease

It is possible that the first work that advocated for a systems-based approach to disease is the one by Goh et al. (2007). They take advantage of the Online Mendelian Inheritance in Man (OMIM) repository to build a bipartite network of disorders linked to their associated genes (see **Figure 2A** middle). Starting from this network, projections are carried out, one to the *disease space* (**Figure 2A** left) and the other to the *gene space* (**Figure 2A** right). In the disease projection, they observe a giant network component, suggesting shared genetic origins of its constituent diseases. The gene projection provides phenotypic relationship between gene pairs and presents a high overlap with a network of high-quality PIs (Goh et al., 2007). Moreover, essential human genes tend to encode hub proteins and are found to be expressed in most tissues. Whereas, disease genes are less connected and possess tissue specificity (Goh et al., 2007).

A similar analysis, focused on the gene projection, was performed considering only autoimmune diseases (Alanis-Lobato et al., 2014). After the application of a community detection algorithm, it was found that genes associated with related diseases clustered together (see **Figure 2B**). This community organization also revealed the presence of clusters disconnected from the main network component, suggesting that the genes forming them are disease specific.

Given a set of proteins associated with a patient's phenotype, Lage and co-workers are able to rank disease-causing proteins as the top candidates with the help of a phenotype similarity score. This also allows them to identify previously unknown disease-causing complexes (Lage et al., 2007). In a similar fashion, a tool named CIPHER scores and prioritizes phenotype-gene pairs, based on an integrated human protein and phenotype network, to reliably predict disease genes (Wu et al., 2008).

In 2014, Zhou and colleagues extracted disease and symptom terms from the Medical Subject Headings (MeSH) in PubMed and linked diseases with symptoms via bibliographic records (**Figure 2A** middle). Instead of simply mapping this network to the disease space, they describe each disease with a vector of symptoms, with entries quantifying the strength of association

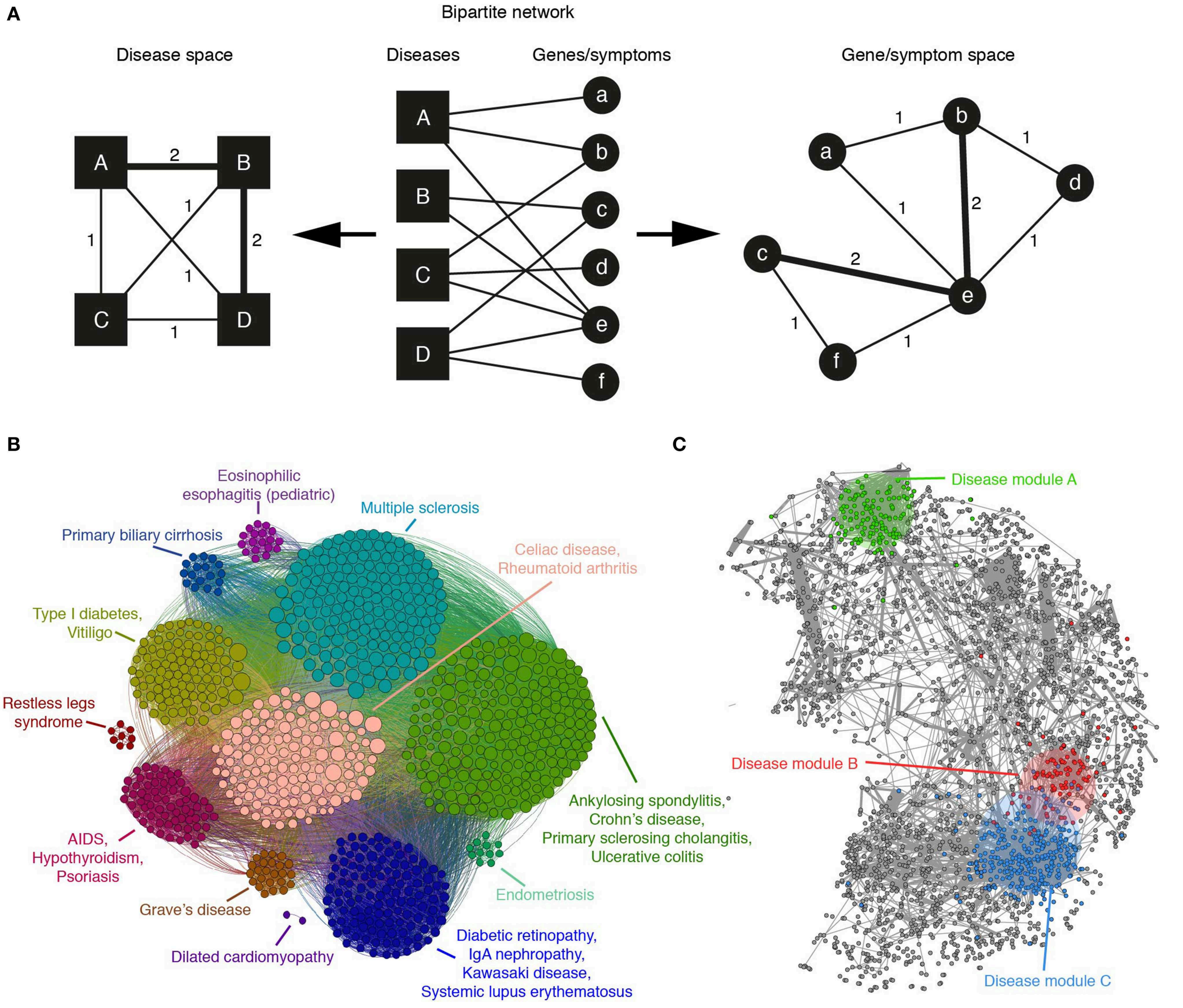

A
Bipartite network
Disease space
Diseases
Genes/symptoms
Gene/symptom space
A
B
C
D
a
b
c
d
e
f
1
2
B
Eosinophilic esophagitis (pediatric)
Multiple sclerosis
Primary biliary cirrhosis
Celiac disease, Rheumatoid arthritis
Type I diabetes, Vitiligo
Restless legs syndrome
Ankylosing spondylitis, Crohn's disease, Primary sclerosing cholangitis, Ulcerative colitis
AIDS, Hypothyroidism, Psoriasis
Endometriosis
Grave's disease
Dilated cardiomyopathy
Diabetic retinopathy, IgA nephropathy, Kawasaki disease, Systemic lupus erythematosus
C
Disease module A
Disease module B
Disease module C

FIGURE 2 | (A) A bipartite network of diseases and their associated genes or symptoms can be mapped to the disease or gene/symptom space by linking nodes of one type that are connected with the same nodes of the other. The weight of the edges in the resulting projection indicates the number of such common nodes. **(B)** The application of a community detection algorithm to the Autoimmune Disease Network, mapped to the gene space, reveals groups of genes associated with similar disorders and high levels of co-morbidity (adapted from Alanis-Lobato et al., 2014). **(C)** An example human protein interactome in which gene products associated with diseases A, B, and C have been labeled with different colors. According to Menche et al. (2015), the topologically closer two diseases are (like B and C), the higher the GO similarity and co-expression of their associated proteins and the higher their co-morbidity and symptom similarity.

between each symptom and the disease. Later, they compute a pairwise cosine-similarity matrix between these vectors and only the most significant similarities are considered to construct a network of weighted links between diseases (Zhou et al., 2014). Analysis of the resulting network shows that disease pairs with high symptom similarity are more likely to share associated genes and PIs. This symptom-based disease network is also organized in highly interconnected communities of similar diseases, which shows that similar symptoms imply similar disorders.

The recent work of Menche and colleagues is quite relevant, as it shows that, despite its incompleteness and biases, the current human PIN can be mined and integrated with disease data to uncover pathobiological relationships between disorders and better understand their etiology. After compiling a network of roughly 140k interactions between more than 13k human proteins, nodes are labeled with their associated diseases with the help of OMIM and a set of 299 disorders defined by MeSH. Although the disease module hypothesis predicts that proteins associated with the same trait should be highly interconnected (Barabási et al., 2011; Loscalzo and Barabasi, 2011), they find that only a few disease-specific proteins form a connected subgraph. Whereas, the rest appear to be randomly distributed in the

PIN because missing links isolate them from their module (Menche et al., 2015). In spite of this result, the small disease subgraphs are significantly larger than the random expectation and their topological properties are biologically meaningful: GO similarity between module members is significantly high and the topologically closer two diseases are, the higher the GO similarity and co-expression of their associated proteins and the higher their co-morbidity and symptom similarity (see **Figure 2C**).

5. Conclusion

Viewing the relationships between cell compartments and their constituting molecules as a complex circuitry of tightly interconnected components is widespread in systems biology (Vidal et al., 2011). This has led to breakthroughs that the study of the individual system components would not have made possible (Takahashi and Yamanaka, 2006; Levine and Oren, 2009; Ravasi et al., 2010). However, available interactomes are far from complete, which makes the production of high quality datasets crucial to unravel the complex relationships between genotype and phenotype (Barabási et al., 2011; Loscalzo and Barabasi, 2011).

Since the identification of biological features to distinguish between interacting and non-interacting proteins is very difficult, mining the topological characteristics of PINs is useful in the reliability assessment and prediction of PIs (Cannistraci et al., 2013b). The best candidates can be integrated with resources of high-confidence PIs to reconstruct well-grounded interactomes (Szklarczyk et al., 2015). Clinical and pathological information can then be superimposed on these networks to detect disease modules, identify co-morbidity and similarities between diseases and even make new protein-disorder associations. All of this by using simple, yet powerful network-based tools (Goh et al., 2007; Alanis-Lobato et al., 2014; Menche et al., 2015).

As the quantity and quality of molecular datasets increase, network science offers a new means to analysing interacting gene products at a systems level (Loscalzo and Barabasi, 2011). This will allow, in the near future, for a redefinition of diseases as sub-networks of a molecular interactome, overlapping with or in close proximity to other similar diseases, rendering a clear picture of the network components whose perturbation has phenotypic impact. Consequently, the integration and holistic analysis of genetic, genomic, chemical, environmental, clinical, and therapeutic data are rapidly driving the development of network medicine, a promising approach aimed at unraveling disease etiology.

Acknowledgments

The author would like to thank Antonia Forshaw for proofreading the article, as well as Prof. Timothy Ravasi and Dr. Carlo Cannistraci for encouraging the preparation of this review.

References

Adamic, L. A., and Adar, E. (2003). Friends and neighbors on the Web. *Soc. Netw.* 25, 211–230. doi: 10.1016/S0378-8733(03)00009-1

Alanis-Lobato, G., Cannistraci, C. V., Eriksson, A., Manica, A., and Ravasi, T. (2015). Highlighting nonlinear patterns in population genetics datasets. *Sci. Rep.* 5:8140. doi: 10.1038/srep08140

Alanis-Lobato, G., Cannistraci, C. V., and Ravasi, T. (2013). Exploitation of genetic interaction network topology for the prediction of epistatic behavior. *Genomics* 102, 202–208. doi: 10.1016/j.ygeno.2013.07.010

Alanis-Lobato, G., Cannistraci, C. V., and Ravasi, T. (2014). "Exploring the genetics underlying autoimmune diseases with network analysis and link prediction," in *Middle East Conference on Biomedical Engineering (MECBME)* (Doha: IEEE), 167–170.

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97. doi: 10.1103/RevModPhys.74.47

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Boguñá, M., Krioukov, D., and Claffy, K. C. (2009). Navigability of complex networks. *Nat. Phys.* 5, 74–80. doi: 10.1038/nphys1130

Cannistraci, C. V., Alanis-Lobato, G., and Ravasi, T. (2013a). From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Rep.* 3, 1–13. doi: 10.1038/srep 01613

Cannistraci, C. V., Alanis-Lobato, G., and Ravasi, T. (2013b). Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* 29, i199–i209. doi: 10.1093/bioinformatics/btt208

Chen, J., Chua, H. N., Hsu, W., Lee, M.-L., Ng, S.-K., Saito, R., et al. (2006a). Increasing confidence of protein-protein interactomes. *Genome Informat.* 17, 284–297.

Chen, J., Hsu, W., Lee, M. L., and Ng, S.-K. (2006b). Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics* 22, 1998–2004. doi: 10.1093/bioinformatics/btl335

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22, 1623–1630. doi: 10.1093/bioinformatics/btl145

Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 98–101. doi: 10.1038/nature06830

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302.

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

Guimerà, R., and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1–6. doi: 10.1073/pnas.0709640104

Higham, D. J., Rasajski, M., and Przulj, N. (2008). Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics* 24, 1093–1099. doi: 10.1093/bioinformatics/btn079

Jaccard, P. (1912). The distribution of flora in the alpine zone. *New Phytol.* 11, 37–50.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* 18, 39–43.

Krioukov, D., Kitsak, M., Sinkovits, R. S., Rideout, D., Meyer, D., and Boguñá, M. (2012). Network cosmology. *Sci. Rep.* 2, 1–6. doi: 10.1038/srep00793

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. (2010). Hyperbolic geometry of complex networks. *Phys. Rev. E* 82:036106. doi: 10.1103/physreve.82.036106

Kuchaiev, O., Rasajski, M., Higham, D. J., and Przulj, N. (2009). Geometric de-noising of protein-protein interaction networks. *PLoS Comput. Biol.* 5:e1000454. doi: 10.1371/journal.pcbi.1000454

Lage, K., Karlberg, E. O., Størling, Z. M., Ólason, P. I., Pedersen, A. G., Rigina, O., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25, 309–316. doi: 10.1038/nbt1295

Lesk, A. M. (2007). "Systems biology," in *Introduction to Genomics, Chapter 7, 1st Edn.* (New York, NY: Oxford University Press), 359–405.

Levine, A. J., and Oren, M. (2009). The first 30 years of p53: growing ever more complex. *Nat. Rev. Cancer* 9, 749–758. doi: 10.1038/nrc2723

Liu, G., Wong, L., and Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics* 25, 1891–1897. doi: 10.1093/bioinformatics/btp311

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2011). Controllability of complex networks. *Nature* 473, 167–173. doi: 10.1038/nature10011

Loscalzo, J., and Barabasi, A.-L. (2011). Systems biology and the future of medicine. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 3, 619–627. doi: 10.1002/wsbm.144

Lü, L., Jin, C.-H., and Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E* 80:046122. doi: 10.1103/PhysRevE.80.046122

Lü, L., and Zhou, T. (2011). Link prediction in complex networks: a survey. *Phys. A* 390, 1150–1170. doi: 10.1016/j.physa.2010.11.027

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, 1257601–1257601. doi: 10.1126/science.1257601

Meyer, M. J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29, 1577–1579. doi: 10.1093/bioinformatics/btt181

Newman, M. (2001). Clustering and preferential attachment in growing networks. *Phys. Rev. E* 64, 1–4. doi: 10.1103/PhysRevE.64.025102

Oliver, S. (2000). Guilt-by-association goes global. *Nature* 403, 601–603. doi: 10.1038/35001165

Papadopoulos, F., Kitsak, M., Serrano, M. A., Boguñá, M., and Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature* 489, 537–540. doi: 10.1038/nature11459

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 3508–3515. doi: 10.1093/bioinformatics/bth436

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. doi: 10.1016/j.cell.2010.01.044

Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 11, 95–130.

Rolland, T., Tas, M., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Saito, R., Suzuki, H., and Hayashizaki, Y. (2002). Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res.* 30, 1163–1168. doi: 10.1093/nar/30.5.1163

Saito, R., Suzuki, H., and Hayashizaki, Y. (2003). Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics* 19, 756–763. doi: 10.1093/bioinformatics/btg070

Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS ONE* 7:e31826. doi: 10.1371/journal.pone.0031826

Schaefer, M. H., Lopes, T. J. S., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J.-F., et al. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.* 9:e1002860. doi: 10.1371/journal.pcbi.1002860

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–D568. doi: 10.1093/nar/gkq973

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676. doi: 10.1016/j.cell.2006.07.024

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Wang, J., Du, Z., Payattakool, R., Yu, P., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164. doi: 10.1038/nbt.2106

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol.* 4:189. doi: 10.1038/msb.2008.27

Yang, H., Nepusz, T., and Paccanaro, A. (2012). Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics* 28, 1383–1389. doi: 10.1093/bioinformatics/bts129

You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S., and Zhou, X. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* 26, 2744–2751. doi: 10.1093/bioinformatics/btq510

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26, 976–978. doi: 10.1093/bioinformatics/btq064

Zhou, T., Lü, L., and Zhang, Y.-C. (2009). Predicting missing links via local information. *Eur. Phys. J. B* 71, 623–630. doi: 10.1140/epjb/e2009-00335-8

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun.* 5:4212. doi: 10.1038/ncomms5212. Available online at: http://www.nature.com/ncomms/2014/140626/ncomms5212/full/ncomms5212.html

Zhu, Y., Zhang, X.-F., Dai, D.-Q., and Wu, M.-Y. (2013). Identifying spurious interactions and predicting missing interactions in the protein-protein interaction networks via a generative network model. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 10, 219–225. doi: 10.1109/TCBB.2012.164

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

ORIGINAL RESEARCH
published: 18 December 2015
doi: 10.3389/fgene.2015.00345

Differential Occurrence of Interactions and Interaction Domains in Proteins Containing Homopolymeric Amino Acid Repeats

Ilaria Pelassa[1] *and Ferdinando Fiumara*[1,2]*

[1] *Department of Neuroscience, University of Torino, Torino, Italy,* [2] *National Institute of Neuroscience (INN), Torino, Italy*

Edited by:
Spyros Petrakis,
Aristotle University of Thessaloniki,
Greece

Reviewed by:
Olwyn Byron,
University of Glasgow, UK
Sandra Macedo-Ribeiro,
Instituto de Biologia Molecular e
Celular, Portugal

***Correspondence:**
Ferdinando Fiumara
ferdinando.fiumara@unito.it

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

Received: *01 September 2015*
Accepted: *20 November 2015*
Published: *18 December 2015*

Citation:
Pelassa I and Fiumara F (2015)
Differential Occurrence of Interactions
and Interaction Domains in Proteins
Containing Homopolymeric Amino
Acid Repeats. Front. Genet. 6:345.
doi: 10.3389/fgene.2015.00345

Homopolymeric amino acids repeats (AARs), which are widespread in proteomes, have often been viewed simply as spacers between protein domains, or even as "junk" sequences with no obvious function but with a potential to cause harm upon expansion as in genetic diseases associated with polyglutamine or polyalanine expansions, including Huntington disease and cleidocranial dysplasia. A growing body of evidence indicates however that at least some AARs can form organized, functional protein structures, and can regulate protein function. In particular, certain AARs can mediate protein-protein interactions, either through homotypic AAR-AAR contacts or through heterotypic contacts with other protein domains. It is still unclear however, whether AARs may have a generalized, proteome-wide role in shaping protein-protein interaction networks. Therefore, we have undertaken here a bioinformatics screening of the human proteome and interactome in search of quantitative evidence of such a role. We first identified the sets of proteins that contain repeats of any one of the 20 amino acids, as well as control sets of proteins chosen at random in the proteome. We then analyzed the connectivity between the proteins of the AAR-containing protein sets and we compared it with that observed in the corresponding control networks. We find evidence for different degrees of connectivity in the different AAR-containing protein networks. Indeed, networks of proteins containing polyglutamine, polyglutamate, polyproline, and other AARs show significantly increased levels of connectivity, whereas networks containing polyleucine and other hydrophobic repeats show lower degrees of connectivity. Furthermore, we observed that numerous protein-protein, -nucleic acid, and -lipid interaction domains are significantly enriched in specific AAR protein groups. These findings support the notion of a generalized, combinatorial role of AARs, together with conventional protein interaction domains, in shaping the interaction networks of the human proteome, and define proteome-wide knowledge that may guide the informed biological exploration of the role of AARs in protein interactions.

Keywords: amino acid repeats, homopolymeric, polyglutamine, polyalanine, protein-protein interactions

INTRODUCTION

Homopolymeric amino acid repeats (AARs) are found in a large number of eukaryotic proteins (Faux, 2012). These repetitions in the primary sequence of proteins have been initially understood simply as unstructured "spacers" between protein domains or even just as "junk" peptides devoid of specific functions (Green and Wang, 1994; Karlin and Burge, 1996; as discussed in Haerty and Golding, 2010), but prone in some cases to misfolding, as in genetic diseases related to the expansion of polyglutamine (polyQ) or polyalanine (polyA) repeats (Almeida et al., 2013). A growing body of evidence is changing these views by showing how at least some of these repeats have defined structural propensities and functional properties. For instance, we have recently found that polyQ and polyA repeats can form coiled coil supersecondary structures which can regulate the oligomerization, interactions, and functions of proteins (Fiumara et al., 2010; Pelassa et al., 2014). Several studies have now explored the functional consequences of the appearance and variation in length of AARs in transcription factors and in other proteins in which they are particularly enriched, showing how these repeats can alter the function of proteins, thus ultimately modulating developmental and post-developmental processes (e.g., Fondon and Garner, 2004; Anan et al., 2007; O'Malley and Banks, 2008; Nasu et al., 2014).

One of the possible mechanisms by which AARs could regulate the function of proteins that contain them is by mediating the interactions of these proteins with other proteins or with other cellular components such as nucleic acids and lipids in membranes. In support of this hypothesis, we have shown for example that polyQ or polyA repeats can mediate interactions between proteins that contain them (e.g., Fiumara et al., 2010; Pelassa et al., 2014), while polyproline (polyP)-II structures and proline-rich sequences can mediate protein-protein interactions by binding to non-repetitive interaction domains (Yu et al., 1994). Evidence exists that some charged AARs may also drive protein-nucleic acid and protein-lipid interactions (Dean, 1983; Nam et al., 2001; DeRouchey et al., 2013).

AARs and conventional protein-protein, -nucleic acid, and -lipid interaction domains, are often found together in the same proteins. Thus, AARs and non-repetitive, conventional interaction domains may work combinatorially in defining the overall specificity and strength of the interactions of their parent proteins with other proteins or with other interaction partners. Initial evidence indicates indeed the possibility that AARs in proteomes, also together with non-repetitive sequences, may participate in the definition of entire protein-protein interaction networks. For example, it has been shown that disease-related and other polyQ proteins could drive the formation of protein-protein interaction networks based on coiled coil-mediated interactions (Fiumara et al., 2010; Petrakis et al., 2012; Schaefer et al., 2012) and this may also be the case for polyA proteins (Pelassa et al., 2014).

It is still unclear, however, to what degree the emerging roles of polyQ AARs in shaping protein-protein interaction networks in proteomes may be generalized to other AARs. Can other AARs also drive the formation of protein-protein interaction networks? And, with which conventional protein interaction domains may AARs cooperate in establishing these interactomes? The answer to these questions must ultimately come from biological experiments. However, given the scale and complexity of the biological problems raised by such questions, proteome-wide bioinformatics screenings may be essential for guiding the informed biological exploration of all the possible roles of AARs in establishing protein-protein interaction networks, also together with conventional protein interaction domains.

Based on these premises, we have undertaken here a quantitative bioinformatics analysis of the protein-protein interaction networks formed by the proteins containing AARs of each one of the 20 amino acids. Furthermore, we have determined whether specific protein-protein, -nucleic acid, and -lipid interaction domains are overrepresented in each one of the 20 AAR-containing protein groups. The results of our analyses overall provide quantitative support to the hypothesis that, together with conventional protein interaction domains, AARs may play a generalized, combinatorial role in establishing protein-protein interaction networks.

RESULTS

Analysis of Interactomes Reveals Differential Connectivity in AAR-Containing Protein Groups

To determine the potential involvement of AARs in establishing protein-protein interaction networks, we first analyzed the interactomes formed by the proteins of each of the 20 groups of proteins of the human proteome containing repeats of at least four units of any one of the 20 amino acids. This AAR length threshold allows one to identify proteins that contain not only long, pure homopolymeric AARs but also more fragmented repeats at a more advanced stage of their "life cycle" (Buschiazzo and Gemmell, 2006; Pelassa et al., 2014). To perform this analysis, we preliminarily scanned the Uniprot complete human proteome in search of proteins containing repeats of the different amino acids. We thus defined 20 protein groups that were identified as "polyX" groups, were X stands for the standard single letter code for one amino acid (i.e., from A to Y). These groups contain variable numbers of proteins ranging from just one, as for the polyW group, to more than 1000 proteins, as for the polyL, polyA, and polyG groups (**Supplementary Table 1**). We then extracted the known interactions of the proteins of each polyX group (represented schematically by *red nodes* in **Figure 1A**) from the whole human protein interactome reported in the BioGrid database (Stark et al., 2006), using the g:Profiler (Reimand et al., 2011) interface (**Figure 1A**), with the exception of the polyW group which contained only one protein. As statistical controls, for each polyX group, we also extracted in the same way the interactions of five groups of proteins selected randomly in the human proteome, each one containing the same number of proteins as the polyX group (*green nodes* in **Figure 1A** indicate schematically one of these control groups). We were thus able to define protein networks formed by either polyX proteins and their interactors, or by the equinumerous, randomly selected proteins and their interactors (schematized in **Figure 1B** as *AAR protein network* and *random protein network*, respectively). These interactomes thus contain two types of interactions, which are schematized in the lower part of **Figure 1B**. One type, that we called "type *a*" interactions, are between two proteins that both contain the AAR or that are both part of the list of randomly selected proteins. The other type of interactions, i.e., "type *b*," is formed by an AAR protein with a protein that does not contain the AAR, or by a protein of the random group with another protein that is not part of the group.

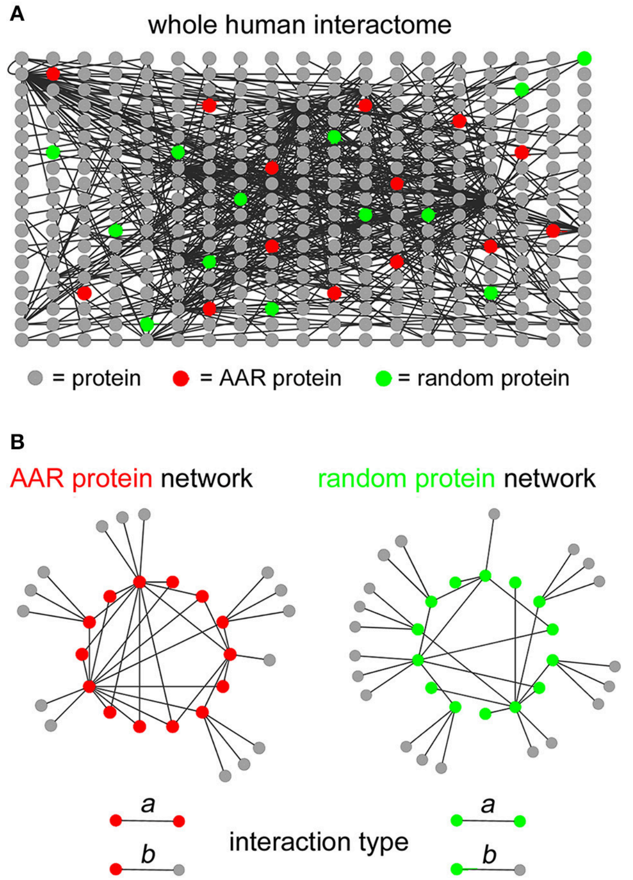


FIGURE 1 | Extraction of interaction networks of AAR proteins from the total human interactome. (A) Schematic simplified representation of the human interactome in graph form. *Gray circles* represent proteins and *black lines* represent binary interactions between proteins as derived from the BioGrid database. *Red circles* represent proteins containing a given AAR. *Green circles* represent proteins selected randomly as a control group for the AAR protein group. **(B)** Simplified schemes representing in graph form (*left scheme*) the interaction network formed by proteins containing a given AAR (*red circles*) and their interactors (*gray circles*), or (*right scheme*) the interaction network formed by randomly selected proteins (*green circles*) and their interactors (*gray circles*). The lower part of the panel shows the two types of interactions that were defined in the interactomes above. Type *a* interactions occur between two AAR-containing proteins or between two proteins of the randomly selected control group. Type *b* interactions occur between an AAR-containing protein and an interactor that does not contain the repeat, or between a protein of the random control group and an interactor that is not part of the random protein group.

To determine whether proteins containing a certain AAR have an increased, decreased, or similar propensity to establish interactions among themselves in comparison with randomly selected proteins, we calculated two quantitative indexes by analyzing the AAR and the random protein networks (**Figures 2A,B**). The first index shows to what extent proteins containing a given AAR tend to interact with proteins containing the same AAR (type *a* interactions) rather than with proteins devoid of it. The second index shows the density of type *a* interactions in each AAR network.

Thus, we first calculated for each AAR protein network the number of type *a* interactions as a proportion of the total number of interactions, i.e., type *a*/(type *a* + type *b*). The same proportion was also calculated for each of the five random control networks. The value of the proportion in the AAR protein network was then normalized to the value of the proportion calculated for each of the five corresponding random protein networks. The resulting five normalized values were then averaged and are plotted in **Figure 2A** (mean ± SEM). This analysis revealed that, for instance, the proportion of type *a* interactions in the interactome of polyQ proteins is on average 1.81 ± 0.06 ($n = 5$) times greater than in the corresponding control networks. This difference in the distribution of type *a* and type *b* interactions between the polyQ network and the average of the control random networks was statistically significant ($p < 0.001$, χ^2 test with Yates' correction), indicating that polyQ proteins tend to establish significantly more interactions with other polyQ proteins than expected by chance. This observation was not unique to polyQ proteins, and in fact similar results were found also for other networks of proteins containing repeats formed by other polar (polyS), charged (poly-D, -E, -K, -R) or small/cyclic (poly-A, -G, -P) amino acids. Conversely, proteins containing polyL or other hydrophobic repeats (poly-C, -M, -Y) tend to have fewer interactions with each other than expected, although this trend is statistically significant only for the polyL group. Other networks of proteins with hydrophobic AARs (poly-I and -V) display non-significant trends toward a slight increase in the proportion of type *a* interactions. The case of polyF networks is difficult to interpret due to the small number of proteins ($n = 61$) containing this repeat and to the consequently higher statistical variability that was observed in the five corresponding control networks. Finally, poly-N, -H, and -T networks did not deviate from what expected by chance in terms of type *a* connectivity.

