A JOURNEY THROUGH 50 YEARS OF STRUCTURAL BIOINFORMATICS IN MEMORIAM OF CYRUS CHOTHIA

EDITED BY: Alfredo Iacoangeli, Paolo Marcatili, Sarah Teichmann and Charlotte Deane

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-705-4
DOI 10.3389/978-2-88974-705-4

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: frontiersin.org/about/contact

A JOURNEY THROUGH 50 YEARS OF STRUCTURAL BIOINFORMATICS IN MEMORIAM OF CYRUS CHOTHIA

Topic Editors:
Alfredo Iacoangeli, King's College London, United Kingdom
Paolo Marcatili, Technical University of Denmark, Denmark
Sarah Teichmann, Wellcome Sanger Institute (WSI), United Kingdom
Charlotte Deane, University of Oxford, United Kingdom

The cover image for this Research Topic was designed by Claire Marks.

Citation: Iacoangeli, A., Marcatili, P., Teichmann, S., Deane, C., eds. (2022). A Journey Through 50 Years of Structural Bioinformatics in Memoriam of Cyrus Chothia. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-705-4

Table of Contents

frontiers | Frontiers in Molecular Biosciences

EDITORIAL
published: 28 April 2022
doi: 10.3389/fmolb.2022.885318

Editorial: A Journey Through 50 Years of Structural Bioinformatics in Memoriam of Cyrus Chothia

Alfredo Iacoangeli[1,2,3]*, *Paolo Marcatili*[4], *Charlotte Deane*[5], *Arthur M. Lesk*[6], *Annalisa Pastore*[2,7] and *Sarah A. Teichmann*[8,9]

[1]Department of Biostatistics and Health Informatics, King's College London, London, United Kingdom, [2]Department of Basic and Clinical Neuroscience, King's College London, Maurice Wohl Clinical Neuroscience Institute, London, United Kingdom, [3]King's College London, National Institute for Health Research Biomedical Research Centre and Dementia Unit at South London and Maudsley NHS Foundation Trust King's College London, London, United Kingdom, [4]Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby, Denmark, [5]Department of Statistics, University of Oxford, Oxford, United Kingdom, [6]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, United States, [7]European Synchrotron Radiation Facility, Grenoble, France, [8]Wellcome Sanger Institute, Hinxton, United Kingdom, [9]Theory of Condensed Matter Group, Cavendish Laboratory/Department of Physics, University of Cambridge, Cambridge, United Kingdom

Keywords: computational biology, bioinforamtics, structural bioinformatic, structural Biology, theoretical biology

Editorial on the Research Topic

A Journey Through 50 Years of Structural Bioinformatics in Memoriam of Cyrus Chothia

Dr Cyrus Chothia FRS was a pioneer and one of the founding figures of theoretical and computational biology, nowadays commonly known as the field of bioinformatics (a term which Cyrus never quite got used to). To cite some of his numerous contributions to the field, the work of Cyrus and co-workers on the relationship between the divergence of sequence and divergence of structure in proteins supported the development of methods of homology modelling (Chothia and Lesk, 1986); his work on mechanisms of conformational change included one of the first characterizations of the structural differences between deoxy- and oxy-haemoglobin (Baldwin and Chothia, 1979; Lesk et al., 1985); his novel taxonomic approach to the study of the relationship between the sequence, structure, function, and evolution of proteins, led to the discovery of canonical structures of complementarity-determining regions (CDRs) of antibodies (Chothia and Lesk, 1987) and the creation of a first hierarchical classification of proteins into subfamilies, families and superfamilies based on structural, and functional similarities (Murzin et al., 1995). His seminal research had major impacts in the field during a long career of over 50 years, and remains a source of inspiration for new generations. We would like to honour his memory and pay tribute to his work with this Research Topic.

As mentioned above, Cyrus' work included diverse computational research areas including antibody canonical loop classification. Two articles in this collection focus on the computational study and design of antibodies. Antibodies can be used to target toxic molecules. In their work, Gilodi et al. coupled experimental methodologies with in silicio design and showed the potential of developing an antibody able to recognize the RNA binding regions of TDP-43. TDP-43 aggregates have been proposed as a potential cause of ALS and their sequestration has been proposed as a therapeutic strategy.

Although the variable domains of an antibody contain the complementarity-determining regions (CDRs) that shape and host the antigen binding site (ABS), the elbow angle and the relative interdomain orientations of the variable and constant domains also influence the shape of ABS (Lesk and Chothia., 1988). Therefore, understanding the link between their dynamics and antigen specificity is crucial for the

OPEN ACCESS

Edited and reviewed by:
Alfonso De Simone,
University of Naples Federico II, Italy

***Correspondence:**
Alfredo Iacoangeli
alfredo.iacoangeli@kcl.ac.uk

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *27 February 2022*
Accepted: *08 March 2022*
Published: *28 April 2022*

Citation:
Iacoangeli A, Marcatili P, Deane C, Lesk AM, Pastore A and Teichmann SA (2022) Editorial: A Journey Through 50 Years of Structural Bioinformatics in Memoriam of Cyrus Chothia. Front. Mol. Biosci. 9:885318. doi: 10.3389/fmolb.2022.885318

modelling and engineering of antibodies. Using molecular dynamics techniques, Fernández-Quintero et al. investigated this relationship and found that CDR loops reveal conformational transitions in the micro-to-millisecond timescale, while the interface and elbow angle dynamics occur on the nanosecond timescale.

The next three studies discuss and attempt to identify general features of protein domains, in the spirit of Cyrus' analyses of protein structures and sequences. Chen et al. used a novel approach based on the study of domain-mediated protein-protein interactions, rather than a traditional focus on individual domains, for the structural profiling of bacterial effectors. These are proteins injected by the bacteria into the host cells that are critical for their virulence and intracellular survival. Their approach led to novel quantitative insights into the structural basis of effectors that might aid the design of effective and selective inhibitors of their pathogenic mechanisms.

As studied by Cyrus particularly in later years of his career, interconnected functional, biophysical, and structural constraints drive the purifying selection leading to variable levels of conservations along protein sequences. In their work, Dubreuil and Levy discussed these constraints while emphasising relevant works of Cyrus. Subsequently, they focused their attention on the evolutionary rate of disordered regions and the role of cellular abundance in their sequence conservation. They found that disordered regions are equivalent to super-accessible surface residues, and they confirmed the strong divergence interdependency between surface and core residues and the weak evolutionary coupling of disordered and domain regions. Finally, they observed that protein abundance impacts the conservation of residues in core, surface and disordered regions with constraints of similar effect size.

In the spirit of Cyrus' global approach to analysing the protein Universe, Konagurthu et al. tried to answer the following question: 'What is the architectural "basis set" of the observed Universe of protein structures?' The authors used an information-theoretic inference method to identify automatically conserved sets of secondary structural elements within any given collection. By applying this method to the ASTRAL SCOP domains, they created an architectural dictionary of 1,493 substructures and used it to dissect the protein data bank (PDB). They made the entire dictionary, associated information and all the concept instances from the analysis of the PDB, publicly available on a webserver (http://lcb.infotech.monash.edu.au/prosodic).

Homology modelling is one of the most established approaches to protein structure prediction and a longstanding tool for Cyrus and co-workers on important structures such as the model of the T cell receptor based on antibody structures (Chothia et al., 1988). Homology models rely on the accurate identification of a suitable structural template based on the sequence of the target protein. Recently, deep learning has shown great potential to mine the coevolutionary information from multiple sequence alignments, leading a substantial improvement in the detection of distant homology. An amusing anecdote is that Cyrus had an antipathy towards the term "coevolution". He correctly pointed out that substitutions are always sequential rather than simultaneous. In a mini-review, Bhattacharya et al. presented the current advances of the protein homology detection field driven by the use of machine learning in Inter-Residue Interaction Map Threading.

Some classes of proteins present a low sequence identity among homologs limiting the use of sequence-based methods for their homology modelling. G protein-coupled receptors (GPCRs) represent one such example. However, GPCR sequences with similar patterns of hydrophobic residues are often structural homologs, even with low sequence identity. In their study, Jabeen et al. designed a method for homology modelling of GPCRs that exploits this biophysical characteristic, as well as other GPCR-specific features. Their method was validated with a number of published benchmarking datasets and a case study on an olfactory receptor is presented in the article. Furthermore, it was implemented in the form of a free tool called Bio-GATS (https://github.com/amara86/Bio-GATS).

Savojardo et al. investigated whether and to what extent single-amino acid pathogenic variants (PVs) could be associated with their solvent exposure. Solvent-Accessible Surface Area (SASA) is indeed a key characteristic of proteins in determining their folding and stability. Savojardo et al. mapped PVs onto a curated set of structures and determined that PVs occur more frequently in residues which are less likely to be accessible by the solvent, and that they are not evenly distributed among the different residue types. Using an in-house deep learning method for the sequence-based prediction of residue SASA, the authors confirmed these results in 12,494 human protein sequences for which no 3-D structure was available.

On a related topic, Di Renzo et al. investigated the interactions between amino acids on the protein surface and the solvent (water) to characterise their solvation properties. Although many descriptors of such properties exist, the local environment of each residue in the context of the protein is complex and often overlooked by existing methods. Based on molecular dynamics simulations, Di Renzo et al. developed a method to characterize the dynamic hydrogen bond network at the interface between protein and solvent, from which they derive the solvation properties of each amino in the protein environment.

Finally, in their review Bordin et al. presented some of Cyrus' accomplishments in the context of the history of protein structure classifications. The authors particularly focused on SCOP and CATH, two major protein structural classifications databases, and the evolutionary insights these two classifications have brought. They conclude their piece by discussing how the growing volume of data, and integration of protein sequences into these structural classifications, is helping to predict new functions in Metazoan organisms.

The articles in this collection cover very diverse areas of Structural Bioinformatics, reflecting the broad impact of Cyrus' research.

As a final remark, Cyrus never forgot that one's life is about the journey and not only the destination, and was ahead of his time in the open-minded way that he collaborated with scientists of all backgrounds and nationalities as well as across disciplines. The

editors and authors of this collection express their deepest gratitude to Cyrus for his enormous contribution to the field of Bioinformatics and for being a generous, supportive, and inspiring colleague and friend.

REFERENCES

Baldwin, J., and Chothia, C. (1979). Haemoglobin: the Structural Changes Related to Ligand Binding and its Allosteric Mechanism. *J. Mol. Biol.* 1292, 175–220. doi:10.1016/0022-2836(79)90277-8

Chothia, C., Boswell, D. R., and Lesk, A. M. (1988). The Outline Structure of the T-Cell Alpha Beta Receptor. *EMBO J.* 712, 3745–3755. doi:10.1002/j.1460-2075.1988.tb03258.x

Chothia, C., and Lesk, A. M. (1986). The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* 54, 823–826. doi:10.1002/j.1460-2075.1986.tb04288.x

Chothia, C., and Lesk, A. M. (1987). Canonical Structures for the Hypervariable Regions of Immunoglobulins. *J. Mol. Biol.* 196, 901–917. doi:10.1016/0022-2836(87)90412-8

Lesk, A. M., and Chothia., C. (1988). Elbow Motion in the Immunoglobulins Involves a Molecular ball-and-socket Joint. *Nature* 335, 6186188–6186190. doi:10.1038/335188a0

Lesk, A. M., Janin, J., Wodak, S., and Chothia, C. (1985). Haemoglobin: The Surface Buried between the $\alpha_1\beta_1$ and $\alpha_2\beta_2$ Dimers in the Deoxy and Oxy Structures. *J. Mol. Biol.* 1832, 267–270. doi:10.1016/0022-2836(85)90219-0

AUTHOR CONTRIBUTIONS

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

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* 2474, 536–540. doi:10.1016/s0022-2836(05)80134-2

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/fmolb.2020.609088

Surprisingly Fast Interface and Elbow Angle Dynamics of Antigen-Binding Fragments

*Monica L. Fernández-Quintero[1], Katharina B. Kroell[1], Martin C. Heiss[1], Johannes R. Loeffler[1], Patrick K. Quoika[1], Franz Waibl[1], Alexander Bujotzek[2], Ekkehard Moessner[3], Guy Georges[2] and Klaus R. Liedl[1]**

[1] Center for Molecular Biosciences Innsbruck (CMBI), Institute of General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innsbruck, Austria, [2] Roche Pharma Research and Early Development, Large Molecule Research, Roche Innovation Center Munich, Penzberg, Germany, [3] Roche Pharma Research and Early Development, Large Molecular Research, Roche Innovation Center Zurich, Schlieren, Switzerland

Fab consist of a heavy and light chain and can be subdivided into a variable (V_H and V_L) and a constant region (C_H1 and C_L). The variable region contains the complementarity-determining region (CDR), which is formed by six hypervariable loops, shaping the antigen binding site, the paratope. Apart from the CDR loops, both the elbow angle and the relative interdomain orientations of the V_H–V_L and the C_H1–C_L domains influence the shape of the paratope. Thus, characterization of the interface and elbow angle dynamics is essential to antigen specificity. We studied nine antigen-binding fragments (Fab) to investigate the influence of affinity maturation, antibody humanization, and different light-chain types on the interface and elbow angle dynamics. While the CDR loops reveal conformational transitions in the micro-to-millisecond timescale, both the interface and elbow angle dynamics occur on the low nanosecond timescale. Upon affinity maturation, we observe a substantial rigidification of the V_H and V_L interdomain and elbow-angle flexibility, reflected in a narrower and more distinct distribution. Antibody humanization describes the process of grafting non-human CDR loops onto a representative human framework. As the antibody framework changes upon humanization, we investigated if both the interface and the elbow angle distributions are changed or shifted. The results clearly showed a substantial shift in the relative V_H–V_L distributions upon antibody humanization, indicating that different frameworks favor distinct interface orientations. Additionally, the interface and elbow angle dynamics of five antibody fragments with different light-chain types are included, because of their strong differences in elbow angles. For these five examples, we clearly see a high variability and flexibility in both interface and elbow angle dynamics, highlighting the fact that Fab interface orientations and elbow angles interconvert between each other in the low nanosecond timescale. Understanding how the relative interdomain orientations and the elbow angle influence antigen specificity, affinity, and stability has broad implications in the field of antibody modeling and engineering.

Keywords: V_H–V_L interface dynamics, C_H1–C_L dynamics, elbow angle, antibody structure design, antibody structure prediction

OPEN ACCESS

Edited by:
Alfredo Iacoangeli, King's College London, United Kingdom

Reviewed by:
Francesco Di Palma, Italian Institute of Technology (IIT), Italy
David Douglas Boehr, Pennsylvania State University (PSU), United States

***Correspondence:**
Klaus R. Liedl
Klaus.Liedl@uibk.ac.at

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *22 September 2020*
Accepted: *21 October 2020*
Published: *24 November 2020*

Citation:
Fernández-Quintero ML, Kroell KB, Heiss MC, Loeffler JR, Quoika PK, Waibl F, Bujotzek A, Moessner E, Georges G and Liedl KR (2020) Surprisingly Fast Interface and Elbow Angle Dynamics of Antigen-Binding Fragments. Front. Mol. Biosci. 7:609088. doi: 10.3389/fmolb.2020.609088

INTRODUCTION

Antibodies are key players as therapeutic agents because of their ability to bind the majority of targets and their suitability for protein engineering (Chiu et al., 2019; Kaplon and Reichert, 2019; Kaplon et al., 2020). Description of the binding properties and characterization of the binding interface is essential for understanding the function of the antibody. The binding ability of antibodies is determined by the antigen-binding fragment (Fab), in particular the variable fragment region (Fv). The Fab consists of a heavy and a light chain and can be subdivided into two types of structurally distinct domains termed the variable (V_H, V_L) and constant domains (C_H1, C_L). The amino acid residues linking V_L to C_L and V_H to C_H1 are called switch residues (Stanfield et al., 2006). In the antigen-binding process, the most important region is the complementarity-determining region (CDR), which consists of six hypervariable loops that shape the antigen-binding site, the paratope (Chothia et al., 1989; Martin and Thornton, 1996; Al-Lazikani et al., 2000; North et al., 2011). Apart from the diversity in length, sequence, and structure of the CDR loops, the relative V_H–V_L interdomain orientation plays an important role in determining the shape of the antigen-binding site (Colman, 1988; Foote and Winter, 1992; Dunbar et al., 2013; Bujotzek et al., 2016). Various studies observed that mutations in the framework regions, in particular in the V_H–V_L interface, can strongly influence the antigen-binding affinity. Thus, mutations in the V_H–V_L interface result in structural changes of the binding site geometry, thereby modifying the relative V_H–V_L orientation (Riechmann et al., 1988; Foote and Winter, 1992; Braden et al., 1994; Banfield et al., 1997; Cauerhff et al., 2004). Numerous studies in literature focused on defining this relative interdomain orientation (Narayanan et al., 2009; Abhinandan, 2010; Almagro et al., 2011; Chailyan et al., 2011). The most commonly used and robust approach to characterize the V_H–V_L pose is ABangle (Dunbar et al., 2013; Teplyakov et al., 2014; Bujotzek et al., 2015, 2016). ABangle is a computational tool to characterize the relative orientations between the antibody variable domains (V_H and V_L) by using five angles and a distance and by comparing it to other known structures (Dunbar et al., 2013; Bujotzek et al., 2015, 2016).

The high variability in the V_H–V_L interdomain orientation is an additional feature of antibodies, which directly increases the size of the antibody repertoire (Chothia et al., 1985; Vargas-Madrazo and Paz-García, 2003; Bujotzek et al., 2016; Knapp et al., 2017; Fernández-Quintero et al., 2020c). This high variability in the V_H–V_L interdomain distribution has been reported for different IL-1β antibody fragments in agreement with the respective NMR ensembles (Fernández-Quintero et al., 2020c). By applying fast Fourier transformation to the interface angles, timescales of 0.1–10 GHz could be assigned to the fastest collective interdomain movements (Fernández-Quintero et al., 2020c). With the increasing number of available Fab X-ray structures, it was noted that these fragments also display a high variability in the elbow angle, which is defined as the angle between the pseudo-two-fold axes relating V_H to V_L and C_H1 to C_L (Sotriffer et al., 2000; Stanfield et al., 2006). The elbow angle has been shown to increase Fab flexibility and thereby to enhance the ability of the same antibody to recognize different antigens (Landolfi et al., 2001; Stanfield et al., 2006; Niederfellner et al., 2011). Additionally, it has been shown that mutations in the Fab elbow region can influence the interdomain conformational flexibility and paratope plasticity (Sotriffer et al., 2000; Henderson et al., 2019).

The C_H1–C_L heterodimer was found to be significantly more stable than the V_H–V_L heterodimer and has been shown to play an essential role for antibody assembly and secretion in the cell (Röthlisberger et al., 2005; Bönisch et al., 2017). Mutual stabilization occurred across both Fab interfaces, and a high degree of cooperation between V_H–V_L and C_H1–C_L could be observed. However, direct interactions among each domain (V_L, C_L/V_H, and C_H1) did not influence the stability of either domain (Röthlisberger et al., 2005).

In this study, we investigate the dynamics of both relative V_H–V_L, C_H1–C_L interface angles and the elbow angle and their respective dependencies on different light-chain types and shifts upon antibody humanization and affinity maturation. The aim is to structurally and mechanistically characterize these interdomain movements and elbow angle flexibilities and assign and estimate timescales to these domain motions.

MATERIALS AND METHODS

Investigated Antibody Fabs

The nine investigated publicly available Fab X-ray structures were chosen to have a representative set of antibodies covering various challenges in antibody engineering and design, as they differ in light-chain types (PDB accession codes: 1PLG, 1NL0, 1BBD, 7FAB, and 1DBA), upon humanization (PDB accession codes: 3L7E, 4PS4) and affinity maturation (1MLB, 2Q76).

Structure Preparation

All Fab X-ray structures were prepared in MOE (Molecular Operating Environment, Montreal, QC, Canada: 2019) (Molecular Operating Environment [MOE], 2020) using the Protonate 3D (Labute, 2009) tool. With the tleap tool of the Amber Tools20 package, the Fab structures were placed into cubic water boxes of TIP3P (Jorgensen et al., 1983) water molecules with a minimum wall distance to the protein of 10 Å (El Hage et al., 2018; Gapsys and de Groot, 2019). Parameters for all antibody simulations were derived from the AMBER force field 14SB (Maier et al., 2015). To neutralize the charges, we used uniform background charges (Darden et al., 1993; Salomon-Ferrer et al., 2013; Hub et al., 2014). Each system was carefully equilibrated using a multistep equilibration protocol (Wallnoefer et al., 2011).