Second, we calculated for each polyX network the average number of type *a* interactions per AAR-containing protein and we normalized this value to the corresponding value calculated for each of the five control random networks. The average of the resulting five normalized values for each AAR group is shown in **Figure 2B**. A One-way ANOVA analysis revealed overall significant differences among the AAR groups [$F_{(18,\ 76)} = 53.69$, $p < 0.001$]. Furthermore, the Dunnett *post-hoc* test, using as a control group the polyM group which has the mean value closest to 1, revealed significant differences ($p < 0.05$ in all instances) from the polyM group for all the AAR protein groups that were also significant in **Figure 2A**, with the addition of the polyY group.

This analysis showed, for example, that polyQ proteins establish 2.39 ± 0.06 ($n = 5$) times more interactions with other polyQ proteins than expected by chance. Also other networks of proteins containing repeats of polar (polyS), charged (poly-D, -E, -K, -R), or small/cyclic (poly-A, -G, -P) amino acids display 1.25–1.7 times more per-node interactions than expected by chance, as also illustrated in the sample networks shown in **Figure 2C**. This figure illustrates how, for instance, networks of polyE or polyA proteins display a higher density of type *a* connections than the corresponding control networks. Again, networks of proteins containing certain hydrophobic AARs (poly-L, -Y) displayed a significantly lesser number of per node connections than expected by chance, as one can also appreciate visually in **Figure 2C** which shows how the density of the polyL network is lower than that of a random network. Finally, the density of type *a* connections in poly-T, -N, -I, -H, -F, -M, and -V networks did not significantly differ from that of the corresponding control networks. Taken together, these findings show a substantial concordance of the two indexes that we used to characterize the type *a* connectivity of the AAR protein networks. In fact, we observed a significant correlation between the two indexes of the 19 AAR groups ($r = 0.69$, $n = 19$, $p < 0.01$). Thus, AAR groups in which the first index is high and statistically significant tend to have also higher values for the second index (e.g., polyQ, polyP, polyE), and this general concordance of the two indexes strengthens the conclusion that these AARs are associated with a higher degree of interactivity among proteins that contain them. Conversely, in some particular cases like that of polyF, even though the first index is high, but not significantly, the second index is close to the value expected by chance, thus indicating overall that the presence of this repeat is not associated with a greater connectivity between the proteins that contain it.

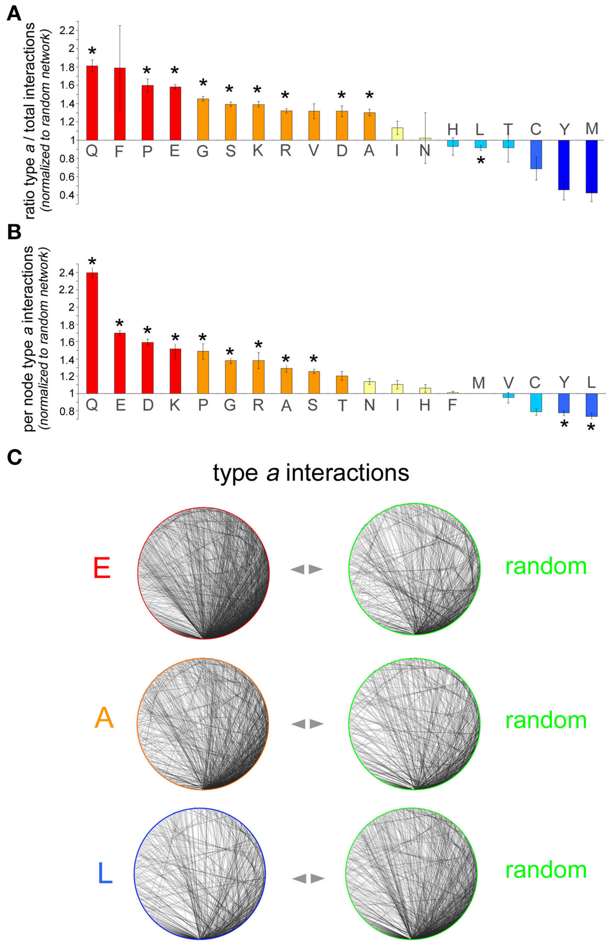


FIGURE 2 | Analysis of interaction networks formed by AAR proteins. (A) Number of type *a* interactions as a proportion of the total number of interactions in the polyX protein network [i.e., type *a*/(type *a* + type *b*)] normalized to the same proportion calculated for each of the five control networks. The bars represent the average of these five normalized ratios ± SEM. Asterisks indicate polyX networks for which the proportion of type *a* interactions differs significantly from that expected by chance ($p < 0.05$, χ^2 test with Yates' correction). **(B)** Sub-networks comprising only type *a* interactions were further analyzed and the number of per node interactions in the AAR protein sub-network was counted and normalized to the number of per node interactions in control sub-networks of random proteins. For each polyX network, this normalized value was calculated for each one of the five control networks. The graph shows, for each polyX protein group, the average normalized number of type *a* interactions per node ± SEM. **(C)** Sample sub-networks (polyE, polyA, and polyL) containing only type *a* interactions such as those that were analyzed in **(B)** (*left graphs*), with one of their respective random networks (*right graphs*). Note how the density of the polyE and polyA networks is higher than that of the corresponding control networks, whereas the density of the polyL network is lower that of the control network.

Taken together, these findings indicate that the presence of certain AARs in protein networks associates with a higher degree of connectivity. These AARs are those formed by certain polar (poly-Q, -S), charged (poly-D, -E, -K, -R), or small/cyclic (poly-A, -G, -P) amino acids, suggesting that these repeats themselves, or protein domains they co-occur with, or they are found within, may promote protein-protein interactions and the formation of interaction networks. Conversely, the presence of certain hydrophobic repeats like polyL and polyY in proteins seems to disfavor the formation of interaction networks, possibly owing to the fact that these repeats are often found in transmembrane domains that sequester proteins in membranes (see Section Discussion).

Possible Roles of AARs in Protein-Protein, -Nucleic Acid, and -Lipid Interactions

In principle, several non-exclusive structural mechanisms (**Figure 3**) may underlie the enhanced mutual interaction propensity of proteins containing certain AARs. Interestingly, some of these possibilities have already been demonstrated experimentally, while others will need to be further investigated in biological experiments. In the simplest case (**Figure 3A**), AARs themselves may be the structural mediators of protein-protein interactions. For instance, polyQ and polyA repeats can mediate protein interactions and oligomerization by forming coiled-coil structures (Fiumara et al., 2010; Pelassa et al., 2014). Another possibility (**Figure 3B**) is that AARs in one protein interact with another structural domain of another protein, as known for the case of proline-rich stretches forming polyproline-II (PP-II) structures which can be bound by SH3 domains (Yu et al., 1994). The enrichment of such AAR-targeting interaction domains in

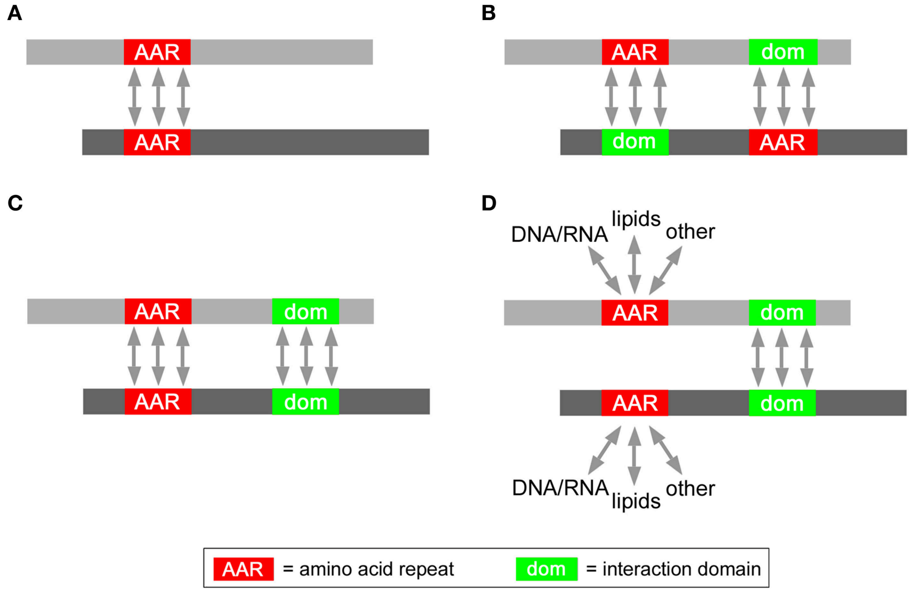


FIGURE 3 | Possible roles of AARs in protein interaction networks. Schematic representation of possible modes of interaction between two AAR-containing proteins (*gray bars*) mediated either by **(A)** homotypic AAR-AAR contacts, or **(B)** by heterotypic AAR-interaction domain contacts, or **(C)** homotypic AAR-AAR and domain-domain contacts, or by **(D)** homotypic domain-domain interactions.

the same proteins containing the AAR may explain the increased tendency of such proteins to interact with each other. In some polyX networks (**Figure 3C**) interactions may also be promoted synergistically by AARs and other conventional protein-protein interaction domains. For instance, polyQ and/or polyA repeats and flanking sequences with coiled coil propensity may co-operate in protein interactions (Fiumara et al., 2010; Pelassa et al., 2014). In principle (**Figure 3D**), certain AARs may even not have a direct role in promoting the interactions between proteins in which they are present (**Figure 3D**). In this case, the interaction would be mediated by conventional protein-protein interaction domains that are overrepresented in the AAR protein group. AARs in this scenario may be involved in interactions with other cellular components like nucleic acids and lipids, or may have other roles unrelated to protein interaction. A possible example of this scenario may be that of proteins containing both charged repeats like polyK and conventional CC domains. In this case, while coiled coils could mediate the protein-protein interactions, the charged repeats may mediate instead interactions with negatively charged surfaces such as the phospolipid bilayer.

Biological experiments will be ultimately necessary for discriminating between these possibilities for the different polyX protein groups. As a first step in this direction, however, it may be important to determine initially, through a systematic proteome-wide analysis, which protein domains are significantly overrepresented in each polyX protein group. These domains may in fact be responsible, together with AARs or by themselves, for the increased mutual interaction propensity of AAR-containing proteins. Thus, this analysis may ultimately guide the biological exploration of the role of AARs in protein interaction networks by indicating which AAR/domain associations are most likely to determine an increase in protein interactivity such as we observe in certain AAR-containing protein groups.

Co-occurrence in Proteins of AARs and Protein-Protein Interaction Domains

To determine whether specific protein domains are overrepresented in the different groups of polyX proteins, we analyzed statistically their domain composition using the DAVID database (Dennis et al., 2003). Specifically, we searched for protein domains which are enriched in each of the polyX protein lists (except for polyW) using a stringent statistical criterion ($p < 0.05$ after applying the Benjamini-Hochberg adjustment). Overall, this analysis revealed the overrepresentation of multiple types of protein domains in most polyX protein groups, with the exception of the poly-C, -M, -N, and -Y groups. An exhaustive list of these domains is reported in the **Supplementary Table 2** and is represented graphically in **Figures 4**, **5** and in **Supplementary Figure 1**. We categorized these domains in four groups, i.e., (i) protein-protein, (ii) protein-nucleic acid, (iii) protein-lipid interaction domains, and (iv) domains involved in other functions or with unclear function. Some domains belong to more than one category as they have been shown to mediate multiple functions (e.g., DNA binding and protein-protein interactions).

Several polyX groups of proteins displayed selective enrichments of domains belonging to these four categories, and individual domains can co-occur with multiple types of AARs. The highest number of significant enrichments of protein-protein interaction domains was observed in the poly-S,

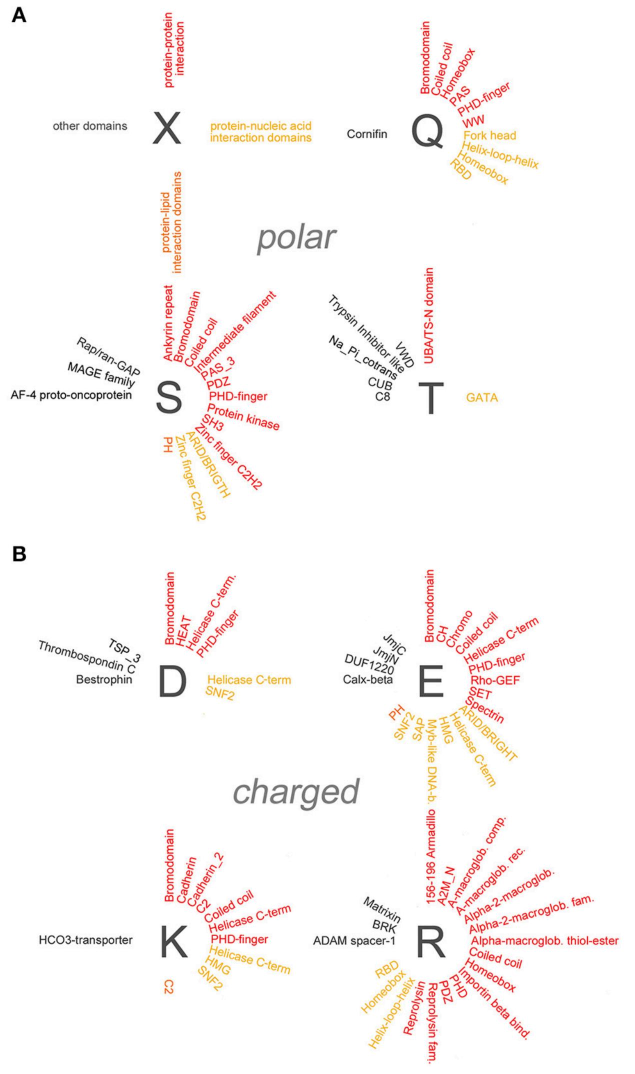


FIGURE 4 | Overrepresentation of protein-protein, -nucleic acid, and -lipid interaction domains in proteins containing polar and charged AARs. (A) Schematic radial representation of the protein domains that are significantly enriched in protein groups containing AARs formed by the polar amino acids indicated by the *dark gray letters*. The legend in the *upper left quadrant* (X) shows how protein-protein interaction domains are indicated in *red*, protein-nucleic acid interaction domains in *dark yellow*, protein-lipid interaction domains in *orange*, and domains with other or unknown functions in *black*. **(B)** As in **(A)** for groups of proteins containing charged AARs.

-E, -K, -R-, -P, and -L groups. Conversely, using the stringent criteria that were adopted, no significant overrepresentation of domains was observed in the poly-C, -M, -N, and -Y groups.

A paradigmatic case of co-occurrence of AARs and protein-protein interaction domains is that of coiled coil domains. These structural domains are indeed significantly overrepresented in proteins containing polar (poly-Q, -S), charged (poly-E, -K, -R), and small/cyclic (poly-A, -G, -P) AARs. Interestingly, in some of these cases (polyQ, polyA) the AARs themselves are known to form coiled coil structures often as part of conventional coiled coil domains (Fiumara et al., 2010; Pelassa et al., 2014), and the same may be in principle possible for short poly-S, -E, -K, and -R stretches when embedded in conventional coiled coil sequences. On the other hand, polyP and polyG stretches form other types of structures (Adzhubei et al., 2013) and their observed co-occurrence with coiled coil domains in the same proteins may not be obviously due to overlap between the AAR and the domain but to other functional reasons. These observations indicate that, at least in certain cases, the observed enrichments of certain

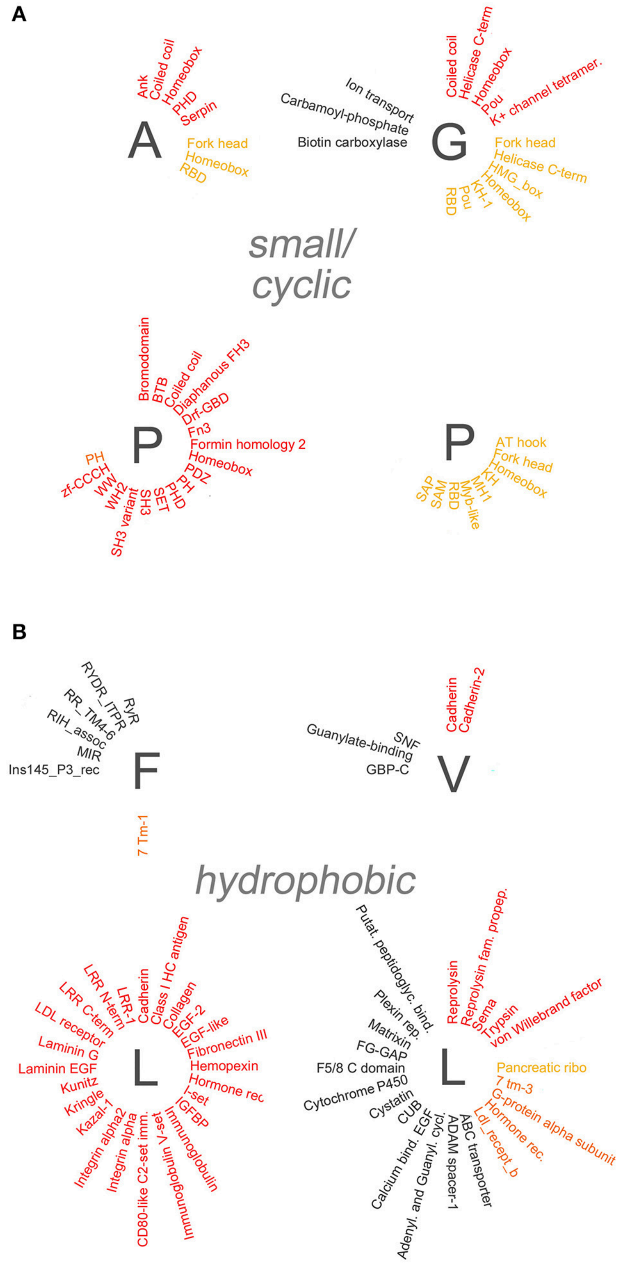


FIGURE 5 | Overrepresentation of protein-protein, -nucleic acid, and -lipid interaction domains in proteins containing small/cyclic and hydrophobic AARs. Schematic radial representation, as in **Figure 4A**, of the protein domains that are significantly enriched in protein groups containing AARs formed by small/cyclic **(A)** or hydrophobic **(B)** amino acids.

domains in some polyX protein groups may result from the at least partial overlap of the repeat and the domains (see the last section of the Results and the **Supplementary Table 3**). Besides coiled coils, other domains such as the bromodomain, which binds acetylated lysine residues on histone proteins, and the PDZ domain, which is commonly found in signaling proteins, are similarly overrepresented in multiple polyX protein groups (poly-D, -E, -K, -P, -Q, -S and poly-P, -R, -S, respectively). Other protein-protein interaction domains tend to co-occur more specifically only with a few polyX protein groups. For example, SH3 domains are enriched only in the poly-P and -S groups, whereas HEAT domains, helical structures involved in intracellular transport, are overrepresented only in the polyD protein group. Interestingly, some nucleic acid-binding domains which also function as protein-protein interaction domains, like the Homeobox domain, are overrepresented especially in protein containing certain polar (polyQ) charged (polyH), or small/cyclic (poly-A, -G, -P) repeats. PolyL proteins represent a quite unique case as they contain a high number of protein-protein interaction domains, mostly associated with trans-membrane or secreted proteins. Such abundance of overrepresented domains in the polyL group may be likely explained by the fact that polyL repeats are often found in signal peptide and transmembrane regions which are characteristic of proteins targeted to the secretory pathway or to cellular membranes with ligand/receptor functions (see Section Discussion). Notably, protein-protein interaction domains are conversely rarely co-occurring with other hydrophobic AARs.

Co-occurrence in Proteins of AARs and Protein-Nucleic Acid Interaction Domains

Domains known to mediate protein-nucleic acid interactions are frequently overrepresented in different polyX protein groups. In particular, multiple DNA-binding domains (e.g., Homeobox, Fork head, helix-loop-helix (HLH), helicase domains) co-occur in proteins with charged, polar, and small/cyclic AARs. The particular enrichment of DNA binding domains in groups of proteins containing charged repeats may reside in the capacity of charged AARs to bind DNA and chromatin components such as the histones (e.g., Dean, 1983; DeRouchey et al., 2013). Thus, charged AARs may have synergistic roles with DNA binding domains in driving the interaction of proteins with the nuclear genetic material. Instead the co-occurrence of this type of domain with hydrophobic AARs is quite exceptional. Interestingly, RNA-binding domains (RBD) are particularly enriched in protein groups containing polar (polyQ) and small/cyclic (poly-A, -G, -P) AARs, but not, at variance with DNA-binding domains, in protein groups with charged AARs, except for the group containing polyR repeats which may favor RNA binding (e.g., Nam et al., 2001).

Co-occurrence in Proteins of AARs, Protein-Lipid Interaction Domains, and Other Domains

We also found evidence for the overrepresentation of some lipid-binding domains in some polyX proteins groups. Rodopsin-like and class C G-protein-coupled receptors are overrepresented in polyI, polyF, and polyL proteins. PolyI repeats also co-occur with synaptobrevin domains. In most cases the hydrophobic repeats lie within the domains themselves as part of transmembrane regions (**Supplementary Table 3**). Two other lipid-binding domains are enriched in non-hydrophobic polyX protein groups. The CH2 domain targets proteins to membranes and is overrepresented in proteins containing polyK repeats, which may indeed also contribute to phospholipid binding (e.g., Reuter et al., 2009), whereas the pleckstrin homology (PH) domains, which bind phosphoinositides, are overrepresented in the poly-E, -S, and -P protein groups.

Taken together, these findings indicate that specific patterns of co-occurrence exist in proteins between AARs and protein domains that mediate interactions with other proteins, nucleic acids, and lipids. These domains, together with the AARs themselves, may contribute to shaping interactomes as illustrated in **Figure 3**.

Overlap of AARs and Protein Domains

As observed for polyQ and polyA repeats in coiled coil domains (Fiumara et al., 2010; Pelassa et al., 2014), the possibility exists that certain repeats may not only co-occur with interaction domains in the same proteins but may also be embedded within these domains. To determine whether this is the case, we verified in the NCBI Conserved Domains Database (CDD) (available at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), for each significant AAR-domain co-occurrence, in which proportion of proteins the AAR and the domain overlap for at least four residues. We found that in 148 out of 189 significant cases of AAR-domain co-occurrence no overlap exists between AARs and the domains that co-occur with them in the same proteins. However, for 41 cases of AAR-domain co-occurrence there is some sign of overlap between repeats and domains. In 21 cases, the overlap is observed in between 25 and 100% of the proteins containing the AAR-domain combination in the CDD database (**Supplementary Table 3**). For example, short polyR repeats were observed within the homeobox domains in 15 out of 22 proteins (i.e., 68%) that contain the polyR-homeobox association, and within the HLH domain in 8 out of 16 proteins (i.e., 50%) containing the polyR-HLH association. These observations indicate that AARs can be part of protein interaction domains and possibly play a functional role in them.

DISCUSSION

The results of our analyses indicate overall that the presence of certain types of AARs in protein networks is associated with a significantly increased protein-protein connectivity, and that significant patterns of co-occurrence, and in some cases overlap, exist between AARs and conventional protein interaction domains. These findings suggest that different types of AARs may play a generalized, combinatorial role in shaping protein interaction networks together with conventional protein-protein interaction domains they co-occur with.

Structural and Functional Roles of AARs in Protein Interactions

We found that proteins that contain a variety of polar (polyQ and polyS), charged (poly-D, -E, -K, -R), and small/cyclic (poly-A, -G, -P) repeats show a greater tendency to interact among themselves, in comparison with proteins devoid of these AARs. This is also paralleled, in the networks formed by proteins bearing these "interaction-enhancing" AARs, by a higher number of AAR-containing interaction partners per AAR protein. The opposite phenomena were observed for the protein groups containing certain hydrophobic AARs, like polyL and polyY, whereas for other AAR protein groups there was no evidence for statistically significant changes in type *a* connectivity. The observed reduction in connectivity among proteins containing polyL and polyY may likely result from the fact that these hydrophobic repeats are mostly part of transmembrane domains and signal peptides (Hikita and Mizushima, 1992; Zhou et al., 2001). In fact, the localization in membranes may limit relatively the possibility of proteins to interact with other proteins in the same membrane compartment, while preserving the possibility of interaction with intra- and extra-cellular (or intra- and extra-luminal, in the case of organelles) proteins. On the other hand, the increased propensity for mutual interactions observed among proteins containing poly-Q, -S, -D, -E, -K, -R, -A, -G, and -P repeats may have several possible, and not mutually exclusive, explanations, as schematized in **Figure 3**. Experimental evidence already indicates that at least polyQ and polyA repeats may directly mediate protein-protein interactions by forming coiled coil structures that can also interact with conventional, non-repetitive coiled coils (Fiumara et al., 2010; Schaefer et al., 2012; Pelassa et al., 2014). This type of interaction between proteins mediated directly by homotypic AAR-AAR contacts may not be a universal phenomenon in the polyX interactomes with enhanced connectivity that we have analyzed. In fact, polyproline-II structures in proline-rich and polyP-containing proteins are known to establish heterotypic interactions with SH3 domains (Yu et al., 1994) which we find being enriched precisely in the same polyP protein group. Thus, the enhanced connectivity observed in some AAR protein networks may result from the enrichment in them of interaction domains capable of AAR binding. While this possibility needs to be tested experimentally for the different AARs, our analyses identified a relatively restricted subset of significantly enriched domains in each polyX protein group that may play a role similar to that of SH3 domains in the interactome of polyP proteins. AARs and conventional protein interaction domains may also cooperate in mediating the binding of proteins to other proteins or to nucleic acids and cellular membranes. This seems particularly plausible for charged AARs. Charged AARs can indeed bind DNA and RNA (e.g., Dean, 1983; Nam et al., 2001; DeRouchey et al., 2013) and may therefore cooperate with sequence-specific DNA- or RNA-binding domains in stabilizing protein interactions with nucleic acids. Charged AARs can also bind histones and may cooperate with DNA-binding domains within the same protein that bind histone-associated DNA. Positively charged AARs can also bind phospholipids (Schwieger and Blume, 2007; Reuter et al., 2009) and we found indeed evidence of a significant overrepresentations of CH2 lipid binding domains in polyK proteins.

Physiological and Pathological Roles of AARs in Shaping Protein Interactomes

Taken together these observations indicate that, given their widespread presence in proteomes and their frequent co-occurrence with protein interaction domains, polyQ -S, -D, -E, -K, -R, -A, -G, and -P AARs may play a significant, generalized role in shaping protein interaction networks. Interestingly, most of these interaction-enhancing AARs can form defined secondary and supersecondary structures like α-helical coiled coil structures in the case of polyQ and polyA repeats (Fiumara et al., 2010; Pelassa et al., 2014), and polyproline II (PP-II) and polyglycine II (PG-II) structures (e.g., Adzhubei et al., 2013) in the case of polyP and polyG repeats, respectively. Also PolyE and polyK repeats can form helical structures in a pH-dependent manner (Inoue et al., 2005; Mirtič and Grdadolnik, 2013), and it is thus conceivable that short repeats of glutamate or other charged amino acids may well be incorporated into defined protein structures. Thus, AARs may favor the formation of protein interactions not only as intrinsically disordered domains through the formation of "fuzzy" complexes (van der Lee et al., 2014) but also through the formation of defined secondary structures, similar to conventional, non-repetitive interaction domains. This conclusion is also supported by our observation that conventional protein interaction domains can contain short AARs within them, which are likely to take part in some aspect of their structure/function.

Different AARs can co-occur in the same protein groups, as observed for polyQ and polyA repeats (Pelassa et al., 2014), and for other AARs (Pelassa and Fiumara, unpublished observations). These observations, together with the existence of specific patterns of co-occurrence of AARs and protein interaction domains, strongly suggest the existence of a combinatorial protein interaction code defined by the variable co-occurrence in different protein groups of multiple types of AARs and of conventional interaction domains. These domains can indeed establish homotypic AAR-AAR and domain-domain interactions, as well as heterotypic AAR-domain and domain-domain interactions. Thus, the combination in one protein of AARs and of various types of interaction domains can finely tune the specificity and stability of the binding of the protein to other proteins, but also to nucleic acids and to phospholipids in membranes. Our observations identify overall a number of potentially relevant AAR-domain co-occurrences whose functional relevance ought to be experimentally tested in different biological contexts, such as transcriptional and translational regulation, protein trafficking, *et cetera*. Biological experiments guided by our findings may ultimately help to define the exact role and the relative contribution of AARs and of co-occurring interaction domains in shaping both the physiological interactomes in the human proteome and the aberrant, pathological protein interaction networks that are established in polyQ or polyA expansion diseases.

Thus, in conclusion, the results of our analyses contribute proteome-wide quantitative evidence supporting the existence of physiological, structural and functional roles of AARs, and pave the way to the informed biological dissection of AAR-mediated protein interaction networks in health and disease.

MATERIALS AND METHODS

Datasets

The complete reference proteome of *Homo sapiens* was retrieved from the Uniprot database (www.uniprot.org) in FASTA format without isoforms. The proteins containing AARs of at least four units were identified using *ad hoc* Perl scripts as in Pelassa et al. (2014). The domain composition of the proteins of interest was derived performing batch searches on the NCBI CCD website (available at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Necessary conversions of the different protein identifiers found in the different databases were performed using the DAVID (https://david.ncifcrf.gov/) or Biomart (www.biomart.org) databases.