All Fabs were simulated twice for 1 μs with different initial velocities, using molecular dynamics as implemented in the AMBER 20 (Case et al., 2020) simulation package. The results for the second 1 μs simulations are summarized in **Supplementary Table 1**, as the conclusions are the same as for the simulations presented in the manuscript. We removed the equilibration and relaxation phase in the respective simulations. Molecular dynamics simulations were performed

using pmemd.cuda (Salomon-Ferrer et al., 2013) in an NpT ensemble to be as close to the experimental conditions as possible and to obtain the correct density distributions of both protein and water. Bonds involving hydrogen atoms were restrained by applying the SHAKE algorithm (Miyamoto and Kollman, 1992), allowing a time step of 2.0 fs. Atmospheric pressure of the system was preserved by weak coupling to an external bath using the Berendsen algorithm (Berendsen et al., 1984). The Langevin thermostat was used to maintain the temperature at 300 K during simulations (Adelman and Doll, 1976).

Interface Angle Calculations

ABangle is a computational tool (Dunbar et al., 2013; Bujotzek et al., 2015, 2016; Fernández-Quintero et al., 2020c) used to characterize the relative orientations between the antibody variable domains (V_H and V_L) using six measurements (five angles and a distance). A plane is projected on each of the two variable domains. To define these planes, the first two components of a principal component analysis of 240 reference coordinates were used for V_H and V_L each. The reference coordinate set consists of Cα coordinates of eight conserved residues for 30 cluster representatives from a sequence clustering of the non-redundant ABangle antibody data set. The planes were then fit through those 240 coordinates, and consensus structures consisting of 35 structurally conserved Cα positions were created for the V_H and V_L domain. Between these two planes, a distance vector C is defined. The six measures are then two tilt angles between each plane (HC1, HC2, LC1, and LC2) and a torsion angle (HL) between the two planes along the distance vector C (dc). The ABangle script can calculate these measures for an arbitrary Fv region by aligning the consensus structures to the found core set positions and fitting the planes and distance vector from this alignment. This online available tool was combined with an in-house python script to reduce computational effort and to visualize our simulation data over time. The in-house script makes use of ANARCI (Dunbar and Deane, 2016) for fast local annotation of the Fv region and pytraj for rapid trajectory processing. The resulting fluctuations in the HL angle (**Supplementary Figure S3**) were further analyzed with a fast Fourier transformation (FFT) (Bergland, 1969) in python to characterize the frequency and timescale of these movements. We applied a frequency filter to assign timescales to movements.

To characterize the relative interdomain C_H1 and C_L orientations (**Supplementary Figure S3**), we defined a torsion angle between the center of mass (COM) of the loops of the C-terminal C_H1 domain, the COM of the C_H1, the COM of the C_L domain, and the COM of the loops of the C-terminal C_L domain.

As measure for the elbow angle (**Supplementary Figure S3**), we calculated a torsion angle between the COM of the variable domain, a defined vector between the COMs of the switch regions (hinge heavy and hinge light) and the COM of the constant region. **Figure 1** depicts all used interface and elbow angle definitions, showing the Fab domains as Lego model.

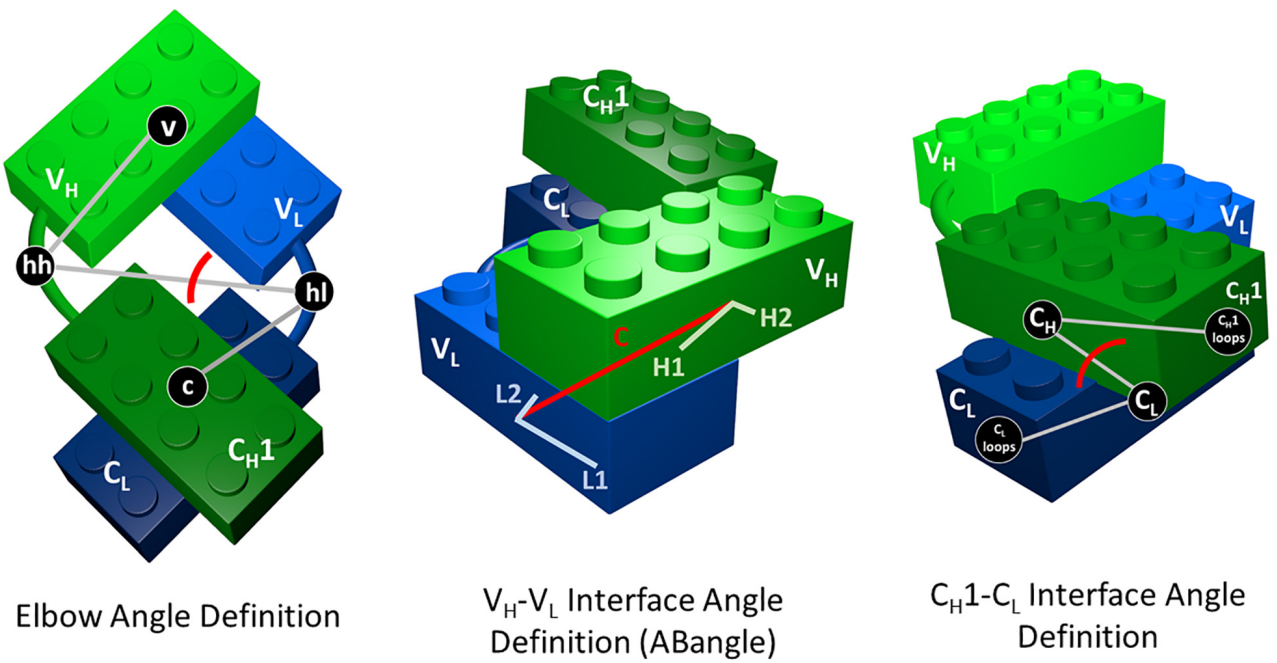


FIGURE 1 | Elbow angle, ABangle (HL angle), and C_H1–C_L interface angle definitions depicted as Lego models. The constant domains are illustrated in dark green (C_H1) and dark blue (C_L), while the variable domains are shown in light green (V_H) and light blue (V_L). To calculate the C_H1–C_L interface angle, we defined a torsion angle between the center of mass of the C-terminal loops of the C_H1 domain, the center of mass of C_H1, the center of mass of the C_L domain, and the center of mass of the C-terminal loops of the C_L domains. The center of masses of the C_H1 and C_L C-terminal loops are depicted in this figure, as C_L and C_H1 loops. For the elbow angle definition, all used centers of mass to define the torsion angle are illustrated. The centers of mass of the switch or hinge regions are abbreviated with hh (hinge heavy) and hl (hinge light).

RESULTS

The first five introduced antigen-binding fragments are part of a study, discussing the influence of different light-chain types (κ and λ light chains) on the resulting elbow angle distributions observed in X-ray structures (Stanfield et al., 2006). While the other six discussed Fabs contribute to a better understanding of the interface and elbow angle flexibilities upon antibody humanization and affinity maturation (Cauerhff et al., 2004; Fransson et al., 2010). By using MD simulations, we investigate the conformational variability of these interface and elbow angle distributions in solution and assign timescales to the dynamics of these movements, which have direct implications in the design of antibody paratopes and molecular recognition. The first investigated antibody is the 10C12 antibody (PDB accession code: 1NL0), inhibiting the human Factor IX calcium-stabilized N-terminal gamma-carboxyglutamic acid-rich (Gla) domain, which is a membrane-anchoring domain found on vitamin K-dependent blood coagulation and regulatory proteins. The 10C12 antibody is a conformation-specific anti-Factor IX antibody to interfere with the Factor IX-membrane interaction (Huang et al., 2004). Same as the 10C12 antibody, the highly resolved IgG1 Fab structure with the PDB accession code 7FAB also contains a λ light chain. The biggest difference between the two Fab structures is the elbow angle orientation.

Figure 2A illustrates the respective distributions and the results of the fast Fourier transformation (FFT) of the two λ light-chain antibodies for both interface angles (V_H–V_L and C_H1–C_L) and the elbow angle. The 10C12 antibody is colored in blue, while the IgG1 7FAB antibody is colored yellow. The fast Fourier transformation shows that all angles of both the 10C12 and IgG1 7FAB antibodies have high variations and allows to assign timescales of 0.1–10 GHz to the fastest collective angle movements. The highest flexibility and variability can be observed for the elbow angle, which fluctuates about ±15° in less than 10 ns, while both interface angles fluctuate around ±5° in less than 1 ns. Especially interesting is that these fast fluctuations in the low nanosecond timescale are substantially faster compared to conformational rearrangements in the antibody paratope, which is in line with previous studies (Fernández-Quintero et al., 2019a,c, 2020a,b). Additionally, also from the histograms (**Figure 2B**) it can be seen that the elbow angle has the highest variability, compared to the interface angle distributions. The starting X-ray structures of the respective antibodies are plotted into the histograms and color-coded, respectively. The third antibody investigated is the highly specific anti-progesterone antibody DB3 (PDB accession code: 1DBA) which can bind progesterone with nanomolar affinity. The DB3 antibody (containing a κ light chain) binds progesterone by forming a hydrophobic pocket by interactions between the three complementarity determining regions L1, H2, and H3 (Arevalo et al., 1993). Another example for a κ light-chain antibody is the IgG2 κ murine monoclonal antibody with high specificity for α-(2→8)-linked sialic acid polymers (PDB accession code 1PLG) (Evans et al., 1995). The fifth studied antigen-binding fragment (IgG2, κ light chain) 8F5, which is obtained by immunization with the native HRV2, neutralizes human rhinovirus serotype 2 and cross-reacts with peptides of the viral capsid protein VP2

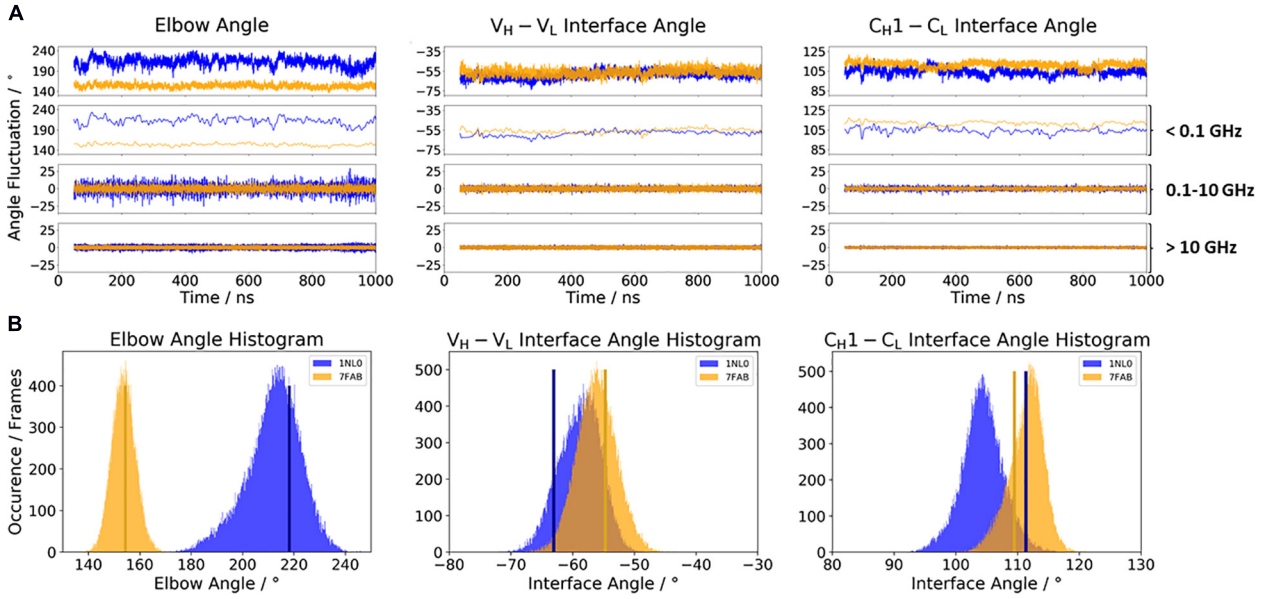


FIGURE 2 | (A) Comparison of the two considered λ light-chain antibodies 10C12 (PDB: 1NL0) and IgG1 Fab (PDB: 7FAB). The biggest difference between these two antibodies is the elbow angle. The distributions of the 10C12 antibody are depicted in blue, while the IgG1 antibody angle distributions are colored yellow. Additionally, the FFT of the respective distributions are displayed, showing the angle variations occurring faster than 1 ns, between 0.1 and 10 ns, and slower than 10 ns. **(B)** Histograms of the respective interface and elbow angle distributions, including the respective X-ray structures of both λ light chain antibodies, which were used as starting structures for molecular dynamics simulations.

(PDB accession code 1BBD) (Tormo et al., 1992). All three κ light-chain antibodies were simulated for two times 1 μs, and the results are depicted in **Figure 3**. In line with the results of the λ light-chain Fabs, the FFT in **Figure 3A** shows that the variability of the interface and elbow angles can be captured in the low nanosecond timescale. The histograms in **Figure 3B** clearly show that compared to **Figure 2**, especially in the elbow angle and the C_H1–C_L interface histograms, the distributions have much more overlay and are also narrower, indicating less variability and diversity in these angles when considering κ light-chain antibodies.

To investigate the effect of antibody humanization, we chose the humanization of a mouse anti-human IL-13 antibody (PDB accession codes: 3L7E and 4PS4) (Fransson et al., 2010; Teplyakov et al., 2011). The antibodies are humanized by the human-framework adaptation method (HFA), which comprises a selection (human framework selection), and an optimization (specificity-determining residue optimization) step. IL-13 is an important member of the growth-hormone-like cytokine family and is involved in the development of asthma (Grünig et al., 2012). **Figure 4** shows the comparison of the c836 antibody with the humanized Specificity Determining Residue Optimization (SDRO) optimized m1295 Fab to investigate if the relative interdomain orientations and the elbow angle distributions are shifted upon antibody humanization. While the relative interdomain V_H–V_L angle distributions are slightly shifted, the C_H1–C_L interface angle distribution for the m1295 variant completely overlaps with the c836 Fab and is much narrower, as a result of the specificity optimization process (**Figure 4A**). The elbow angle distribution for the chimeric c836 Fab is shaped bimodally, while m1295 has only one dominant elbow angle minimum in solution (**Figure 4B**). Again, the variability of the interface and elbow angle movements can be captured, as their fluctuations occur in the 0.1–10 GHz timescale.

Another unique ability of antibodies is to evolve in response to antigens and undergo cycles of mutation and selection leading to an enhanced affinity and specificity (Wabl and Steinberg, 1996; Acierno et al., 2007; Mishra and Mariuzza, 2018). To understand and characterize the underlying biophysical mechanism of affinity maturation, we investigated the maturation of an anti-chicken egg-white lysozyme antibody D44.1 (PDB accession codes 1MLB and 2Q76) (Braden et al., 1994; Cauerhff et al., 2004). Both D44.1 and the matured F10.6.6 Fab are murine monoclonal antibodies, which are related in sequence and structure as they origin from the same gene rearrangement. The affinity matured F10.6.6 antibody (K_A = 1.02*10^{10} M^{-1}) was reported to have a 700-times higher-affinity constant compared to D44.1 (K_A = 1.44*10^7 M^{-1}), due to a higher surface complementarity to the antigen (Acierno et al., 2007). The D44.1 Fab differs from the affinity matured variant F10.6.6 in twenty mutations, seven of them located in the CDR loops, while the other mutations can be found in both the V_H–V_L and C_H1–C_L interface. As the majority of mutations occur in the framework, already on the structural level a stabilization of the V_H–V_L interface has been reported (Braden et al., 1994; Cauerhff et al., 2004). **Figure 5** shows the angle distributions of the D44.1 antibody compared to the further matured F10.6.6 antibody. Upon affinity maturation, we observe a rigidification in the V_H–V_L angle and elbow angle distributions

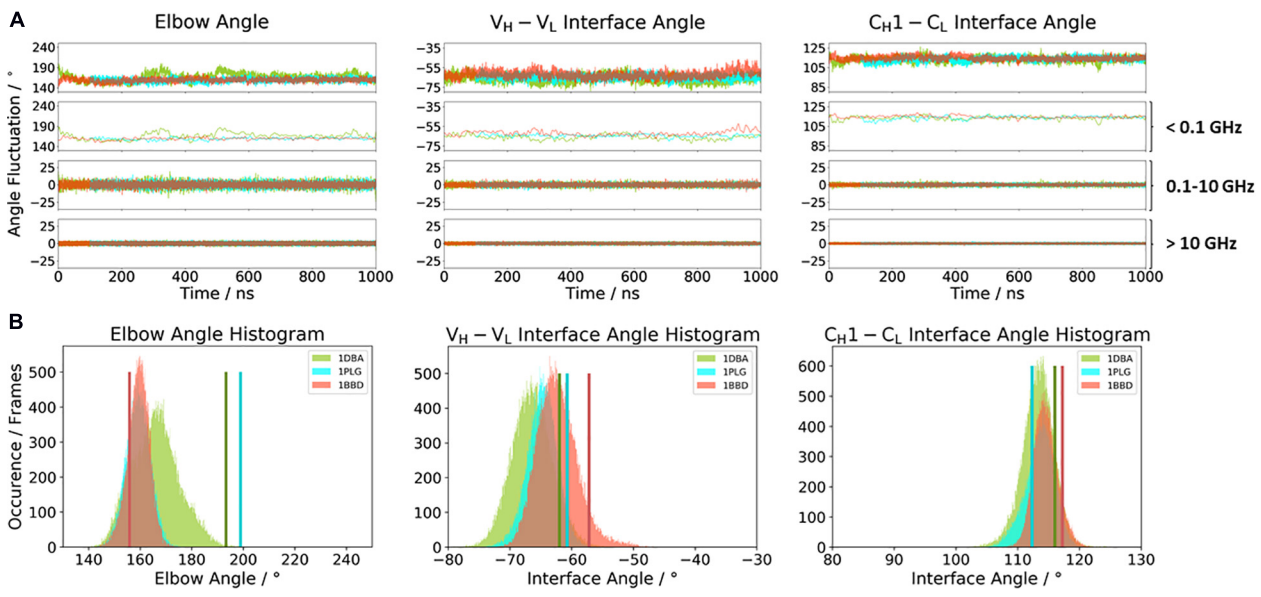


FIGURE 3 | (A) Comparison of the three κ light chain antibodies, which also show a big spread in the elbow angle in the crystal structures. The specific anti-progesterone antibody DB3 (PDB: 1DBA) is colored in green. The 8F5 antibody (PDB: 1BBD) is colored in red, and the X-ray structure of the murine antibody which has a high specificity toward α-(2→8)-linked sialic acid polymers (PDB: 1PLG) is colored light blue. Additionally, the FFT of the respective distributions are displayed, showing the angle variations occurring faster than 1 ns, between 0.1 and 10 ns, and slower than 10 ns. **(B)** Histograms of the respective interface and elbow angle distributions, including the respective X-ray structures of all three κ light chain antibodies, which were used as starting structures for molecular dynamics simulations.