Definition and Analysis of PolyX Protein Interactomes and Control Random Interactomes

Protein interaction networks formed by proteins containing a given AAR and their interactors were extracted form the BioGrid database of protein-protein interactions using the g:Profiler web interface (available at http://biit.cs.ut.ee/gprofiler/), deselecting the "significant only" option so that all interactions could be downloaded in a tab-delimited text files. Control networks formed by proteins selected at random in the human proteome and their interactors were obtained in the same way. Randomness in the selection of the proteins was achieved by using a random number generator to select protein IDs from a complete list of all the human protein IDs ordered as elements of an array. In particular, we reiteratively used the Perl "rand" function to select sets of random elements of the desired numerosity from the elements of this array. The files derived from g:Profiler for both AAR and random networks were then analyzed with *ad hoc* Perl scripts in order to identify and quantify "type *a*" and "type *b*" interactions (see Section Results).

Analysis of the Overrepresentation of Protein Domains in PolyX Protein Groups

The protein domains that are overrepresented in the polyX protein groups were identified using the DAVID database. We searched, using the "Protein domains" selection menu, for "Pfam" domains enriched with a Benjamini score <0.05. Coiled coil domains were identified, using the "Functional categories" selection menu, searching for "SP_PIR_KEYWORDS."

Analysis of the Overlap Between AARs and Protein Domains

The overlap between AARs and conventional domains in proteins was determined using *ad hoc* Perl scripts. These scripts compared for each protein the relative positions of the AARs and of protein domains whose positions were derived from the NCBI CDD database (see the Section Datasets above). The overlap of coiled coil domains with polyQ and polyA repeats was shown previously (Fiumara et al., 2010; Schaefer et al., 2012; Pelassa et al., 2014), and here we did not analyze further the coiled coil/AAR overlap.

Graphs

Bar graphs were generated using Excel software (Microsoft). Network graphs were generated using CytoScape software (available at www.cytoscape.org) selecting the "degree sorted circle layout." Other schematic representations and figures were generated using Photoshop Elements 11 software (Adobe).

Data Analysis and Statistics

Data were processed and analyzed statistically using Excel (Microsoft) and SPSS 21 (IBM) software. Appropriate statistical tests were performed as indicated in the text and $p < 0.05$ was considered as statistically significant in all instances.

ACKNOWLEDGMENTS

This work was supported by Fondazione Telethon (grant GGP11223).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00345

Supplementary Table 1 | Number of proteins in the 20 polyX groups.

Supplementary Table 2 | Protein domains overrepresented in polyX protein groups.

Supplementary Table 3 | Overlap between AARs and protein domains.

Supplementary Figure 1 | Overrepresentation of protein-protein, -nucleic acid, and -lipid interaction domains in proteins containing polyH and polyI AARs.

REFERENCES

Adzhubei, A. A., Sternberg, M. J., and Makarov, A. A. (2013). Polyproline-II helix in proteins: structure and function. *J. Mol. Biol.* 425, 2100–2132. doi: 10.1016/j.jmb.2013.03.018

Almeida, B., Fernandes, S., Abreu, I. A., and Macedo-Ribeiro, S. (2013). Trinucleotide repeats: a structural perspective. *Front. Neurol.* 4:76. doi: 10.3389/fneur.2013.00076

Anan, K., Yoshida, N., Kataoka, Y., Sato, M., Ichise, H., Nasu, M., et al. (2007). Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Mol. Biol. Evol.* 4, 281–287. doi: 10.1093/molbev/msl161

Buschiazzo, E., and Gemmell, N. J. (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays* 28, 1040–1050. doi: 10.1002/bies.20470

Dean, J. (1983). Decondensation of mouse sperm chromatin and reassembly into nucleosomes mediated by polyglutamic acid *in vitro*. *Dev. Biol.* 99, 210–216. doi: 10.1016/0012-1606(83)90269-5

Dennis, G. Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* 4:P3. doi: 10.1186/gb-2003-4-5-p3

DeRouchey, J., Hoover, B., and Rau, D. C. (2013). A comparison of DNA compaction by arginine and lysine peptides: a physical basis for arginine rich protamines. *Biochemistry* 52, 3000–3009. doi: 10.1021/bi4001408

Faux, N. (2012). Single amino acid and trinucleotide repeats: function and evolution. *Adv. Exp. Med. Biol.* 769, 26–40. doi: 10.1007/978-1-4614-5434-2_3

Fiumara, F., Fioriti, L., Kandel, E. R., and Hendrickson, W. A. (2010). Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell* 143, 1121–1135. doi: 10.1016/j.cell.2010.11.042

Fondon, J. W. III, and Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U.S.A.* 101, 18058–18063. doi: 10.1073/pnas.0408118101

Green, H., and Wang, N. (1994). Codon reiteration and the evolution of proteins. *Proc. Natl. Acad. Sci. U.S.A.* 91, 4298–4302. doi: 10.1073/pnas.91.10.4298

Haerty, W., and Golding, G. B. (2010). Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences. *Genome* 53, 753–762. doi: 10.1139/G10-063

Hikita, C., and Mizushima, S. (1992). Effects of total hydrophobicity and length of the hydrophobic domain of a signal peptide on *in vitro* translocation efficiency. *J. Biol. Chem.* 267, 4882–4888.

Inoue, K., Baden, N., and Terazima, M. (2005). Diffusion coefficient and the secondary structure of poly-L-glutamic acid in aqueous solution. *J. Phys. Chem. B* 109, 22623–22628. doi: 10.1021/jp052897y

Karlin, S., and Burge, C. (1996). Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. U.S.A.* 93, 1560–1565. doi: 10.1073/pnas.93.4.1560

Mirtič, A., and Grdadolnik, J. (2013). The structure of poly-L-lysine in different solvents. *Biophys. Chem.* 175–176, 47–53. doi: 10.1016/j.bpc.2013.02.004

Nam, Y. S., Petrovic, A., Jeong, K. S., and Venkatesan, S. (2001). Exchange of the basic domain of human immunodeficiency virus type 1 Rev for a polyarginine stretch expands the RNA binding specificity, and a minimal arginine cluster is required for optimal RRE RNA binding affinity, nuclear accumulation, and trans-activation. *J. Virol.* 75, 2957–2971. doi: 10.1128/JVI.75.6.2957-2971.2001

Nasu, M., Yada, S., Igarashi, A., Sutoo, D., Akiyama, K., Ito, M., et al. (2014). Mammalian-specific sequences in pou3f2 contribute to maternal behavior. *Genome Biol. Evol.* 6, 1145–1156. doi: 10.1093/gbe/evu072

O'Malley, K. G., and Banks, M. A. (2008). A latitudinal cline in the Chinook salmon (*Oncorhynchus tshawytscha*) Clock gene: evidence for selection on PolyQ length variants. *Proc. Biol. Sci.* 275, 2813–2821. doi: 10.1098/rspb.2008.0524

Pelassa, I., Corà, D., Cesano, F., Monje, F. J., Montarolo, P. G., and Fiumara, F. (2014). Association of polyalanine and polyglutamine coiled coils mediates expansion disease-related protein aggregation and dysfunction. *Hum. Mol. Genet.* 23, 3402–3420. doi: 10.1093/hmg/ddu049

Petrakis, S., Raskó, T., Russ, J., Friedrich, R. P., Stroedicke, M., Riechers, S. P., et al. (2012). Identification of human proteins that modify misfolding and proteotoxicity of pathogenic ataxin-1. *PLoS Genet.* 8:e1002897. doi: 10.1371/journal.pgen.1002897

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* 39, W307–W315. doi: 10.1093/nar/gkr378

Reuter, M., Schwieger, C., Meister, A., Karlsson, G., and Blume, A. (2009). Poly-l-lysines and poly-l-arginines induce leakage of negatively charged phospholipid vesicles and translocate through the lipid bilayer upon electrostatic binding to the membrane. *Biophys. Chem.* 144, 27–37. doi: 10.1016/j.bpc.2009.06.002

Schaefer, M. H., Wanker, E. E., and Andrade-Navarro, M. A. (2012). Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. *Nucleic Acids Res*. 404, 273–87. doi: 10.1093/nar/gks011

Schwieger, C., and Blume, A. (2007). Interaction of poly(L-lysines) with negatively charged membranes: an FT-IR and DSC study. *Eur. Biophys. J.* 36, 437–450. doi: 10.1007/s00249-006-0080-8

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 34, D535–D539. doi: 10.1093/nar/gkj109

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., et al. (2014). Classification of intrinsically disordered regions and proteins. *Chem. Rev*. 114, 6589–6631. doi: 10.1021/cr400525m

Yu, H., Chen, J. K., Feng, S., Dalgarno, D. C., Brauer, A. W., and Schreiber, S. L. (1994). Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell* 7, 933–945. doi: 10.1016/0092-8674(94)90367-0

Zhou, F. X., Merianos, H. J., Brunger, A. T., and Engelman, D. M. (2001). Polar residues drive association of polyleucine transmembrane helices. *Proc. Natl. Acad. Sci. U.S.A.* 98, 2250–2255. doi: 10.1073/pnas.041593698

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 August 2015
doi: 10.3389/fgene.2015.00265

Detecting modules in biological networks by edge weight clustering and entropy significance

Paola Lecca[1]† and *Angela Re*[2]*

[1] *Centre for Integrative Biology, University of Trento, Italy,* [2] *Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento, Trento, Italy*

OPEN ACCESS

Edited by:
Spyros Petrakis,
Aristotle University
of Thessaloniki, Greece

Reviewed by:
Monika Heiner,
Brandenburg Technical University
Cottbus-Senftenberg, Germany
Qiangfeng Cliff Zhang,
Stanford University, USA

Correspondence:
Angela Re,
Laboratory of Translational Genomics,
Centre for Integrative Biology,
University of Trento,
Via Sommarive 9 - 38123 Povo,
Trento, Italy
angela.re@unitn.it

†Present Address:
Paola Lecca,
Member of Association for Computing
Machinery, New York, USA

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

Received: *25 March 2015*
Accepted: *30 July 2015*
Published: *27 August 2015*

Citation:
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

Detection of the modular structure of biological networks is of interest to researchers adopting a systems perspective for the analysis of omics data. Computational systems biology has provided a rich array of methods for network clustering. To date, the majority of approaches address this task through a network node classification based on topological or external quantifiable properties of network nodes. Conversely, numerical properties of network edges are underused, even though the information content which can be associated with network edges has augmented due to steady advances in molecular biology technology over the last decade. Properly accounting for network edges in the development of clustering approaches can become crucial to improve quantitative interpretation of omics data, finally resulting in more biologically plausible models. In this study, we present a novel technique for network module detection, named WG-Cluster (Weighted Graph CLUSTERing). WG-Cluster's notable features, compared to current approaches, lie in: (1) the simultaneous exploitation of network node and edge weights to improve the biological interpretability of the connected components detected, (2) the assessment of their statistical significance, and (3) the identification of emerging topological properties in the detected connected components. WG-Cluster utilizes three major steps: (i) an unsupervised version of k-means edge-based algorithm detects sub-graphs with similar edge weights, (ii) a fast-greedy algorithm detects connected components which are then scored and selected according to the statistical significance of their scores, and (iii) an analysis of the convolution between sub-graph mean edge weight and connected component score provides a summarizing view of the connected components. WG-Cluster can be applied to directed and undirected networks of different types of interacting entities and scales up to large omics data sets. Here, we show that WG-Cluster can be successfully used in the differential analysis of physical protein–protein interaction (PPI) networks. Specifically, applying WG-Cluster to a PPI network weighted by measurements of differential gene expression permits to explore the changes in network topology under two distinct (normal vs. tumor) conditions. WG-Cluster code is available at https://sites.google.com/site/paolaleccapersonalpage/.

Keywords: protein–protein network, weighted network, node weight, edge weight, clustering, connected component, entropy

1. Introduction

With biology increasingly becoming a data-rich field, objectives of systems biology research include organizing molecular interactions as networks and characterizing their structure, dynamics, and controllability. Since the turn of the century, high-throughput interaction mapping has emerged as an extremely useful approach to identify the constituents and connections of these networks. For instance the systematic identification of pairwise protein interactions (Rual et al., 2005; Petschnigg et al., 2014) or protein associations into complexes (Havugimana et al., 2012) has been enormously valuable both for understanding the function of individual proteins and for elucidating the organizing principles of the cellular physical architecture. Additional types of interactions have been charted including protein-DNA (Chen et al., 2015), protein-RNA (Moore et al., 2014; Re et al., 2014) and kinase-substrate (Linding et al., 2007; Varjosalo et al., 2013) interactions. Many of the molecular interaction data generated find their way into database resources that are available online (Turner et al., 2010; Horn et al., 2014; Orchard et al., 2014). The ability to generate, process and integrate omics data is instrumental to increasingly faithful reconstructions of the information flow in biological systems. In this vein, the conceptualization of biological systems as networks and the subsequent reconstruction of their modular organization acquire great interest (Barabási and Oltvai, 2004; Barabási et al., 2011; Ideker and Krogan, 2012). The notion of a module refers to a discrete entity whose constituent elements are similar in some quantifiable (e.g., chemical, physical, or functional) property and/or in the profile of their relationships. Biology displays many examples of modules which generally accomplish relatively separable functions such as nucleic acid synthesis, DNA replication, mitotic spindle assembly and protein degradation (Hartwell et al., 1999; Barabási and Oltvai, 2004).

In recent years, a rich collection of computational approaches has emerged for module detection in weighted networks, where weights can be constrained by topological or alternative numerical properties of nodes (for example, node molecular activity extracted from transcriptomics profiling) and edges (for example, edge confidence). Aside from weight assignment either to nodes (Ideker et al., 2002; Bader and Hogue, 2003) or to edges (Tanay et al., 2004; Liu et al., 2009; Pandey et al., 2014), clustering algorithms differ in the procedures for finding modules including, for example, simulated annealing (Ideker et al., 2002), greedy (Chuang et al., 2007; Nacu et al., 2007), genetic (Klammer et al., 2007), and network propagation (Vandin et al., 2011; TCGA Research Network, 2013) algorithms. Despite all of this exciting research in network clustering, some limitations stand out as remarkable. First, processing tens of thousands of nodes and the edges among them is hard to accomplish in fast timescales. Second, albeit equally interesting properties, it remains unclear how to meaningfully account for both node and edge weights in a module detection procedure.

Here, we present a novel algorithm for modular structure detection, named WG-Cluster (Weighted Graph CLUSTERing), which seeks to address previous shortcomings to detect modules. Within WG-Cluster, a module is defined as a connected component where nodes are characterized by homogeneous weights and are connected by edges of homogeneous weights. To this aim, WG-Cluster combines an edge-based network clustering with a fast-gready algorithm. The treatment of network edge weights within WG-Cluster represents a novelty compared to most clustering algorithms since, by the initial edge-based network clustering, network edge weights underlie the subsequent detection and prioritization of the connected components. Furthermore, the procedural choice adopted by WG-Cluster permits to obtain modules, homogeneous not only in node weights but also in edge weights, without discernible additional cost in computational efficiency. Module prioritization can become particularly useful in applications related to differential network analysis where the primary goal is to identify modules changing across different conditions. Finally, it is worth mentioning here also the introduction of a measure of the significance of the returned connected components which is based on node weights. WG-Cluster is here applied for the analysis of a differential network, i.e., a network where node and edge weights are defined by the changes observed in node and edge numerical properties between two conditions. Differential network analysis is useful to tackle the dynamic nature of molecular interactions, for instance as a consequence of environmental shifts. Computational integration of a network with molecular profiles acquired in different contexts has shown a popular approach to extract context-dependent responsive modules, which mark strikingly changed regions of the network. The input network for the current WG-Cluster application is a differential network, which was obtained by integrating a physical protein–protein interaction (PPI) network with changes in gene expression between a normal and tumor conditions. Our analysis showed that WG-Cluster is useful for comprehensively analysing the quantitative changes affecting nodes or interactions in the network and for recognizing modules which link to functional properties.

2. Materials and Methods

2.1. Data Description and Pre-processing

We gathered multi-assay omics data to define the weighted network which is the primary input to WG-cluster. We collected PPIs from the open-access IntAct database which adopts a merging algorithm and a scoring system to provide richly annotated molecular interaction data. IntAct PPIs are described in the controlled vocabulary specified by the Proteomics Standards Initiative for Molecular Interaction (PSI-MI) data (Hermjakob et al., 2004) and adhere to the guidelines (Orchard et al., 2007) about the Minimum Information required for reporting a Molecular Interaction Experiment, which were supplied by the International Molecular Exchange (IMEx) consortium. PPIs involving human protein entities were selected and downloaded along with their confidence scores. Protein identifiers defined by the Universal Protein Resource (Uniprot) protein accessions (http://www.uniprot.org/) were mapped to gene identifiers defined by the HUGO Gene Nomenclature Committee (HGNC) gene symbols (http://www.genenames.

org/). We next integrated the IntAct PPIs with tumor-dependent changes in messenger RNA (mRNA) expression profiles. Processed gene expression data related to colon adenocarcinoma were downloaded from The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/). mRNA profiles were generated from 155 tumor and 19 normal tissue samples. Processed data were lowess normalized and collapsed by gene symbol ($\log_2$ scale). A differential co-expression score was computed for each gene pair, by subtracting the pairwise Pearson's correlation coefficient in the tumor condition from that in the normal condition. Next, the IntAct PPI confidence scores were multiplied by the differential co-expression scores to estimate the change in the interaction strength resulting from the differential co-expression of the mRNAs encoding the interacting proteins. The product between the IntAct score and the differential co-expression score defines the final weight of an edge in the differential network. The weight of a node in the differential network was obtained by computing the ratio between the average values of mRNA expression across samples in the normal and tumor conditions (mRNA fold change). This differential network, where both nodes and edges were weighted, was the primary input to the WG-Cluster algorithm.

2.2. WG-Cluster

The WG-Cluster algorithm is implemented in R (R software available at http://www.r-project.org), which provides one of the most widely used, most flexible and mature open source environments. For the most computationally intense tasks WG-Cluster employs built-in R functions implemented as a C(++) or Fortran code, that are optimized and faster than functions coded in R from scratch. The input data consist of the network edges reported in Simple Interaction File (SIF) format (Cytoscape, 2015) and of node weights reported in tabular format (node, weight). The algorithm sequentially executes three computational modules. First, it estimates the optimal number of clusters (sub-graphs) that split up the graph (i.e., network) and executes a Lloyd's K-means clustering (Du et al., 2006) of the edge weights to detect sub-graphs with edges of similar weights. Second, a fast-greedy modularity optimization procedure (Clauset et al., 2004) finds (if any) the connected components (i.e., modules) in each sub-graph. An entropy score is computed for each connected component and is used as a measure of the statistical significance of the connected component. Finally, an analysis of the convolution between sub-graph mean edge weight and connected component entropy allows for a summarizing view of both properties in the detected connected components (**Figure 1**). In the following, we give the details about each computational module of WG-Cluster. Hereafter, we will denote with V the number of vertices and with NE the number of edges in the input graph.

2.2.1. Detection of Sub-graphs

The optimal number of sub-graphs which partition the input graph is estimated by minimizing the total within-clusters sum of squares (WCSS) obtained with a K-means procedure. For a set of edge weights $\mathbf{w} = (w_1, w_2, \ldots, w_{NE})$, K-means clustering tries to find a set of K sub-graphs $S = (S_1, S_2, \ldots, S_K)$ that is a solution to the minimization problem:

$$WCSS = \sum_{i=1}^{K} \sum_{\mathbf{w} \in S_i} ||\mathbf{w} - \mu_i||^2$$

where μ_i is the mean of the edge weights $\mathbf{w}$ in the sub-graph S_i.

An elbow in the curve interpolating the points ($n_{\text{sub-graphs}}$, WCSS) suggests the appropriate number of sub-graphs n_{optimal}. In our implementation, n_{optimal} is estimated as the minimum value of n_{clusters} at which the first derivative of WCSS w.r.t. $n_{\text{sub-graphs}}$ is null within a tolerance $0 < \epsilon \ll 1$, i.e.,

$$\left| \frac{d\ WCSS}{dn_{\text{sub-graphs}}} \right| \leq \epsilon.$$

The first derivative of the curve ($n_{\text{sub-graphs}}$, WCSS) is calculated by the Stineman algorithm (Johannesson and Bjornsson, 2012). Algorithm 1 reports the pseudo-code of the first module of WG-Cluster.

The problem of WCSS minimization is known to be NP-hard, implying long running times, that can become unacceptable in case of biological networks with thousands of nodes and tens of thousands of edges. Furthermore, if the input data do not have a strong clustering structure, the procedure may not converge. For this reason, WG-Cluster adopts the Lloyd's algorithm whose complexity is linear in the number of edges and number of sub-graphs, and is recommended in case of data poorly clustered (Du et al., 2006). Algorithm 2 presents the pseudo-code of the Lloyd's K-means. Those iterations are repeated until the centroids stop changing, within a tolerance quantified by the parameter `threshold` (see the pseudo-code 2).

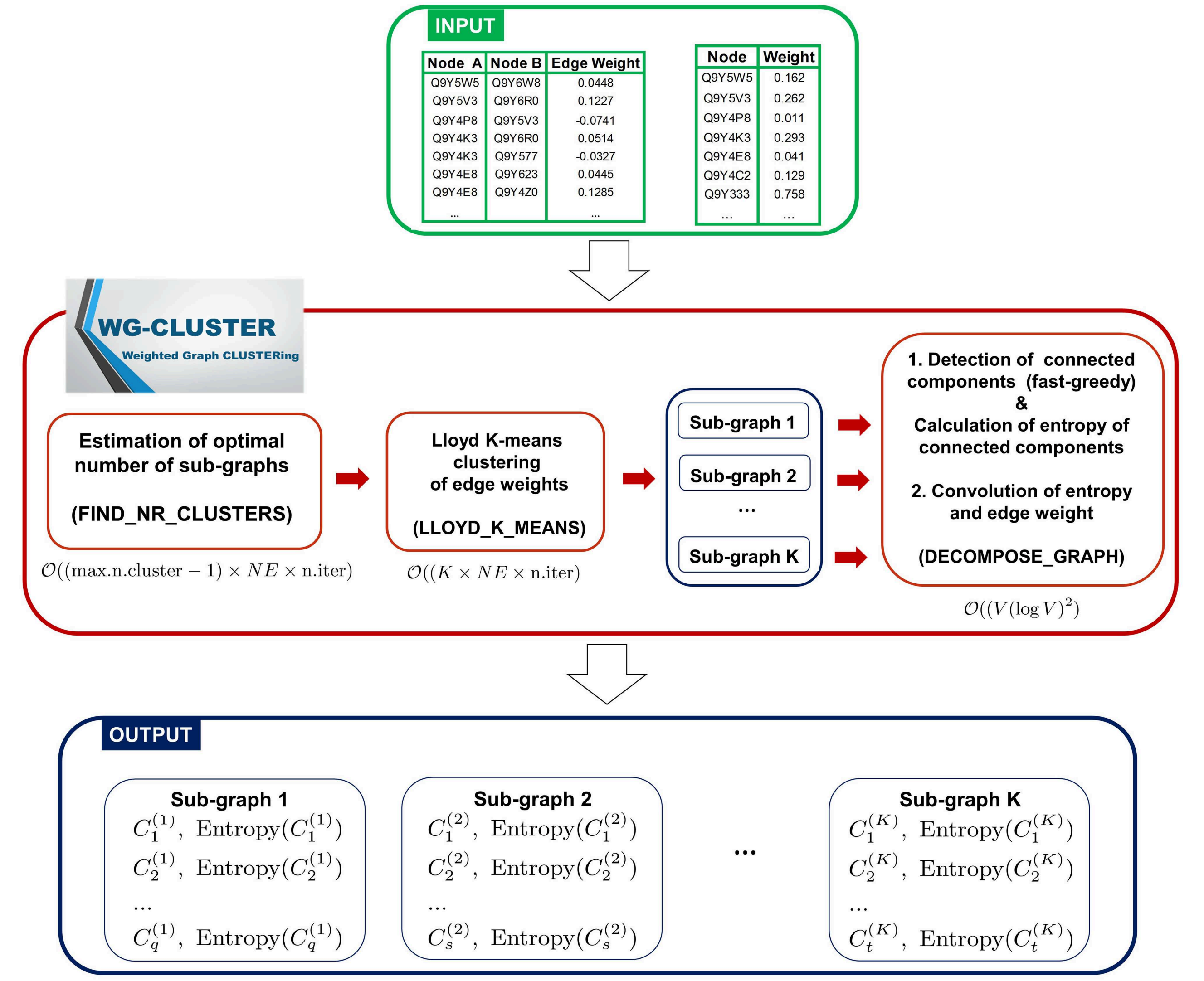
In Supplementary Material (Section 1.1) we present the exploratory analysis of other clustering approaches and the motivation of the choice of the K-means algorithm in WG-Cluster.

2.2.2. Detection of Connected Components

Each sub-graph S_i ($i = 1, \ldots, K$) returned by the K-means clustering is decomposed into connected components $C_l^{(i)}$ (with $l = 1, 2, \ldots, L_i$, where L_i is the number of connected components in the sub-graph S_i) via a fast-greedy optimization procedure (Clauset et al., 2004), as illustrated in **Figure 2**. The entropy of each connected component is calculated as follows:

$$E_{C_l^{(i)}} = -\sum_{j=1}^{N(C_l^{(i)})} \frac{p_j \log_2 p_j}{d_j} \qquad (1)$$

where $N(C_l^{(i)})$ is the number of nodes in the connected component $C_l^{(i)}$, p_j is the fold change of the expression level (from normal to tumor condition) of gene j (normalized between 0 and 1) and d_j is the sum of the weights of the edges adjacent to the node representing gene j (known as *node strength*). Denoting



Node A	Node B	Edge Weight
Q9Y5W5	Q9Y6W8	0.0448
Q9Y5V3	Q9Y6R0	0.1227
Q9Y4P8	Q9Y5V3	-0.0741
Q9Y4K3	Q9Y6R0	0.0514
Q9Y4K3	Q9Y577	-0.0327
Q9Y4E8	Q9Y623	0.0445
Q9Y4E8	Q9Y4Z0	0.1285
...		...

Node	Weight
Q9Y5W5	0.162
Q9Y5V3	0.262
Q9Y4P8	0.011
Q9Y4K3	0.293
Q9Y4E8	0.041
Q9Y4C2	0.129
Q9Y333	0.758
...	...

WG-CLUSTER
Weighted Graph CLUSTERing
Estimation of optimal number of sub-graphs
(FIND_NR_CLUSTERS)
$\mathcal{O}((\text{max.n.cluster} - 1) \times NE \times \text{n.iter})$
Lloyd K-means clustering of edge weights
(LLOYD_K_MEANS)
$\mathcal{O}((K \times NE \times \text{n.iter})$
Sub-graph 1
Sub-graph 2
...
Sub-graph K
1. Detection of connected components (fast-greedy) & Calculation of entropy of connected components
2. Convolution of entropy and edge weight
(DECOMPOSE_GRAPH)
$\mathcal{O}((V(\log V)^2)$
OUTPUT
Sub-graph 1
$C_1^{(1)}$, Entropy$(C_1^{(1)})$
$C_2^{(1)}$, Entropy$(C_2^{(1)})$
...
$C_q^{(1)}$, Entropy$(C_q^{(1)})$
Sub-graph 2
$C_1^{(2)}$, Entropy$(C_1^{(2)})$
$C_2^{(2)}$, Entropy$(C_2^{(2)})$
...
$C_s^{(2)}$, Entropy$(C_s^{(2)})$
...
Sub-graph K
$C_1^{(K)}$, Entropy$(C_1^{(K)})$
$C_2^{(K)}$, Entropy$(C_2^{(K)})$
...
$C_t^{(K)}$, Entropy$(C_t^{(K)})$

FIGURE 1 | Algorithmic modules of WG-Cluster. WG-Cluster takes as input the SIF file of the network edges and a text file reporting node labels in the first column and node weights in the second one. If the second file is not available, WG-Cluster by default assigns an equal weight to all nodes. WG-Cluster implements three computational modules: (i) an unsupervised version of the K-means algorithm identifies sub-graphs with similar edge weights, (ii) a fast-greedy algorithm detects the connected components of each sub-graph utilizing similarity in node topological properties, (iii) the estimates of the convolution of connected component entropy and sub-graph mean edge weight guide the selection of significant connected components representative of global trends in the network. Complexity of modules for estimating the optimal number of sub-graphs and for running the Lloyd's K-means is linear in the number of edges *NE* and number of iterations; the complexity of the module for detecting connected components is $\mathcal{O}(V(\log V)^2)$, where *V* is the number of vertices.

with $D^{(j)}$ the number of nodes directly connected to node j, d_j is thus defined as

$$d_j = \sum_{h=1}^{D^{(j)}} w_{jh}.$$

where w_{jh} is the edge weight between the node j and its directly connected node h.

The entropy is used as a measure of significance of the connected components. In order to establish a threshold on the entropy significance, we generated for each connected component $C_l^{(i)}$ an ensemble of 100 random connected components with the same degree distribution of the reference connected component $C_l^{(i)}$.

A connected component is considered significant, and retained, if its entropy value is more than three standard deviations far from the mean entropy of the corresponding ensemble of random connected components. Let denote with $\{C_{l'}^{(i')}\}$, where $l' \in \{1, 2, \ldots, L_i'\}$ with $L_i' \leq L_i$, and $i' \in \{1, 2, \ldots, K'\}$ with $K' \leq K$.