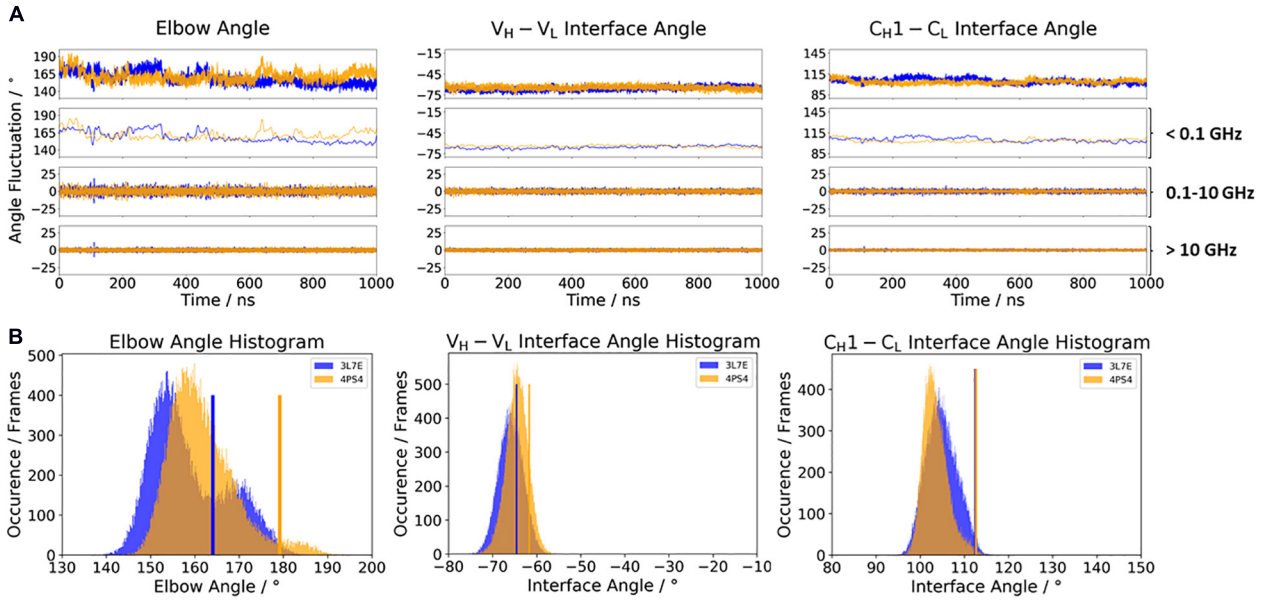


FIGURE 4 | (A) Influence of different framework origins on the relative interface and elbow angle distributions. The chimeric c836 antibody (PDB: 3L7E) is illustrated in blue, while the optimized humanization variant m1295 (PDB: 4PS4) is depicted in orange. The calculations of the FFT of the respective distributions are displayed, showing the angle variations occurring faster than 1 ns, between 0.1 and 10 ns, and slower than 10 ns. **(B)** Histograms of the respective interface and elbow angle distributions, including the respective X-ray structures of the chimeric and the humanized Fabs, which were used as starting structures for molecular dynamics simulations.

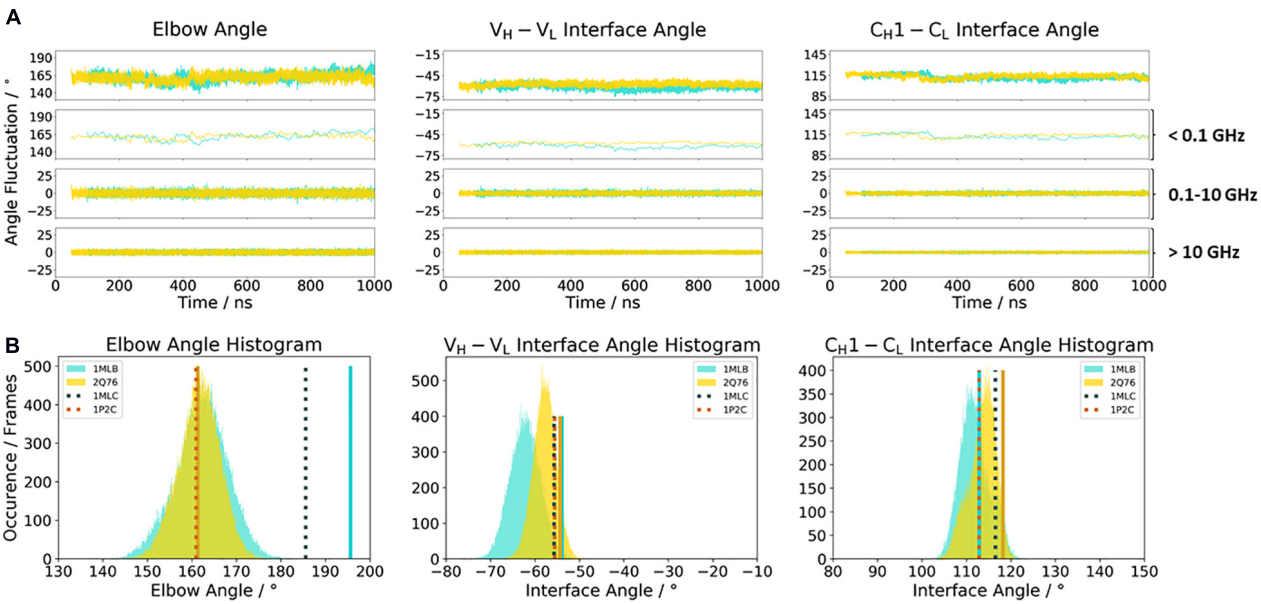


FIGURE 5 | (A) Influence of affinity maturation on the relative interface and elbow angle distributions. The D44.1 Fab (colored in cyan) reveals a broader interface and elbow angle distribution compared to the matured F10.6.6 antibody. The calculations of the FFT of the respective distributions are displayed, showing the angle variations occurring faster than 1 ns, between 0.1 and 10 ns, and slower than 10 ns. **(B)** Histograms of the respective interface and elbow angle distributions, including the respective X-ray structures of the affinity maturation pair D44.1 and F10.6.6, which were used as starting structures for molecular dynamics simulations.

(**Figure 5A**). This rigidification can also be confirmed by the narrower histograms of the matured F10.6.6 Fab illustrated in **Figure 5B**. In agreement with previous results, the FFT of both the D44.1 and F10.6.6 antibodies shows that also in this example the dynamics and flexibility of the interface and elbow angle distributions occur in the low nanosecond timescale. We used the X-ray structures crystallized without antigen as starting structure for the simulations to identify whether the binding competent

relative interdomain and elbow angle orientations are preexisting without the presence of the antigen. We clearly see that for the D44.1 antibody the relative interdomain orientation of the crystal structure binding to the antigen is present and more favorable in solution compared to the X-ray structure without the antigen. The resulting elbow angle distribution (**Figure 5B**) in solution shows that none of the two available crystal structures of the D44.1 antibody is actually favored in solution. Upon maturation, the relative interdomain orientations, especially the V_H–V_L orientation, in the X-ray structures do not change anymore upon binding, which is in line with the observed rigidification already on the X-ray structural level. The fact that we sample all binding competent V_H–V_L interface orientations supports the idea of a preexisting conformational ensemble out of which the binding competent state is selected and therefore follows the paradigm of conformational selection (Ma et al., 1999; Tsai et al., 1999; Fernández-Quintero et al., 2019a,b, 2020e).

DISCUSSION

In this present study, we characterize and quantify the relative interdomain and elbow angle orientations between antibodies bearing κ or λ light chains and between antibodies before and after humanization, upon affinity maturation. By using FFT, we were able to assign timescales to these fast interface and elbow angle movements in the low nanosecond timescale, which has direct implications in the field of antibody structure engineering and design.

Various studies already investigated the influence of different light chains (κ or λ light chains) on phenotypic differences, e.g., conformational flexibility, half-life, and propensity to alter antibody specificity (Montaño and Morrison, 2002; Wardemann et al., 2004; Stanfield et al., 2006; Townsend et al., 2016). Thus, the differences in κ and λ light chains result in distinct binding specificities. In line with previous observations, we observe that κ and λ light chains differ in their conformational flexibility. While the distributions in interface and elbow angles of the κ light-chain antibodies—independently of their starting geometries—overlap with each other and result in similar favorable orientations in solution, the Fabs consisting of a λ light chain reveal shifts and a higher diversity in possible elbow angles and interface orientations (**Figures 2**, **3**). We can clearly see from the FFT that the fast interface and elbow angle movements take place in the low nanosecond timescale (0.1–10 GHz) independent of the light chain (**Figures 2A**, 3A). We particularly chose the antibodies to have the biggest spread in the elbow angle orientations, ranging from 127° to 220° (**Supplementary Figure S1**). The 10C12 antibody (**Figure 2A—blue**) shows overall much more variability in all interface and elbow angles in the 0.1–10 GHz timescale, compared to the IgG1 7FAB antibody.

The free energy surfaces of the interface and elbow angle movements are shaped parabolically. Thus, if the fast movements of the interface and elbow angle are approximated by a harmonic potential, the force constants by fitting the free energy curves to quadratic functions and calculated the characteristic frequencies of the domain movements by using classical mechanics. As observed by the FFT, the majority of the interdomain and elbow angle dynamics occurs in the low nanosecond timescale (**Figure 6** and **Supplementary Figure S4**). **Figure 6** illustrates the respective free energy surface with the fitted quadratic functions. The fluctuations of these interdomain and elbow angles occur in the 0.1–10 GHz timescale and interconvert between each other in the 0.1–10 GHz timescale. The fact that these interface and elbow angles fluctuate ±5°/±10° within this single minimum in solution introduces a new view on these interfaces which directly influences the design and structure prediction of antibodies. Compared to the fast interdomain and elbow angle dynamics, the loop rearrangements occur in the high micro-to-millisecond timescale. Therefore, changes in the CDR loop conformations might be responsible for the dynamics slower than 10 ns. Thus, also conformational changes of the paratope directly influence the relative interdomain orientations and the elbow angle (Sotriffer et al., 1998; Sotriffer et al., 2000; Fernández-Quintero et al., 2020c,f).

In the context of antibody humanization (Zhang et al., 2013; Margreitter et al., 2016), apart from the CDR loop length and sequence, the relative V_H–V_L interdomain orientation has already been discussed to directly influence antigen binding (Bujotzek et al., 2016). Modulation of the V_H–V_L orientation diversifies antibody paratopes and thereby allows to accommodate diverse antigenic shapes that antibodies are confronted with (Teplyakov et al., 2011; Bujotzek et al., 2016). **Figure 4** shows the humanization of a mouse anti-human IL-13 antibody, which after the humanization and SDRO process showed a higher specificity compared to the murine (Fransson et al., 2010). This step-by-step antibody humanization has already been shown to result in a reduced conformational diversity, reflected by a substantial decrease in conformational space (Fernández-Quintero et al., 2020a). Our results are perfectly in line with these observations, as the C_H1–C_L interface angle and the elbow angle rigidify upon humanization. Additionally, we were able to identify a small

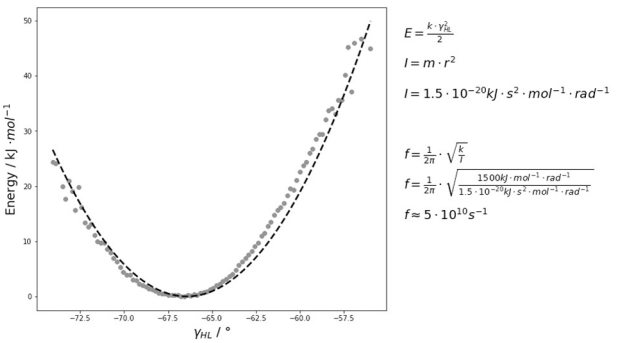


FIGURE 6 | An exemplary free energy surface of the DB3 antibody with the fitted quadratic function is illustrated and shows that these interdomain and elbow angle fluctuations interconvert with each other in the 0.1–10 GHz timescale. We estimated the force constants k and included the respective equations used for the frequency f calculations. The variable I represents the moment of inertia which was used to calculate the frequencies.

shift in the V_H–V_L interface distribution in solution for the m1295, which might be more favorable and contribute to better recognition and binding of the antigen.

Elucidating the affinity maturation process has been the focus of numerous studies (Cauerhff et al., 2004; Cho et al., 2005; Acierno et al., 2007; Li et al., 2010; Wong et al., 2011; Adhikary et al., 2015; Jeliazkov et al., 2018; Mishra and Mariuzza, 2018; Fernández-Quintero et al., 2019b; Shehata et al., 2019; Chan et al., 2020). Upon affinity maturation (**Figure 5**), we observe for the matured F10.6.6 antibody in both V_H–V_L interface and elbow angle histograms (**Figure 5B**) a narrower distribution, compared to the broader surface of the D44.1 Fab (**Supplementary Table S1**). A structural ensemble for both antibodies before and after affinity maturation is illustrated in **Supplementary Figure S2**, and also the rigidification upon affinity maturation is reflected in a lower number of clusters. Even though rigidification might only be one of the various consequences of affinity maturation, it still represents a fundamental mechanism resulting in an increase in specificity (Thorpe and Brooks, 2007; Li et al., 2015; Di Palma and Tramontano, 2017). Therefore, understanding the interface and elbow angle flexibility and dynamics upon affinity maturation is a prerequisite for all other affinity increasing changes, e.g., improved interfacial interactions, increased buried surface area, and improved shaped complementarity (Fernández-Quintero et al., 2020d,e). This observed rigidification, not only in the CDR loops but also in the V_H–V_L and elbow angle dynamics, clearly confirms the role of the interdomain dynamics in tailoring antibody specificity. All binding competent interface and elbow angle orientations preexist in solution, without the presence of the antigen. Thus, the relative interdomain and elbow angles clearly follow the concept of conformational selection (Ma et al., 1999).

CONCLUSION

For all investigated antibodies, we observe that changes in the sequences (e.g., different light-chain types, humanization, and affinity maturation) can influence and shift the interface and elbow angle distributions. Our results show that antibodies with a λ light chain do not only have broader X-ray angle distributions but also have higher variations in their relative interface angle distributions, especially in the C_H1 and C_L distributions. Upon humanization of a mouse anti-human IL-13 antibody, we observe small shifts in the V_H–V_L distributions and a rigidification in C_H1 and C_L and elbow angle distributions. In line with the rigidification as a consequence of the specificity optimization process, we also observe a rigidification in the V_H–V_L and elbow angle distributions upon affinity maturation. The rigidification upon affinity maturation might only be one of various consequences; however, understanding the flexibilities of the antibody interfaces is prerequisite for all other specificity-increasing changes. Both Fab interfaces and the elbow angle show movements occurring in the 0.1–10 GHz timescale (fluctuations around ±5°/±10°, respectively), which directly influence the binding site geometry. Thus, the understanding of these fast dynamics has broad implications in the field of antibody structure prediction and design.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

The manuscript was discussed and written through contributions of all authors. All authors have given approval to the final version of the manuscript.

FUNDING

This work was supported by the Austrian Science Fund (P30565, P30737, P30402, and DOC 30). Furthermore this project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 764958.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.609088/full#supplementary-material

REFERENCES

Abhinandan, K. R. (2010). ACR Martin. Analysis and prediction of VH/VL packing in antibodies. *Protein Eng. Des. Select.* 23, 689–697. doi: 10.1093/protein/gzq043

Acierno, J. P., Braden, B. C., Klinke, S., Goldbaum, F. A., and Cauerhff, A. (2007). Affinity maturation increases the stability and plasticity of the Fv domain of anti-protein antibodies. *J. Mol. Biol.* 374, 130–146. doi: 10.1016/j.jmb.2007.09.005

Adelman, S. A., and Doll, J. D. (1976). Generalized Langevin equation approach for atom/solid-surface scattering: general formulation for classical scattering off harmonic solids. *J. Chem. Phys.* 64, 2375–2388. doi: 10.1063/1.432526

Adhikary, R., Yu, W., Oda, M., Walker, R. C., Chen, T., and Stanfield, R. L. (2015). Adaptive mutations alter antibody structure and dynamics during affinity maturation. *Biochemistry* 54, 2085–2093. doi: 10.1021/bi501417q

Al-Lazikani, B., Lesk, A. M., and Chothia, C. (2000). Canonical structures for the hypervariable regions of T cell αβ receptors. *J. Mol. Biol.* 295, 979–995. doi: 10.1006/jmbi.1999.3358

Almagro, J. C., Beavers, M. P., Hernandez-Guzman, F., Maier, J., Shaulsky, J., Butenhof, K., et al. (2011). Antibody modeling assessment. *Proteins Struct. Funct. Bioinform.* 79, 3050–3066. doi: 10.1002/prot.23130

Arevalo, J. H., Stura, E. A., Taussig, M. J., and Wilson, I. A. (1993). Three-dimensional structure of an anti-steroid fab' and progesterone-fab' complex. *J. Mol. Biol.* 231, 103–118. doi: 10.1006/jmbi.1993.1260

Banfield, M. J., King, D. J., Mountain, A., and Brady, R. L. (1997). VL:VH domain rotations in engineered antibodies: crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. *Proteins Struct. Funct. Bioinform.* 29, 161–171. doi: 10.1002/(sici)1097-0134(199710)29:2<161::aid-prot4>3.0.co;2-g

Berendsen, H., van Postma, J. P. M., van Gunsteren, W., DiNola, A., and Haak, J. R. (1984). Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684. doi: 10.1063/1.448118

Bergland, G. D. (1969). A guided tour of the fast Fourier transform. *IEEE Spectrum* 6, 41–52. doi: 10.1109/mspec.1969.5213896

Bönisch, M., Sellmann, C., Maresch, D., Halbig, C., Becker, S., Toleikis, L., et al. (2017). Novel CH1:CL interfaces that enhance correct light chain pairing in heterodimeric bispecific antibodies. *Protein Eng. Des. Select.* 30, 685–696. doi: 10.1093/protein/gzx044

Braden, B. C., Ouchon, H. S., Eiselé, J.-L., Bentley, G. A., Bhat, T. N., and Navaza, J. (1994). Three-dimensional structures of the free and the antigen-complexed Fab from monoclonal anti-lysozyme antibody D44.1. *J. Mol. Biol.* 243, 767–781. doi: 10.1016/0022-2836(94)90046-9

Bujotzek, A., Dunbar, J., Lipsmeier, F., Schäfer, W., Antes, I., Deane, C. M., et al. (2015). Prediction of VH–VL domain orientation for antibody variable domain modeling. *Proteins Struct. Funct. Bioinform.* 83, 681–695. doi: 10.1002/prot.24756

Bujotzek, A., Lipsmeier, F., Harris, S. F., Benz, J., Kuglstatter, A., and Georges, G. (2016). VH-VL orientation prediction for antibody humanization candidate selection: a case study. *mAbs* 8, 288–305. doi: 10.1080/19420862.2015.1117720

Case, D. A., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., and Cheatham, T. E. (2020). *AMBER 2020*. San Francisco, CA: University of California.

Cauerhff, A., Goldbaum, F. A., and Braden, B. (2004). Structural mechanism for affinity maturation of an anti-lysozyme antibody. *Proc. Natl. Acad. Sci. U.S.A.* 101:3539. doi: 10.1073/pnas.0400060101

Chailyan, A., Marcatili, P., and Tramontano, A. (2011). The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS J.* 278, 2858–2866. doi: 10.1111/j.1742-4658.2011.08207.x

Chan, D. T. Y., Jenkinson, L., Haynes, S. W., Austin, M., Diamandakis, A., and Burschowsky, D. (2020). Extensive sequence and structural evolution of Arginase 2 inhibitory antibodies enabled by an unbiased approach to affinity maturation. *Proc. Natl. Acad. Sci. U.S.A.* 117:16949. doi: 10.1073/pnas.1919565117

Chiu, M. L., Goulet, D. R., Teplyakov, A., and Gilliland, G. L. (2019). Antibody structure and function: the basis for engineering therapeutics. *Antibodies* 8:55. doi: 10.3390/antib8040055

Cho, S., Swaminathan, C. P., Yang, J., Kerzic, M. C., Guan, R., and Kieke, M. C. (2005). Structural basis of affinity maturation and intramolecular cooperativity in a protein-protein interaction. *Structure* 13, 1775–1787. doi: 10.1016/j.str.2005.08.015

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., and Air, G. (1989). Conformations of immunoglobulin hypervariable regions. *Nature* 342, 877–883. doi: 10.1038/342877a0

Chothia, C., Novotný, J., Bruccoleri, R., and Karplus, M. (1985). Domain association in immunoglobulin molecules: the packing of variable domains. *J. Mol. Biol.* 186, 651–663.

Colman, P. M. (1988). "Structure of antibody-antigen complexes: implications for immune recognition," in *Advances in Immunology*, ed. F. J. Dixon (Cambridge, MA: Academic Press).

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092.

Di Palma, F., and Tramontano, A. (2017). Dynamics behind affinity maturation of an anti-HCMV antibody family influencing antigen binding. *FEBS Lett.* 591, 2936–2950. doi: 10.1002/1873-3468.12774

Dunbar, J., and Deane, C. M. (2016). ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* 32, 298–300.