Algorithm 1 | Compute the optimal number of sub-graphs K

```
procedure FIND_NR_SUB_GRAPHS(edge.weights, max.n.sub.graphs, seed)

    $NE$ ← Number of edges of the graph

1. Calculate the within-cluster sum of squares (wcss) via a K-means solution.

    wcss[1] ←(NE - 1) × Variance(edge.weights)

    set.seed(seed)
    for (i in 2:max.n.sub.graphs) do
        wcss[i] ← $\sum_1^i$ ( calculate.wcss ( K-means(edge.weights, centroids = i) ) )
    end for


2 Estimate $d\ WCSS/dn.sub.graphs$ with the Stineman algorithm.

    n.sub.graphs ← 1:max.n.sub.graphs
    wcss.derivative ← Stineman.derivative(n.sub.graphs, wcss)

3. Set a tolerance value.

    tolerance ← $\epsilon$

4. Find the first local minimum of $d\ WCSS/dn.sub.graphs$

    wcss.derivative.null ← $\{-\epsilon \leq$ wcss.derivative $\leq \epsilon\}$
    K ← wcss.derivative.null[1]

5. Return the optimal number of sub-graphs $K$.
    Return $K$

end procedure
```

Algorithm 2 | Lloyd's K-means algorithm

```
procedure LLOYD_K_MEANS(edge.weights, K, distance)

    1. Randomly choose $K$ items from the edge weights vector and use these as the initial means.

    2. Iterations of assignments and centroid recalculation.
    while distance(centroids, edge.weights) > threshold do
        a. Assign edge weights to the centroids
        for i $\leq NE$ do
            Assign edge.weights[i] to closest sub-graph according to the distance measure.
        end for
        b. Recalculate centroids.
    end while
end procedure
```

2.2.3. Convolution of Mean Edge Weight and Entropy

Both the connected component entropy and the mean weight of the edges of the sub-graph to which a connected component belongs are considered to classify the connected components.

The convolution of the entropy of selected connected components (E_{selected}) with the mean edge weight MW of the sub-graphs to which they belong is performed as follows:

$$E_{\text{selected}}[h] * MW[h] = \sum_q E_{\text{selected}}[q] \cdot MW[q-h] \qquad (2)$$

where $E_{\text{selected}} = \{E_{C^{(i')}_{l'}}\}$ is the vector of the entropies of the significant connected components, and $MW = \{(1/NE^{(i')}) \sum_{l=1}^{NE^{(i')}} w_l\}$ is the mean edge weight of the sub-graph to which they belong.

The convolution in Equation (2) calculates the area overlap between the probability distributions of the entropy and of the mean edge weight as a function of the amount by which one of the distribution is translated. The area of the overlap of the two distribution measures the similarity between the entropy and mean edge weight distribution. The density of the convolution is a spectrum of the frequency of this similarity score and offers a way to classify the connected components by their membership to intervals of frequency corresponding to local maxima or minima of the convolution density. Maxima of the convolution density correspond to the most frequent values of similarity between entropy and mean edge weight, whereas local minima correspond to the least frequent values of similarity. Then, connected components can be classified according to the frequency of the convolution between their entropy and the mean edge weight of the sub-graph to which they belong. Algorithm 3 provides the steps of the pseudo-code implementing the procedure of detection and selection of significant connected components.

2.3. Functional Analysis of Connected Components

Enrichment analysis based on the generic Gene Ontology (GO) slim (http://geneontology.org/), a cut-down version of the Gene Ontology annotations, was conducted for each retained connected component (hypergeometric test). GO enrichment p-values were transformed in Benjamini-Hochberg false discovery rate (FDR) values and retained at the significance level of 0.05.

3. Results

3.1. Performances on synthetic data

We evaluated the performances of WG-Cluster in processing Erdös-Rényi random graphs, consisting of 500 nodes and an increasing number of edges, in terms of user CPU running time.

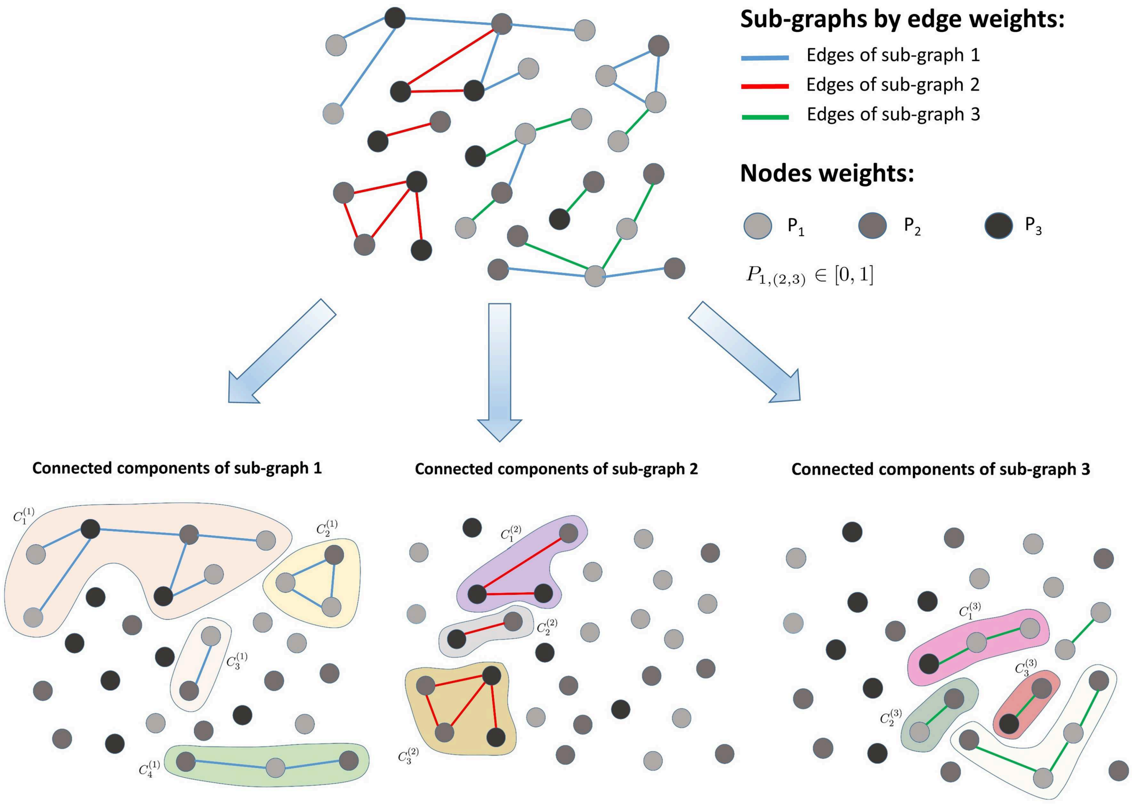


FIGURE 2 | Sub-graph decomposition into connected components. The algorithm first clusters the input graph into sub-graphs consisting of similar edge weights and next detects the connected components present within each sub-graph.

Algorithm 3 | Detection and selection of connected components

```
procedure DECOMPOSE_GRAPH(sub-graphs, node.weights)

1. Detection of connected components and calculation of their entropy.

    for (i in 1:K) do

        a. Fast-greedy decomposition of the i-th sub-graph into connected components.
        connected.components[[i]] ← fast.greedy.decomposition(sub-graph[i])

        b. Entropy calculation.

        for (l in 1:$L_i$) do
            (b.1) entropy of connected components with Equation (1).

            connected.components.entropy[l] ← entropy(connected.components[[i]][l], node.weights)

            (b.2) Generate an ensemble of random weighted Erdös-Renyi connected components.

            for (v in 1:100) do
                random.cc.component ← erdos.renyi.graph(nr.of.nodes=$N(C_l^{(i)})$, nr.of.edges=$NE(C_l^{(i)})$)
                edge.weights.random.cc.component ← Unif(0, 1)
                node.weights.random.cc.component ← Unif(0, 1)
                random.cc.entropies[v] ← calculate.entropy(random.cc.component,node.weights.random.cc.component,edge.weights.random.cc.component)
            end for

            (b.3) Calculate the mean of the entropies of the ensemble of random connected components.
            random.cc.entropy[l] ← calculate.mean.entropy(random.cc.entropies)
            (b.4) Select connected components.
            if cc.entropy[l] $\notin \left[-3\,\sigma + \text{random.cc.entropy[l]}\,, +3\,\sigma + \text{random.cc.entropy[l]}\right]$ then
                selected.connected.components $\leftarrow C_l^{(i)}$
            else
                discard $C_l^{(i)}$
            end if
        end for
    end for

2. Convolution of Entropy ($E$) and mean edge weight ($MW$).

    (a) Calculate the density of the convolution estimated by Equation (2).
    density.of.convolution $\leftarrow$ density(convolve (E, MW))
    (b) Detect the maxima of the convolution density.
    (c) Select and return the connected components whose values of convolution of (E, MW) fall under convolution density maxima.
end procedure
```

TABLE 1 | Summary of widely used hierarchical methods for module detection.

Method	Type of graph	Weighted edges	Weighted nodes
Edge-Betweenness (Girvan and Newman, 2001)	Directed and undirected	True	False
Fast-greedy (Clauset et al., 2004)	Directed and undirected	True	False
InfoMap (Rosvall and Bergstrom, 2008)	Directed and undirected	True	True

"True" and "False" in the two last columns stand for "the method can process also" and "the method does not process," respectively. For instance, edge-betweenness clustering method can process and take into account edge weights, but it does not handle information about node weights.

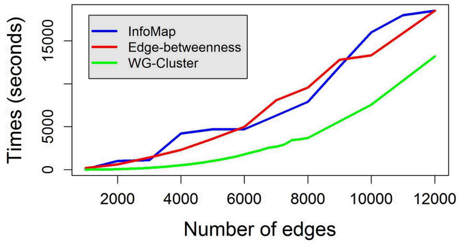


FIGURE 3 | Running times to cluster random weighted graphs with increasing number of edges. WG-Cluster running time on a random weighted graph of 500 nodes and an increasing number of edges is compared with that achieved by the edge betweenness graph clustering algorithm (Girvan and Newman, 2001) and that of InfoMap (Rosvall and Bergstrom, 2008). Each algorithm was utilized in its R implementation on a desktop Windows 8.1 PC with a 3.1 GHz CPU. WG-Cluster ensured faster running time and a RAM usage inferior to 3Gb.

Edge weights were drawn from a uniform distribution between 0 and 1 and clustered in 10 groups. A uniform distribution between 0 and 1 was also used to obtain node weights. We compared WG-Cluster running times to the running times of three widely used deterministic hierarchical approaches to graph clustering: (i) edge betweenness based clustering, (ii) label propagation, and (iii) InfoMap, which were selected because they handle directed (as well as undirected) and weighted networks as WG-Cluster does (see **Table 1** for a summary of the currently available deterministic clustering methods implemented in R). Non-deterministic clustering algorithms [e.g., Walktrap (Pons and Latapy, 2005), Spinglass (Reichardt and Bornholdt, 2006), and label propagation (Raghavan et al., 2007)] were left out of this comparative analysis since they require the determination of the number of runs needed to build a consensus partition. This parameter often depends on the topological structure of the graph and can remarkably affect the performances (that are usually satisfactory on single runs). We also excluded from the comparison the algorithms that do not handle the processing of undirected networks [e.g., Leading eigenvectors, (Newman, 2006)]. From this analysis, WG-Cluster showed to outperform the alternative algorithms (**Figure 3**).

In Supplementary Material, Section 1.2, we provide a more comprehensive analysis of the time complexity of WG-Cluster applied to random graphs of increasing number of edges and number of nodes.

Finally, further improvements in efficiency will be tested in the next version of WG-Cluster by the usage of recent libraries developed specifically to perform an optimized memory-efficient management of large datasets. The input/output and data rearrangement operations on large datasets are computationally time consuming, and their speeding is one of the main research topic engaging the developers of the majority of programming languages. R proposed two major solutions to optimize the efficiency of massive dataset processing (Kane and Emerson, 2013; Adler et al., 2014). Using these solutions, WG-Cluster could take advantage of the benefits of R (i.e., interactive data analysis and rich, flexible statistical programming environment), and, at the same time, of the benefit of C(++) language, i.e., an optimized memory-efficient management of big datasets.

3.2. Application

Biological systems are highly dynamical entities by depending on environment, tissue type, disease state or development. Nonetheless, relatively little effort has been spent in differential network analysis, i.e., the analysis of the changes occurring in a network in response to different conditions. Even though an increasing number of studies seek to analyse the dynamics of networks directly, through experimental mapping of networks across multiple conditions (Grossmann et al., 2015; Martin et al., 2015), a longstanding approach in differential network biology is to construct differential networks by integrating static (at standard laboratory conditions) molecular interaction networks (e.g., PPI networks) with changes observed in messenger RNA expression in different biological conditions (de Lichtenberg et al., 2005). The resulting differential network is a weighted network where node weights reflect the changes in mRNA expression levels and where edge weights reflect the changes in interaction strengths due to differential mRNA co-expression levels under the two conditions. It is worth noting that the strongest differential interactions are not necessarily the strongest ones in the static networks. Since both node and edge properties are deeply ingrained in the clustering procedure, WG-Cluster can provide a unique view of the differences in network topology between any two biological conditions.

As a proof-of-principle, we applied the WG-Cluster approach to analyse the differential PPI network that arises when tumor and normal conditions are contrasted. The current application focused on the colorectal cancer which stands among the most common cancers with more than 1.2 million new cases and about 600,000 deaths per year worldwide (Jemal et al., 2011). Messanger RNA expression data were obtained from The Cancer Genome Atlas which, importantly, provides samples from tumor tissues and from matched normal tissues. We acquired PPIs from the IntAct database because it provides a heuristic scoring system which relies on the available annotation evidences associated with an interacting pair of proteins. The differential network was constructed as follows: a node was weighted by the mRNA fold change and an edge was weighted by multiplying the IntAct PPI confidence score with the difference of the mRNA co-expression scores between the normal/tumor conditions. In the Supplementary Material (Sections 1.3 and 1.4), we provide a full description of this network edge weight model. Furthermore, we show that this network edge weighting approach leads to improved clustering quality compared to the classical approach which is based only on differential co-expression. The differential network consisted of 5569 nodes and 18,078 edges, out of which 8880 were strengthened and 9198 weakened in the tumor condition compared to the normal one.

Applying WG-cluster to the differential network detected 6215 connected components which were arranged in 29 sub-graphs of distinct mean edge weights. Upon connected components detection, WG-Cluster allows the estimation of the statistical significance of the entropy of each connected component by comparing the observed value against the distribution of entropies obtained from appropriately randomized connected components. The rate of connected component exclusion appeared stably moderate when we incremented the number of standard deviations from the expected entropy value; setting this number at three resulted in the exclusion of 26.87% of connected components (**Figure 4A**). Statistically significant connected components can be prioritized by any sort of network property such as mean edge weight of the sub-graph, or entropy or number of nodes by connected component. It is noteworthy that the numerical features associated with each connected component provide complementary information. For instance, correlation between mean edge weight and entropy values was not statistically significant (Spearman's coefficient $= -0.02$, $P = 0.15$). Since the mean weight of the edges in a sub-graph reflects the mean change in interactions strength and the entropy of a connected component reflects the mRNA expression changes, the observed lack of correlation is interesting because it is in agreement with previous data showing that the strongest differential interactions do not necessarily involve the strongest differential genes (Ideker and Krogan, 2012).

The last WG-Cluster step implements the convolution of the probability distribution of the connected component entropy with that of the sub-graph mean edge weight. This operation offers an appealing way to classify connected components in terms of both of those properties which, in our vision, are of equal interest. Since we were interested into obtaining a summarizing view of the network clustering, we selected the connected components yielding the most frequent convolution values (**Figure 4B**). We then interpreted those convolution values in terms of the corresponding sub-graph mean edge weight and connected component entropy values.

The number of the connected components obtained was found to increase in sub-graphs yielding lower mean edge weight (**Figure 4D**); conversely, no trend was detectable by analysing the mean entropy of the selected connected components resulting from each sub-graph (**Figure 4E**). Since the edge scores in the differential network result from the product of the IntAct scores with the differential co-expression scores, we verified that a low mean edge weight depended on low differential co-expression score, which resulted to be the case; indeed, the percentage of interactions where the differential co-expression score was higher than the IntAct score positively correlated with the sub-graph mean edge weight (**Figure 4C**).

In summary, by a general survey of WG-Cluster outcome, the majority of the detected connected components were found to consist of moderately changing interactions. More interestingly, the arrangement of the detected connected components by decreasing sub-graph mean edge weight (as shown in **Figure 4D**), which is inherent to WG-Cluster, streamlined the identification of connected components of markedly changing interactions. Those connected components, albeit limited in number, are undoubtedly the most interesting for unveiling the most striking changes in network topology between tumor/normal conditions (**Figure 4D**). Gene Ontology enrichment analysis was conducted to broadly assess the functional significance of module selection since exploring the fine details of specific modules is out of the scope of our study. This analysis showed that sub-graph clustering by mean edge weight broadly corresponded to a clustering of GO biological processes (**Figure 5**). Genes involved in cell cycle, cell death, mRNA processing and protein modification processes were found to be overrepresented in modules of weakened interactions in the tumor compared to the normal condition (sub-graph positive mean edge weight). On the other hand, genes acting in cell adhesion, extracellular matrix organization and cell-cell signaling resulted overrepresented in modules of interactions which were found strengthened in the tumor vs normal condition (sub-graph negative mean edge weight). It is reassuring that the GO categories overrepresented in the connected components were largely found in agreement with a previous survey of pathways consistently overrepresented in a large collection of signatures of differentially expressed genes of prognostic value in colorectal cancer (Lascorz et al., 2011). This case study showed that WG-Cluster allows shedding light into the network organization by fast and statistically robust module detection. In the context of a differential network analysis, it

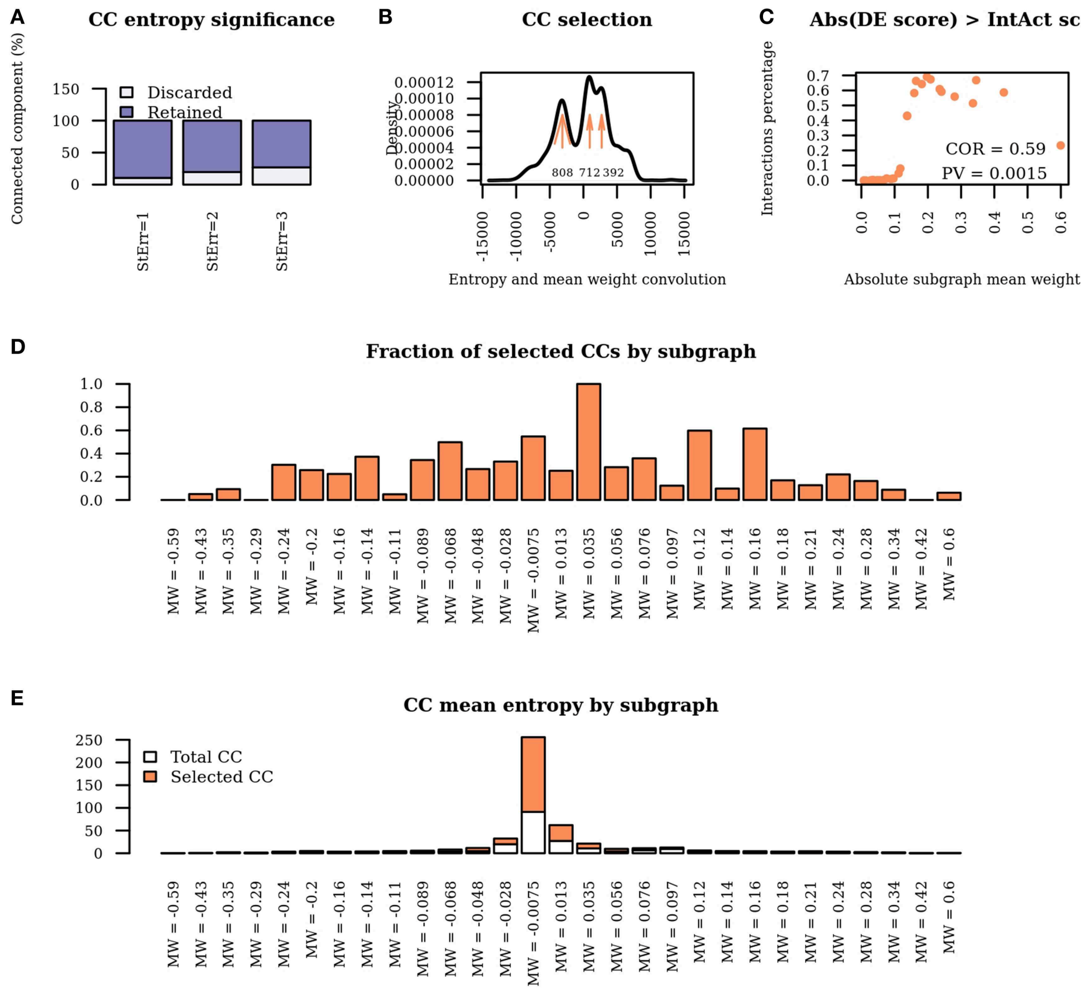


FIGURE 4 | Network properties of WG-Cluster reconstructed modules. (A) Bar plot displaying the fraction of connected components which are discarded / retained according to the number of standard deviations of the entropy from the mean value of the distribution of entropy derived from randomized connected components. **(B)** Density plot of the convolution between the connected component entropy and mean edge weight of the respective sub-graph. Maximum points in the density plot are highlighted by arrows. The number at each arrow denotes the number of selected connected components, i.e., connected components whose entropy and mean edge weight correspond to convolution intervals at the maxima of the density plot. **(C)** Dot plot displaying the percentage of interactions yielding differential co-expression scores higher than IntAct scores as a function of subgraph mean edge weight. Scores are taken in absolute value. **(D)** Bar plot showing the fraction of connected components retained in each sub-graph. **(E)** Bar plot showing the mean entropy of connected components selected solely on the basis of entropy significance or on the basis of convolution analysis in each sub-graph.

delivers emergent information about the quantitative changes of interaction strength and gene mRNA abundance between two conditions, and allows the user to pursue specific modules on the basis of any available biological rationale, including the extent of changes in interaction strength, the extent of mRNA fold change or the functional characterization of modules.

4. Discussion

Molecule interconnectivity in human cells is daunting with ~ 20,000 protein-coding genes and ~ 87,000 protein isoforms. Consequently, a network formalization of cellular processes is extremely useful to analyse the growing amount of data on many types of interactions, which include but are not limited to physical PPIs. A rich array of methods is currently available to detect network modular organization (Andreopoulos et al., 2009; Chen et al., 2014). Major limitations of most clustering methods, in very general terms, include the high computational cost and the inefficiency in exploiting the knowledge on edge strength (Toubiana et al., 2013). These aspects appear increasingly limiting in the light of the steady increase in the size of interaction maps and of the efforts to achieve interaction scoring standards (Villaveces et al., 2015). In this work, a new algorithm for network clustering has been developed that leverages existing information on both network nodes and edges to efficiently provide statistically significant modules. The detected modules are allowed to overlap, which reflects a common biological scenario, where, for instance,

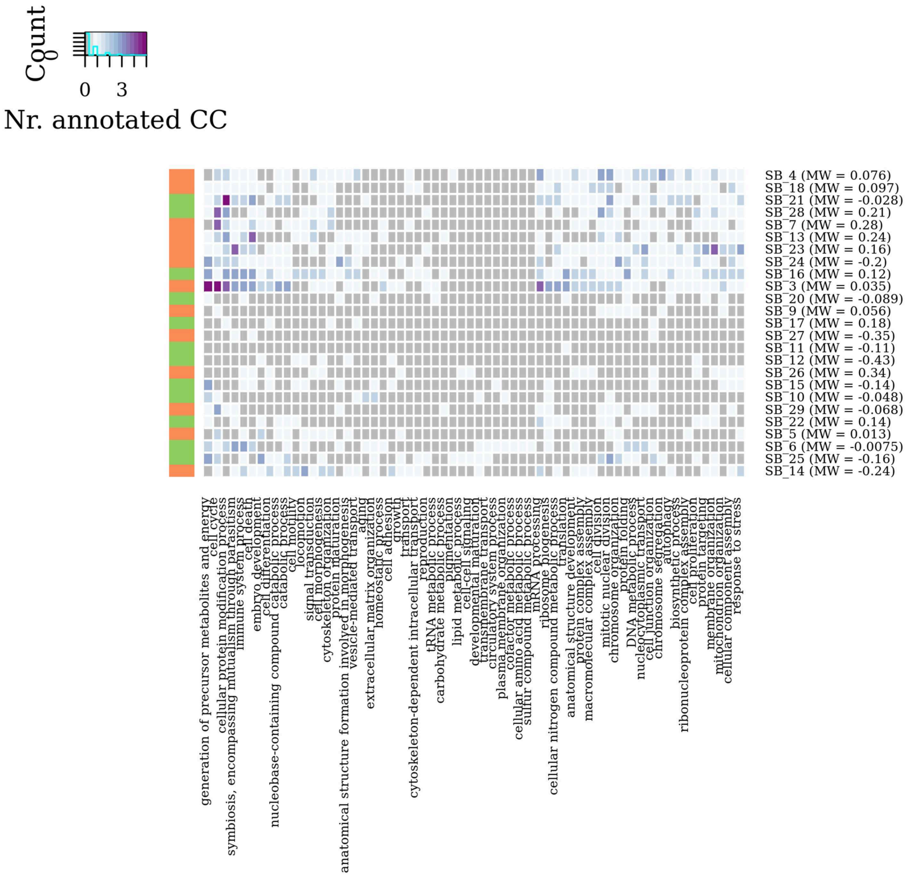


FIGURE 5 | Module enrichment in Gene Ontology categories. Heat map showing in each sub-graph the number of connected components which resulted statistically significant enriched in GO Biological Process categories (adjusted P-value < 0.05). Vertical bar colors denote the sign of sub-graph mean edge weights.

proteins can participate in multiple functions by participating to multiple functional modules. Within WG-Cluster, the connected components are homogeneous in terms both of node weights and of edge weights. We required homogeneity to extend to the numerical properties assigned both to network nodes and edges as both of them are expected to be biologically informative and useful to prioritize the study of the clustering results. To detect the modules, not only the reachability among nodes but also the homogeneity in the edges connecting the nodes has to be verified. To avoid the simultaneous verification of both requirements, which is highly time-consuming, WG-Cluster separates the two operations, firstly by identifying sub-graphs of homogeneous edge weights and, secondly, by detecting modules within each sub-graph. This procedural choice ensures, in an efficient way, connected components to be homogeneous in edge weights by construction. Furthermore, an entropy score is assigned to each connected component, which reflects the weights of nodes included in the connected component. The entropy score is utilized to measure the statistical significance of each component. Although not submitting node weights is allowed in WG-Cluster, it is worth noting that this choice invariably leads to entropy estimates which only depend on purely structural node properties. Therefore, partial input data limits the richness of information which could be made available by WG-Cluster. Finally, a convolution analysis of the entropy of the connected components with the mean edge weight of the sub-graphs was introduced to provide a global overview of the returned connected components and inform downstream analysis.

WG-Cluster is a method to cluster weighted networks into connected components, where nodes are homogeneous in their weights and are connected to each other by edges of homogeneous weights, and therefore WG-Cluster is suitable for many applications. A prominent applicative context is related to differential network analysis, which can discern cellular processes differently active under different conditions, such as with or without treatment by a pharmacological agent, with or without disease. Differential approaches have begun to drive considerable efforts in network biology, through the development of experimental assays to directly capture condition-specific networks (Ochoa and Beltrao, 2015) or through the integration of networks with condition-specific molecular profiles (Ideker et al., 2002; Jansen et al., 2002; Guo et al., 2007).

The case study presented here suggests WG-Cluster as a possible method for differential network analysis. A network of physical PPI interactions, which are scored utilizing community standards and are deposited in the IntAct database, was integrated with mRNA expression data acquired from colon adenocarcinoma tumor samples or from normal samples. Our integrative approach relied on the rationale that the strength of a protein–protein interaction depends on the extent of congruent protein levels and on their protein affinity. Under the assumption that protein expression can be approximated with mRNA expression and that the interaction score in IntAct reflects the interaction affinity, we specified nodes and edge weights of the differential network as follows. Node weights were defined by the mRNA level fold changes while edge weights were defined by the product of the IntAct scores with the differential mRNA co-expression scores between the two conditions. Applying WG-Cluster to the differential network permitted to prioritize modules in the PPI network representing regions of progressively decreasing changes between the tumor and normal conditions. Despite the fact that the majority of interactions changed moderately between the two conditions, the organization of the detection of weighted connected components by sub-graph, which is implemented in WG-Cluster, permitted to streamline the identification of modules of markedly changing interactions. Furthermore, it was possible to discern modules of interactions which get weakened or strengthened in the tumor compared to the normal condition. Interestingly, separating the modules by average increase or decrease in the strength of their interactions reflected also on their functional enrichment into distinct GO categories.

WG-Cluster is available as an open-source tool at https://sites.google.com/site/paolaleccapersonalpage/ for the community of computational biologists to encourage its further development and/or its integration in general analytical workflows.

Author Contributions

PL and AR equally contributed to the conception, design and testing of the WG-Cluster algorithm; both the authors equally contributed also to the selection of the data and the case study for WG-Cluster application and to the interpretation of the algorithm output. Both the authors contributed to writing the manuscript.