Dunbar, J., Fuchs, A., Shi, J., and Deane, C. M. (2013). ABangle: characterising the VH–VL orientation in antibodies. *Protein Eng. Des. Select.* 26, 611–620. doi: 10.1093/protein/gzt020

El Hage, K., Hédin, F., Gupta, P. K., Meuwly, M., and Karplus, M. (2018). Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size. *eLife* 7:e35560.

Evans, S. V., Sigurskjold, B. W., Jennings, H. J., Brisson, J.-R., To, R., and Altman, E. (1995). Evidence for the extended helical nature of polysaccharide epitopes. The 2.8 .ANG. Resolution structure and thermodynamics of ligand binding of an antigen binding fragment specific for .alpha.-(2.fwdarw.8)-Poly(sialic acid). *Biochemistry* 34, 6737–6744. doi: 10.1021/bi00020a019

Fernández-Quintero, M. L., Heiss, M. C., and Liedl, K. R. (2020a). Antibody humanization—the Influence of the antibody framework on the CDR-H3 loop ensemble in solution. *Protein Eng. Des. Select.* 32, 411–422. doi: 10.1093/protein/gzaa004

Fernández-Quintero, M. L., Heiss, M. C., Pomarici, N. D., Math, B. A., and Liedl, K. R. (2020b). Antibody CDR loops as ensembles in solution vs. canonical clusters from X-ray structures. *mAbs* 12:1744328. doi: 10.1080/19420862.2020.1744328

Fernández-Quintero, M. L., Hoerschinger, V. J., Lamp, L. M., Bujotzek, A., Georges, G., and Liedl, K. R. (2020c). VH-VL interdomain dynamics observed by computer simulations and NMR. *Proteins* 88, 830–839. doi: 10.1002/prot.25872

Fernández-Quintero, M. L., Kraml, J., Georges, G., and Liedl, K. R. (2019a). CDR-H3 loop ensemble in solution – conformational selection upon antibody binding. *mAbs* 11, 1077–1088. doi: 10.1080/19420862.2019.1618676

Fernández-Quintero, M. L., Loeffler, J. R., Bacher, L. M., Waibl, F., Seidler, C. A., and Liedl, K. R. (2020d). Local and global rigidification upon antibody affinity maturation. *Front. Mol. Biosci.* 7:182. doi: 10.3389/fmolb.2020.00182

Fernández-Quintero, M. L., Loeffler, J. R., Kraml, J., Kahler, U., Kamenik, A. S., and Liedl, K. R. (2019b). Characterizing the diversity of the CDR-H3 loop conformational ensembles in relationship to antibody binding properties. *Front. Immunol.* 9:3065. doi: 10.3389/fimmu.2018.03065

Fernández-Quintero, M. L., Loeffler, J. R., Waibl, F., Kamenik, A. S., Hofer, F., and Liedl, K. R. (2020e). Conformational selection of allergen-antibody complexes—surface plasticity of paratopes and epitopes. *Protein Eng. Des. Select.* 32, 513–523. doi: 10.1093/protein/gzaa014

Fernández-Quintero, M. L., Math, B. F., Loeffler, J. R., and Liedl, K. R. (2019c). Transitions of CDR-L3 loop canonical cluster conformations on the micro-to-millisecond timescale. *Front. Immunol.* 10:2652. doi: 10.3389/fimmu.2019.02652

Fernández-Quintero, M. L., Pomarici, N. D., Loeffler, J. R., Seidler, C. A., and Liedl, K. R. (2020f). T-Cell Receptor CDR3 loop conformations in solution shift the relative Vα-Vβ domain distributions. *Front. Immunol.* 11:1440. doi: 10.3389/fimmu.2020.01440

Foote, J., and Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.* 224, 487–499. doi: 10.1016/0022-2836(92)91010-m

Fransson, J., Teplyakov, A., Raghunathan, G., Chi, E., Cordier, W., and Dinh, T. (2010). Human framework adaptation of a mouse anti-human IL-13 antibody. *J. Mol. Biol.* 398, 214–231. doi: 10.1016/j.jmb.2010.03.004

Gapsys, V., and de Groot, B. L. (2019). Comment on "Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size.". *bioRxiv* [Preprint]. doi: 10.7554/eLife.44718

Grünig, G., Corry, D. B., Reibman, J., and Wills-Karp, M. (2012). Interleukin 13 and the evolution of asthma therapy. *Am. J. Clin. Exp. Immunol.* 1, 20–27.

Henderson, R., Watts, B. E., Ergin, H. N., Anasti, K., Parks, R., and Xia, S.-M. (2019). Selection of immunoglobulin elbow region mutations impacts interdomain conformational flexibility in HIV-1 broadly neutralizing antibodies. *Nat. Commun.* 10:654.

Huang, M., Furie, B. C., and Furie, B. (2004). Crystal structure of the calcium-stabilized human factor IX gla domain bound to a conformation-specific anti-factor IX antibody. *J. Biol. Chem.* 279, 14338–14346. doi: 10.1074/jbc.m314011200

Hub, J. S., de Groot, B. L., Grubmüller, H., and Groenhof, G. (2014). Quantifying artifacts in ewald simulations of inhomogeneous systems with a net charge. *J. Chem. Theory Comput.* 10, 381–390. doi: 10.1021/ct400626b

Jeliazkov, J. R., Sljoka, A., Kuroda, D., Tsuchimura, N., Katoh, N., and Tsumoto, K. (2018). Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification. *Front. Immunol.* 9:413. doi: 10.3389/fimmu.2018.00413

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869

Kaplon, H., Muralidharan, M., Schneider, Z., and Reichert, J. M. (2020). Antibodies to watch in 2020. *mAbs* 12:1703531. doi: 10.1080/19420862.2019.1703531

Kaplon, H., and Reichert, J. M. (2019). Antibodies to watch in 2019. *mAbs* 11, 219–238. doi: 10.1080/19420862.2018.1556465

Knapp, B., Dunbar, J., Alcala, M., and Deane, C. M. (2017). Variable regions of antibodies and T-cell receptors may not be sufficient in molecular simulations investigating binding. *J. Chem. Theory Comput.* 13, 3097–3105. doi: 10.1021/acs.jctc.7b00080

Labute, P. (2009). Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* 75, 187–205. doi: 10.1002/prot.22234

Landolfi, N. F., Thakur, A. B., Fu, H., Vásquez, M., Queen, C., and Tsurushita, N. (2001). The integrity of the ball-and-socket joint between V and C domains is essential for complete activity of a humanized antibody. *J. Immunol.* 166:1748. doi: 10.4049/jimmunol.166.3.1748

Li, B., Zhao, L., Wang, C., Guo, H., Wu, L., and Zhang, X. (2010). The protein-protein interface evolution acts in a similar way to antibody affinity maturation. *J. Biol. Chem.* 285, 3865–3871. doi: 10.1074/jbc.m109.076547

Li, T., Tracka, M. B., Uddin, S., Casas-Finet, J., Jacobs, D. J., and Livesay, D. R. (2015). Rigidity emerges during antibody evolution in three distinct antibody systems: evidence from QSFR analysis of Fab fragments. *PLoS Comput. Biol.* 11:e1004327. doi: 10.1371/journal.pcbi.1004327

Ma, B., Kumar, S., Tsai, C.-J., and Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng. Des. Sel.* 12, 713–720. doi: 10.1093/protein/12.9.713

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* 11, 3696–3713. doi: 10.1021/acs.jctc.5b00255

Margreitter, C., Mayrhofer, P., Kunert, R., and Oostenbrink, C. (2016). Antibody humanization by molecular dynamics simulations—in-silico guided selection of critical backmutations. *J. Mol. Recognit.* 29, 266–275. doi: 10.1002/jmr.2527

Martin, A. C. R., and Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J. Mol. Biol.* 263, 800–815. doi: 10.1006/jmbi.1996.0617

Mishra, A. K., and Mariuzza, R. A. (2018). Insights into the structural basis of antibody affinity maturation from next-generation sequencing. *Front. Immunol.* 9:117. doi: 10.3389/fimmu.2018.00117

Miyamoto, S., and Kollman, P. A. (1992). Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 13, 952–962. doi: 10.1002/jcc.540130805

Molecular Operating Environment [MOE] (2020). *1010 Sherbrooke St. West, Suite #910*. Montreal, QC: Molecular Operating Environment.

Montaño, R. F., and Morrison, S. L. (2002). Influence of the isotype of the light chain on the properties of IgG. *J. Immunol.* 168:224. doi: 10.4049/jimmunol.168.1.224

Narayanan, A., Sellers, B. D., and Jacobson, M. P. (2009). Energy-based analysis and prediction of the orientation between light- and heavy-chain antibody variable domains. *J. Mol. Biol.* 388, 941–953. doi: 10.1016/j.jmb.2009.03.043

Niederfellner, G., Lammens, A., Mundigl, O., Georges, G. J., Schaefer, W., and Schwaiger, M. (2011). Epitope characterization and crystal structure of GA101 provide insights into the molecular basis for type I/II distinction of CD20 antibodies. *Blood* 118, 358–367. doi: 10.1182/blood-2010-09-305847

North, B., Lehmann, A., and Dunbrack, R. L. Jr. (2011). A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* 406, 228–256. doi: 10.1016/j.jmb.2010.10.030

Riechmann, L., Clark, M., Waldmann, H., and Winter, G. (1988). Reshaping human antibodies for therapy. *Nature* 332, 323–327. doi: 10.1038/332323a0

Röthlisberger, D., Honegger, A., and Plückthun, A. (2005). Domain interactions in the fab fragment: a comparative evaluation of the single-chain Fv and FAB format engineered with variable domains of different stability. *J. Mol. Biol.* 347, 773–789. doi: 10.1016/j.jmb.2005.01.053

Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., and Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.* 9, 3878–3888. doi: 10.1021/ct400314y

Shehata, L., Maurer, D. P., Wec, A. Z., Lilov, A., Champney, E., and Sun, T. (2019). Affinity maturation enhances antibody specificity but compromises conformational stability. *Cell Rep.* 28, 3300.e4–3308.e4.

Sotriffer, C. A., Liedl, K. R., Linthicum, D. S., Rode, B. M., and Varga, J. M. (1998). Ligand-induced domain movement in an antibody fab: molecular dynamics studies confirm the unique domain movement observed experimentally for fab NC6.8 upon complexation and reveal its segmental flexibility11Edited by I. Wilson. *J. Mol. Biol.* 278, 301–306. doi: 10.1006/jmbi.1998.1684

Sotriffer, C. A., Rode, B. M., Varga, J. M., and Liedl, K. R. (2000). Elbow flexibility and ligand-induced domain rearrangements in antibody Fab NC6.8: large effects of a small hapten. *Biophys. J.* 79, 614–628. doi: 10.1016/s0006-3495(00)76320-x

Stanfield, R. L., Zemla, A., Wilson, I. A., and Rupp, B. (2006). Antibody elbow angles are influenced by their light chain class. *J. Mol. Biol.* 357, 1566–1574. doi: 10.1016/j.jmb.2006.01.023

Teplyakov, A., Luo, J., Obmolova, G., Malia, T. J., Sweet, R., and Stanfield, R. L. (2014). Antibody modeling assessment II. Structures and models. *Proteins* 82, 1563–1582. doi: 10.1002/prot.24554

Teplyakov, A., Obmolova, G., Malia, T., and Gilliland, G. (2011). Antigen recognition by antibody C836 through adjustment of V-L/V-H packing. *Acta Crystallogr. Sect. F Struct. Biol. Crystal. Commun.* 67, 1165–1167. doi: 10.1107/s1744309111027746

Thorpe, I. F., and Brooks, C. L. (2007). Molecular evolution of affinity and flexibility in the immune system. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8821–8826. doi: 10.1073/pnas.0610064104

Tormo, J., Stadler, E., Skern, T., Auer, H., Kanzler, O., Betzel, C., et al. (1992). Three-dimensional structure of the Fab fragment of a neutralizing antibody to human rhinovirus serotype 2. *Protein Sci.* 1, 1154–1161. doi: 10.1002/pro.5560010909

Townsend, C. L., Laffy, J. M. J., Wu, Y.-C. B., Silva, O., Hare, J., Martin, V., et al. (2016). Significant differences in physicochemical properties of human immunoglobulin kappa and lambda CDR3 regions. *Front. Immunol.* 7:388. doi: 10.3389/fimmu.2016.00388

Tsai, C.-J., Kumar, S., Ma, B., and Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Sci.* 8, 1181–1190. doi: 10.1110/ps.8.6.1181

Vargas-Madrazo, E., and Paz-García, E. (2003). An improved model of association for VH–VL immunoglobulin domains: asymmetries between VH and VL in the packing of some interface residues. *J. Mol. Recognit.* 16, 113–120. doi: 10.1002/jmr.613

Wabl, M., and Steinberg, C. (1996). Affinity maturation and class switching. *Elsevier* 8, 89–92. doi: 10.1016/s0952-7915(96)80110-5

Wallnoefer, H. G., Liedl, K. R., and Fox, T. (2011). A challenging system: free energy prediction for factor Xa. *J. Comput. Chem.* 32, 1743–1752. doi: 10.1002/jcc.21758

Wardemann, H., Hammersen, J., and Nussenzweig, M. C. (2004). Human autoantibody silencing by immunoglobulin light chains. *J. Exp. Med.* 200, 191–199. doi: 10.1084/jem.20040818

Wong, S. E., Sellers, B. D., and Jacobson, M. P. (2011). Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins* 79, 821–829. doi: 10.1002/prot.22920

Zhang, D., Chen, C.-F., Zhao, B.-B., Gong, L.-L., Jin, W.-J., and Liu, J.-J. (2013). A novel antibody humanization method based on epitopes scanning and molecular dynamics simulation. *PLoS One* 8:e80636. doi: 10.1371/journal.pone.0080636

Conflict of Interest: AB, EM, and GG were Roche employees: Roche has an interest in developing antibody-based therapeutics. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 07 January 2021
doi: 10.3389/fmolb.2020.626363

Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences

Castrense Savojardo[1], *Matteo Manfredi*[1], *Pier Luigi Martelli*[1]* *and Rita Casadio*[1,2]

[1] Biocomputing Group, Department of Pharmacy and Biotechnologies, University of Bologna, Bologna, Italy, [2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council, Bari, Italy

Solvent accessibility (SASA) is a key feature of proteins for determining their folding and stability. SASA is computed from protein structures with different algorithms, and from protein sequences with machine-learning based approaches trained on solved structures. Here we ask the question as to which extent solvent exposure of residues can be associated to the pathogenicity of the variation. By this, SASA of the wild-type residue acquires a role in the context of functional annotation of protein single-residue variations (SRVs). By mapping variations on a curated database of human protein structures, we found that residues targeted by disease related SRVs are less accessible to solvent than residues involved in polymorphisms. The disease association is not evenly distributed among the different residue types: SRVs targeting glycine, tryptophan, tyrosine, and cysteine are more frequently disease associated than others. For all residues, the proportion of disease related SRVs largely increases when the wild-type residue is buried and decreases when it is exposed. The extent of the increase depends on the residue type. With the aid of an in house developed predictor, based on a deep learning procedure and performing at the state-of-the-art, we are able to confirm the above tendency by analyzing a large data set of residues subjected to variations and occurring in some 12,494 human protein sequences still lacking three-dimensional structure (derived from HUMSAVAR). Our data support the notion that surface accessible area is a distinguished property of residues that undergo variation and that pathogenicity is more frequently associated to the buried property than to the exposed one.

Keywords: solvent accessible surface area, relative solvent accessibility, protein variations, prediction of solvent accessible surface, pathogenic protein variations

OPEN ACCESS

Edited by:
Sarah Teichmann,
Wellcome Sanger Institute (WT),
United Kingdom

Reviewed by:
Joost Schymkowitz,
VIB & KU Leuven Center for Brain & Disease Research, Belgium
Carlo Travaglini-Allocatelli,
Sapienza University of Rome, Italy

***Correspondence:**
Pier Luigi Martelli
pierluigi.martelli@unibo.it

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *05 November 2020*
Accepted: *07 December 2020*
Published: *07 January 2021*

Citation:
Savojardo C, Manfredi M, Martelli PL and Casadio R (2021) Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences. Front. Mol. Biosci. 7:626363. doi: 10.3389/fmolb.2020.626363

INTRODUCTION

In structural bioinformatics, Solvent Accessible Surface Area (SASA) [or briefly Accessible Surface Area (ASA)] of proteins has always been considered a main feature for determining protein folding and stability. Early computational studies (Lee and Richards, 1971; Chothia, 1976; Miller et al., 1987, and references therein) emphasized the role of solvent exposed vs. non-exposed amino acid residues in determining the protein structure. Typically, ASA is defined as the polar solvent accessible area of a given protein, and it is computed by means of a solvent molecule, which probes the protein surface beyond the van der Waals radius. After the first rolling ball algorithm

(Shrake and Rupley, 1973), many alternatives became available for computing ASA from the atomic coordinates of the protein in its folded and unfolded state [for review see Ali et al. (2014)]. Evidently, ASA is a function of the three dimensional structure of the protein and, based on ASA values, amino acid residues of a protein can be classified as buried or exposed (Kabsch and Sander, 1983), a property that is conserved through evolution in protein families (Rost and Sander, 1994). ASA is routinely computed as an absolute value or as Relative Solvent Accessibility (RSA), when the ASA value is divided by the maximum possible solvent accessible surface area of the residue (Tien et al., 2013). ASA gained also a pivot role in detecting protein-protein interfaces of molecular complexes in the Protein Data Bank (PDB) [for review see Savojardo et al. (2020), and references therein].

With the advent of machine and deep learning-based approaches (Baldi, 2018), many methods became available for predicting RSA and ASA. They differ mainly in the machine learning approach, the volume of the database of protein structures and the predicted output (ASA, RSA, or binary classification) (Rost and Sander, 1994; Pollastri et al., 2002; Drozdetskiy et al., 2015; Ma and Wang, 2015; Fan et al., 2016; Wu et al., 2017; Kaleel et al., 2019; Klausen et al., 2019).

Surface accessible area of residues can be important also for functional annotation of disease related protein variants. However, this property has been rarely included into the physico-chemical characteristics adopted to describe the residues undergoing variations (Chen and Zhou, 2005; Martelli et al., 2016; Savojardo et al., 2019).

In this study, we investigate the relation between the pathogenicity of human protein variations and the solvent exposure of the residues undergoing variation (wild-type residues). To this aim, we provide an updated version of a highly curated dataset of Single Residue Variations (SRVs) occurring in human proteins that can be mapped in high-quality structures deposited in the Protein Data Bank (PDB). The dataset, here referred to as HVAR3D-2.0, is generated from data available at the HUMASVAR database and builds on top of data previously analyzed in a different study (Savojardo et al., 2019). On this structural dataset, we explore the relationship between pathogenicity of SRVs and the solvent accessibility of the corresponding wild-type residues. In particular, we determine that the majority (67%) of disease-related SRVs occur in buried positions whereas neutral SRVs occur mostly (64.3%) in exposed residues. Moreover, SRVs targeting specific residue types such as glycine, tryptophan, tyrosine, and cysteine, are more frequently associated with disease than others are. Finally, for all residues, and in particular for asparagine, glutamine, histidine, and lysine, the proportion of disease related SRVs largely increases when the wild-type residue is buried, and decreases when it is exposed, confirming that, among other factors, the context can be associated to the pathogenicity of the variations (Casadio et al., 2011).

We extended the above analysis to a larger set of variations included in HUMSAVAR and collected in a dataset called HVARSEQ. In order to estimate the solvent accessibility of all residues undergoing disease-related or neutral SRVs in human proteins, we developed an in-house method based on deep-learning for predicting solvent exposure from sequence. Our method performance is comparable to state-of-the-art methods. We apply it to all the residues of human protein sequences, undergoing pathogenic and neutral SRVs in HVARSEQ.

Results of the large-scale analysis on protein sequences support what observed in protein structures and confirm the different distribution buried/exposed wild-type residues in disease-related and neutral SRVs. Our data suggest that solvent accessibility is a distinguished property of wild type residues undergoing pathogenic variations.

MATERIALS AND METHODS

Variation Databases

All human Single-Residue Variations (SRVs) were collected from HUMASVAR version 2020_04 (Aug 2020). As a first filtering step, we retained variations labeled as "Disease" and "Polymorphism," neglecting all variations labeled as "Unclassified." Disease-related SRVs not associated with OMIM diseases were excluded. After this procedure we ended up with a large set of SRVs occurring on human protein sequence. Here this dataset is referred to as HVARSEQ (Human VARiations in SEQuences)

In order to build the structural dataset (here referred to as HVAR3D-2.0, Human VARiations in three Dimensional structures), we firstly identified, among all the sequences included in HVARSEQ, the subset of proteins endowed with a PDB structure meeting the following criteria:

- Coverage of the corresponding UniProtKB sequence is $\geq$70%;
- Experimental method is X-ray crystallography;
- Resolution is $\leq$ 3Å.

The mapping of SRV positions on protein structure was performed using data from the Structure Integration with Function, Taxonomy and Sequence (SIFTS) project[1]. Protein structures having ambiguous or wrong SIFTS mapping files were excluded from the dataset.

Computing Solvent Exposure

The absolute Accessible Surface Area (ASA) of each wild-type residue undergoing variation has been computed using the DSSP program (Kabsch and Sander, 1983). Relative Solvent Accessibility (RSA) values were then obtained dividing absolute ASA values in Å^2 by residue-specific maximal accessibility values, as extracted from the Sander and Rost scale (Rost and Sander, 1994). Finally, each residue has been classified as buried (B) if its RSA was below 20%, and exposed (E) otherwise.

Computing P_D, $P_{D|R}$, $P_{D|B,R}$, and $P_{D|E,R}$

In this study, the background probability of a wild-type residue to be disease associated in a dataset of wild-type residues is computed as follows:

$$P_D = \frac{n_D}{N} \tag{1}$$

[1]https://www.ebi.ac.uk/pdbe/docs/sifts/.

where n_D and N are the number of wild-type residues undergoing disease-related variations and the total number of wild-type residues undergoing variations (disease related or not) in the dataset, respectively.

The conditional probability of being disease related when variated, given a wild-type residue R, is computed as follows:

$$P_{D|R} = \frac{n_{DR}}{n_R} \quad (2)$$

where n_{DR} and n_R are the number of wild-type residues of a given R type, which are disease related upon variations, and the total number of R residues in the whole dataset, respectively.

The conditional probability of a wild-type residue R to be disease related upon variation when buried is computed as:

$$P_{D|B,R} = \frac{n_{DBR}}{n_{BR}} \quad (3)$$

where n_{DBR} and n_{BR} are the number of buried wild type R residue in the set of wild type disease related upon variation and the total number of buried R wild type residues, respectively.

Similarly, the conditional probability of a wild-type residue R to be disease related upon variation when exposed is computed as:

$$P_{D|E,R} = \frac{n_{DER}}{n_{ER}} \quad (4)$$

where n_{DER} and n_{ER} are the number of exposed wild type R residue in the set of wild-type disease related upon variation and the total number of exposed R wild type residues, respectively.

All the above probabilities are estimated considering the structural dataset HVAR3D-2.0, and by computing the residue solvent accessibility with the DSSP program. Moreover, we extended the analysis to the whole HVARSEQ sequence dataset, by estimating the residue exposure state (buried and exposed) with a predictor implemented in-house and described in the following section.

Predicting Solvent Accessibility From the Protein Sequence

The method implements a deep-learning architecture processing an input based on the following descriptors:

- The residue one-hot encoding, representing primary sequence information;
- Evolutionary information encoded with a protein sequence profile, as extracted from multiple sequence alignment generated using the HHblits version 3 program (Steinegger et al., 2019). We performed two search iterations with default parameters against the Uniclust30 database (Mirdita et al., 2017).

Our deep architecture processes the input using three cascading Bidirectional Long-Short Term Memory (BLSTM) layers (Graves and Schmidhuber, 2005). BLSTMs belong to the class of LSTM (Hochreiter and Schmidhuber, 1997), a special recurrent neural network architecture well-suited for processing protein sequence data and extracting significant sequential relations between elements of the sequence. BLSTMs are an extension of LSTMs performing a double scanning of the input sequence, from left to right and vice versa, in order to better capture the sequential relations among sequence positions. The adoption of the recurrent BLSTM allows the method to take into consideration the local sequence context without the explicit use of a fixed-size window centered on each residue.

The output of the third recurrent layer is then provided as input to a time-distributed fully connected layer adopting a sigmoid activation function. This layer is responsible for the final, binary classification of each residue in the sequence into buried or exposed classes. In particular, the numerical output value in the range [0, 1] attached to each residue is interpreted as a probability p of being exposed: all residues with $p \geq 0.5$ are predicted as exposed while those with $p < 0.5$ are classified as buried.

The dataset adopted to train and test the predictor presented in this study has been extracted from the Protein Data Bank (interrogated Oct 15, 2019) (Berman, 2000). Overall, the dataset comprises 2532 non-redundant, author-declared functional monomeric PDB structures, obtained with X-ray crystallography at $<$ 2.5 Å resolution and covering more than 70% of corresponding UniProtKB sequences. All proteins in the dataset share <30% sequence identity. This dataset was then randomly split into a training set, comprising 2,352 sequences, and an independent blind test set including 200 sequences. Proteins in the training set were further split into 10 equally-sized sets for setting the values of hyperparameters with a cross-validation procedure.

Solvent exposure for training/testing data has been computed using DSSP as detailed in Section: Computing solvent exposure. The residues were classified into buried and exposed using a RSA threshold of 20%. Using this threshold, the set of residues is roughly divided into equally sized subsets comprising 52% and 48% of buried and exposed residues, respectively, providing balanced datasets for training and testing.

RESULTS

HVAR3D-2.0: A Dataset of Variations Covered by 3D Structure

The structural dataset collected in this work, here referred to as HVAR3D-2.0, is an updated version of the dataset described in a previous study (Savojardo et al., 2019). The dataset has been derived by mapping on PDB structures OMIM-related and neutral SRVs annotated in the HUMSAVAR database[2], release 2020_08 (Aug, 2020). Only structures determined with X-ray crystallography with resolution ≤3Å and covering ≥70% of the corresponding UniProtKB sequences were selected. After this stringent filtering, we ended-up with a high-quality dataset comprising 10,760 human SRVs occurring on 1,255 PDB entries (corresponding to 1,285 protein chains). The set includes 6,778 and 3,982 disease-related and neutral SRVs, respectively. **Table 1** lists a summary of the HVAR3D-2.0 content. The HVAR3D-2.0 dataset is available in **Supplementary Table 1** in TSV format.

[2]https://www.uniprot.org/docs/humasavar

TABLE 1 | Statistics of HVAR3D 2.0 dataset.

Description	Counts (#)
PDB structures	1,255
PDB chains	1,285
Distinct SRV positions	9,379
SRVs	10,760
Disease-related SRVs	6,778
Neutral SRVs	3,982

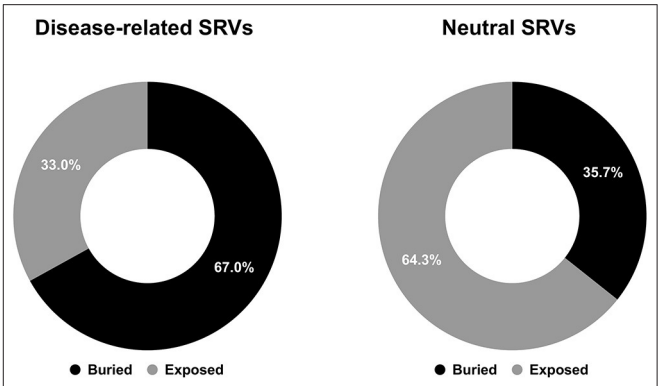


FIGURE 1 | Pie charts showing the fractions of buried/exposed wild-type residues undergoing disease-related (left) and neutral (right) SRVs in the HVAR3D-2.0 dataset, respectively.

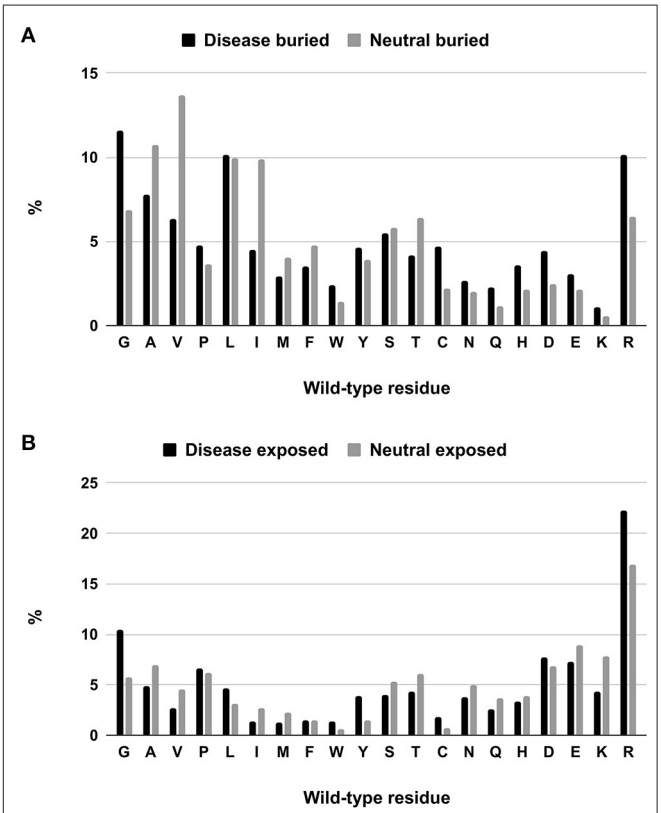


FIGURE 2 | Composition of buried **(A)** and exposed **(B)** wild-type residues undergoing disease-related and neutral variations in the HVAR3D-2.0 dataset.

In the present study, we are interested in investigating the relation between the pathogenicity of SRVs and the solvent accessibility of the residue undergoing variation. For this reason, we firstly computed Accessible Surface Area (ASA) values for all 1,285 protein chains included in the HVAR3D dataset using the DSSP program (Kabsch and Sander, 1983). Raw ASAs were then converted into Relative Solvent Accessibility (RSA) values using the Rost and Sander maximal accessibility scale (Rost and Sander, 1994). Finally, all residues with RSA $\geq$ 20% were labeled as exposed (E) or buried (B) otherwise. This threshold (or similar ones, in the range of 15–25% RSA) is routinely adopted for computing the protein surfaces and deriving classification datasets in many studies (Thompson and Goldstein, 1996; Mucchielli-Giorgi et al., 1999; Pollastri et al., 2002; Kaleel et al., 2019), since it roughly divides the set of residues in a protein in two equally-sized subsets. In HVAR3D, using a 20% RSA threshold, we obtain 55% and 45% of residues classified as buried and exposed, respectively, corresponding to a realistic characterization of the protein interior (accounting for completely and partially buried residues) and surface (Miller et al., 1987). Preliminary analysis highlighted that the choice of the RSA threshold (in the reasonable range of 15–25% RSA) only minorly affects the conclusions drawn in this study (data not shown). For this reason, all the subsequent analyses were performed using the aforementioned threshold.

Focusing our attention to structure positions undergoing SRVs, we firstly computed the different proportions of buried and exposed wild-type residues associated to disease-related and neutral SRVs. As shown in **Figure 1**, 67% of wild-type residues undergoing disease-related variations are located in buried positions and about 64% of wild-type residues involved in neutral variations are exposed. This conclusion corroborates, on a much larger structural database, results partially reported in previous studies (Martelli et al., 2016; Savojardo et al., 2019). The relative abundance of disease-related variations in buried positions of the protein and of neutral ones in exposed positions suggests that the solvent accessibility of the variated position is a further property to consider when determining the pathogenicity of a variation.

Analyzing Distributions of Variated Wild-Type Residues in the Structure Database

We tackle the problem of associating solvent exposure to a specific wild-type residue as a characteristic feature to be associated to its variation type (neutral or disease related). We compute the relative frequency of occurrence in the buried and exposed sets of each residue undergoing a disease related or neutral variation (**Figures 2A,B**). It is evident that while some residue types are more often disease related when variated in the buried state (Q, H, D, E, K), others (including G, W, C, and R) are disease related upon variation in either state.

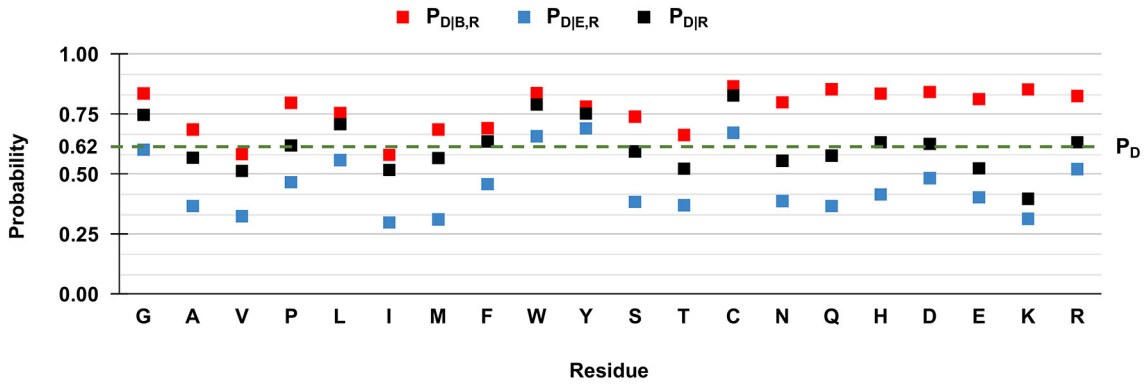


FIGURE 3 | Probabilities of the 20 wild-type residues undergoing disease-related variations, depending on the wild type residue and the exposure state in HVAR3D-2.0. Buried and exposure state of each residue position are estimated with DSSP as described in Section: Analyzing distributions of variated wild-type residues in the structure database. $\mathbf{P_D}$: the probability of a wild-type residue (position) to be disease associated in the HVAR3D-2.0 dataset [see Equation (1)]. $\mathbf{P_{D|R}}$: the conditional probability of being disease related when variated, given a wild-type residue [see Equation (2)]. $\mathbf{P_{D|B,R}}$: the conditional probability of a wild-type residue to be disease related upon variation when buried [see Equation (3)]. $\mathbf{P_{D|E,R}}$: the conditional probability of a wild-type residue to be disease related upon variation when exposed [see Equation (4)].

However, when we compute the conditional probabilities per residue type, clearly the tendency of the majority of the wild-type residues is that of being disease-related upon variation when buried (red squares in **Figure 3**). Indeed, in **Figure 3** we show to which extent the knowledge of the solvent exposure changes the *a priori* probability of a given residue type to be associated with disease. For each residue type R, we report the conditional probability of being associated to disease ($P_{D|R}$, black squares) and how the two conditional probabilities ($P_{D|B,R}$ and $P_{D|E,R}$ in red and blue squares, respectively) change, given that the variated residue is buried or exposed. We contrast these values to the baseline frequency of disease related variations in the HVAR3D-2.0 dataset, referred to as P_D and equal to 0.62.

In **Figure 3**, when comparing $P_{D|R}$ of each residue R (black squares) with the baseline value P_D, it is evident that not all the residues are equally likely to be associated with disease when variated. Residues like glycine (G), leucine (L) tryptophan (W), tyrosine (Y), and cysteine (C) show values of $P_{D|R}$ that are higher than the baseline, indicating that their variations are frequently associated to disease in the database. Furthermore, for all residues the relation $P_{D|B,R} > P_{D|R} > P_{D|E,R}$ holds. This means that for all residue types, the probability that SRVs are related to disease is higher when the wild-type residue is buried (red squares) than when it is exposed (blue squares). The extent of this difference depends on the residue type and it is remarkable for asparagine (N), glutamine (Q), histidine (H), and lysine (K). All these residues are polar and abundant on the protein surface (data not shown). On average, when variated, they are associated to disease with a frequency comparable or lower than the baseline 0.62. However, when variations of these residue types occur in buried positions, the frequency of disease related variations raises to values around 0.8, reaching 0.85 in the case of glutamine (Q) and lysine (K). Remarkably, for three residues [tryptophan (W), tyrosine (Y) and cysteine (C)] the frequency of disease-related variation is higher than the baseline, rather independently of the exposure state. Conversely, the fraction of disease-related variations of valine (V) and isoleucine (I) is lower than the baseline, independently of their accessibility.

Overall, these findings highlight a relation between the pathogenicity of the variation and the solvent accessibility of the wild-type residue and show that the extent of the association depends on the residue type. In all cases, variations occurring in buried positions are more likely to be disease-related. This is particularly so for charged residues, for polar residues such as asparagine (N), glutamine (Q) and histidine (H), and for proline (P), cysteine (C), and tryptophan (W).

HVARSEQ: A Dataset of Protein Sequences With Variations

Here we make use of computational prediction of solvent accessibility to extend our analysis to all the positions undergoing variations contained in HUMSAVAR. From the HUMSAVAR database, release 2020_08 (Aug, 2020), we collected all polymorphisms and all OMIM-related SRVs occurring in protein sequences. Unclassified SRVs were filtered-out from the set. Overall, 69,385 SRVs were collected. 29,949 and 39,436 SRVs are disease-related and neutral, respectively, occurring on 12,494 protein sequences. Here, this extended set of protein sequences is referred to as HVARSEQ. In **Table 2** we summarize the basic statistics of the dataset. The HVARSEQ dataset is available in **Supplementary Table 2** in TSV format.

Predicting Solvent Accessibility

For computing solvent accessibility from protein sequences, we implemented an in-house method for predicting solvent exposure from sequence. The method is based on deep-learning processing of several input features, which encode the protein sequence and the sequence profile (see Materials and Methods for more details on the method). Our method classifies each residue of the sequence into two classes: buried (B), corresponding

TABLE 2 | Statistics of HVARSEQ dataset.

Description	Counts (#)
UniProtKB sequences	12,494
Distinct SRV positions	64,869
SRVs	69,385
Disease-related SRVs	29,949
Neutral SRVs	39,436

TABLE 3 | Performance of our deep learning-based method for predicting solvent exposure from protein sequence.

Scoring index	Dataset		
	Cross-validation	Blind test	HVAR3D-2.0
MCC	0.63	0.63	0.60
Q2 (accuracy)	81%	82%	80%
F1	81%	82%	80%

TABLE 4 | Performance of different methods for solvent accessibility prediction on the blind test set described in this study comprising 200 protein sequences.

Method	MCC	Q2 %	F1 %
PaleAle 5.0	0.65	82	84
NetSurfP-2.0	0.67	83	81
Our method	0.63	82	82

to residues whose RSA is lower than 20%, and exposed (E), corresponding to residues with RSA $\geq$ 20%.

Performances are listed in **Table 3** and are evaluated adopting three different testing sets (by adopting a cross validation procedure (leftmost column); on the blind test (central column); on our HVAR3D-2.0 dataset, for which solvent exposure can be directly computed using DSSP). Comparing the first two columns, it is evident that our method is robust, achieving generalization performances that are as good and even better than cross-validation results. Overall, our method is able to discriminate buried from exposed residues with Q_2 (accuracy), MCC (Matthew Correlation Coefficient) and F1 equal to 82%, 0.63 and 82%, respectively. When scored on the HVAR3D-2.0 dataset, the performance is almost unchanged, suggesting that our method is quite stable across different datasets.

We also performed a side-by-side comparison between our method and two state-of-the-art approaches, namely PaleAle5.0 (Kaleel et al., 2019) and NetSurfP-2.0 (Klausen et al., 2019). Results are reported in **Table 4**. All methods perform quite well, with comparable scoring indexes. It is worth mentioning that the testing set used in this benchmark is non-redundant only with respect to our training set: this condition is not guaranteed for the other two methods evaluated, which adopt different training sets. In general, we can conclude that our method well-compares with recent tools at the state-of-the-art.