Acknowledgments

The authors thank Alessandro Quattrone of the Laboratory of Translational Genomics of Centre for Integrative Biology, University of Trento (Italy) for his useful suggestions and comments to this work.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00265

References

Adler, D., Glser, C., Nenadic, O., Oehlschlgel, J., and Zucchini, W. (2014). *ff: Memory-Efficient Storage of Large Data on Disk and Fast Access Functions*. Availble online at: http://cran.r-project.org/web/packages/ff/index.html

Andreopoulos, B., An, A., Wang, X., and Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinf.* 10, 297–314. doi: 10.1093/bib/bbn058

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Barabási, 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

Chen, B., Fan, W., Liu, J., and Wu, F. X. (2014). Identifying protein complexes and functional modules–from static ppi networks to dynamic ppi networks. *Brief Bioinf.* 15, 177–194. doi: 10.1093/bib/bbt039

Chen, H., Li, H., Liu, F., Zheng, X., Wang, S., Bo, X., et al. (2015). An integrative analysis of tfbs-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Sci. Rep.* 5:8465. doi: 10.1038/srep08465

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140. doi: 10.1038/msb4100180

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

Cytoscape. (2015). *SIF Simple Interaction Format*. Available online at: http://wiki.cytoscape.org/Cytoscape_User_Manual/Network_Formats.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science* 307, 724–727. doi: 10.1126/science.1105103

Du, Q., Emelianenko, M., and Ju, L. (2006). Convergence of the lloyd algorithm for computing centroidal voronoi tesellation. *SIAM J. Numer. Anal.* 44, 102–119. doi: 10.1137/040617364

Girvan, M., and Newman, M. E. J. (2001). Community structure in social and biological networks. *Proc, Natl. Acad, Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Grossmann, A., Benlasfer, N., Birth, P., Hegele, A., Wachsmuth, F., Apelt, L., et al. (2015). Phospho-tyrosine dependent protein-protein interaction network. *Mol. Syst. Biol.* 11:794. doi: 10.15252/msb.20145968

Guo, Z., Wang, L., Li, Y., Gong, Y., Yao, C., Ma, W., et al. (2007). Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics* 23, 2121–2128. doi: 10.1093/bioinformatics/btm294

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402(6761 Suppl.), C47–C52. doi: 10.1038/35011540

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., et al. (2012). A census of human soluble protein complexes. *Cell* 150, 1068–1081. doi: 10.1016/j.cell.2012.08.011

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., et al. (2004). The hupo psi's molecular interaction format - a community standard for the representation of protein interaction data. *Nat. Biotechnol.* 22, 177–183. doi: 10.1038/nbt926

Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., et al. (2014). Kinomexplorer: an integrated platform for kinome biology studies. *Nat. Methods* 11, 603–604. doi: 10.1038/nmeth.2968

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8:565. doi: 10.1038/msb.2011.99

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

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12, 37–46. doi: 10.1101/gr.205602

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Johannesson, T., and Bjornsson, H. (2012). *Stineman, A Consistently Well Behaved Method of Interpolation.* Availble online at: http://rpackages.ianhowson.com/cran/stinepack/

Kane, M. J., and Emerson, J. W. (2013). *Bigmemory: Manage Massive Matrices with Shared Memory and Memory-Mapped File.* Availble online at: http://cran.r-project.org/web/packages/bigmemory/index.html

Klammer, M., Godl, K., Tebbe, A., and Schaab, C. (2007). Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC Bioinformatics* 11:351. doi: 10.1186/1471-2105-11-351

Lascorz, J., Chen, B., Hemminki, K., and Föersti, A. (2011). Consensus pathways implicated in prognosis of colorectal cancer identified through systematic enrichment analysis of gene expression profiling studies. *PLoS ONE* 6:e18867. doi: 10.1371/journal.pone.0018867

Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jørgensen, C., Miron, I. M., et al. (2007). Systematic discovery of *in vivo* phosphorylation networks. *Cell* 129, 1415–1426. doi: 10.1016/j.cell.2007.05.052

Liu, G., Wong, L., and Chua, H. N. (2009). Complex discovery from weighted ppi networks. *Bioinformatics* 25, 1891–1897. doi: 10.1093/bioinformatics/btp311

Martin, H., Shales, M., Fernandez-Pinar, P., Wei, P., Molina, M., Fiedler, D., et al. (2015). Differential genetic interactions of yeast stress response mapk pathways. *Mol. Syst. Biol.* 11:800. doi: 10.15252/msb.20145606

Moore, M. J., Zhang, C., Gantman, E. C., Mele, A., Darnell, J. C., and Darnell, R. B. (2014). Mapping argonaute and conventional rna-binding protein interactions with rna at single-nucleotide resolution using hits-clip and cims analysis. *Nat. Protoc.* 9, 263–293. doi: 10.1038/nprot.2014.012

Nacu, S., Critchley-Thorne, R., Lee, P., and Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics* 23, 850–858. doi: 10.1093/bioinformatics/btm019

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

Ochoa, D., and Beltrao, P. (2015). Kinase-two-hybrid: towards the conditional interactome. *Mol. Syst. Biol.* 11, 798. doi: 10.15252/msb.20156107

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. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stüempflen, V. et al. (2007). The minimum information required for reporting a molecular interaction experiment (mimix). *Nat. Biotechnol.* 25, 894–898. doi: 10.1038/nbt1324

Pandey, G., Arora, S., Manocha, S., and Whalen, S. (2014). Enhancing the functional content of eukaryotic protein interaction networks. *PLoS ONE* 9:e109130. doi: 10.1371/journal.pone.0109130

Petschnigg, J., Groisman, B., Kotlyar, M., Taipale, M., Zheng, Y., Kurat, C. F., et al. (2014). The mammalian-membrane two-hybrid assay (mamth) for probing membrane-protein interactions in human cells. *Nat. Methods* 11, 585–592. doi: 10.1038/nmeth.2895

Pons, P., and Latapy, M. (2005). Computing communities in large networks using random walks. *Lect. Notes Comput. Sci.* 3733, 284–293. doi: 10.1007/11569596_31

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76:036106. doi: 10.1103/PhysRevE.76.036106

Re, A., Joshi, T., Kulberkyte, E., Morris, Q., and Workman, C. T. (2014). Rna-protein interactions: an overview. *Methods Mol. Biol.* 1097, 491–521. doi: 10.1007/978-1-62703-709-9_23

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

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

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173-1178. doi: 10.1038/nature04209

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

TCGA Research Network. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49. doi: 10.1038/nature12222

Toubiana, D., Fernie, A. R., Nikoloski, Z., and Fait, A. (2013). Network analysis: tackling complex data to study plant metabolism. *Cell* 31, 29–36. doi: 10.1016/j.tibtech.2012.10.011

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). irefweb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010:baq023. doi: 10.1093/database/baq023

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Varjosalo, M., Sacco, R., Stukalov, A., van Drogen, A., Planyavsky, M., Hauri, S., et al. (2013). Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized ap-ms. *Nat. Methods* 10, 307–314. doi: 10.1038/nmeth.2400

Villaveces, J. M., Jimnez, R. C., Porras, P., Toro, N., Duesbury, M., Dumousseau, M., et al. (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database* pii: bau131. doi: 10.1093/database/bau131

Conflict of Interest Statement: 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.

ORIGINAL RESEARCH
published: 28 July 2015
doi: 10.3389/fgene.2015.00252

Identifying critical differentiation state of MCF-7 cells for breast cancer by dynamical network biomarkers

Pei Chen[1], *Rui Liu*[2], *Luonan Chen*[3,4]* *and Kazuyuki Aihara*[3]*

[1] School of Computer Science, South China University of Technology, Guangzhou, China, [2] School of Mathematics, South China University of Technology, Guangzhou, China, [3] Collaborative Research Center for Innovative Mathematical Modelling, University of Tokyo, Tokyo, Japan, [4] Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Spyros Petrakis,
Aristotle University of Thessaloniki,
Greece

Reviewed by:
Guanglong Jiang,
Indiana University School of Medicine,
USA
Kyle Bergin Gustafson,
University of Maryland - College Park,
USA
Yong Wang,
Academy of Mathematics and
Systems Science, Chinese Academy
of Sciences, China

****Correspondence:***
Luonan Chen,
Chinese Academy of Sciences, 320
Yueyang Road, Shanghai 20031,
China
lnchen@sibs.ac.cn;
Kazuyuki Aihara,
University of Tokyo, 4-6-1 Komaba,
Meguro-ku, Tokyo 153-8505, Japan
aihara@sat.t.u-tokyo.ac.jp

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

Received: *14 May 2015*
Accepted: *13 July 2015*
Published: *28 July 2015*

Citation:
Chen P, Liu R, Chen L and Aihara K
(2015) Identifying critical differentiation
state of MCF-7 cells for breast cancer
by dynamical network biomarkers.
Front. Genet. 6:252.
doi: 10.3389/fgene.2015.00252

Identifying the pre-transition state just before a critical transition during a complex biological process is a challenging task, because the state of the system may show neither apparent changes nor clear phenomena before this critical transition during the biological process. By exploring rich correlation information provided by high-throughput data, the dynamical network biomarker (DNB) can identify the pre-transition state. In this work, we apply DNB to detect an early-warning signal of breast cancer on the basis of gene expression data of MCF-7 cell differentiation. We find a number of the related modules and pathways in the samples, which can be used not only as the biomarkers of cancer cells but also as the drug targets. Both functional and pathway enrichment analyses validate the results.

Keywords: cell differentiation, dynamical network biomarker (DNB), pre-transition state, critical transition, early-warning signal, breast cancer

Introduction

Breast cancer, one of the most common cancers, is clearly a heterogeneous, complex, interrelated disease involving multi-factorial etiologies. The tumorgenesis of breast cancer is typically characterized by a combination of the interactions between environmental (external) factors and a genetically susceptible host (Ou et al., 2010). The prevalence of breast cancer as well as the growing economic and societal burden of the treatment is making it urgently necessary to implement interventions to prevent or at least delay the occurrence of breast cancer. However, it is still a challenging task to detect breast cancer in its early stage since it is usually silent and without clear symptoms in its initial stages, while irreversible complications may develop rapidly before the implementation of effective treatment (Saini et al., 2011). Many studies of breast cancer are based on MCF-7 cells. MCF-7 is the acronym of Michigan Cancer Foundation-7. The MCF-7 cells are cancer cells that are classified as invasive breast ductal carcinoma. Although the underlying molecular mechanism of the progression for MCF-7 cells is far from clear, it has been found that heregulin (HRG) and epidermal growth factor (EGF) are involved in inducing the critical transition of cell differentiation or proliferation (Normanno et al., 1994; Suzuki et al., 2004; Nagashima et al., 2007; Saeki et al., 2009). In this work, we quantitatively analyze time-course microarray data of MCF-7 cells, and identify the key genes, i.e., dynamical network biomarker (DNB), which may indicate the imminent critical transition of the cancer cells during cell differentiation or proliferation.

We previously hypothesized that a complex biological process (e.g., disease progression) can be divided into three stages or states (**Figures 1A,B**): (A) a before-transition stage (or a normal state) with high resilience and robustness to perturbations; (B) a pre-transition stage (or a pre-disease state), just before the critical transition to the disease state, i.e., occurring before an imminent phase transition point is reached, therefore, with low resilience and robustness due to its dynamical structure; (C) an after-transition stage (or a disease state), representing a seriously deteriorated stage possibly with high resilience and robustness again, because it is generally difficult for the system at this state to recover or return to the normal state even after treatment (Chen et al., 2012; Liu et al., 2012a). This classification is supported by the observations that there exist catastrophic shifts during the progression of many chronic diseases, i.e., the sudden deterioration of diseases (Litt et al., 2001; McSharry et al., 2003; Venegas et al., 2005; Hirata et al., 2010; He et al., 2012). A drastic or qualitative transition in a focal system or network, from a normal state to a disease state, corresponds to a so-called bifurcation point in dynamical systems theory (Gilmore, 1993; Murray, 2002). When the system is near a bifurcation point, or a critical point, there exists a dominant group, which we called as the DNB. The DNB can be defined by the following three conditions (Chen et al., 2012): The correlation between any pair of members in DNB becomes very strong; The correlation between one member of DNB and any other molecule of non-DNB becomes very weak; Any member of DNB becomes highly fluctuating. The DNB is not only a theoretical concept, but also has been successfully applied to real biological data, and used to identified the early-warning signals of sudden deterioration of several complex diseases (Li et al., 2013; Liu et al., 2013a,b, 2014a; Zeng et al., 2014; Tan et al., 2015).

In this work, by applying the DNB approach to the datasets of MCF-7 breast cancer cell line (GSE13009, GSE6462, and GSE10145), we identify the DNB members composed by a group of genes that may indicate the imminent critical transition during the progression of breast cancer cells.

Methods

We first describe the theoretical basis, i.e., the DNB theory, and then provide the procedures used to preprocess input datasets and implement the detail DNB score algorithm.

Theoretical Basis

As explained in Section Introduction, a biological process can be generally divided into the three stages, i.e., (A) the before-transition state (or normal state in complex diseases), (B) the pre-transition state (or pre-disease state in complex diseases) and (C) the after-transition state (or disease state in complex diseases) (**Figure 1A**). The before-transition state is a stable state representing a stable stage with high resilience, during which the state may change gradually. The pre-transition state is a state defined as the limit of the before-transition state just before a critical transition. This state is considered to be still reversible to the before-transition state since appropriate external interventions can drive it back to the before-transition state relatively easily. However, further progression beyond the pre-transition state will result in a drastic transition to the after-transition state, another stable state, and it is difficult to return to the before-transition state even with intensive interventions. The after-transition state represents a seriously ill stage in complex diseases.

Different from the traditional biomarkers, e.g., molecular biomarkers and network biomarkers (Liu et al., 2012b; Wen et al., 2014; Zhang et al., 2014, 2015), which are designed to distinguish the disease samples from normal samples and thus reflect the severity or presence of the illness at the disease state, the DNB theory aims to distinguish the pre-disease samples from normal samples according to the critical dynamical behavior of DNB molecules (Liu et al., 2014b). In other words, the DNB method is designed to identify a group of strongly correlated and significantly fluctuating molecules, which are also called "the leading network" because those genes may lead the transition of the whole system from the normal state to the disease state (Liu et al., 2012a).

Although elucidating the critical transition at the network level holds the key to understand the fundamental mechanism of disease development or cell differentiation, it is notably hard to reliably identify the pre-transition state because there may be little apparent difference between the before-transition and pre-transition states. This is also the reason why diagnosis based on traditional biomarkers may fail to indicate the pre-transition state. The theoretical basis for detecting DNB is summarized by the following conditions (**Figures 1C,D**), which have been proven to hold simultaneously when the system approaches the pre-transition state (Chen et al., 2012):

1. Deviations of a group of molecules called DNB among the whole population of molecules, drastically increases (the fluctuation condition);
2. Correlation between any two molecules among DNB increases (the internal correlation condition);
3. Correlation between any molecule in DNB and another in non-DNB decreases (the external correlation condition);
4. There are no drastic changes for deviations and correlations of molecules among the remaining molecules of the system, i.e., non-DNB.

Dynamics satisfying the preceding conditions can be viewed as locally herding behavior, i.e., members in a DNB subnetwork act together with strongly correlated fluctuation. These conditions imply an imminent regime shift or a phase transition, and therefore, can be used to signal the impending emergence of the critical transition. Such a phenomenon can also be described as the DNB molecules get dynamically correlated or connected so that the system can be reorganized in a different way.

Data Processing and Algorithm

Three gene expression profiling datasets were downloaded from the NCBI GEO database (ID: GSE13009, GSE6462, and GSE10145) (www.ncbi.nlm.nih.gov/geo). In these datasets, probe sets without corresponding gene symbols were not considered in our analysis. The expression values of probe sets mapped to the same gene were averaged. Genes in the DNBs for the

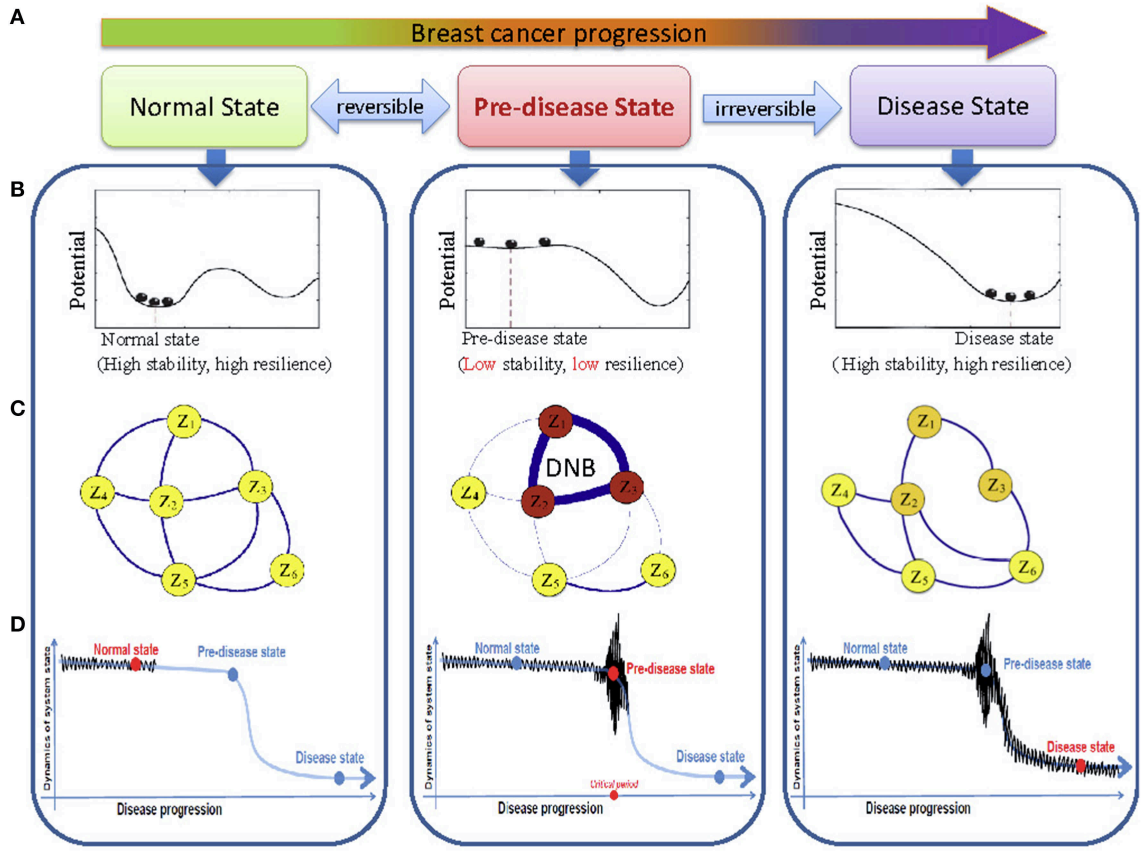


FIGURE 1 | The outline for identifying the transition state by DNB based on time-course data. (A) The progression of breast cancer cells can be divided into three states, i.e., the before-transition state, the pre-transition state, and the after-transition state. **(B)** A system at the before-transition state or the after-transition state is stable with high resilience, while it is unstable with low resilience when it is at the pre-transition state. **(C)** In the pre-transition state, large fluctuations of the DNB members are correlated strongly. This critical phenomenon do not appear at the before-transition and the after-transition states. **(D)** The DNB members show large fluctuations in their expressions at the pre-transition state, compared with smaller fluctuations of the expressions at the before-transition and the after-transition states.

three datasets were linked and correlated by the combined functional couplings among them from various databases of protein-protein interactions of STRING, FunCoup, and BioGrid. In each disease dataset, the expression profiling information was mapped to the integrated networks individually for identifying the corresponding DNB. For each species, we downloaded the biomolecular interaction networks from various databases, including BioGrid (http://www.thebiogrid.org), TRED, KEGG (http://www.genome.jp/kegg), and HPRD (http://www.hprd.org). First, the available functional linkage information for Mus musculus and Homo sapiens was downloaded from these databases and combined. For instance, after removing any redundancy in dataset GSE13009, we obtained 37,950 linkages in 13785 human genes. Next, the genes evaluated in these microarray datasets were mapped individually to their integrated functional linkage networks. In order to trigger critical changes, MCF-7 cells were exposed to growth factors heregulin (HRG) for up to 6 h and the temporal expression of transcription factors was monitored (Saeki et al., 2009). There were the case group and the control group for the experiment. For the case group, the gene expressions were recorded respectively in 17 time points (10 min, 15 min, 20 min, 30 min, 45 min, 1 h, 1 h 30 min, 2 h, 3 h, 4 h, 6 h, 8 h, 12 h, 24 h, 36 h, 48 h, and 72 h). The networks were visualized using Cytoscape (www.cytoscape.org) and a part of the functional analysis was based on Integrate and understand complex omics data (IPA). The detailed algorithm is given in the Supplementary Materials.

Results

The Identified DNB and the Pre-transition State

Applying the DNB method to dataset GSE13009, the DNB containing 104 genes was identified for HRG-induced differentiation of cancer cells. We listed all of the identified DNB members in Supplementary Table S1 "Detail description of the identified DNB." The process of identifying the DNB can be found in "The algorithm for identifying the DNB" of Supplementary Materials. During the progression of cancer cells, we also identified the pre-transition state between the before-transition state and the after-transition state (**Figure 2**), which is the critical stage when the progression of MCF-7 cells is just before the differentiation triggered by HRG (Nagashima et al., 2007). Actually, based on **Figure 2**, the sharp increase of the DNB score (the red curve) represents an indicative early-warning signal 1–1.5 h after the expose to HRG, and thus before the differentiation detected by molecular markers. In fact, the original assay showed that the AP-1 complex in HRG-treated MCF-7 cells contains c-JUN, c-FOS, and FRA-1, although the association of c-JUN in the complex is transient (Saeki et al., 2009). Besides, the stimulation of MCF-7 breast cancer cells with EGF and HRG resulted in very similar early transcription profiles up to 90 min; however, subsequent cellular phenotypes differed after 3 h (Saeki et al., 2009), which suggests that the differentiation is around 3 h (the 9th sampling time point). The bootstrap validation (the blue curves) is also known in **Figure 2**, which exhibits that the randomly chosen groups containing the same number of genes with DNB are insensitive to the critical transition.

Figure 3 shows the dynamical evolution in the whole feature network based on the case data. It can be seen from **Figure 3** that the selected 104 genes (the top right corner in each network) are strongly correlated with large fluctuations 1–1.5 h before the critical transition, which provides a significant signal indicating the pre-transition stage of cell differentiation, while other genes show no significant signal. Clearly, when the differentiation is impending, these selected genes form a special subnetwork, the so-called DNB, which makes the first move from the before-transition state toward the after-transition state during the transition. Interestingly, members of the DNB behaved similarly to other genes after the system moved to the after-transition state. It can be seen that, on the other hand, neither the whole gene network nor the DNB presents a signal before or after the transition, which shows the sensitivity of the DNB method only at the pre-transition state. In fact, the DNB method reveals the existence of the pre-transition state, which, however, may not be detected by molecules such as EGR4, FOSL-1, FHL2, and DIPA, although these four transcription factors are proved to be effective for indicating the differentiation of breast cancer cells (Saeki et al., 2009). In other words, the molecular biomarkers cannot provide early-warning signals before the cell differentiation (at 3 h, or the 9th sampling time point). Therefore, the benefits brought by the DNB method in signaling the pre-transition state make the identification and management of high-risk cases effective.

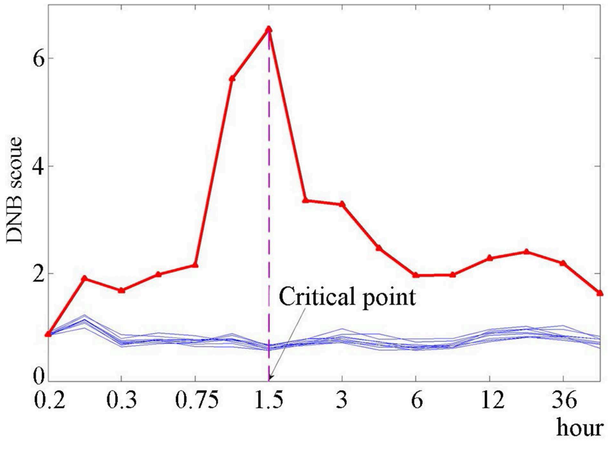


FIGURE 2 | The DNB scores for the identified dominant group and bootstrap groups. The DNB scores are shown for the identified dominant group (red curve) and 10 groups from bootstrap (blue curves). It can be seen that for the red curve, the DNB score increases sharply from the 6th point (1 h) and reaches the peak at the 7th point (1.5 h). For the bootstrap analysis, we randomly selected 10 gene sets, each of which is composed of the same number of members as the dominant group. Then the DNB score was calculated for each randomly chosen group.

Validation

Hereto we have shown the sensitivity and effectiveness of the identified DNB. **Figures 4A,B** respectively show the DNB scores based on independent datasets GSE6462 and GSE10145. From dataset GSE6462, it can be seen from **Figure 4A** that the identified DNB also showed a signal for large dose (1 and 10 nM) HRG expose at the 4th sampling point (30 min), while there is no clear signal for small dose (0.1 and 0.5 nM) HRG expose. It agrees with the original experiments (Nagashima et al., 2007) that HRG-induced cellular differentiation of MCF-7 cells is observed around 60 min. From **Figure 4B**, it can be seen that the signal is detected by the DNB score at the 4th time point, which also agrees with the observations and shows the sensitivity of the identified DNB. The bootstrap analysis for both datasets is shown in Figure S1 of Supplementary Materials.

Functional Analysis

Heregulin (HRG) can induce dose-dependent transient and sustained intracellular signaling, proliferation, and differentiation of MCF-7 breast cancer cells (Barlund et al., 2002; Huang et al., 2009). In the infected host, some metabolic pathways responded to these interruptions and became increasingly disordered. The following results show that some reported phenomena were consistent with our investigations, which also provides novel insights into the biological processes.

The identified DNB module is related to the regulation of an apoptotic process (GO:0042981) with the significant *P*-value (2.93E-06), the regulation of the programmed cell death (GO:0043067) with the significant *P*-value (4.10E-05) and the regulation of the cell death (GO:0010941) with the significant *P*-value (7.41E-04) by the website tool DAVID Bioinformatics Resource (Huang et al., 2009). By the pathway analysis in the KEGG database, we found that seven genes (CEBPA, SMAD3, GSK3B, LAMC2, MMP1, PIK3R3, and RXRA) in this DNB module participate in cancer pathways, and many genes of this module also take part in other cancer-related pathways, e.g., the Wnt signaling pathway with *P*-value (9.10E-03), the p53 signaling pathway with *P*-value (1.10E-04), and the ECM-receptor interaction with *P*-value (2.30E-03).

Many genes in this DNB module have been proved to be related to a cancer or tumor process, and in particular, some of these genes are associated with breast cancer. For example, BCAS4 is an important gene for breast tumor development

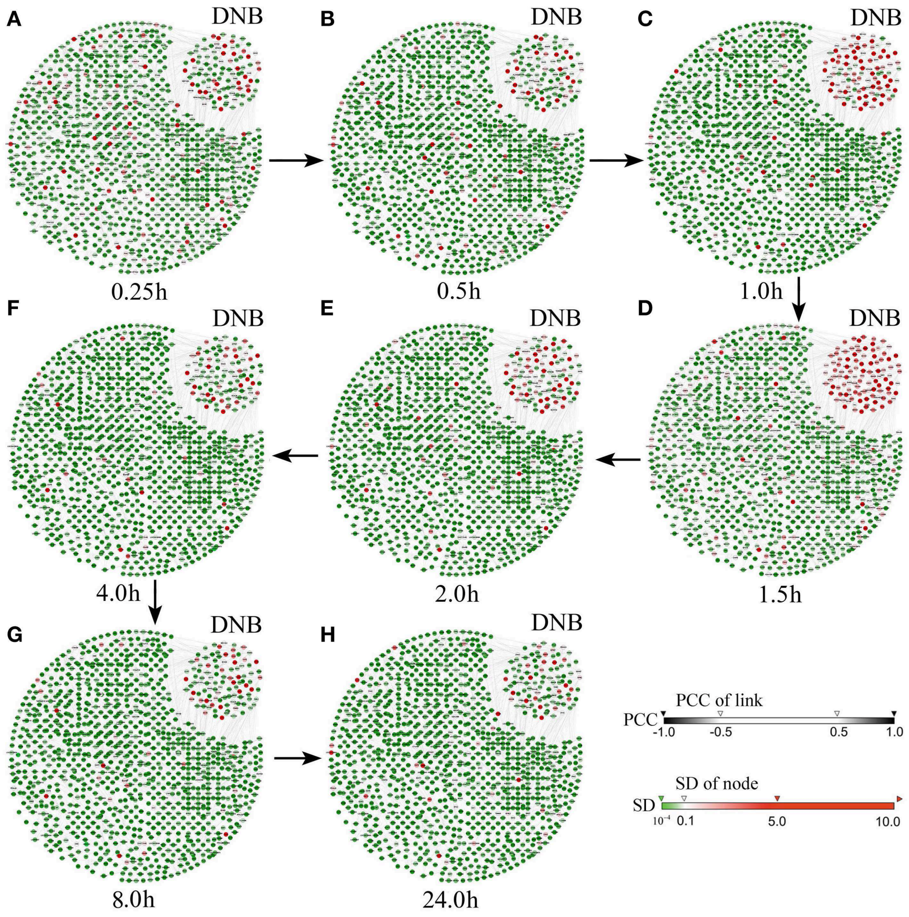


FIGURE 3 | Dynamical changes in the network including the selected DNB during the progression of HRG-induced breast cancer. The figures show the dynamical changes of the molecular network at **(A)** 0.25 h, **(B)** 0.5 h, **(C)** 1 h, **(D)** 1.5 h, **(E)** 2 h, **(F)** 4 h, **(G)** 8 h, and **(H)** 24 h. It can be seen that, the DNB members are correlated strongly while each member shows large fluctuation in its expression during 1–1.5 h. This critical phenomenon does not appear before or after this period, i.e., the before-transition or the after-transition state. Thus, the pre-transition stage is around 1–1.5 h, just before the cell differentiation triggered by HRG (7).

and progression (Barlund et al., 2002). ARID3B is one of genes which regulates cell motility and actin cytoskeleton organization (Casanova et al., 2011) and is found to be associated with breast cancer onset (Akhavantabasi et al., 2012). TNFRSF21 encodes a tumor necrosis factor receptor, which can regulate the NF-kappaB and mediate an apoptosis process (Kasof et al., 2001). LAMC2 encodes the gamma chain isoform laminin, which is involved in many biological processes, and LAMC2 is also proved to be related to the breast cancer process (Sathyanarayana et al., 2003; Koshikawa et al., 2005). Therefore, DNB for HRG-induced breast cancer can mainly induce cancer by affecting the processes of regulation of apoptosis, regulation of programmed cell death and regulation of cell death.

Discussion

Breast cancer is a progressive disease and its deterioration course is primarily characterized by cancer cell differentiation or proliferation, which significantly damages the health of women all over the world. Detecting the early-warning signal of the cell differentiation of cancer cells provides an opportunity to interrupt and prevent the continuing costly cycle of managing breast cancer and its complications. The critical transition of cancer cells involving proliferation or differentiation can be induced by a ligand of the ErbB family receptor, heregulin, which evokes kinase activity of MCF-7 cells. Actually, in MCF-7, HRG induced graded signaling and early transcription, followed by auto-induction of multiple positive/negative feedback mechanisms, and prolongation of signaling activity might switch cells irreversibly (Saeki et al., 2009). It is an important future problem to analyses whether the HRG-induced critical transition is reversible in the pre-transition state.

In this work, we applied the DNB method to the identification of the pre-transition state on the basis of a composition of microarray data from the breast cancer cell line. First, we introduced the DNB approach which aims at detecting the critical signals of the cell differentiation and indicating the pre-transition

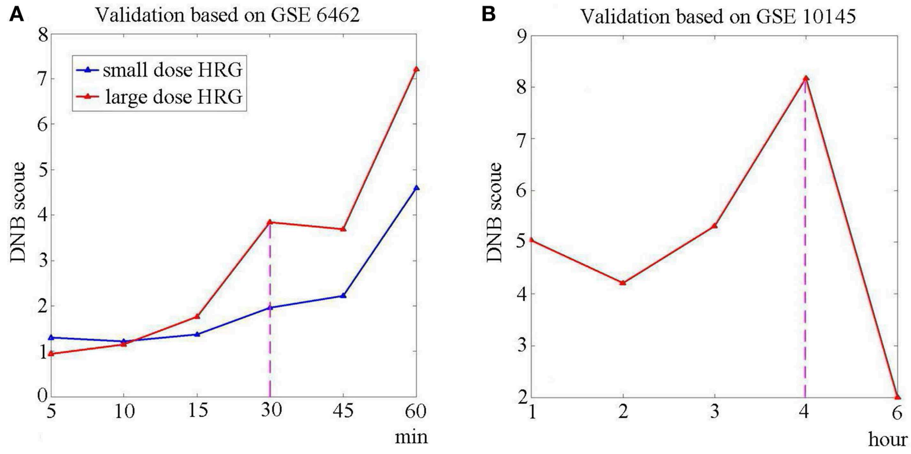


FIGURE 4 | The validation of DNB based on independent datasets. To validate the sensitivity and effectiveness, we calculated the DNB score using the identified genes, based on two independent dataset. **(A)** The DNB scores based on GSE6462. The red curve represents the case of large dose HRG usage (1and 10 nM), while the blue curve stands for the case of small dose HRG expose (0.1 and 0.5 nM). It can be seen that there is a signal at the 4th sampling point (30 min) when the MCF-7 cells are exposed to large dose of HRG. **(B)** The DNB score based on GSE10145. The curve shows that a peak of DNB score is at the 4th sampling point (4 h).

state or stage. Second, based on the cell line data, we identified the pre-transition stage right before the cell differentiation induced by heregulin (HRG) during the progression of cancer cells. Actually, an indicative early-warning signal is presented by DNB at 1 h after the expose to HRG. The validation based on bootstrap (**Figure 2**) and other two datasets (**Figure 4**) demonstrated the sensitivity and effectiveness of the identified DNB for the HRG triggered differentiation. Besides, we showed that some metabolic pathways responded to the HRG-induced interruptions and became increasingly disordered during the biological process. Therefore, the DNB method provides a new way to pry into the underlying mechanism of cell differentiation and thus is helpful to achieve the timely intervention. This is the main value in the potential applications of the DNB method from a network point of view.

On the other hand, there are limitations of this work. First, the validity of the identified pre-transition state and the DNB needs further supports from biological experiments and clinical studies. Second, the method is insensitive when the genes are not differentially expressed (see the algorithm stated in the Supplementary Material). The algorithm is also needed to be further improved on the aspects of both sensitivity and accuracy. Although this work is merely a step toward detecting the early-warning signals of critical transition during cancer cell progression of breast cancer and the algorithm is expected to be improved in both time saving and capacity efficient ways, it opens a window of an opportunity for experimental and clinical study on the early-warning system of breast cancer.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant numbers 9143920024, 61134013, 91439103, 11326035, and 11401222); Fundamental Research Funds for the Central Universities (Grant number 2014ZZ0064); the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant Number KSCX2-EW-R-01) and 863 project (Grant number 2012AA020406); JST's "Super Highway," the accelerated research to bridge university IPs and practical use.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00252

References

Akhavantabasi, S., Sapmaz, A., Tuna, S., and Erson-Bensan, A. E. (2012). miR-125b targets ARID3B in breast cancer cells. *Cell Struct. Funct.* 37, 27–38. doi: 10.1247/csf.11025

Barlund, M., Monni, O., Weaver, J. D., Kauraniemi, P., Sauter, G., Heiskanen, M., et al. (2002). Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosomes Cancer* 35, 311–317. doi: 10.1002/gcc.10121

Casanova, J. C., Uribe, V., Badia-Careaga, C., Giovinazzo, G., Torres, M., and Sanz-Ezquerro, J. J. (2011). Apical ectodermal ridge morphogenesis in limb development is controlled by Arid3b-mediated regulation of cell movements. *Development* 138, 1195–1205. doi: 10.1242/dev.057570

Chen, L., Liu, R., Liu, Z., Li, M., and Aihara, K. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* 2:342. doi: 10.1038/srep00342

Gilmore, R. (1993). *Catastrophe Theory for Scientists and Engineers*. New York, NY: Dover Publications. doi: 10.1038/srep00342

He, D., Liu, Z. P., Honda, M., Kaneko, S., and Chen, L. (2012). Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. *J. Mol. Cell Biol.* 4, 140–152. doi: 10.1093/jmcb/mjs011

Hirata, Y., Bruchovsky, N., and Aihara, K. (2010). Development of a mathematical model that predicts the outcome of hormone therapy for prostate cancer. *J. Theor. Biol.* 264, 517–527. doi: 10.1016/j.jtbi.2010.02.027

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

Kasof, G. M., Lu, J. J., Liu, D., Speer, B., Mongan, K. N., Gomes, B. C., et al. (2001). Tumor necrosis factor-alpha induces the expression of DR6, a member of the TNF receptor family, through activation of NF-kappaB. *Oncogene* 20, 7965–7975. doi: 10.1038/sj.onc.1204985

Koshikawa, N., Minegishi, T., Sharabi, A., Quaranta, V., and Seiki, M. (2005). Membrane-type matrix metalloproteinase-1 (MT1-MMP) is a processing enzyme for human laminin gamma 2 chain. *J. Biol. Chem.* 280, 88–93. doi: 10.1074/jbc.M411824200

Li, M., Tao, Z., Liu, R., and Chen, L. (2013). Detecting tissue-specific early-warning signals for complex diseases based on dynamical network biomarkers: study of type-2 diabetes by cross-tissue analysis. *Brief. Bioinform*. 15, 229–243. doi: 10.1093/bib/bbt027

Litt, B., Esteller, R., Echauz, J., D'Alessandro, M., Shor, R., Henry, T., et al (2001). Epileptic seizures may begin hours in advance of clinical onset: a report of five patients. *Neuron* 30, 51–64. doi: 10.1016/S0896-6273(01)00262-8

Liu, R., Aihara, K., and Chen, L. (2013a). Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes. *Quant. Biol.* 1, 105–114. doi: 10.1007/s40484-013-0008-0

Liu, R., Li, M., Liu, Z. P., Wu, J., Chen, L., and Aihara, K. (2012a). Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci. Rep*. 2:813. doi: 10.1038/srep00813

Liu, R., Wang, X., Aihara, K., and Chen, L. (2014a). Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med. Res. Rev.* 34, 455–478. doi: 10.1002/med.21293

Liu, R., Yu, X., Liu, X., Xu, D., Aihara, K., and Chen, L. (2014b). Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics* 30, 1579–1586. doi: 10.1093/bioinformatics/btu084

Liu, X. P., Liu, R., Zhao, X.-M., and Chen, L. (2013b). Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med. Genomics* 6(Suppl. 2):S8. doi: 10.1186/1755-8794-6-S2-S8

Liu, X. P., Liu, Z. P., Zhao, X.-M., and Chen, L. (2012b). Identifying disease genes and module biomarkers by differential interactions. *J. Am. Med. Inform. Assoc.* 19, 241–248. doi: 10.1136/amiajnl-2011-000658

McSharry, P. E., Smith, L. A., and Tarassenko, L. (2003). Prediction of epileptic seizures: are nonlinear methods relevant? *Nat. Med.* 9, 241–242. doi: 10.1038/nm0303-241

Murray, J. D. (2002). *Mathematical Biology, 3rd Edn*. New York, NY: Springer. doi: 10.1038/srep00342

Nagashima, T., Shimodaira, H., Ide, K., Nakakuki, T., Tani, Y., Takahashi, K., et al. (2007). Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation. *J. Biol. Chem.* 282, 4045–4056. doi: 10.1074/jbc.M608653200

Normanno, N., Ciardiello, F., Brandt, R., Salomon, D. S. (1994). Epidermal growth factor-related peptides in the pathogenesis of human breast cancer. *Breast Cancer Res. Treat.* 29, 11–27.

Ou, K.-W., Hsu, K.-F., Cheng, Y.-L., Hsu, G.-C., Hsu, H.-M., and Yu, J.-C. (2010). Asymptomatic pulmonary nodules in a patient with early-stage breast cancer: Cryptococcus infection. *Int. J. Infect. Dis.* 14, e77–e80. doi: 10.1016/j.ijid.2009.03.007

Saeki, Y., Endo, T., Ide, K., Nagashima, T., Yumoto, N., Toyoda, T., et al. (2009). Ligand-specific sequential regulation of transcription factors for differentiation of MCF-7 cells. *BMC Genomics* 10:545. doi: 10.1186/1471-2164-10-545

Saini, K. S., Taylor, C., Ramirez, A. J., Palmieri, C., Gunnarsson, U., Schmoll, H. J., et al. (2011). Role of the multidisciplinary team in breast cancer management: results from a large international survey involving 39 countries. *Ann. Oncol.* 23, 853–859. doi: 10.1093/annonc/mdr352

Sathyanarayana, U. G., Padar, A., Huang, C. X., Suzuki, M., Shigematsu, H., Bekele, B. N., et al. (2003). Aberrant promoter methylation and silencing of laminin-5-encoding genes in breast carcinoma. *Clin. Cancer Res.* 9, 6389–6394.

Suzuki, H., Okunishi, R., Hashizume, W., Katayama, S., Ninomiya, N., Osato, N., et al. (2004). Identification of region-specific transcription factor genes in the adult mouse brain by medium-scale real-time RT-PCR. *FEBS Lett.* 573, 214–218. doi: 10.1016/j.febslet.2004.07.068

Tan, Z., Liu, R., Zheng, L., Hao, S., Fu, C., Li, Z., et al. (2015). Cerebrospinal fluid protein dynamic driver network: at the crossroads of brain tumorigenesis. *Methods* 83, 36–43. doi: 10.1016/j.ymeth.2015.05.004

Venegas, J. G., Winkler, T., Musch, G., Vidal Melo, M. F., Layfield, D., Tgavalekos, N., et al (2005). Self-organized patchiness in asthma as a prelude to catastrophic shifts. *Nature* 434, 777–782. doi: 10.1038/nature03490

Wen, Z., Zhang, W., Zeng, T., and Chen, L. (2014). MCentridFS: a tool for identifying module biomarkers for multi-phenotypes from high-throughput data. *Mol. BioSyst.* 10, 2870–2875. doi: 10.1039/C4MB00325J

Zeng, T., Zhang, C., Zhang, W., Liu, R., Liu, J., and Chen, L. (2014). Deciphering early development of complex diseases by progressive module network. *Methods* 67, 334–343. doi: 10.1016/j.ymeth.2014.01.021

Zhang, W., Zeng, T., and Chen, L. (2014). EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. *J. Theor. Biol.* 362, 35–43. doi: 10.1016/j.jtbi.2014.05.041

Zhang, W., Zeng, T., Liu, X., and Chen, L. (2015). Diagnosing phenotypes of single-sample individuals by edge biomarkers. *J. Mol. Cell Biol.* 7, 231–241. doi: 10.1093/jmcb/mjv025

Conflict of Interest Statement: 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 August 2015
doi: 10.3389/fgene.2015.00260

Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types

Martin H. Schaefer[1,2]*, *Luis Serrano*[1,2,3] *and Miguel A. Andrade-Navarro*[4,5]

[1] Systems Biology Research Unit, Centre for Genomic Regulation – European Molecular Biology Laboratory, Barcelona, Spain, [2] Universitat Pompeu Fabra, Barcelona, Spain, [3] Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain, [4] Faculty of Biology, Johannes Gutenberg University of Mainz, Mainz, Germany, [5] Institute of Molecular Biology, Mainz, Germany

OPEN ACCESS

Edited by:
Josselin Noirel,
Conservatoire National des Arts et Métiers, France

Reviewed by:
Andrei Ivanov,
Emory University, USA
Giuseppe Gallone,
University of Oxford, UK

Correspondence:
Martin H. Schaefer,
Systems Biology Research Unit, Centre for Genomic Regulation – European Molecular Biology Laboratory, Aiguader 88, 08003 Barcelona, Spain
martin.schaefer@crg.eu

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

Received: *10 February 2015*
Accepted: *21 July 2015*
Published: *04 August 2015*

Citation:
Schaefer MH, Serrano L and Andrade-Navarro MA (2015) Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. Front. Genet. 6:260. doi: 10.3389/fgene.2015.00260

Protein–protein interaction (PPI) networks are associated with multiple types of biases partly rooted in technical limitations of the experimental techniques. Another source of bias are the different frequencies with which proteins have been studied for interaction partners. It is generally believed that proteins with a large number of interaction partners tend to be essential, evolutionarily conserved, and involved in disease. It has been repeatedly reported that proteins driving tumor formation have a higher number of PPI partners. However, it has been noticed before that the degree distribution of PPI networks is biased toward disease proteins, which tend to have been studied more often than non-disease proteins. At the same time, for many poorly characterized proteins no interactions have been reported yet. It is unclear to which extent this study bias affects the observation that cancer proteins tend to have more PPI partners. Here, we show that the degree of a protein is a function of the number of times it has been screened for interaction partners. We present a randomization-based method that controls for this bias to decide whether a group of proteins is associated with significantly more PPI partners than the proteomic background. We apply our method to cancer proteins and observe, in contrast to previous studies, no conclusive evidence for a significantly higher degree distribution associated with cancer proteins as compared to non-cancer proteins when we compare them to proteins that have been equally often studied as bait proteins. Comparing proteins from different tumor types, a more complex picture emerges in which proteins of certain cancer classes have significantly more interaction partners while others are associated with a smaller degree. For example, proteins of several hematological cancers tend to be associated with a higher number of interaction partners as expected by chance. Solid tumors, in contrast, are usually associated with a degree distribution similar to those of equally often studied random protein sets. We discuss the biological implications of these findings. Our work shows that accounting for biases in the PPI network is possible and increases the value of PPI data.

Keywords: protein–protein interactions, study bias, network analysis, degree distribution, cancer genes

Introduction

Protein–protein interaction (PPI) networks are important models of the functional organization of the cell. To date many small and large scale studies exist mapping PPIs in human (the integrated database HIPPIE; Schaefer et al., 2012, hosts PPIs from 34,625 different studies). However, we are still far from the complete knowledge of the human interactome (Venkatesan et al., 2009), especially when its (spatial and temporal) dynamics and context-dependence are taken into account (Ideker and Krogan, 2012; Schaefer et al., 2013). High error rates associated with the experimental methods applied to measure PPIs have been recognized as a major burden for completing this goal (Von Mering et al., 2002). However, besides experimental error, other biases pose problems on the analysis of PPI networks.

Protein–protein interaction networks are associated with two types of biases: technical biases caused by limitations inherent to the experimental techniques applied to generate the PPI networks and study biases driven by the research interests guiding the selection of bait proteins tested for interaction partners. Examples for technical biases are the tendency of tandem affinity purification followed by mass spectrometry (TAP/MS) to detect interactions between highly abundant proteins (Von Mering et al., 2002; Björklund et al., 2008; Ivanic et al., 2009) and interactions involving small proteins under 15 kDa (Gavin et al., 2002). Yeast two-hybrid (Y2H) tends to detect interactions between protein pairs located in the nucleus (Jensen and Bork, 2008).

The study bias arises due to the fact that proteins are studied an uneven amount of times: some proteins (e.g., with higher biomedical relevance) are studied more often than proteins with unknown biological function. In yeast, the more GO terms a protein is annotated to the more likely it is to be studied (Gillis and Pavlidis, 2011; Gillis et al., 2014). This type of bias is particularly strong in aggregated networks (Gillis et al., 2014) as are commonly used in network biology. Not surprisingly, highly studied proteins are associated with a higher number of known PPI partners (their degree; Hakes et al., 2008). This poses a major challenge on the analysis and interpretation of PPI networks: it might misleadingly suggest a correlation between the biological relevance of a protein and network properties as, for example, the degree of a protein. Indeed, several studies reported a higher degree for essential proteins (Coulomb et al., 2005) and for disease proteins such as cancer proteins (Wachi et al., 2005; Jonsson and Bates, 2006; Rambaldi et al., 2008). It is unclear to which extent the reported higher degree of disease proteins reflects biological properties of disease proteins in networks and how much their degree is influenced by the fact that disease proteins are studied more often than other proteins.

The observation that disease proteins have more interaction partners than non-disease proteins led to numerous computational studies using directly or indirectly the degree of a protein as a predictor for its function or disease relation (e.g., Xu and Li, 2006; Nie and Yu, 2013) that thereby might only reveal highly studied proteins that are more likely to be associated to the studied function anyway.

To avoid misleading conclusions from biased PPI networks, it was repeatedly proposed to rely on non-biased large scale screens for the analysis of network properties of distinct protein classes (Zotenko et al., 2008; Rolland et al., 2014). However, the experimental coverage of the protein set of interest is usually low when only a single or few large scale studies are considered. To our knowledge, there is only one study that addressed the bias directly with a normalization strategy for the analysis of properties of HIV targets (Dickerson et al., 2010).

Here, we first aim to quantify the impact of the study bias on the observed degree distribution in a large integrated PPI network. We then investigate if one of the most frequently made claims with respect to network properties of disease proteins, the higher degree of cancer proteins, holds when we take into account the higher number of times these proteins have been tested for PPI partners. Surprisingly, we find that a much more complex picture of the degree-disease relation emerges when correcting for the study bias, with a high heterogeneity across different cancer types.

Materials and Methods

Protein–Protein Interaction Data

Protein–Protein Interactions were retrieved from HIPPIE version 1.5 (Schaefer et al., 2012). HIPPIE is an integrated PPI resource aggregating all PPIs from various expert-curated databases. HIPPIE implements a confidence score, which reflects the amount and type of evidence supporting an interaction (such as the number of studies reporting an interaction). However, for the purpose of this analysis we considered all 122,755 PPIs in HIPPIE as we reasoned that filtering for experimental evidence would further increase the study bias in the resulting subnetwork. Bait usage statistics were extracted from the PPI databases Mint (Chatr-aryamontri et al., 2007), IntAct (Kerrien et al., 2007), and iRefWeb (Turner et al., 2010). We annotated the number of studies in which a protein was used as a bait.

Statistical Analyses

Statistical hypothesis testing was performed with the *R* statistical computing environment. For estimating the significance of the Pearson correlation, the test statistic was based on Pearson's product moment correlation coefficient. The confidence interval was based on Fisher's *Z* transform. The randomization test was performed by replacing each cancer protein by a non-cancer protein that had been equally often tested as a bait. To obtain reasonably distinct random protein sets we included proteins with similar bait usage when there were fewer than four proteins that had been tested as a bait equally often. Therefore, we successively extended a random set with similarly often studied proteins until the size of the set exceeded four proteins. First, we included proteins tested as baits 20 times more or 20 times less often than the original protein. If there were still less than four proteins in the range we successively increased the range to 150 times tested and then to 250 times tested more or less than the original proteins.

Cancer Data

A recent study analyzed almost 5000 different human cancer exomes and their matched normal-tissue samples to detect significantly mutated genes in a representative selection of 21 tumor types under a unified statistical framework (Lawrence et al., 2014). From this study, we extracted the enrichment of somatic point mutations for each gene and tumor type. We considered a gene a cancer gene if the enrichment q-value (the false discovery rate adjusted equivalent to the p-value) was below 0.1 for the respective tumor type. From the 21 different cancer types, we analyzed 15 that were associated with at least seven genes.

Gene Ontology Enrichment

For the GO term enrichment analysis we used the tool ConsensusPathDB (Kamburov et al., 2011). For the analysis of highly studied proteins, only terms below a q-value threshold of 0.01 were considered. For the analysis of functions associated with highly connected cancer genes, we applied the same q-value threshold to select terms enriched among all genes of the respective cancer and additionally tested for the resulting terms if they were significantly more associated with highly connected proteins (as compared to lowly connected).

Results

Highly Studied Proteins have More Protein Interaction Partners

To quantify the relation between the number of times a protein has been studied and the reported number of PPI partners, we computed the degree of each protein from the integrated PPI database HIPPIE (Schaefer et al., 2012). Next, we recorded how many times each protein has been studied as a bait in studies reporting PPIs (**Figure 1A** displays the fraction of proteins for which we had information on how often they had been tested as bait proteins). Finally, we annotated the number of PubMed abstracts linked to each protein (as a proxy for the number of studies reporting the protein; provided by the PubMed FTP server; downloaded on January 8, 2015). In **Figure 1B** the number of interaction partners of a protein is plotted against the number of studies in which the protein has been tested for interaction partners (**Figure 1B** shows the relation in log–log space, **Supplementary Figure S1** on linear scale). **Figure 1C** visualizes the relation between the number of interaction partners and both the number of all studies and of studies testing the protein as a bait for interaction partners after grouping the number of interaction partners into quartiles. As expected, the correlation between the number of times a protein has been tested for interaction partners as a bait protein and the interaction degree of a protein (Pearson correlation of 0.520) is higher than the correlation between the total number of times a protein has been studied (including studies not focused on PPIs) and the degree (Pearson correlation of 0.334). However, both variables are significantly correlated with the protein interaction degree ($p < 10^{-16}$; see Materials and Methods).

Properties of Highly Studied Proteins

Using ConsensusPathDB (Kamburov et al., 2011) we evaluated the enrichment of functions and pathways in the set of 7114 bait proteins in terms of q-values (see Materials and Methods for details). In accordance with a previous study that investigated functional categories enriched among entire networks (Futschik et al., 2007), we found a strong enrichment of proteins with nuclear localization, or with functions in cell cycle and metabolism ($q < 10^{-4}$) among the proteins used as baits. When calculating the enrichment of functional terms and pathways among the 173 proteins most frequently used as a bait (at least 20 times) relative to that of the full bait list, most strongly enriched were "pathways related to cancer" ($q < 10^{-39}$). Other strongly enriched protein classes were related to viral infection [Hepatitis B ($q < 10^{-28}$), Epstein–Barr ($q < 10^{-21}$), HIV ($q < 10^{-17}$), and Herpes simplex ($q < 10^{-15}$)] and signaling pathways [TNFalpha ($q < 10^{-28}$), TGFbeta ($q < 10^{-24}$), and Leptin signaling ($q < 10^{-23}$)]. While the enrichment of nuclear proteins in the entire bait set might be caused by a technical detection bias of the still predominantly used Y2H assay, which requires nuclear localization of the bait and prey proteins, the strong enrichment for cancer pathways in the frequently studied bait set clearly indicates a selection bias toward proteins with high biomedical relevance.

Correcting for the Bait Usage Bias

To reconfirm the previously reported (Wachi et al., 2005; Jonsson and Bates, 2006; Rambaldi et al., 2008) difference in the degree distribution between cancer and non-cancer proteins, we retrieved and pooled somatically mutated cancer genes from 21 different tumor types (Lawrence et al., 2014). We compared the number of PPIs of cancer proteins to the number of PPIs of non-cancer proteins. We observed that the cancer proteins have a significantly higher number of PPI partners ($p < 10^{-16}$; Wilcoxon Mann–Whitney test; **Figure 2A**) but we suspected that this difference could be an artifact caused by the largely different number of times the two protein classes have been studied for interaction partners.

To investigate this artifact, we randomly generated sets of non-cancer proteins equivalent (in terms of having been studied as baits) to the set of cancer proteins. This was done by replacing each cancer protein by a randomly selected protein used the same number of times as a bait protein than the cancer protein (or similar number of times if no protein existed that was tested the exact same number of times). For each of the 10,000 generated random sets, we calculated the mean number of interaction partners (**Figure 2B**). We found that cancer proteins tend to be involved in more PPIs than non-cancer proteins used as baits as often as cancer proteins. However, we did not observe a significant difference (a p-value computed as the fraction of times the mean degree of the randomized set was larger than the observed mean degree for cancer proteins; $p = 0.0626$). The lack of a significant difference between cancer proteins and equally often studied random protein sets (as compared to the highly significant difference between cancer proteins and all non-cancer proteins) suggests that previous observations on

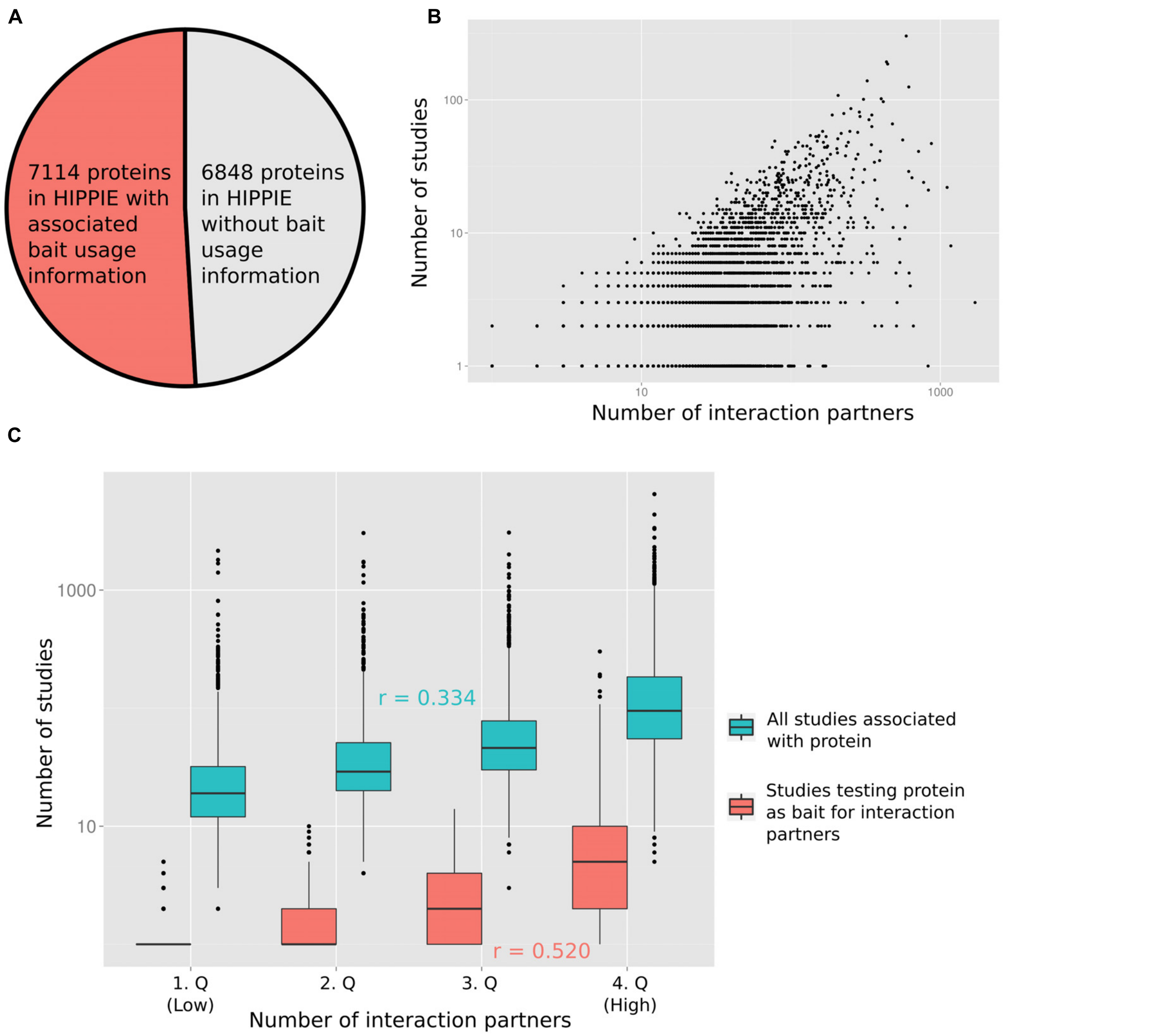


FIGURE 1 | (A) For slightly more than half of the proteins in HIPPIE we had information of how often they had been tested as bait proteins for interaction partners. **(B)** The number of studies in which a protein has been tested as a bait for interaction partners is positively correlated with the number of reported interactions (Pearson correlation of 0.520). **(C)** The protein–protein interaction (PPI) degree distribution has been split into four equally sized quartiles (1.Q–4.Q). The distribution of numbers of all studies and only those studies testing the associated proteins as baits for interaction partners are shown for the quartiles. Pearson correlation values (*r*) are indicated.

particular network characteristics of cancer proteins are biased by the differential research interest in disease versus non-disease proteins.