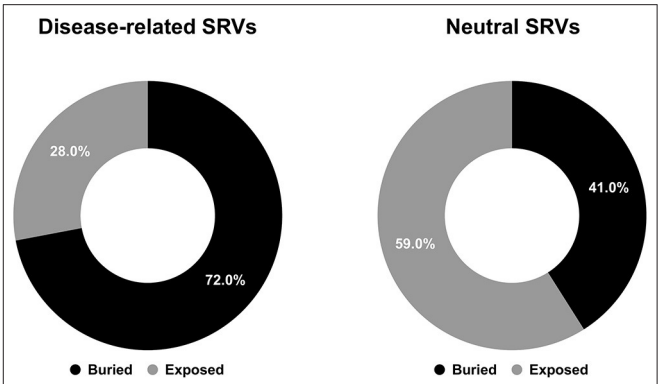


FIGURE 4 | Pie charts showing the fractions of predicted buried/exposed positions disease-related (left) and neutral (right) upon variations in the HVARSEQ dataset, respectively.

Analyzing Distributions of Variated Wild-Type Residues in the Sequence Dataset

After computing solvent accessibility over HVARSEQ, we assessed the proportions of buried and exposed predictions separately on the subsets of residues undergoing disease-related and neutral variations. Results are in **Figure 4**.

As to the prediction, 72% of disease related SRVs occurs in buried positions and 58% of neutral SRVs affect exposed residues. Interestingly, the proportions of buried/exposed positions for disease and neutral SRVs are in agreement with those assessed on the structural dataset (67% and 64.3%, respectively: compare **Figures 1**, **4**). The result further corroborates the notion that residues undergoing disease-related variations are mainly in buried positions.

We then evaluated $P_{D|R}$, $P_{D|B,R}$, and $P_{D|E,R}$ for all the residue types and results are reported in **Figure 5**. We also show the baseline probability P_D (0.43), which represents the proportion of positions that undergo disease-related variations in the HVARSEQ dataset.

The comparison between $P_{D|R}$ and $P_{D,}$ which are both independent from predictions, confirms the finding obtained on the HVAR3D-2.0 dataset: residues such as glycine (G), tryptophan (W), tyrosine (Y), and cysteine (C), when undergoing variation, are more frequently associated to disease than expected from the baseline. In the sequence set, this behavior characterizes also arginine (R) and aspartic acid (D).

Similarly to the structural case, for all residues we have that $P_{D|B,R} > P_{D|R} > P_{D|E,R}$, highlighting that for all residue types, SRVs are more frequently associated to disease when occurring in buried positions than in exposed ones. The tendency is remarkable for the majority of residues, already identified from HVAR3D-2.0 and including asparagine (N), lysine (K), and histidine (H). The analysis on HVARSEQ highlights a difference between $P_{D|B,R}$ and $P_{D|E,R}$ for tryptophan (W) and cysteine (C). However, this discrepancy can be due to prediction errors on these two less abundant (rare) residues in the database. Similarly, to what described for HVAR3D-2.0 (**Figure 3**), the frequency

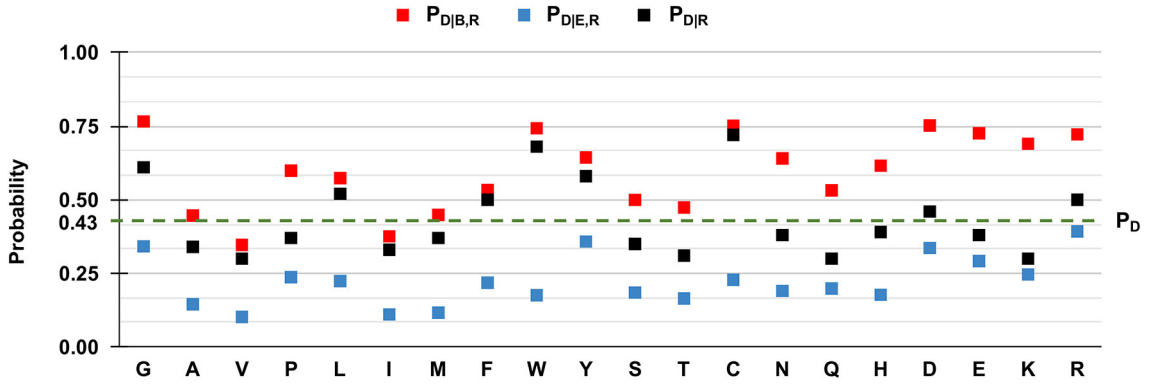


FIGURE 5 | Frequency of disease-related SRVs, depending on the wild type residue and the exposure state in HVARSEQ. Here, buried and exposure states of each residue position have been predicted using the method described in Section Analyzing distributions of variated wild-type residues in the sequence database. $\mathbf{P_D}$: the probability of a wild-type residue (position) to be disease associated in the HVARSEQ dataset [see Equation (1)]. $\mathbf{P_{D|R}}$: the conditional probability of being disease related when variated, given a wild-type residue [see Equation (2)]. $\mathbf{P_{D|B,R}}$: the conditional probability of a wild-type residue to be disease related upon variation when buried [see Equation (3)]. $\mathbf{P_{D|E,R}}$: the conditional probability of a wild-type residue R to be disease related upon variation when exposed [see Equation (4)].

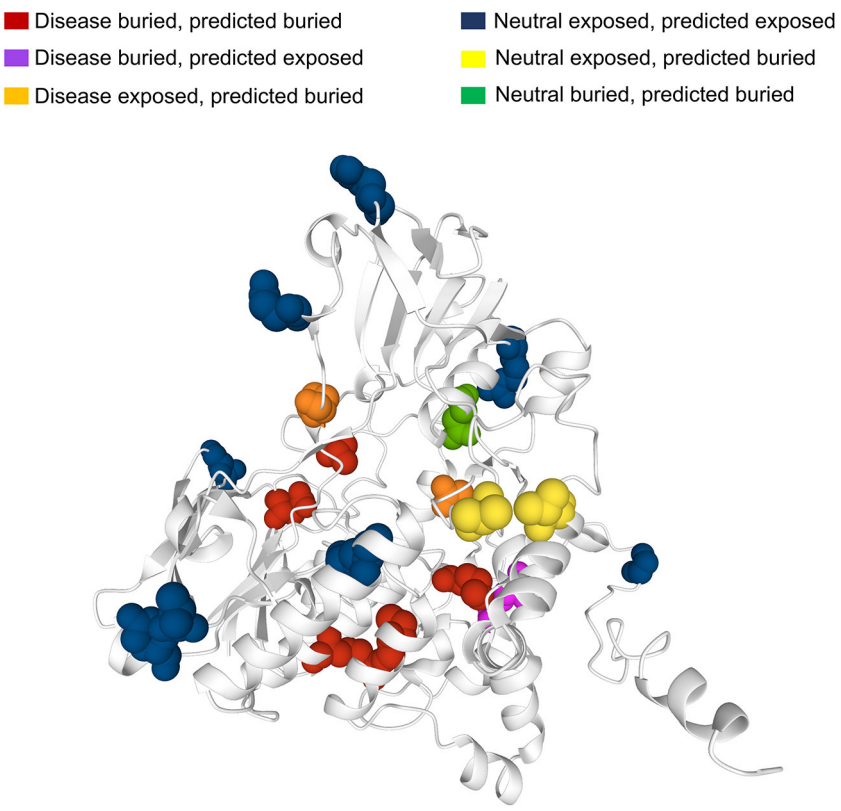


FIGURE 6 | Mapping SASA predictions on a protein model. The model is that of human Dimethylaniline monooxygenase 3 (UniProtKB: P31513) derived from the SWISS-MODEL Repository. Solvent exposure is computed from the available 3D protein model using DSSP. Variation (SVR) positions are highlighted using the spacefill view. In red, buried positions associated to disease-related SRVs and correctly predicted as buried by our method. In magenta, buried disease-related positions wrongly predicted as exposed. In orange, exposed disease-related positions wrongly predicted as buried. In blue, exposed neutral SRV positions correctly predicted as exposed. In yellow, exposed neutral positions wrongly predicted as buried. In green, buried neutral positions correctly predicted as buried.

of disease-related SRVs occurring at valine (V) and isoleucine (I) residues is lower than the baseline, independently of the exposure state.

Case Study

Many human protein sequences, without any associated three-dimensional (3D) structure, are endowed with models that can be derived from the SWISS-MODEL Repository[3], directly linked to the protein UniProtKB file. For sake of curiosity, we took advantage of an example to show the 3D location of our sequence-based prediction. In particular, in **Figure 6** we show the model of the human Dimethylaniline monooxygenase 3 protein (UniProtKB: P31513)[4]. This protein has 19 SRVs in HVARSEQ, eight of which are disease-related and 11 are neutral. Disease-related SRVs are all associated to Trimethylaminuria (OMIM:602079)[5], a disease condition resulting from the abnormal presence of large amounts of volatile and malodorous trimethylamine within the body. In **Figure 6**, we map all the solvent exposure predictions for all SRV positions into the 3D model.

It is evident that the vast majority of disease-related SRVs (6 out of 8) are in buried positions. Of these, five are correctly predicted as buried by our method (in red) while only one is wrongly predicted as exposed (in magenta). Neutral SRVs are mostly exposed (10 out of 11): eight of these are correctly predicted in exposed regions (in blue).

Results illustrate the general trend of what we observed in the structural data set and are consistent with the accuracy of the prediction method.

CONCLUSION AND PERSPECTIVE

In this paper, we focus on the solvent accessible surface area, a property of protein residues, firstly described and computed in several biophysical studies, to which Cyrus Chothia contributed (Chothia, 1976). The property, which nowadays can be computed with machine learning based methods, is here exploited in relation to another important problem: the annotation of variations in human proteins as disease related or not. We took advantage of an ample set of human protein structures to observe that indeed disease related variations occur more frequently in buried regions of the proteins than in solvent accessible surfaces. In turn, neutral polymorphisms are characterized by a more frequent solvent exposure. We then proved that with a deep learning method performing at the state of art, the tendency is observable also in the majority of all the wild-type residues undergoing variations that are presently listed in HUMSAVAR. We suggest that the solvent accessible surface area of wild type residues is a distinguished property to be included among those necessary to annotate pathogenic from non-pathogenic variations.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

RC, PM, and CS: conceptualization and writing. RC, PM, CS, and MM: methodology. MM and CS: software. CS, MM, and PM: data curation and visualization. RC and PM: supervision. All authors contributed to the article and approved the submitted version.

FUNDING

The work was supported by the PRIN2017 grant (project 2017483NH8_002), delivered to CS from the Italian Ministry of University and Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.626363/full#supplementary-material

[3]https://swissmodel.expasy.org/repository
[4]https://www.uniprot.org/uniprot/P31513.
[5]https://www.omim.org/entry/60207

REFERENCES

Ali, S., Hassan, M. D., Islam, A., and Ahmad, F. (2014). A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states. *Curr. Protein Pept. Sci.* 15, 456–476. doi: 10.2174/1389203715666140327114232

Baldi, P. (2018). Deep learning in biomedical data science. *Annu. Rev. Biomed. Data Sci.* 1, 181–205. doi: 10.1146/annurev-biodatasci-080917-013343

Berman, H. M. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Martelli, P. L. (2011). Correlating disease related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* 32, 1161–1170. doi: 10.1002/humu.21555

Chen, H., and Zhou, H.-X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.* 33, 3193–3199. doi: 10.1093/nar/gki633

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105, 1–12. doi: 10.1016/0022-2836(76)90191-1

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43, W389–394. doi: 10.1093/nar/gkv332

Fan, C., Liu, D., Huang, R., Chen, Z., and Deng, L. (2016). PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility. *BMC Bioinform.* 17:S8. doi: 10.1186/s12859-015-0851-2

Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Kaleel, M., Torrisi, M., Mooney, C., and Pollastri, G. (2019). PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning. *Amino Acids* 51, 1289–1296. doi: 10.1007/s00726-019-02767-6

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., et al. (2019). NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.* 87, 520–527. doi: 10.1002/prot.25674

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400. doi: 10.1016/0022-2836(71)90324-X

Ma, J., and Wang, S. (2015). AcconPred: predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. *BioMed Res. Int.* 2015:678764. doi: 10.1155/2015/678764

Martelli, P. L., Fariselli, P., Savojardo, C., Babbi, G., Aggazio, F., and Casadio, R. (2016). Large scale analysis of protein stability in OMIM disease related human protein variants. *BMC Genomics* 17:397. doi: 10.1186/s12864-016-2726-y

Miller, S., Lesk, A. M., Janin, J., and Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. *Nature* 328, 834–836. doi: 10.1038/328834a0

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* 45, D170–D176. doi: 10.1093/nar/gkw1081

Mucchielli-Giorgi, M. H., Hazout, S., and Tufféry, P. (1999). PredAcc: prediction of solvent accessibility. *Bioinformatics* 15, 176–177. doi: 10.1093/bioinformatics/15.2.176

Pollastri, G., Baldi, P., Fariselli, P., and Casadio, R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. *Proteins Struct. Funct. Genet.* 47, 142–153. doi: 10.1002/prot.10069

Rost, B., and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins Struct. Funct. Bioinforma.* 20, 216–226. doi: 10.1002/prot.340200303

Savojardo, C., Babbi, G., Martelli, P., and Casadio, R. (2019). Functional and structural features of disease-related protein variants. *Int. J. Mol. Sci.* 20:1530. doi: 10.3390/ijms20071530

Savojardo, C., Martelli, P. L., and Casadio, R. (2020). Protein–protein interaction methods and protein phase separation. *Annu. Rev. Biomed. Data Sci.* 3, 89–112. doi: 10.1146/annurev-biodatasci-011720-104428

Shrake, A., and Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. *Lysozyme and insulin. J. Mol. Biol.* 79, 351–371. doi: 10.1016/0022-2836(73)90011-9

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* 20:473. doi: 10.1186/s12859-019-3019-7

Thompson, M. J., and Goldstein, R. A. (1996). Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins* 25, 38–47. doi: 10.1002/(SICI)1097-0134(199605)25:1<38::AID-PROT4>3.0.CO;2-G

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., and Wilke, C. O. (2013). Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* 8:e80635. doi: 10.1371/journal.pone.0080635

Wu, W., Wang, Z., Cong, P., and Li, T. (2017). Accurate prediction of protein relative solvent accessibility using a balanced model. *BioData Min.* 10:1. doi: 10.1186/s13040-016-0121-5

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fmolb.2021.626837

Characterizing Hydropathy of Amino Acid Side Chain in a Protein Environment by Investigating the Structural Changes of Water Molecules Network

Lorenzo Di Rienzo[1†], *Mattia Miotto*[1,2†], *Leonardo Bò*[1], *Giancarlo Ruocco*[1,2], *Domenico Raimondo*[3*] and *Edoardo Milanetti*[1,2*]

[1]Center for Life Nanoscience, Istituto Italiano di Tecnologia, Rome, Italy, [2]Department of Physics, Sapienza University, Rome, Italy, [3]Department of Molecular Medicine, Sapienza University, Rome, Italy

OPEN ACCESS

Edited by:
Alfredo Iacoangeli,
King's College London,
United Kingdom

Reviewed by:
Alejandro Giorgetti,
University of Verona, Italy
Daniele Di Marino,
Polytechnic University of Marche, Italy

***Correspondence:**
Edoardo Milanetti
edoardo.milanetti@uniroma1.it
Domenico Raimondo
domenico.raimondo@uniroma1.it

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: 06 November 2020
Accepted: 04 January 2021
Published: 26 February 2021

Citation:
Di Rienzo L, Miotto M, Bò L, Ruocco G, Raimondo D and Milanetti E (2021) Characterizing Hydropathy of Amino Acid Side Chain in a Protein Environment by Investigating the Structural Changes of Water Molecules Network. Front. Mol. Biosci. 8:626837. doi: 10.3389/fmolb.2021.626837

Assessing the hydropathy properties of molecules, like proteins and chemical compounds, has a crucial role in many fields of computational biology, such as drug design, biomolecular interaction, and folding prediction. Over the past decades, many descriptors were devised to evaluate the hydrophobicity of side chains. In this field, recently we likewise have developed a computational method, based on molecular dynamics data, for the investigation of the hydrophilicity and hydrophobicity features of the 20 natural amino acids, analyzing the changes occurring in the hydrogen bond network of water molecules surrounding each given compound. The local environment of each residue is complex and depends on the chemical nature of the side chain and the location in the protein. Here, we characterize the solvation properties of each amino acid side chain in the protein environment by considering its spatial reorganization in the protein local structure, so that the computational evaluation of differences in terms of hydropathy profiles in different structural and dynamical conditions can be brought to bear. A set of atomistic molecular dynamics simulations have been used to characterize the dynamic hydrogen bond network at the interface between protein and solvent, from which we map out the local hydrophobicity and hydrophilicity of amino acid residues.

Keywords: hydropathy, molecular dynamics simulation, hydrophobicity, local structural environment, water molecules network

1 INTRODUCTION

Hydration water molecules play a crucial role in living organisms as most biological processes occur in an aqueous environment (Rothschild and Mancinelli, 2001), which actively influences the structure and function of biomolecules and their interactions (Levy and Onuchic, 2006; Ball 2008). Compounds immersed in water display different behaviors depending on their chemical characteristics. In particular, the arrangement of the water molecules that hydrate compounds changes according to their properties (Vagenende and Trout, 2012; Tomobe et al., 2017). So we can extract information on the chemical nature and function of the solute by studying the attraction and repulsion of chemical compounds toward the water (Chothia, 1976). In general, both hydrophobic and hydrophilic effects are dominant driving forces for several biochemical processes, such as protein

folding, nucleic acid stability, molecular recognition, and binding (Tanford, 1972; Brooks et al., 1998; Aftabuddin and Kundu, 2007; Moret and Zebende, 2007; Miotto et al., 2018).

In light of this, solvation water should be considered an integral part of biological macromolecules. In particular, water molecules in solutions are divided into 1) internal water molecules that occupy cavities in the biomolecule structure and can be identified in crystallography; 2) water molecules that interact with the molecular surface and 3) bulk water. Depending on the category, the organization of the water molecules is associated with different time scales. The relaxation times for internal waters range from tens of ns to ms since they require local rearrangement of the protein to occur. On the other hand, the motion of bulk water has the time scale of the picoseconds. In between, there is the motion of surface water molecules that are characterized by residence times on the order of tens of picoseconds (Tarek and Tobias, 2000; Qvist et al., 2009; Mondal et al., 2017).

In general, the investigation of the behavior of water in the hydration shells of organic compounds is a fundamental analysis to better understand most biological processes both from a theoretical and practical point of view (Raschke, 2006).

An effective measure of the interaction between water and amino acids, the hydropathy index (a number representing the hydrophobic or hydrophilic properties of its side chain), was firstly proposed in 1982 by Kyte *et al.* (Kyte and Doolittle, 1982). Indeed, in the computational biology field, attributing a single number, the hydropathy index, to each amino acid is very useful for studying the chemical-physical and structural properties of proteins. Over the past few decades, many hydrophobicity and hydrophilicity scales, based on both experimental and theoretical approaches, have been defined, and these schematizations have proven their usefulness in the characterization of protein regions and the development of computational methods (Chothia, 1974; Jones, 1975; Kyte and Doolittle, 1982; Sweet and Eisenberg, 1983; Rose et al., 1985; Wilce et al., 1995). For instance, one of the typical use of the hydrophobicity and hydrophilicity values for the 20 amino acids is the prediction of transmembrane regions in protein structure modeling (Deber et al., 2001).

Recently we have developed a new theoretical-computational method analyzing the orientation of water molecules surrounding a small organic compound, as computed from molecular dynamics simulations (Bonella et al., 2014). The procedure is based on the calculation of the conditional probability density of finding a water molecule with a specific orientation, given its distance from the nearest atom of the solute (Babiaczyk et al., 2010; Bonella et al., 2014).

We thus applied this method to the 20 natural amino acids defining the *WOPHS* (Water Orientation Probability Hydropathy Scale) hydropathy scale, the first scale to be *vectorial* as it associates three indices for each amino acid (Bonella et al., 2014). In fact, we argued that assigning a single number is not enough to characterize the solvation properties of amino acids, in particular when both hydrophobic and hydrophilic regions are present in the same residue. In this respect, our characterization can be used to understand some of the known ambiguities in the ranking of amino acids in the current scales available in the literature. This method presents several advantages over previously developed computational and experimental approaches: it is sensitive to the specific environment of the amino acids and can be applied to unnatural and modified amino acids, as well as to other small organic molecules (Bonella et al., 2014; Leopizzi et al., 2017). In particular, analyzing the structural changes of the dynamic hydrogen bond network, we studied both the *trans*-membrane passive permeation properties for a set of neutral drugs (Milanetti et al., 2016) and the properties of non-steroidal anti-inflammatory drugs to predict the extraction recovery of NSAIDs from biological fluids set by solid-phase extraction (Milanetti et al., 2019). When amino acids solvation properties are studied, the main limitation of this method relied on considering a single amino acid in solution instead of inserting it in a functional protein chain. Moreover, the method was developed uniquely for the TIP4 water model, limiting its use to most molecular dynamics simulations (Babiaczyk et al., 2010).

Since the characteristics of the neighboring residues influence the hydropathy of the examined amino-acid, in this work we define the hydropathy properties of each amino acid taking into account the structural environment that surrounds it. In this way, we incorporate the effects of the own characteristics of each amino acid, as well as the chemical and structural properties induced by the surrounding environment.

Furthermore, the method has atomic resolution (Leopizzi et al., 2017), meaning that, given a protein, it is possible characterizing not only a single residue or a set of residues, but we can also quantify the hydrophobic and hydrophilic properties of a set of atoms that contribute to the formation of a portion of the molecular surface. This perspective is particularly important for the improvement of predictive methods of protein-protein interactions (Nicolau et al., 2014). In addition, we have also extended the method to other models of water molecules, especially those typically used for molecular dynamics simulations of proteins, enabling the application of our approach also to the trajectories of simulations already performed.

In particular, we have selected a representative set of experimentally solved protein structures and for each of them, we performed an extensive molecular dynamics simulation. We thus studied the hydropathy profile of the amino acid when they are in different protein structural environments, underlining that, especially for some residues, the solvation properties can sensibly differ according to the characteristics of the different neighborhoods. The analysis of our results allows us to define different regions in a plane describing the hydrophobicity and hydrophilicity properties: each residue belonging to the proteins in our dataset is a point on this plane and its position is not only due to its own chemical properties but also to the nature of the residues closest in structure.

The goodness of the characterization proposed here was evaluated considering the average positions of the residues on the two planes, classifying them by amino acids. These results are in perfect agreement with the hydrophobicity measurement of a biological experimental scale, which is considered the state of the art in this field (Hessa et al., 2005). Furthermore, the dispersion of the residue set for each amino acid was analyzed to underline how

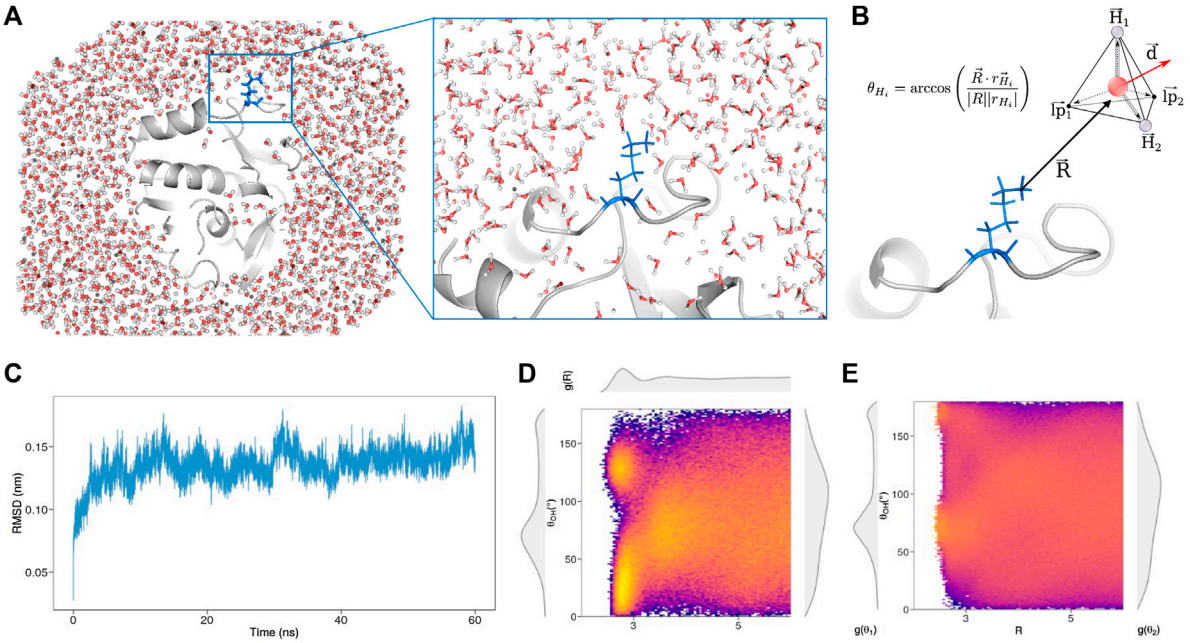


FIGURE 1 | (A) Snapshot taken from the molecular dynamics simulation of Concanavalin B (PDB id: 1CNV) performed with explicit solvent. The protein structure is represented in grey, while blue sticks (also zoomed on the right) highlight the position and orientation of an explicate residue, Lys 258, with respect to the surrounding water molecules **(B)** The disposition of each water molecule around a given residue is described representing each solvent molecule as a tetrahedron and evaluating the angles, θ, formed by each vertex of the tetrahedron with the vector, $\vec{R}$ that joins the nearest heavy atom of the residue with the water oxygen atom. The oxygen atom is used as both center of the tetrahedron and origin for the angle definition. The water dipole, $\vec{d}$ is depicted as a red arrow **(C)** Root mean square deviation (RMSD) of the protein-heavy atoms as a function of time, using the initial structure as a reference **(D)** Joined probability distribution, $P_j(R,\theta)$, of finding a water molecule at distance R from the residue with one angle, θ of the corresponding tetrahedron. Top, left and right side plots show the marginal probabilities, $P(R)$, $P_j(\theta_1)$, and $P_j(\theta_2)$, respectively **(E)** Conditional probability distribution, $P_c(\theta|R)$. Left and right side plots show the marginal probabilities, $P_c(\theta_1)$, and $P_c(\theta_2)$, respectively.

the nature of the residues belonging to the structural neighborhood has an important effect on the single residue characterization.

2 RESULTS AND DISCUSSION

2.1 Hydropathy Profile for Single Residue in a Specific Protein Environment

In this section we explain the idea we adopted for the calculation of the amino acid solvation properties, studying the distance and the orientation of water molecules with respect to a solute molecule. We investigated the hydropathy of residues in their natural environment, i.e. inserted in a functional and folded protein chain.

To do so, we selected 20 proteins of known structure from the dataset collected by Hensen *et al.* (Hensen et al., 2012) (see Methods for details), searching very different proteins in terms of structural features to make the analysis as general as possible. In this perspective, we analyzed the SCOP class (Andreeva et al., 2014; Andreeva et al., 2020; of each of the selected protein, demonstrating as our dataset covers several different folds and therefore ensuring the generality of our findings (See **Supplementary Table S1**). For each of these proteins, a molecular dynamics simulation of 60 ns was performed, studying the behavior of the explicit solvent molecules around the solute (**Figure 1A**), after the equilibration time (**Figure 1C**). To testify that we sampled configuration only after the equilibration in all the simulations we performed, we reported in Supporting Information the Root Mean Square Deviation and the Solvent Accessible surface as a function of time for all the proteins (See **Supplementary Figures S1–S2**).

We note that the explored time span allows us to well grasp the organization of surface waters, while much longer simulations would be needed to consider also the effect of structural water molecules.

According to our method, each solvent molecule can be schematized as a tetrahedron, with the water oxygen in the center and the vertices constituted by the two hydrogen atoms and the two lone pair electrons (**Figure 1B**), so as each water molecule can form up to four hydrogen bonds (HB). In particular, we associate any water molecule to the closest atoms of the solute focusing only on the first hydration shells, i.e. water molecules closer to any solute atoms than 6 *Å*. Since each water molecule is assigned to one solute atom, for each water molecule the solvent behavior is represented by three quantities representing the position and the orientation with respect to the solute: the distance R between the oxygen atom and the closest heavy atom of the solute, the *hydrogen bond angle* θ and the *dipole angle* ϕ. Each hydrogen bond angle is defined as the angle formed between the R and each vertex of the tetrahedron using the

oxygen atom as the origin. Similarly, the dipole angle is built using the vector R and the dipole moment $\vec{d}$ (see **Figure 1B** for a sketch). In this work, we focus on R and θ, since these quantities allow a complete characterization of the solute hydropathy. Indeed, a non-polar (hydrophobic) molecule in an aqueous solution interacts with the solvent only through van der Waals forces. Since the Coulombic interaction among H_2Os is strong, water molecules privilege their internal HBs contacts. Alternatively, the interplay between polar or charged molecules and solvent occurs mainly via Coulombic forces, attracting one of the hydrogens or one of the lone pair electrons toward the solvent atom. Therefore, when a hydrophobic solute is examined, water molecules place one of the faces of the tetrahedron toward the solute in order to leave all possible HBs available; on the contrary, a water molecule close to a polar or a charged solute reorients itself to point toward him one of its lone pairs or hydrogens.

In a nutshell, given the set of atoms composing an amino acid, we carry out statistical analysis of the orientations of the water molecules that hydrate them. In **Figure 1D** we show a colormap reporting the joint probability to observe a water molecule with a given R and θ in the surroundings of the Lys 258 belonging to Concanavalin B (PDB id: 1CNV). As we can see also from the marginal distributions on the panel sides, well-defined peaks reflect the solvation properties of the residue in the protein environment.

On top of **Figure 1D**, we report $P(R)$, the probability density distribution of finding a water molecule at a distance R from the solute, where j is the subscript indicating that the probability density is extracted from the joint probability. The curve is characterized by two maxima (this happens for almost all the amino acids), and it is, therefore, possible to identify the first and the second shell of hydration, after which there is the bulk water. On the right and left part of **Figure 1D** we show $P_j(\theta_1)$ and $P_j(\theta_2)$, the probability density distribution of finding a water molecule with a certain HB orientation in first or second shell respectively, that is having a R in the shells defining interval (see Methods).

It has been demonstrated that, in order to improve the resolution of the description of first and second solvation shells and to achieve a better characterization of the solute features, the adoption of the *conditional probabiliy* represent a powerful tool (Babiaczyk et al., 2010). Indeed in this formalism, we report the probability of having a certain θ, conditional on the solvent locating at a distance R from the solute atom (See Methods for further details). **Figure 1E** shows the colormap of the conditional probabilities related to Lys 258 and the corresponding probability densities will be indicated with the subscript c.

2.2 Joint and Conditional Probability for Residue Characterization

For each solvent-exposed residue in our dataset, we built an hydropathy profile juxtaposing their $P(R)$, $P_j(\theta_1)$ and $P_j(\theta_2)$. In this way, each residue is statically characterized by the positions and the orientations of the water molecules surrounding it during the simulation. We obtained a very interesting separation of the amino acid hydropathy by applying a Principal Component Analysis (PCA), where the system is rotated to go into the reference system which maximizes the variance of the data. In **Figure 2A**, we show the two principal components (percentage of explained variance equal to 88%): each point in this plot represents a given residue explored in its protein environment at physiological pH, and the 20 natural amino acids are colored differently. In particular, charged residues are colored in shades of blue, the non-charged polar residues in red while the hydrophobic residues are depicted in shades of yellow. Interestingly, PCA analysis reveals that residues with similar features are clearly grouped together. In particular, the negatively charged residues, Glu and Asp, form an isolated group, underlining their peculiar behavior in solvent interaction, while in the main cluster of residues each region is characterized by a preference for a certain type of residues.

We also performed a PCA analysis considering separately $P(R)$, $P_j(\theta_1)$ and $P_j(\theta_2)$. Results are reported in **Figures 2B–D** respectively. We can notice that the two PCA analyses gave very similar results. According to us, this could mean that, when the joint probability is used to build the profile, the dominating signal is related to the water molecules position, while the information about its orientation gets mainly overwhelmed.

To obtain a finer representation of the all water molecule "signals", we decided to use the conditional probability to amplify the angular aspect of the hydropathy profile.

To this aim, we performed the same PCA analysis using the $P_c(\theta_1)$ and $P_c(\theta_2)$ as obtained from the conditional probabilities together with p(R). The result is reported in **Figure 3A**. We can identify four macro-regions: the negatively charged (blue dots) amino acid region ($PC1 \simeq 1$, $PC2 \simeq 0.8$), the positively (cyan dots) charged region ($PC1 \simeq -1$, $PC2 \simeq -0.5$), hydrophobic (red dots) amino acid portion ($PC1 \simeq 0.8$, $PC2 \simeq -0.2$) and the polar non charged (yellow dots) residue zone ($PC1 \simeq -0.2$, $PC2 \simeq 0$).

Next, we performed hierarchical clustering of the residues based separately on the two angular density distributions (see **Figure 3B**). The high values achieved by the silhouette analysis (see **Figure 3C**) indicate that different subdivisions of residues are possible. For different types of groupings of residues, we note that both $P_c(\theta_1)$ and $P_c(\theta_2)$ are able to separate amino acids in several clusters composed of amino acids with different biochemical features.

It is worth noting that $P_c(\theta_1)$ well isolates a group of hydrophobic (red) residues from the charged residues (both the positively and negatively charged) but this separation is even more clear by using the $P_c(\theta_2)$ parameter.

2.3 Hydrophobic and Hydrophilic Properties of Amino Acid Side Chains in the Native Structure

The PCA plane we obtained using conditional probabilities (**Figure 3A**), is a schematic and meaningful description of the solvation properties of the amino acids when they are studied in the native environment. In fact, it is a clever representation of the

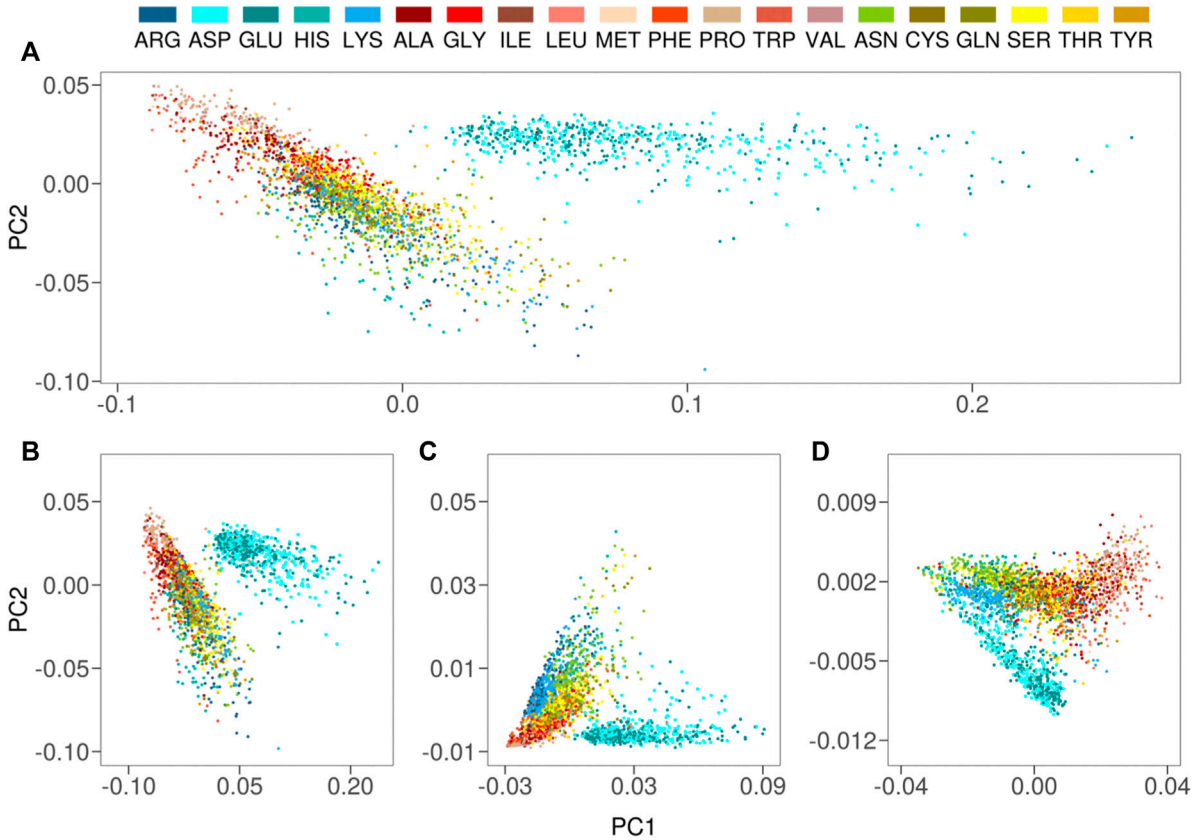


FIGURE 2 | (A) Projection along the first two principal components of the residues in the Protein dataset as obtained by a PCA analysis using $P(R)$, $P_j(\theta_1)$, and $P_j(\theta_2)$ as descriptors for each residue. Each dot in the plane represents a residue, with different colors corresponding to different amino acids **(B)** Same as **A)** but using only the $P(R)$s as descriptors for each residue **(C)** Same as **A)** but using only the $P_j(\theta_1)$s as descriptors for each residue **(D)** Same as **A)** but using only the $P_j(\theta_2)$s as descriptors for each residue.

behavior of the solvent molecules that hydrate protein residues. In **Figure 4** we depicted in the PCA plane the points regarding each of the 20 natural amino acids of our dataset using different colors. This way to measure hydropathy characteristics, reporting them as "explored regions" with different chemico-physical features by the amino acid rather than single values assumed by the molecule itself, allowed us to better illustrate the results we obtained. In fact, we demonstrate in this way that some amino acids explore peculiar regions in this plane while other amino acids like Arg, Tyr, Trp, and Thr, clearly populate overlapping regions of the plane. According to us, this may reflect the plasticity of some residues, to emphasize differently hydrophobic or hydrophilic aspects of their atomic structure in different protein local environments due to different biological contests. We summarize this concept of "hydropathy explored regions" in **Figure 4** where we defined four portions of the PCA plane according to the kind of residues that explores these areas. We identified the explored hydrophobic area ("Hb" area, depicted in red in **Figure 4**) in which Ile, Leu, Phe, Val, Pro, and Met residues are very well focused and in good qualitative agreement with previous hydrophobic scales. Then we mapped a clear negative charge explored area ("Neg" area depicted in cyan) where Asp and Glu clusterize. A third portion of PCA plane was defined as positive charge explored area ("Pos" area, depicted in blue in **Figure 4**) where almost all Lysines of our dataset well converge and Arginine side chain is present for half of the observed configurations; according to us, Lysine explores in few cases the Hb area probably due to the long aliphatic chain, that in some cases outweighs the hydrophilic character.