Studying the Degree Distributions of Different Cancers

Next we investigated if the deviation between observed and expected degree distributions differs across cancer types. Therefore, we applied the same randomization strategy as before to correct the study bias in the degree distributions of cancer proteins from 15 different tumor types (Lawrence et al., 2014). An interesting picture emerged: while proteins from several cancer types had close to random expectation degree distributions, most cancers of the hematological system had the highest deviation between mean of the observed degree distribution and the mean of the degree distribution of randomly sampled protein sets studied similarly often for interaction partners (**Figure 3**). The highest deviations between observed and expected degree distribution were for chronic lymphocytic leukemia (CLL; $p = 0.0248$; randomization test), diffuse large B-cell lymphoma (DLBCL; $p = 0.0354$; randomization test) and acute myeloid leukemia (LAML; $p = 0.0525$; randomization test). Interestingly, the higher degree distribution of hematological cancer proteins is achieved by distinct protein sets and not an artifact of overlapping cancer protein sets: no protein was associated to these three cancers and just three proteins appeared in association with two (see **Supplementary Table S1**).

To investigate possible functional reasons for the higher than expected by chance degree distribution of hematological

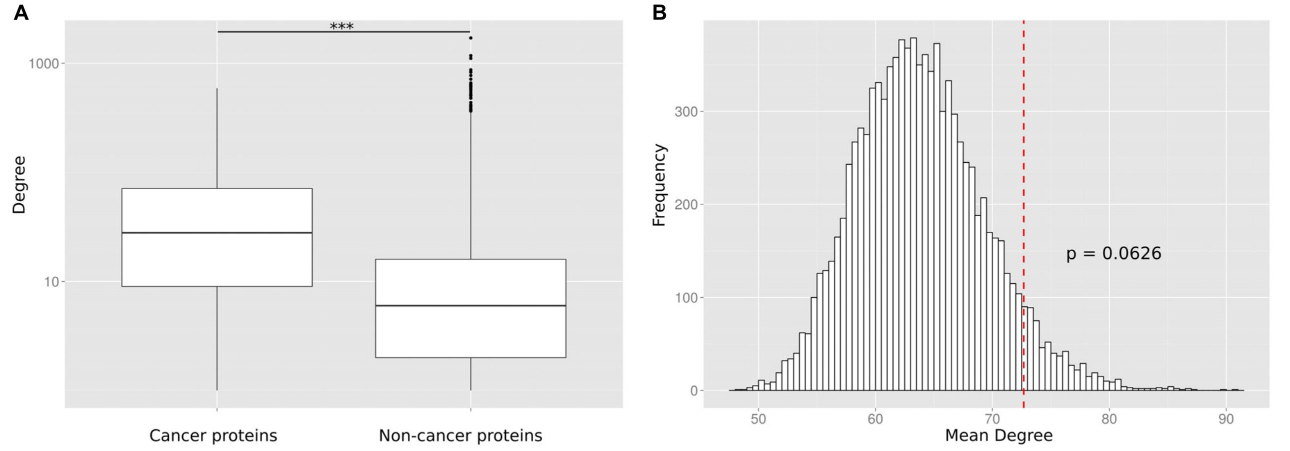


FIGURE 2 | (A) Without correcting for the study bias the PPI degree difference between cancer proteins and non-cancer proteins is highly significant (*** $p < 10^{-16}$). **(B)** The number of PPIs is not significantly enriched for cancer proteins as compared to equally frequently studied random protein sets. The histogram shows the mean degree of 10,000 random gene sets with the same bait usage distribution as the cancer protein set. The position of the red dotted line indicates the mean degree of the cancer protein set. The *p*-value is computed as the fraction of times the mean degree of the randomized set was larger than the observed mean degree and is not significant ($p = 0.0626$).

cancer proteins, we computed for the proteins from those cancer classes the ratio between the degree and the number of times a protein had been tested as a bait protein (as a proxy for a bias normalized degree estimate; **Supplementary Table S1**). For each of the three cancer types, we focused on the 50% of the proteins with the highest ratio. Interestingly, in two of the cases the most highly connected proteins were indicative of cancerogenesis processes specific to the respective hematological tumor.

Two (RPS15 and XPO1) of the three CLL proteins with the highest ratio of degree to experiments (out of six CLL proteins for which we have PPI experimental data) are involved in the establishment of ribosome localization (while none are from the proteins with a lower ratio). The third of the highest ratio proteins (SF3B1) is also a ribonucleoprotein. The higher degree of the ribosome-related proteins ($p = 0.05$; Fisher test) is not surprising as 100s of closely interacting proteins are involved in the biogenesis and transport of the ribosomal subunits (Fromont-Racine et al., 2003; Altvater et al., 2012). Interestingly, CLL cells show impaired assembly of ribosomes (Rubin, 1971), which likely reduces their metabolic activity and helps them to avoid cell death (Defoiche et al., 2010).

Of the seven DLBCL proteins with the highest ratio, five were involved in the activation of leukocytes (of a total of 13 DLBCL proteins with bait usage information). From the six proteins with lower ratio none was associated with this function ($p < 0.05$; Fisher test). Interestingly, many lymphomas resemble gene expression patterns of activated B cells (Alizadeh et al., 2000). Leukocyte activation has been shown to be driven by a large and highly interconnected protein network (Calvano et al., 2005).

The examples of ribonucleoproteins in CLL and leukocyte activators in DLBCL illustrate how selection for tumor-specific functions modify the observed degree distribution of each tumor. In conclusion, there is no generally elevated connectivity of cancer proteins. Only in some cancer types groups of proteins tend to be mutated that belong to highly interconnected cellular networks.

To estimate how robust our observations are with respect to variations in the computation of the test statistic, we repeated the randomization procedure computing the median degree of the original and randomized protein sets instead of the mean. The overall observation remained unchanged: random proteins with bait usage similar to that of cancer proteins have higher degree than random proteins without any constraints on the bait usage (both for mean and median; **Supplementary Figure S2**). However, using the median we observed a significant degree enrichment for cancer proteins ($p < 0.01$; randomization test) and this time CLL, LAML, and BRCA had significantly higher number of PPIs as compared to random sets (all $p < 0.05$; randomization test).

Discussion

Here, we quantify how the frequency with which a protein has been studied for interaction partners affects its reported degree distribution. We estimate that the resulting bias is higher than previously quantified biases resulting from technical limitations. For example, the correlation between protein abundance and degree ranges for different TAP/MS networks from 0.21 to 0.46 (Ivanic et al., 2009) while we observe a correlation >0.5 between the number of times a protein has been tested as a bait and its degree.

Our findings have a dramatic impact on the common understanding of the relation between protein function and degree. Specifically, we challenge the previous finding that cancer proteins tend to have more interaction partners than non-cancer proteins (Wachi et al., 2005; Jonsson and Bates, 2006; Rambaldi

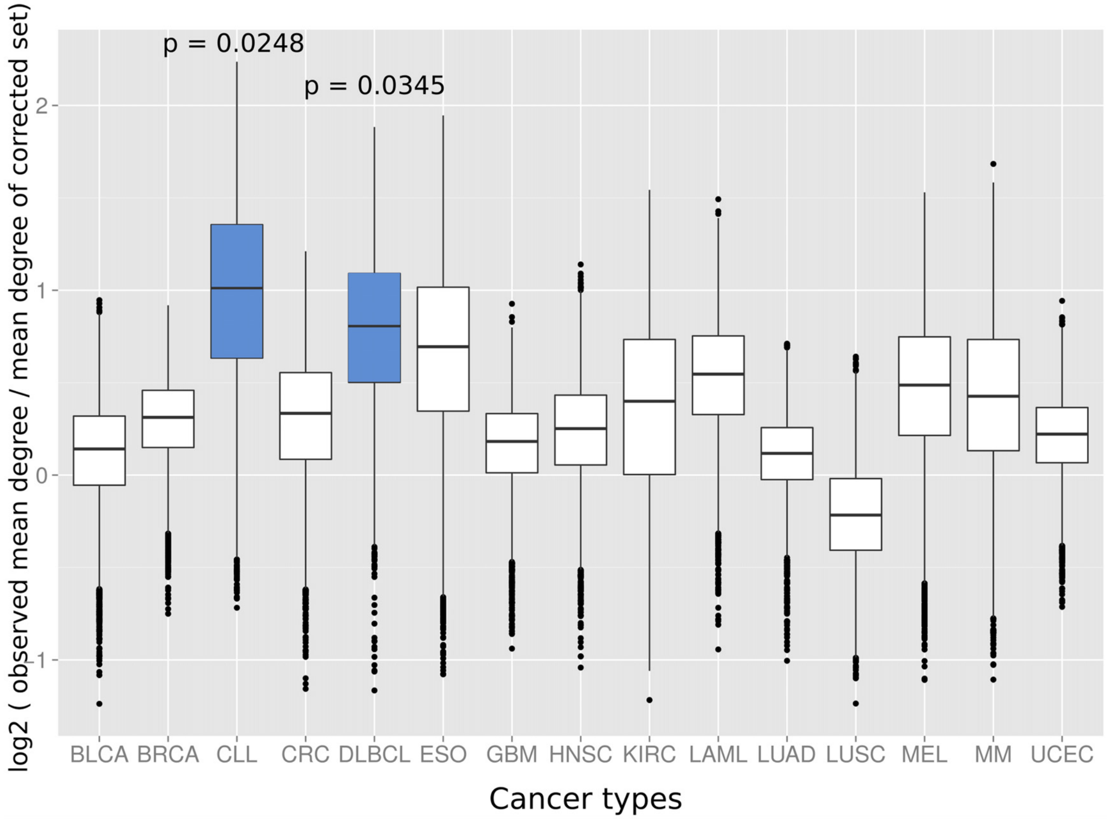


FIGURE 3 | The distributions show the log2 of the mean degree of proteins of the specific cancer type divided by the mean degree of 10,000 randomized protein sets with the same bait usage distribution (a value of 0 would therefore signify that the mean degree of the cancer protein set equals the observed mean degree of the random protein set, positive values that the mean degree of the proteins of the respective cancer type is higher than for the random set and, vice versa, negative values that the mean degree of the random set proteins is higher as for the proteins of the cancer type). Blue boxes indicate that the mean of the original degree distribution of the respective cancer type is significantly higher ($p < 0.05$; randomization test) as those of randomized protein sets with the same bait usage distribution. The cancer types on the *x*-axis are: BLCA, bladder cancer; BRCA, breast cancer; CLL, chronic lymphocytic leukemia; CRC, colorectal cancer; DLBCL, diffuse large B-cell lymphoma; ESO, esophageal adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck cancer; KIRC, kidney clear cell carcinoma; LAML, acute myeloid leukemia; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MEL, melanoma; MM, multiple myeloma, and UCEC, endometrial cancer.

et al., 2008). In fact a more complex picture emerges: while some cancer types are associated with proteins of lower or average connectivity, others are associated with promiscuous proteins. The different degree distributions correlate with functional specificities of the tumor types. Interestingly, the higher degree distribution of hematological cancer genes is driven by largely different protein sets with distinct functions (the proteins with the highest ratio between degree and bait usage are ribonucleoproteins for CLL and proteins involved in leukocyte activation for DLBCL).

From our analysis it follows that many cancer gene prediction approaches might have overestimated their performance as they directly or indirectly use the PPI degree as a feature for classification. A classifier that preferentially selects proteins with high degree will therefore favor highly studied proteins, which in turn are more likely to be cancer proteins. This should be taken into consideration by either using less biased networks from proteome-scale screens or by omitting degree-related features for classification.

One limitation of the presented method is that the reported number of times a protein has been tested as a bait gives only a rough and a lower bound estimate as for many experiments this information is not available in the public databases. Also, the distinction between bait and prey protein might not apply to all types of experimental methods (as for example for crystallization of complexes). As described in the Results section, our method shows a certain sensitivity with respect to the chosen statistics. However, the overall tendency in the results stayed the same when the median instead of the mean was computed: randomly sampled proteins that have been studied as often as cancer proteins are more similar in their degree distribution to cancer proteins as to arbitrarily often studied proteins.

In summary, we argue for the crucial importance of taking into account the number of times a protein has been studied when analyzing PPI networks. Ignoring the resulting degree distribution bias is not just leading to wrong biological assumptions on the relation between network topology and protein function but also introduces circularity into network-based disease gene prediction.

To come to reliable conclusions regarding degree differences between protein classes, it would be generally favorable if rarely studied proteins would be increasingly often tested for PPI partners to eliminate the differences in the very uneven bait usage distribution. While these sharp differences persist, the here presented methods can help to reduce the impact of the study bias when comparing degree distributions and could be applied to other disease protein classes.

Author Contributions

MS and MA conceived and designed the analyses. MS analyzed the data. MS, LS, and MA wrote the paper.

Acknowledgments

The research leading to these results has received funding from the Severo Ochoa award, the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° HEALTH-F4-2011-278568 (PRIMES), from the Spanish Ministerio de Economía y Competitividad (Plan Nacional BIO2012-39754) and the European Fund for Regional Development (EFRD).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00260

FIGURE S1 | The number of studies in which a protein has been tested as a bait for interaction partners is plotted against the number of reported interactions in linear scale.

FIGURE S2 | We randomly sampled 100 protein sets of size 10 (a) from the cancer proteins, (b) equally often studied (as bait) non-cancer proteins, and (c) non-cancer proteins without any constraints on the bait usage. We computed both the mean and the median for each of the resulting 300 protein sets. The resulting mean/median degree distributions are shown. Although with this sampling strategy all distributions are pairwise dissimilar (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$), the random proteins that have been studied as often as the cancer proteins have a much more similar degree distribution to the cancer proteins as compared to randomly sampled background proteins (even though the similarity is higher when the mean is computed than when the median is computed).

TABLE S1 | The table shows hematological cancer proteins for which PPI and bait usage information was available. Gene name and entrez gene id, tumor type and the ratio between the degree and the number of times the protein has been tested as a bait are indicated.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511. doi: 10.1038/35000501

Altvater, M., Chang, Y., Melnik, A., Occhipinti, L., Schütz, S., Rothenbusch, U., et al. (2012). Targeted proteomics reveals compositional dynamics of 60S pre-ribosomes after nuclear export. *Mol. Syst. Biol.* 8, 628. doi: 10.1038/msb.2012.63

Björklund, Å. K., Light, S., Hedin, L., and Elofsson, A. (2008). Quantitative assessment of the structural bias in protein–protein interaction assays. *Proteomics* 8, 4657–4667. doi: 10.1002/pmic.200800150

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V, Cho, R. J., et al. (2005). A network-based analysis of systemic inflammation in humans. *Nature* 437, 1032–1037. doi: 10.1038/nature03985

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V, Castagnoli, L., et al. (2007). MINT: the molecular INTeraction database. *Nucleic Acids Res.* 35, D572–D574. doi: 10.1093/nar/gkl950

Coulomb, S., Bauer, M., Bernard, D., and Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks. *Proc. Biol. Sci.* 272, 1721–1725. doi: 10.1098/rspb.2005.3128

Defoiche, J., Zhang, Y., Lagneaux, L., Willems, L., and Macallan, D. C. (2010). In vivo ribosomal RNA turnover is down-regulated in leukaemic cells in chronic lymphocytic leukaemia. *Br. J. Haematol.* 151, 192–195. doi: 10.1111/j.1365-2141.2010.08334.x

Dickerson, J., Pinney, J., and Robertson, D. (2010). The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. *BMC Syst. Biol.* 4:80. doi: 10.1186/1752-0509-4-80

Fromont-Racine, M., Senger, B., Saveanu, C., and Fasiolo, F. (2003). Ribosome assembly in eukaryotes. *Gene* 313, 17–42. doi: 10.1016/S0378-1119(03)00629-2

Futschik, M. E., Chaurasia, G., and Herzel, H. (2007). Comparison of human protein–protein interaction maps. *Bioinformatics* 23, 605–611. doi: 10.1093/bioinformatics/btl683

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147. doi: 10.1038/415141a

Gillis, J., Ballouz, S., and Pavlidis, P. (2014). Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *J. Proteomics* 100, 44–54. doi: 10.1016/j.jprot.2014.01.020

Gillis, J., and Pavlidis, P. (2011). The impact of multifunctional genes on "guilt by association" analysis. *PLoS ONE* 6:e17258. doi: 10.1371/journal.pone.0017258

Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008). Protein-protein interaction networks and biology–what's the connection? *Nat. Biotechnol.* 26, 69–72. doi: 10.1038/nbt0108-69

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8, 565. doi: 10.1038/msb.2011.99

Ivanic, J., Yu, X., Wallqvist, A., and Reifman, J. (2009). Influence of protein abundance on high-throughput protein-protein interaction detection. *PLoS ONE* 4:e5815. doi: 10.1371/journal.pone.0005815

Jensen, L. J., and Bork, P. (2008). Not Comparable, but complementary. *Science* 322, 56–57. doi: 10.1126/science.1164801

Jonsson, P. F., and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297. doi: 10.1093/bioinformatics/btl390

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.* 39, D712–D717. doi: 10.1093/nar/gkq1156

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., et al. (2007). IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* 35, D561–D565. doi: 10.1093/nar/gkl958

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Nie, Y., and Yu, J. (2013). Mining breast cancer genes with a network based noise-tolerant approach. *BMC Syst. Biol.* 7:49. doi: 10.1186/1752-0509-7-49

Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A., and Ciccarelli, F. D. (2008). Low duplicability and network fragility of cancer genes. *Trends Genet.* 24, 427–430. doi: 10.1016/j.tig.2008.06.003

Rolland, T., Tşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Rubin, A. D. (1971). Defective control of ribosomal RNA processing in stimulated leukemic lymphocytes. *J. Clin. Invest.* 50, 2485–2497. doi: 10.1172/JCI106749

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS ONE* 7:e31826. doi: 10.1371/journal.pone.0031826

Schaefer, M. H., Lopes, T. J. S., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J.-F., et al. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.* 9:e1002860. doi: 10.1371/journal.pcbi.1002860

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* 2010:ba9023. doi: 10.1093/database/baq023

Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., et al. (2009). An empirical framework for binary interactome mapping. *Nat. Methods* 6, 83–90. doi: 10.1038/nmeth.1280

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., et al. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403. doi: 10.1038/nature750

Wachi, S., Yoneda, K., and Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21, 4205–4208. doi: 10.1093/bioinformatics/bti688

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* 4:e1000140. doi: 10.1371/journal.pcbi.1000140

Conflict of Interest Statement: 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.

frontiers in Genetics

MINI REVIEW
published: 19 August 2015
doi: 10.3389/fgene.2015.00257

Human protein interaction networks across tissues and diseases

Esti Yeger-Lotem[1]* *and Roded Sharan*[2]*

[1] *Department of Clinical Biochemistry and Pharmacology, Ben-Gurion University of the Negev, Beer-Sheva, Israel,* [2] *Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel*

Protein interaction networks are an important framework for studying protein function, cellular processes, and genotype-to-phenotype relationships. While our view of the human interaction network is constantly expanding, less is known about networks that form in biologically important contexts such as within distinct tissues or in disease conditions. Here we review efforts to characterize these networks and to harness them to gain insights into the molecular mechanisms underlying human disease.

Keywords: protein interaction network, tissue-specific network, disease-specific network, network perturbation, gene expression

OPEN ACCESS

Edited by:
Miguel Andrade,
Johannes-Gutenberg University of Mainz, Germany

Reviewed by:
Matteo Barberis,
University of Amsterdam, Netherlands
Oksana Sorokina,
The University of Edinburgh, UK

Correspondence:
Esti Yeger-Lotem,
Department of Clinical Biochemistry and Pharmacology, Ben-Gurion University of the Negev,
PO Box 653,
Beer-Sheva 84105, Israel
estiyl@bgu.ac.il;
Roded Sharan,
Blavatnik School of Computer Science, Tel Aviv University,
PO Box 39040, Tel Aviv 69978, Israel
roded@post.tau.ac.il

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

Received: *28 May 2015*
Accepted: *17 July 2015*
Published: *19 August 2015*

Citation:
Yeger-Lotem E and Sharan R (2015) Human protein interaction networks across tissues and diseases. Front. Genet. 6:257. doi: 10.3389/fgene.2015.00257

Introduction

Protein molecules constitute the main building blocks of cells and mediate most cellular processes. In human, they are encoded by over 22,000 different genes, which give rise to many more proteins through alternative splicing mechanisms. These numerous proteins do not work in isolation: instead, they interact with each other and with other types of molecules to form complex cellular machines and to pass signals within cells and across tissues. In recognition of the fundamental role of these molecular interactions, much effort has been invested in the last two decades in their mapping. From small-scale experiments that measure interactions between a few proteins, mapping has changed to large-scale screens using high-throughput techniques such as yeast two-hybrid and co-immunoprecipitation (e.g., Rual et al., 2005; Stelzl et al., 2005; Ewing et al., 2007; Rolland et al., 2014). Owing to these mapping efforts, our current view of the physical interactions between human proteins encompasses over 200,000 interactions among over 20,000 proteins, and is continuously expanding. The resulting network of all known protein-protein interactions (PPIs), known as the human interactome, has become a key framework for studying protein function, cellular processes, and genotype-to-phenotype relationships, as reviewed elsewhere (Barabási et al., 2011; Vidal et al., 2011). However, this broad network is also limited. PPIs have rarely been measured in the context of distinct cell types, tissues, or in disease conditions, making it difficult to model and understand context-related phenotypes.

While knowledge of human context-specific PPIs is limited, we are witnessing a rapid accumulation of context-specific molecular expression profiles. The human body consists of tens of tissues, sub-tissues, and cell types that differ from one another in morphology and function. In a seminal study published more than a decade ago, Su et al. (2004) opened a window into their molecular characteristics by profiling the transcriptomes of 79 human tissues via DNA microarrays. Other studies profiled the transcriptomes of human tissues by techniques such as massively parallel signature sequencing (Jongeneel et al., 2005), expressed sequence tags (EST) (Hillier et al., 1996), and next generation RNA sequencing (e.g., Illumina's BodyMap 2.0). Most recent is the RNA sequencing of multiple human tissues from a number of individuals by the Genotype Tissue Expression project (Mele et al., 2015). The proteomes of human tissues have also been profiled by immunohistochemistry (Pontén et al., 2009; Uhlén et al., 2015) and mass-spectrometry techniques (Kim et al., 2014; Wilhelm et al., 2014). In addition to efforts to profile normal tissues, profiling techniques have also been employed to characterize different diseases. One of the more prominent initiatives is The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), which is actively mapping genomic, transcriptomic, proteomic, and epigenomic changes in cancerous tissues compared to normal tissues. These measurements shed light on the parts of the interactome that are active in these diverse contexts, although direct experimentation is required to reveal the actual PPI changes, in particular the formation of novel interactions (Ideker and Krogan, 2012). Below we discuss efforts to harness these context-specific molecular expression profiles to elucidate network properties of human tissues and to identify interaction-based disease mechanisms.

Features of Tissue and Cell-type Specific Networks

Given the lack of context-specific PPIs that were measured in different tissues and cell types, many studies revert to identifying PPIs that are feasible in these contexts. Their underlying assumption is that a PPI is feasible within a specific context if the corresponding proteins are expressed in that context. Of course not all feasible interactions actually take place, as they depend on many other factors such as localization and conformation of the two proteins, yet co-expression is necessary. Additionally, co-expression has often been based on RNA levels, as protein expression levels were rarely available. This approach had been used previously in model organisms to analyze their network dynamics in response to stimuli (Luscombe et al., 2004) or during cell cycle (de Lichtenberg et al., 2005), and has been used extensively for analyzing tissue interactomes (e.g., Lopes et al., 2011; Barshir et al., 2013; Song et al., 2014). Some differences in the sets of PPIs that are feasible within tissues and involve tissue-specific (TS) proteins and globally expressed (GE) "housekeeping" proteins are exemplified in **Figure 1**.

One of the first questions that had been asked was whether genes and PPIs that appear to be TS or GE have distinct topological features relative to the generic human interactome or to each other. Dezso et al. (2008) complied transcriptome profiles of 31 tissues, and found that the set of GE genes was larger than previously assumed. They showed that the topology of the GE PPI network was characterized by higher connectivity and shorter paths between proteins relative to the generic interactome. Lin et al. (2009) analyzed the number of interactions (degree), closeness, and betweenness centralities of GE and TS proteins within the generic PPI network. They found that GE genes were more central and may form a core, while clusters of TS genes attach to the core at more peripheral positions in the network. Using the data of Su et al. (2004), Bossi and Lehner (2009) found extensive direct interactions between GE and TS proteins, and suggested a model for the evolution of TS functions through the modification of core cellular processes. Souiai et al. (2011) used EST data across 45 tissues to test whether tissue-specificity is encoded in the interactome. They also found that GE genes were located at the topological center of the interactome. Denoting interactions occurring at a subset of tissues as TS interactions (TSI), they found that TSI involved in regulatory and developmental functions were also central, whereas TSI involved in organ physiological functions were peripheral. Kiran et al. (Kiran and Nagarajaram, 2013) analyzed features of highly connected proteins, namely hubs, in tissue interactomes. They showed that, among other features, TS hubs were associated with a lower degree of interactome centrality as compared with GE hubs. Waldman et al. (2010) analyzed translation efficiency, and showed that genes that were translated more efficiently in a specific tissue encode proteins that tend to have more interactions in that tissue, relative to other proteins in the same tissue.

The application of RNA-sequencing to human tissues revealed that many more transcripts were expressed per tissue than previously acknowledged (Ramsköld et al., 2009). Emig and Albrecht (2011) were among the first to harness RNA-sequencing data to the analysis of tissue interactomes. They showed that, in contrast to previous studies based on microarray profiles, TSI were less common, and were mainly involved in transmembrane transport and receptor activation. They also suggested that a considerable part of tissue-specificity is likely to be achieved by alternative splicing and interactions involving protein isoforms (further discussed in Buljan et al., 2012). In accordance with this suggestion, Ellis et al. (2012) demonstrated experimentally that neural-regulated exons can remodel PPIs by stimulating and repressing different partner interactions. Another study showed that proteins enriched with splice variants tend to occupy central positions in tissue interactomes (Sinha and Nagarajaram, 2014). Recently, it was claimed that splicing play mostly a complementary role in driving cellular specificity, except for the brain, which exhibits a more divergent splicing program (Mele et al., 2015).

Another technological breakthrough that is taking place in recent years is the profiling of proteomes at large scale. Since the correlation between transcript and protein levels is partial (Schwanhausser et al., 2011), proteome profiling opens a more direct way to identify feasible PPIs. Liu et al. (2014) used proteomic data (Kim et al., 2014) to analyze tissue interactomes. They showed that, relative to the generic interactome, tissue interactomes are smaller, sparser, and that hubs may have more important roles. Barshir et al. (2013) combined transcript and protein measurements to create 16 extensive tissue interactomes. Their comparative analysis (Barshir et al., 2014) revealed that each tissue interactome is dominated by a core sub-network that is common to all tissues, with only a small fraction being TS. Most tissue hubs were GE and retained their large PPI degree across tissues, and were enriched in regulatory functions. Lastly, they found in each tissue a significant correlation between transcript expression level and number of PPIs involving the encoded protein.

An important application of tissue interactomes is to shed light on disease mechanisms. Lage et al. (2008) systematically mapped over 1000 heritable diseases to the tissues in which they manifest clinically by using text-mining. They showed that

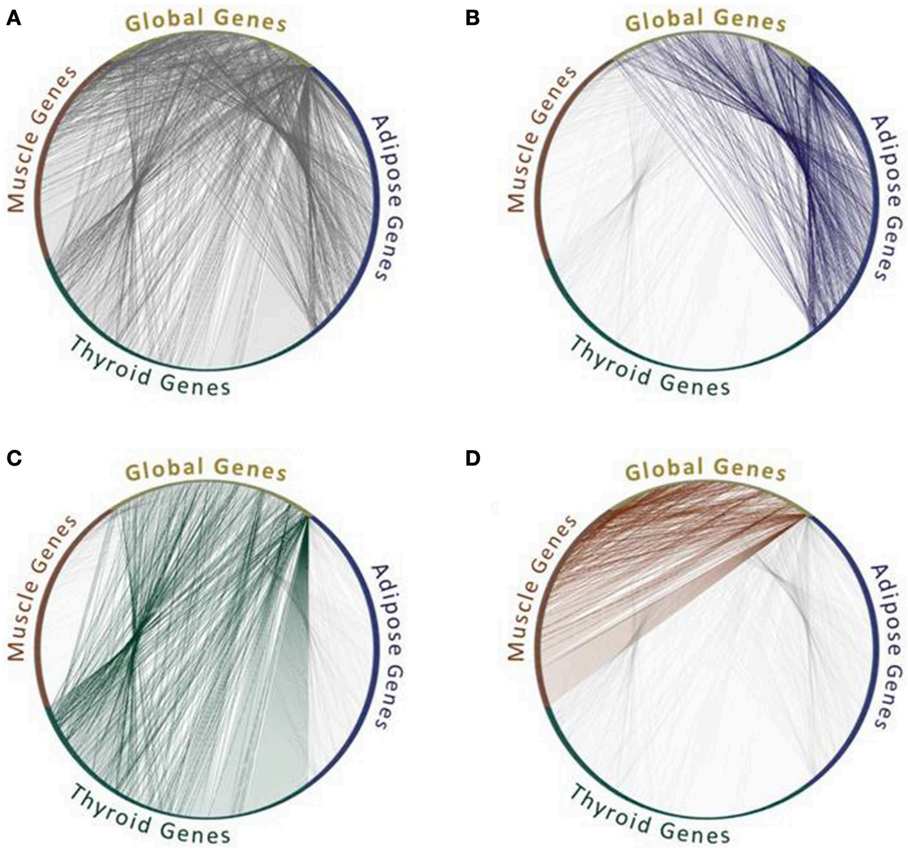


FIGURE 1 | Feasible protein interactions change between tissues. All protein interactions **(A)** and feasible protein interactions that connect "global genes," which are expressed in all three tissues, with tissue-specific genes that are expressed in one tissue out of adipose **(B)**, or thyroid **(C)**, or muscle **(D)**. Data of the genes expressed per tissue were extracted from GTEx Portal (Mele et al., 2015) and limited to genes with 50 counts and above. Data of protein interactions were extracted using MyProteinNet (Basha et al., 2015) from BioGrid (Chatr-Aryamontri et al., 2015), DIP (Xenarios et al., 2002), IntAct (Kerrien et al., 2012), and MINT (Licata et al., 2012) databases. Only global genes that have tissue-specific interactions in each of the three tissues are shown.

proteins and complexes that were linked to diseases tend to be over-expressed in the tissue where defects cause pathology, with the exception of proteins and complexes associated with cancers. Magger et al. (2012) showed that the usage of tissue interactomes, created from a generic interactome by removing or penalizing interactions involving non-expressed proteins, considerably improved the prioritization of disease genes. Li et al. (2014) assessed tissue interactomes weighted by DNA methylation data, and showed that they enhance prediction of disease genes. Barshir et al. (2014) focused on genes causing hereditary diseases and found that they tend to have PPIs that occur exclusively in the tissue where defects cause pathology. They demonstrated that these tissue-exclusive PPIs can highlight disease mechanisms, and, owing to their small number, suggested that they constitute an efficient filter for interrogating disease etiologies.

Perturbed Networks in Disease

Protein networks are perturbed in disease due to sequence mutations and expression changes. Zhong et al. (2009) were the first to systematically probe the effect of sequence (disease-causing) mutations on PPIs. They focused on known mutations causing Mendelian disorders and categorized them according to whether they have a truncation effect ("truncating," including nonsense mutations, out-of-frame indels, or defective splicing) or not ("in-frame," including missense mutations and in-frame indels). They showed that truncating mutations seem to lead to node-removal effects in the PPI network, while in-frame mutations are associated with edge-specific perturbations.

In a later study, Wang et al. (2012) examined the effect of disease-causing mutations using a structurally resolved PPI network, consisting of interactions and their atomic-resolution interfaces. They found that in-frame mutations tend to occur on the interaction interfaces of causal proteins and no similar enrichment was detected in non-interacting domains. This suggests that PPI perturbations play an important role in disease. Additionally, they found that the disease specificity for different mutations on the same gene can be explained by their location within the interface, further underscoring the importance of PPIs for the study of disease mechanisms.

On the technological side, Wei and Yu (Wei et al., 2014) developed an experimental pipeline to examine the consequences of different mutations on protein stability and interactions. They used the pipeline to show that disease causing mutations on interactions interfaces are more likely to perturb the corresponding interactions than mutations away from interfaces. Lambert et al. (2013) developed an experimental pipeline to score modulated interactions. The pipeline couples affinity purification to data-independent mass-spectrometric acquisition. The authors used it to identify interaction changes following disease-associated mutations and drug exposure.

Recently, Rolland et al. (2014) compared the impact of mutations associated with human disorders to that of common variants with no reported phenotypic consequences on PPIs. They focused on 32 genes with 115 disease and common variants, testing up to four disease and four common variants per disease gene for their impact on the ability of the corresponding proteins to interact with known interaction partners. They found that disease variants were 10-fold more likely to perturb interactions than common variants; more than 55% of the 107 interactions tested were perturbed by at least one disease-associated variant. In a follow-up study, Sahni et al. (2015) investigated the consequences of 2890 disease-causing missense mutations in 1140 genes. Out of 197 mutations covering 89 proteins with at least two PPI partners (in the HI-II-14 map of Rolland et al., 2014), 26% were found to cause a complete loss of interactions, 31% resulted in specific loss of some interactions, and 43% did not change the interaction partners. Disease mutations were shown to perturb interactions that are functionally relevant in the particular tissue affected by the specific disease. Sahni et al. further conclude that gain of interactions is a rare event in human disease, finding very little evidence for it.

The Road Ahead

Network biology in the past decade was focused on general networks per species, representing the interaction potential of every two proteins. It is becoming clear that these networks, while providing important insights, do not materialize in all conditions. Rather, different sub-networks are formed in different contexts depending on protein expression, structure, and more. In the future, when interactome measurements become as standard and inexpensive as genome sequencing, one can envision the construction of patient-specific networks that could dramatically improve our understanding of human disease and its treatment. With the accumulation of more individual-specific network data, statistical techniques that are currently limited to sequence data, such as association studies, could be generalized to the network world, ever refining our views of cells and organisms.

Acknowledgments

We thank Idan Hekselman for his help in preparing the figure. This study was supported by the Israel Science Foundation administered by the Israel Academy of Sciences and Humanities through grant number 860/13 to EYL. RS was supported by a Naomi Kadar Award.

References

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Barshir, R., Basha, O., Eluk, A., Smoly, I. Y., Lan, A., and Yeger-Lotem, E. (2013). The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Res.* 41, D841–D844. doi: 10.1093/nar/gks1198

Barshir, R., Shwartz, O., Smoly, I. Y., and Yeger-Lotem, E. (2014). Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput. Biol.* 10:e1003632. doi: 10.1371/journal.pcbi.1003632

Basha, O., Flom, D., Barshir, R., Smoly, I., Tirman, S., and Yeger-Lotem, E. (2015). MyProteinNet: build up-to-date protein interaction networks for organisms, tissues and user-defined contexts. *Nucleic Acids Res.* 43, W258–W263. doi: 10.1093/nar/gkv515

Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 5, 260. doi: 10.1038/msb.2009.17

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., et al. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell* 46, 871–883. doi: 10.1016/j.molcel.2012.05.039

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* 43, D470–D478. doi: 10.1093/nar/gku1204

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science* 307, 724–727. doi: 10.1126/science.1105103

Dezso, Z., Nikolsky, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., et al. (2008). A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.* 6:49. doi: 10.1186/1741-7007-6-49

Ellis, J. D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J. A., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell* 46, 884–892. doi: 10.1016/j.molcel.2012.05.037

Emig, D., and Albrecht, M. (2011). Tissue-specific proteins and functional implications. *J. Proteome Res.* 10, 1893–1903. doi: 10.1021/pr101132h

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Bio.* 3, 89. doi: 10.1038/msb4100134

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6, 807–828. doi: 10.1101/gr.6.9.807

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8, 565. doi: 10.1038/msb.2011.99

Jongeneel, C. V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C. D., Khrebtukova, I., et al. (2005). An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* 15, 1007–1014. doi: 10.1101/gr.4041005

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846. doi: 10.1093/nar/gkr1088

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., et al. (2014). A draft map of the human proteome. *Nature* 509, 575–581. doi: 10.1038/nature13302

Kiran, M., and Nagarajaram, H. A. (2013). Global versus local hubs in human protein-protein interaction network. *J. Proteome Res.* 12, 5436–5446. doi: 10.1021/pr4002788

Lage, K., Hansen, N. T., Karlberg, E. O., Eklund, A. C., Roque, F. S., Donahoe, P. K., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20870–20875. doi: 10.1073/pnas.0810772105

Lambert, J. P., Ivosev, G., Couzens, A. L., Larsen, B., Taipale, M., Lin, Z. Y., et al. (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat. Methods* 10, 1239–1245. doi: 10.1038/nmeth.2702

Li, M., Zhang, J., Liu, Q., Wang, J., and Wu, F. X. (2014). Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation. *BMC Med. Genomics* 7(Suppl. 2):S4. doi: 10.1186/1755-8794-7-S2-S4

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., et al. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40, D857–D861. doi: 10.1093/nar/gkr930

Lin, W. H., Liu, W. C., and Hwang, M. J. (2009). Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. *BMC Syst. Biol.* 3:32. doi: 10.1186/1752-0509-3-32

Liu, W., Wang, J., Wang, T., and Xie, H. (2014). Construction and analyses of human large-scale tissue specific networks. *PLoS ONE* 9:e115074. doi: 10.1371/journal.pone.0115074

Lopes, T. J., Schaefer, M., Shoemaker, J., Matsuoka, Y., Fontaine, J. F., Neumann, G., et al. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* 27, 2414–2421. doi: 10.1093/bioinformatics/btr414

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308–312. doi: 10.1038/nature02782

Magger, O., Waldman, Y. Y., Ruppin, E., and Sharan, R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.* 8:e1002690. doi: 10.1371/journal.pcbi.1002690

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Pontén, F., Gry, M., Fagerberg, L., Lundberg, E., Asplund, A., Berglund, L., et al. (2009). A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* 5, 337. doi: 10.1038/msb.2009.93

Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* 5:e1000598. doi: 10.1371/journal.pcbi.1000598

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178. doi: 10.1038/nature04209

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161, 647–660. doi: 10.1016/j.cell.2015.04.013

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global quantification of mammalian gene expression control. *Nature* 473, 337–342. doi: 10.1038/nature10098

Sinha, A., and Nagarajaram, H. A. (2014). Nodes occupying central positions in human tissue specific PPI networks are enriched with many splice variants. *Proteomics* 14, 2242–2248. doi: 10.1002/pmic.201400249

Song, J., Wang, Z., and Ewing, R. M. (2014). Integrated analysis of the Wnt responsive proteome in human cells reveals diverse and cell-type specific networks. *Mol. Biosyst.* 10, 45–53. doi: 10.1039/C3MB70417C

Souiai, O., Becker, E., Prieto, C., Benkahla, A., De las Rivas, J., and Brun, C. (2011). Functional integrative levels in the human interactome recapitulate organ organization. *PLoS ONE* 6:e22051. doi: 10.1371/journal.pone.0022051

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6062–6067. doi: 10.1073/pnas.0400782101

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419. doi: 10.1126/science.1260419

Vidal, M., Cusick, M. E., and Barabási, A. L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Waldman, Y. Y., Tuller, T., Shlomi, T., Sharan, R., and Ruppin, E. (2010). Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Res.* 38, 2964–2974. doi: 10.1093/nar/gkq009

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164. doi: 10.1038/nbt.2106

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F. M., Lee, H. R., et al. (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* 10:e1004819. doi: 10.1371/journal.pgen.1004819

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587. doi: 10.1038/nature13319

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305. doi: 10.1093/nar/30.1.303

Zhong, Q., Simonis, N., Li, Q. R., Charloteaux, B., Heuze, F., Klitgord, N., et al. (2009). Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* 5, 321. doi: 10.1038/msb.2009.80

Conflict of Interest Statement: 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.

frontiers in Genetics

OPINION
published: 15 September 2015
doi: 10.3389/fgene.2015.00289

Quo vadis[1] computational analysis of PPI data or why the future isn't here yet

Konstantinos A. Theofilatos [1], Spiros Likothanassis [1, 2] and Seferina Mavroudi [1, 2, 3]*

[1] InSyBio Ltd., London, UK, [2] Pattern Recognition Laboratory, Department of Computer Engineering and Informatics, University of Patras, Patras, Greece, [3] Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece, Patras, Greece

Keywords: protein–protein interactions, protein–protein interaction networks, protein clusters, protein functions, biomarker discovery, drug discovery

OPEN ACCESS

Edited by:
Spyros Petrakis,
Aristotle University of Thessaloniki, Greece

Reviewed by:
Jin Chen,
Michigan State University, USA
Miguel Andrade,
Johannes-Gutenberg University of Mainz, Germany

****Correspondence:***
Konstantinos A. Theofilatos,
k.theofilatos@insybio.com

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

***Received:** 02 June 2015*
***Accepted:** 31 August 2015*
***Published:** 15 September 2015*

Citation:
Theofilatos KA, Likothanassis S and Mavroudi S (2015) Quo vadis computational analysis of PPI data or why the future isn't here yet. Front. Genet. 6:289. doi: 10.3389/fgene.2015.00289

Introduction

Proteins have been proven to be among the most significant cellular molecules as they participate in most cellular functionalities. Researchers have deployed a variety of experimental methods for the identification of Protein–Protein Interactions (PPIs). The emergence of high-throughput experimental techniques for the prediction of PPIs, revealed a wide range of PPIs in many organisms. This information alongside with information from small scale experimental techniques has been stored in public available databases and repositories. It is well-known that experimental data include many false positive predictions and provide only low coverage on the full interactomes. This fact has led to the design and development of many computational methods for the prediction of PPIs (Theofilatos et al., 2011).

The experimental PPI data have been extensively used in many studies during the last decades and their availability gave a significant boost in training new algorithmic models for the prediction of PPIs and in the overall analysis of PPI data.

Despite the promising results of algorithmic solutions for PPIs' analysis which fostered molecular biology research, in our opinion the research on computational methods for analyzing PPI data has been recently stagnated. Using the online tool MLtrends introduced by Palidwor and Andrade-Navarro (2010) in a preliminary investigation we have observed that the publications related to the search term *"protein–protein interactions" AND analysis* (abstract and title were searched for this term) present a constant increase in absolute numbers. However, when applying normalization by dividing with the total number of annual publications, we observe a relatively stable percentage of publications related PPI analysis in the last decade. In contrast, systems biology publications present a big positive slope in the last decade even when normalized by the total number of annual publications. This diversification shows that even if the actual total number of publications related to PPIs analysis is increasing as the total number of scientific journals is increasing in the last decade, their total impact on the systems biology domain is decreasing. Additionally, only a few PPI based research works have been published lately with significant impact in clinical research and translational bioinformatics.

In this paper, first we summarize the developments on computational analysis of PPI data and second, we present our belief about the future of PPI data analysis emphasizing in presenting the constraints that have delayed the transition from the current methodologies to a holistic bioinformatics approach, for linking biological and clinical data. Specific solutions are also

[1] Latin phrase meaning ″Where are you going?

proposed for all these constraints in order to achieve the optimal exploitation of PPI bioinformatics' approaches.

State-of-the-art and Recent Advancements of the Computational Analysis of PPI Data

A wide range of computational methodologies and tools have been proposed during the last decade for the analysis of PPI data. These methods are emphasizing on algorithmic solutions for the problems of predicting and scoring PPIs, the construction of PPI networks, the prediction of protein complexes, and the functional annotation of proteins. The results of these solutions have been uploaded to public available databases and many tools have been supplied to the molecular research scientific community enabling the analysis of the PPIs from a single organism in a few days. State of the art computational methods for the prediction of PPI combine information from different sources and have presented adequate classification performance. Recent approaches (Zhang et al., 2012; Saha et al., 2014; Theofilatos et al., 2014) have attempted to overcome the bottlenecks in this PPIs prediction, namely the definition of negative datasets, the feature selection, the class imbalance, the tradeoff between classification performance and interpretability, missing features values and the calculation of a confidence score for every PPI. The advancements on the computational prediction and scoring of PPI enabled the construction of binary PPI networks with increased coverage on the full interactome. Many tools have been developed so far offering efficient and interactive visualization of large PPI networks (Smoot et al., 2011; Li et al., 2013; Tripathi et al., 2014). As a next step, significant results have been extracted from the analysis of PPI graphs. This analysis includes methods for predicting protein complexes (Chen et al., 2014) and the functional and topological characterization of proteins (Ma et al., 2014). Recent approaches (Nepusz et al., 2012; Hanna and Zaki, 2014; Theofilatos et al., 2015) have attempted to face with some of the remaining challenges for the prediction of protein complexes such as the prediction of overlapping clusters and the management of weighted PPI graphs thus increasing the accuracy having limited prior knowledge for known protein complexes. One of the ultimate goals of PPI data analysis is to functionally characterize proteins and their interactions. The main limitation of protein function prediction until now is that a combined framework to characterize the full proteome functionally of a single organism having a meaningful confidence score for every annotation does not exist. However, with the continued development of new algorithms and the improvement of experimental techniques we strongly believe that this will be achieved in the next few years.

Bottlenecks in Computational Analysis of PPI Data and Reasons for its Reduced Impact in pharmacy and Medicine

Computational methodologies for the analysis of PPIs undoubtedly contributed to the advancement of systems biology research. In our opinion, however we have reached the point where research on novel computational methods has stagnated and further advancements in systems biology research cannot be achieved solely through the development of more sophisticated algorithmic solutions. Recently, many researchers have suggested that advancements in the field of PPIs research will be facilitated by improved integration of clinical and molecular data, introducing new clinical phenotype data, such as the ones coming from integrated data using the technologies of smart sensors and personalized medicine, in a format manageable to computational approaches (Tiffin et al., 2009). The recent advances in other fields of molecular biology, such as the next generation sequencing data and their analysis, has enabled the transformation of traditional bioinformatics to translation bioinformatics. Despite the large availability of works describing specific combinations of datasets to develop tools suitable for disease genes prioritization, "our understanding of how to perform useful predictions using multiple data sources or across biological networks is still rudimentary" (Nabhan and Sarkar, 2014), and in particular, to our knowledge, only a few systematic studies focused on the exploitation of integrated network methods in medicine applications (Schaefer et al., 2012; Vinayagam et al., 2014).

For all these reasons, a new approach and more ambitious objectives should be set for the analysis of PPI data in order to overcome all these limitations, to meet clinical needs and cover the lost space in translation bioinformatics analysis which has been gained by genomics and transcriptomics analysis. The main challenges for PPI analysis according to our opinion, except from the already mentioned data integration task and the linkage of PPI data with clinical data, are the incorporation of environmental information in the PPI data analysis, the extended study of PPIs among different organisms, e.g., host–pathogen interactions, the three dimensional reconstruction of 2D PPI networks for better representation of protein structure, isoforms and spatial information, the design of new methods for biomarker discovery using PPI data and the development of new methods which will facilitate drug discovery using PPI data. In addition, as firstly proposed by Lopes et al. (2011) and adopted by Schaefer et al. (2012) and Furlong (2013), the availability of condition-specific interactomes that are more representative of the interactions of the proteins in a given tissue or under certain conditions will improve the significance of such analysis. This could be done by providing more realistic results, especially for the exploration of human diseases, where the network topology properties of proteins encoded by disease genes in interactomes should be reassessed with spatiotemporal resolution in healthy and disease states.

Traditional PPI analysis' approaches study physical and functional PPIs without taking into account environmental influences which may strongly affect a PPI or even the formation of a protein complex. Specifically, it is known that the post translational modifications of proteins, which play an important role in enabling them to interact with other proteins, are significantly affected by environmental changes. Moreover, most complex diseases are attributed to generalized disturbances in genetic and proteomic level in cooperation with environmental causes. In order to exploit PPI analysis in medicine application, it should be combined with environmental information and one way to achieve this is the integration of metabolomics data.

Another field of PPI analysis, which has not yet been thoroughly explored, is the study of interactions between proteins from various organisms with a striking example being the interactions of proteins from host and pathogen organisms which play a significant role in the viability of the affected cells. Until recently there existed a lack of large scale efforts to analyze host–pathogen interactions (Krishnadev and Srinivasan, 2008). However, these data are now available and a few methods, such as (Kleftogiannis et al., 2015), for their analysis have already been published presenting the potential of this field.

One of the most significant new ideas is the one proposed by Garbutt et al. (2014). This idea refers to the prevalent two dimensional format of the PPI graphs which is oversimplified and may lead to loss of information. To take advantage of the dynamic nature of PPI data, a new three dimensional representation should be stated integrating protein structure, conformation, isoforms and spatial information. Several recent research works take advantage of this idea to incorporate atomic-level protein structure information in PPI networks (Das et al., 2014) in order to examine the structural principles of disease mutations over a PPI network, or even to elucidate the genetic and molecular mechanisms of underlying human diseases (Wang et al., 2012).

One of the ultimate goals of PPI analysis should be the biomarkers' discovery. PPI networks contain significant information for the cellular mechanisms and functionalities which should be exploited to uncover disturbances in a network level. The traditional methods which attempt to uncover biomarkers from genetic variations or differences in the expression level have limited applicability as they export a large number of biomarkers without being able to locate the cause of the disturbance. When studying diseases in a network level, the variations are smaller and network based biomarkers are most likely to represent the cause of the disease. For this reason, more emphasis should be given to methods for comparing networks and locating biomarkers from the disturbed proteins, protein interactions, and protein complexes. Preliminary reports on methods for biomarkers' discovery through PPI networks comparison, revealed a new controversial issue (Wang et al., 2011). There are some arguing that hubs in PPI networks are most likely to be found as biomarkers and others arguing against. This issue should be further studied and clarified in order to uncover the network metrics which are adequate to be used for biomarker discovery using PPI graphs.

Another field of PPI analysis which should be further reinforced is drug discovery through PPI data. PPI data analysis has a variety of applications in drug discovery so far (Engin et al., 2014). An interesting idea to reduce the possible complications of a potential drug is to target proteins which are interacting with the target protein but have reduced significance in the overall network topology or even are leafs of the PPI graph. Even more such ideas are required to be implemented in a novel way to exploit PPI networks and their topological characteristics in the drug design process.

In the last 5 years significant initiatives, such as ELIXIR-Data for Life (Crosswell and Thornton, 2012) and Global Alliance for Data Sharing (Hayden, 2013), have attempted to promote biological data sharing, provide the adequate infrastructures and bring together molecular biologists, bioinformaticians and clinicians in order to translate life science research mainly to medicine and bioindustries. These initiatives should be even more re-enforced and promoted in order to integrate productively the knowledge and experience of so different fields toward the realization of personalized medicine. These efforts will be eased by the expected universal adoption of electronic medical records standardization and omics translation to clinical medicine (Issa et al., 2014). However, the full clinical potential of these initiatives will still remain unexplored until they are formed in a network perspective that place them within the systems medicine context. Protein–protein interaction analysis will by nature play a significant role in this network-perspective formation.

Conclusions

In this opinion article we have presented our belief about the future of PPI data analysis emphasizing in presenting the constraints that delayed the transition from the current methodologies to a holistic bioinformatics approach, for linking biological and clinical data. The main constraints that should be surpassed are the incorporation of environmental information, the host–pathogen PPI data analysis and the expansion of the traditional 2D representation of PPI networks with a more flexible and informative 3D one. These constraints are of equal importance and most of them should be surpassed in order to ease the exploitation of PPI analysis in clinical applications. Moreover, we have stated the most significant areas of clinical applications of PPI data analysis which are biomarkers and drug discovery, and we have proposed certain ideas for advancing PPI analysis in these fields. The next few years, a new boost of clinical data is expected through the new electronic health records and data coming from the developing technologies of smart sensors and personalized medicine (Groves et al., 2013; Yang et al., 2015) and the computational analysis of PPI data should be ready to exploit this boost.

References

Chen, B., Fan, W., Liu, J., and Wu, F. X. (2014). Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. *Brief. Bioinformatics* 15, 177–194. doi: 10.1093/bib/bbt039

Crosswell, L. C., and Thornton, J. M. (2012). ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.* 30, 241–242. doi: 10.1016/j.tibtech.2012.02.002

Das, J., Fragoza, R., Lee, H. R., Cordero, N. A., Guo, Y., Meyer, M. J., et al. (2014). Exploring mechanisms of human disease through structurally resolved protein interactome networks. *Mol. Biosyst.* 10, 9–17. doi: 10.1039/C3MB70225A

Engin, H. B., Gursoy, A., Nussinov, R., and Keskin, O. (2014). Network-based strategies can help mono-and poly-pharmacology drug discovery: a systems biology view. *Curr. Pharm. Des.* 20, 1201–1207. doi: 10.2174/13816128113199990066

Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends Genetics* 29, 150–159. doi: 10.1016/j.tig.2012.11.004

Garbutt, C. C., Bangalore, P. V., Kannar, P., and Mukhtar, M. S. (2014). Getting to the edge: protein dynamical networks as a new frontier in plant–microbe interactions. *Front. Plant Sci.* 5:312. doi: 10.3389/fpls.2014.00312

Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S. (2013). *The 'Bigdata' Revolution in Healthcare*. McKinsey Quarterly.

Hanna, E. M., and Zaki, N. (2014). Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. *BMC Bioinformatics* 15:204. doi: 10.1186/1471-2105-15-204

Hayden, E. C. (2013). Geneticists push for global data-sharing. *Nature* 498, 16–17. doi: 10.1038/498017a

Issa, N. T., Byers, S. W., and Dakshanamurthy, S. (2014). Big data: the next frontier for innovation in therapeutics and healthcare. *Expert Rev. Clin. Pharmacol.* 7, 293–298. doi: 10.1586/17512433.2014.905201

Kleftogiannis, D., Wong, L., Archer, J. A., and Kalnis, P. (2015). Hi-Jack: a novel computational framework for pathway-based inference of host-pathogen interactions. *Bioinformatics* 31, 2332–2339. doi: 10.1093/bioinformatics/btv138

Krishnadev, O., and Srinivasan, N. (2008). A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite. *In Silico Biol.* 8, 235–250.

Li, W., Kinch, L. N., and Grishin, N. V. (2013).Pclust: protein network visualization highlighting experimental data. *Bioinformatics* 29, 2647–2648. doi: 10.1093/bioinformatics/btt451

Lopes, T. J., Schaefer, M., Shoemaker, J., Matsuoka, Y., Neumann, G., Andrade-Navarro, M. A., et al. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* 27, 2414–2421. doi: 10.1093/bioinformatics/btr414

Ma, X., Chen, T., and Sun, F. (2014).Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. *Brief. Bioinformatics* 15, 685–698. doi: 10.1093/bib/bbt041

Nabhan, A. R., and Sarkar, I. N. (2014).Structural network analysis of biological networks for assessment of potential disease model organisms. *J. Biomed. Inform.* 47, 178–191. doi: 10.1016/j.jbi.2013.10.011

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9, 471–472. doi: 10.1038/nmeth.1938

Palidwor, G. A., and Andrade-Navarro, M. A. (2010). MLTrends: graphing MEDLINE term usage over time. *J. Biomed. Discov. Collab.* 5, 1–6. doi: 10.5210/disco.v5i0.2680

Saha, I., Zubek, J., Klingström, T., Forsberg, S., Wikander, J., Kierczak, M., et al. (2014). Ensemble learning prediction of protein–protein interactions using proteins functional annotations. *Mol. Biosyst.* 10, 820–830. doi: 10.1039/c3mb70486f

Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS ONE* 7:e31826. doi: 10.1371/journal.pone.0031826

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Theofilatos, K., Dimitrakopoulos, C., Likothanassis, S., Kleftogiannis, D., Moschopoulos, C., Alexakos, C., et al. (2014). The Human Interactome Knowledge Base (HINT-KB): an integrative human protein interaction database enriched with predicted protein–protein interaction scores using a novel hybrid technique. *Artif. Intell. Rev.* 42, 427–443. doi: 10.1007/s10462-013-9409-8

Theofilatos, K., Dimitrakopoulos, C., Tsakalidis, A., Likothanassis, S., Papadimitriou, S., and Mavroudi, S. (2011). Computational approaches for the prediction of protein-protein interactions: a survey. *Curr. Bioinform.* 6, 398–414. doi: 10.2174/157489311798072981

Theofilatos, K., Pavlopoulou, N., Papasavvas, C., Likothanassis, S., Dimitrakopoulos, C., Georgopoulos, E., et al. (2015). Predicting protein complexes from weighted protein–protein interaction graphs with a novel unsupervised methodology: evolutionary enhanced Markov clustering. *Artif. Intell. Med.* 63, 181–189. doi: 10.1016/j.artmed.2014.12.012

Tiffin, N., Andrade-Navarro, M. A., and Perez-Iratxeta, C. (2009). Linking genes to diseases: it's all in the data. *Genome Med.* 1:77. doi: 10.1186/gm77

Tripathi, S., Dehmer, M., and Emmert-Streib, F. (2014). NetBioV: an R package for visualizing large network data in biology and medicine. *Bioinformatics* 30, 2834–2836. doi: 10.1093/bioinformatics/btu384

Vinayagam, A., Zirin, J., Roesel, C., Hu, Y., Yilmazel, B., Samsonova, A. A., et al. (2014). Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nat. Methods* 11, 94–99. doi: 10.1038/nmeth.2733

Wang, X., Gulbahce, N., and Yu, H. (2011). Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* 10, 280–293. doi: 10.1093/bfgp/elr024

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164. doi: 10.1038/nbt.2106

Yang, J. J., Li, J., Mulder, J., Wang, Y., Chen, S., Wu, H., et al. (2015). Emerging information technologies for enhanced healthcare. *Comp. Industry* 69, 3–11. doi: 10.1016/j.compind.2015.01.012

Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., and Honig, B. (2012). PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res.* 41, D828–D833. doi: 10.1093/nar/gks1231

Conflict of Interest Statement: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read,
for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous
– yet also collaborative,
fair and constructive

FAST PUBLICATION
Average 85 days from
submission to publication
(across all journals)

COPYRIGHT TO AUTHORS
No limit to article
distribution and re-use

TRANSPARENT
Editors and reviewers
acknowledged by name
on published articles

SUPPORT
By our Swiss-based
editorial team

IMPACT METRICS
Advanced metrics
track your article's impact

GLOBAL SPREAD
5'100'000+ monthly
article views
and downloads

LOOP RESEARCH NETWORK
Our network
increases readership
for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on