The presence of Arginine even in the Hb area is biologically very relevant because our result is connecting biological and biophysical principles of Arginine behavior in native proteins: this trend may be impossible to explain by using a just single hydropathy value. In fact, according to us, Arginine hydropathy can vary drastically within a protein environment and so we could define it as a Janus-headed side chain. This observation agrees with experimental data related to this amino acid. In fact, previous experiments by C. Preston Moon and Karen G. Fleming *et al.* (Moon and Fleming, 2011) clearly demonstrated that a membrane protein can accommodate an Arginine side-chain placed near the apolar middle of a lipid bilayer with much less cost in energy than has been previously predicted (Dorairaj and Allen, 2007; MacCallum et al., 2007). In fact, the guanidino group of Arginine could interact with non-polar aromatic and aliphatic side chains above and below the guanidinium plane

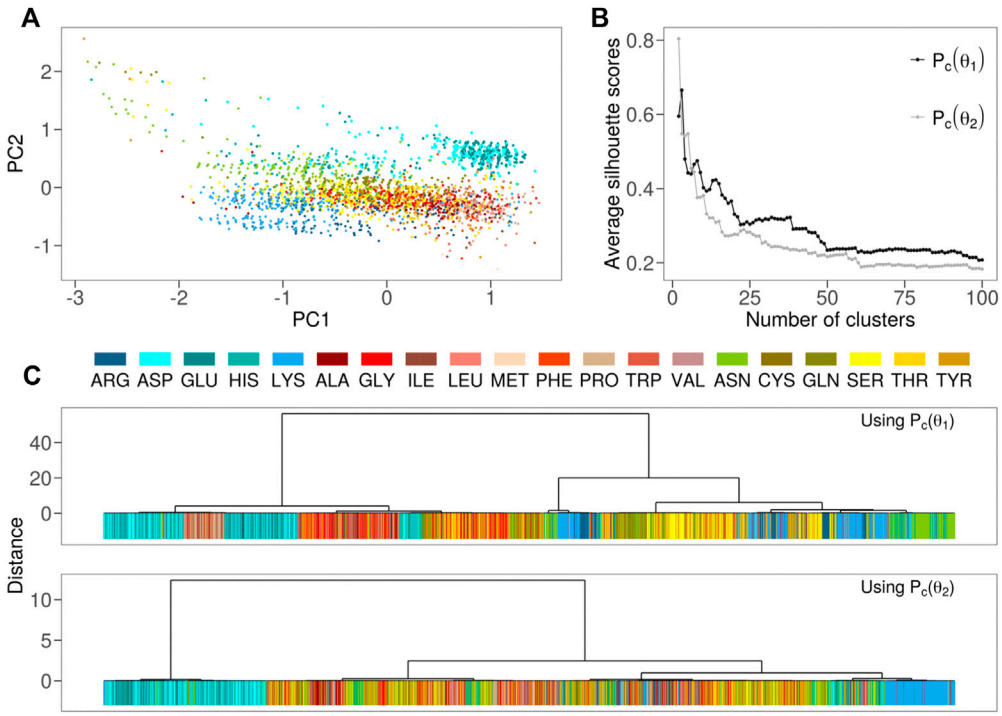


FIGURE 3 | (A) Projection along the first two principal components of the residues in the Protein dataset as obtained by a PCA analysis using $P(R)$, $P_c(\theta_1)$, and $P_c(\theta_2)$ as descriptors for each residue. Each dot in the plane represents a residue, with different colors corresponding to different amino acids **(B)** Cluster of the residue forming the Protein dataset using the $P_c(\theta_1)$ **(top)** or $P_c(\theta_2)$ **(bottom)** as descriptors for each residue **(C)** Average silhouette score as a function of the number of clusters considered in **(B)**.

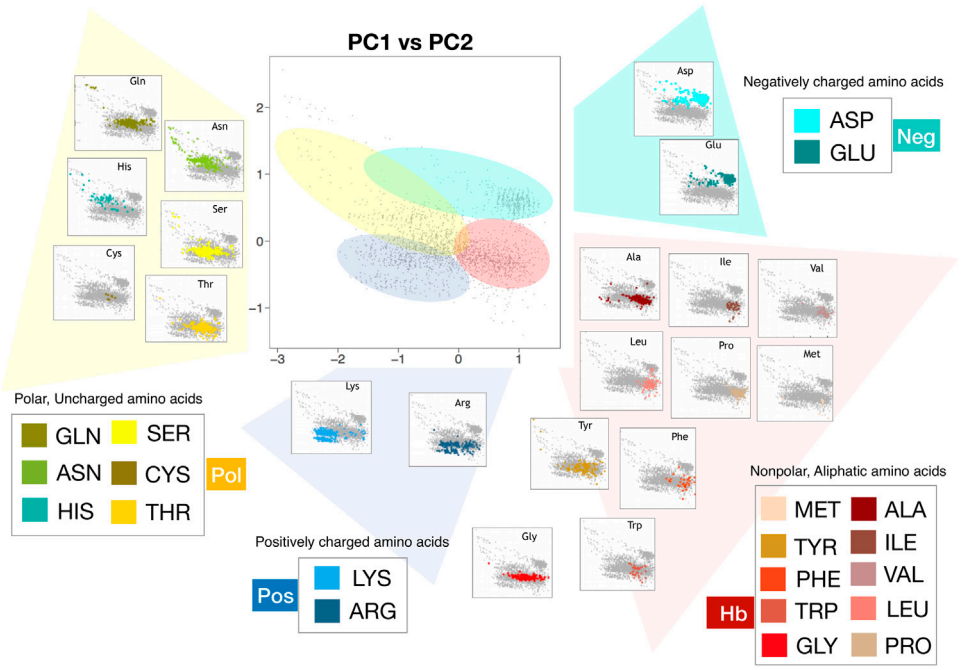


FIGURE 4 | Representations in the plane identified by the first and second principal components of all the residues comprising the 20 proteins of the Protein dataset (grey dots). The PCA analysis has been carried out using for each residue the observed $P(R)$, $P_c(\theta_1)$, and $P_c(\theta_2)$ computed as described in the Methods. In each panel, dots corresponding to the same kind of amino acid are highlighted with different colors.

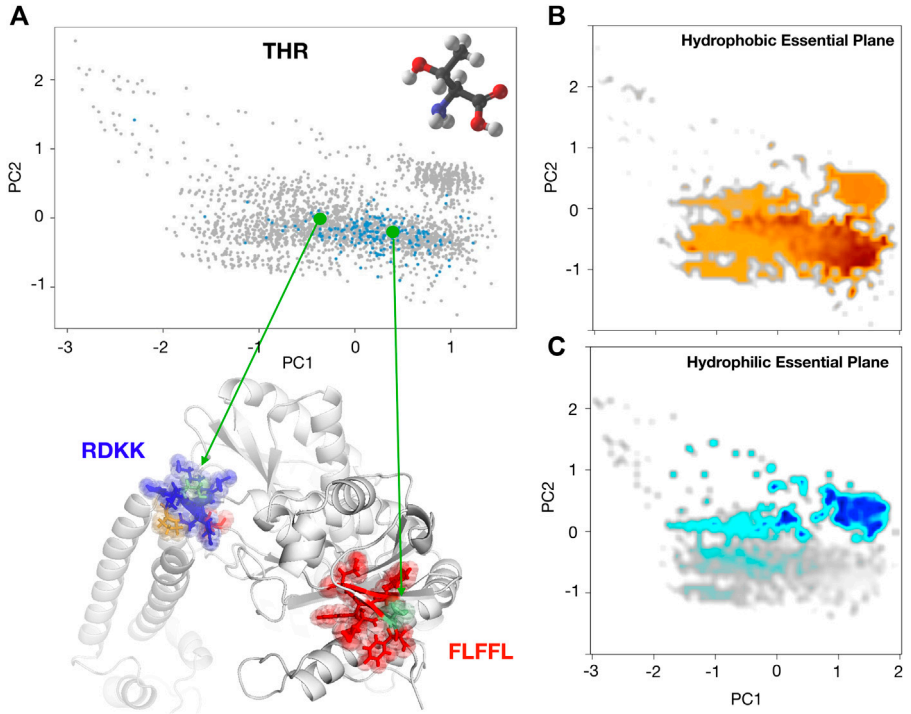


FIGURE 5 | (A) Projection along the first two principal components of the residues in the Protein dataset as obtained by a PCA analysis using $P(R)$, $P_c(\theta_1)$, and $P_c(\theta_2)$ as descriptors for each residue. Each dot in the plane represents a residue. Blue dots highlight the positions in the plane of all Threonine (Thr). Green dots correspond to two cases where the surroundings the considered Threonine is composed of hydrophilic (blue) or hydrophobic (red) amino acids. The average value of the hydrophobicity index **(B)** and of the hydrophobicity index **C)** of the WOPHS scale have been calculated for each element of the grid with which the plan is partitioned.

while hydrogen bonding with polar side chains is restricted to in-plane positions. Related to this point we would like to remember that the first solved structure of a voltage-gated potassium channel (Schow et al., 2011), gave rise to many discussions about the energetics of the interactions between Arginines and lipids, as the structure suggested a gating mechanism in which charged Arginines were exposed to the hydrophobic bilayer interior.

We further observed on the left side of the PCA plane and located between Neg and Pos areas, a region we defined polar explored region ("Pol" area, depicted in yellow in **Figure 4**) were polar, uncharged amino acids, at physiological pH, are positioned: the location of the area qualitatively agrees with the residue group features of these amino acids that are more hydrophilic than those of the Hb area because they contain functional groups that form hydrogen bonds with water. This class of amino acids includes Ser, Thr, Cys, Asp, and Gln. The presence of this polar area agrees with studies of Peters *et al.* about the assessment of the most accurate hydrophobicity scale (Peters and Elofsson, 2014). They demonstrated that better hydrophobic scales rank the polar amino acids Gln and (in particular) Asn as less hydrophobic. It is interesting to underline that even this polar area overlaps with the Hb area, in agreement with the concept of the ability of amino acids to explore several hydrophilicity-hydrophobicity regions.

To better point up this concept, we would like to report the case of the Threonine (**Figure 5A**) hydropathy analysis in two different contexts. We selected two Threonine residues, Thr 599 and Thr 302 both belonging to the same proteins (PDB:1xwl), characterized by different positions on the PCA plane. The reason for this different behavior in terms of solvent interaction has to be sought in the neighbor residues. In particular, the Thr within the polar region is surrounded by three charged residues (RDKK, reported in blue in the Figure) that inevitably influence his hydrophilic behavior; on the other hand, the Thr within the non-polar zone is enclosed in a set of non-polar residues (FLFFL, in red in the Figure), thus forming an overall hydrophobic region.

Another interesting example is represented by Threonine and Tryptophan. They are straddling the polar and hydrophobic areas and this behavior confirms that our approach is correct. In fact, Tryptophan and Tyrosine can be involved in interactions with ligands that contain aromatic groups via stacking interactions. However, tryptophan has nitrogen in its side chain and Tyrosine has oxygen, allowing hydrogen bonding interactions to be made with other residues or even solvent molecules, commonly seen in polar amino acids like Serine, which has oxygen in its side chain. But we should also keep in mind that Tryptophan has an indole function, but its lone pair of nitrogen is involved in the aromatic system. Thus, it makes only weal H-bonding, which could be not good enough to categorize as "polar". All these observations are in

TABLE 1 | Results of the analysis of the essential plane shown in **Figure 3A**. For each amino acid, we report the number of cases in which it is found solvent-exposed in simulation and the percentage with respect to all the solvent exposed residues (Occurrence); the hydrophobicity values we obtained with our geometrical characterization and the gyration radius, a measure of the dispersion of the points regarding each residues.

Res	Occurrence	H_r	Gyration radius
ALA	148 (5.4%)	0.47	0.57
ARG	168 (6.1%)	1.60	0.64
ASN	237 (8.6%)	2.03	0.68
ASP	300 (10.9%)	3.20	0.59
CYS	7 (0.3%)	0.59	0.20
GLN	205 (7.5%)	1.26	0.67
GLU	266 (9.7%)	3.32	0.55
GLY	166 (6.1%)	0.80	0.52
HIS	60 (2.2%)	2.29	0.76
ILE	38 (1.4%)	0.00	0.33
LEU	59 (2.2%)	0.79	0.41
LYS	268 (10.4%)	2.76	0.61
MET	12 (0.4%)	0.38	0.76
PHE	34 (1.2%)	0.24	0.52
PRO	97 (3.5%)	0.55	0.27
SER	254 (9.3%)	1.46	0.67
THR	197 (7.2%)	0.66	0.53
TRP	38 (1.4%)	0.65	0.37
TYR	129 (4.7%)	0.83	0.77
VAL	39 (1.4%)	0.46	0.35

agreement with the fact that Tyrosine and Triptophane side chains are the typical cases for which numerical values obtained for characterization of the hydrophobicity are controversial, being identified as hydrophobic in some studies (Levitt, 1976; Sweet and Eisenberg, 1983) but hydrophilic in others (Ooi et al., 1987; Oobatake and Ooi, 1988) and our concept of "explored region" should be the right approach.

At the end of this qualitative analysis, we decided to support our speculations by introducing also quantitative data relative to the side chain hydropathy characterization in the native protein context. Although it was not our aim, as proof of the significance of our hydrophobicity/hydrophilicity representation, we developed a mean hydrophobicity measure for each residue (H_r) (see methods for details). We achieved a very good agreement with the biological hydrophobicity scale (or the Hessa scale), which is based on *in vitro* experiments where the recognition of artificial helices by the Sec translocon was measured (Hessa et al., 2005). However, it can be noted that, in this case, the local microenvironment is not known. For example, residues in the helical segment might be interacting with other parts of the protein rather than interacting with lipids or water. The insertion by the translocon might also be a non-equilibrium process. In particular, in order to highlight the mean properties obtainable from this plot, we calculated the centroids of the points regarding each of the 20 natural residues. Using as reference the position of Isoleucine, indicated as the most hydrophobic residue here (Hessa et al., 2005), we calculate the radial and the angular distance of each centroid with the Isoleucine centroid (see Methods for details). In this framework, the higher is the distance with Isoleucine higher is the hydrophilicity of the residues. Notably, we found a strong linear correlation (R = 0.84) between the ΔG of amino acids side chains in the translocon scale and their values of mean hydrophobicity, H_r of our native-protein scale (**Figure 4** and in **Table 1**). meaning that our solvation analysis greatly reproduces one of the best performing hydrophobicity scales (Peters and Elofsson, 2014).

Indeed, it is interesting to note that even if the mean properties of the 20 residues can be successfully described using this representation, looking at the plots in **Figure 4** it emerges clearly that points belonging to the same amino acid category can spread a lot on this plane, meaning that even the same amino acid can be characterized by very different hydropathy when it is inserted in different environments. Quantitatively, as a measure of the dispersion of the points regarding the various residues, we calculated the amino acid gyration radius (see Methods). We report the results in **Table 1**.

It results that residues with a well known hydrophobic tendency, such as Proline, Isoleucine, Valine, experience a low variability since they repel water very strongly. On the other hand, residues with a less defined solvent preference, such as Asparagine, Tyrosine, Methionine, are characterized by higher gyration radius values, meaning that they can modify their features influenced by the surroundings.

In light of all these considerations, using the hydrophobicity and hydrophilicity scales presented here (Bonella et al., 2014), we built two maps of these characteristics on the conditional probabilities PCA plane reported in **Figure 3A**. In particular, by placing a square grid on it we can collect all the points inside each square pixel: since each of these points represents a residue with its hydrophobicity and hydrophobicity values, we can mediate these values obtaining a colormap with the hydrophobicity and hydrophilicity observed in that region of the plane. After a smoothing procedure, we obtain the maps depicted in **Figures 5B**,C. From this perspective, the evaluation of the hydropathy properties of a given amino acid, located in a specific protein sequence and structure, depends on the position it assumes on this plane, and this position surely depends on their own chemico-physical features but also on the characteristics of its structural neighborhood.

An additional analysis showing the correlation between the secondary structure of a residue and its hydration properties is reported in Supporting Information (See **Supplementary Figures S3–S5**). Using DSSP Touw et al. (2015); Kabsch and Sander. (1983) we labeled each residue with its secondary structure and we evaluated how the different secondary structures are located in the plane reported in **Figure 4**. It is worth noting that some non-polar residue, such as ALA and LEU, are usually characterized by a low value of the Hydrophobicity index, but when they are found in loops they can exhibit even high value of the index, probably because of the usual high solvent exposure of this secondary structure.

3 CONCLUSIONS

Investigating the properties of the hydrogen bond network at the interface between hydration water molecules and solute plays a

crucial role in the characterization of the physico-chemical properties of the latter. Here, we presented a completely *in-silico* method capable of analyzing the positions and the orientations of water molecules around any residue of protein structures. This allows us to emphasize the contribution to the solvation properties caused by the local structural environment, underlining that not only the nature of single amino acid determines its hydropathy features, but also the types of residues close to it.

In particular, we analyzed the motion of the water molecules belonging to the first two hydration shells for a set of proteins, defining a new description of both the hydrophilicity and hydrophobicity properties. Studying the probability of water molecule's orientation conditional to the distance to the solute, we built an essential plane of hydrophilicity and hydrophobicity, through a dimensionality reduction of the probability density distribution. On average, the location of each amino acid on this plane is in perfect agreement with its biochemical properties, in fact, an index defined considering the average position of each amino acid has an excellent correlation with one of the state-of-art hydrophobicity scales.

This notwithstanding, the dispersion of each amino acid (considering all the occurrences of a given residue in the proteins of our dataset) is a good marker of its variability in terms of solvation features. Indeed, this dispersion well classifies amino acids with marked properties, such as strong, from amino acids with less pronounced or intermediate hydropathy properties, meaning that the local structural environment in these cases plays a predominant role in modifying their interaction with the solvent.

4 MATERIALS AND METHODS

4.1 Protein Dataset and Residue Selection

We consider the dataset proposed by Hensen *et al.* (Hensen et al., 2012), where a collection of 112 representative proteins for each family were reported. From this initial set, we selected the 20 proteins, having 1) longer sequences and 2) no missing or incomplete residues. Considering all proteins together, we ended up with a total of 6,745 residues. For each protein, a molecular dynamics simulation with explicit solvent was performed. Since we were interested in characterizing solvation-related features, we consider only residues found in interaction with more than 50,000 water molecules during the whole analyzed simulation. An interaction between a residue and a water molecule is established if the distance between the oxygen atom of the water and any of the residue heavy atom is smaller than 6 *Å*. We ended up with 2,775 residues.

4.2 Molecular Dynamics Simulation

The following protocol was used for each of the 20 simulations. We used Gromacs 2020 (Spoel et al., 2005) and built the system topology using the CHARMM-27 force field (Brooks et al., 2009). The protein was placed in a dodecahedric simulative box, with periodic boundary conditions, filled with TIP3P water molecules (Jorgensen et al., 1983). We checked that each atom of the protein was at least at a distance of 1.1 nm from the box borders. The system was then minimized with the steepest descent algorithm. Next, a relaxation of water molecules and thermalization of the system was run in NVT and NPT environments each for 0.1 ns at 2 fs time-step. The temperature was kept constant at 300 K with v-rescale algorithm (Bussi et al., 2007); the final pressure was fixed at 1 bar with the Parrinello-Rahman algorithm (Parrinello and Rahman, 1980). LINCS algorithm Hess et al. (1997) was used to constraint h-bonds. A cut-off of 12 Å was imposed for the evaluation of short-range non-bonded interactions and the Particle Mesh Ewald method Cheatham et al. (1995) for the long-range electrostatic interactions. Finally, we performed 60 ns of molecular dynamics with a time step of 2 fs, saving configurations every 2 ps. We considered the last 20 ns (10,000 frames) for the analyzes.

4.3 Evaluation of Solvent-Residue Geometrical Descriptors

Molecular dynamics simulation data were used to characterize the geometrical disposition of the water molecules around protein residues. In particular, for each protein of the Protein dataset, we sampled 10,000 configurations (one each 2 ps) from the corresponding molecular dynamics simulation. For every water molecule in each frame, we evaluate the minimum distance, *R*, between the water oxygen and the heavy atoms of each protein residue.

Solvent molecules whose oxygen atom had a distance bigger than 6 Å were discarded from all subsequent analyses. All remaining water molecules were then assigned to their nearer residue, again on the basis of the distance *R*.

Then, for each water molecule, we build the tetrahedron having the oxygen atom as the center and the two hydrogen atoms occupying two of the four vertexes. In this way, we ensure that the tetrahedron is always well defined. We indicate with $\vec{r}_{H_1}$ and $\vec{r}_{H_2}$ the vectors originating in the tetrahedron center and pointing to the hydrogen atoms; while we refer to the vectors linking the center with the other to vertex as $\vec{r}_{\mathrm{lp}_1}$ and $\vec{r}_{\mathrm{lp}_2}$ (where lp stands for *lone pairs*). Finally, we define also the vector joining the nearest heavy atom of the residue with the oxygen atom of the water molecule, $\vec{R}$, and the dipole moment vector, $\vec{d}$ (see **Figure 1** for a sketch).

Once we know the set of six vectors $[\vec{R}, \vec{d}, \vec{r}_{H_1}, \vec{r}_{H_2}, \vec{r}_{\mathrm{lp}_1}, \vec{r}_{\mathrm{lp}_2}]$, we can compute the five angles that efficiently summerize the disposition of the water molecule with respect to the protein residue. In particular,

$$\theta_{H_i} = \arccos\left(\frac{\vec{R} \cdot \vec{r}_{H_i}}{|R||r_{H_i}|}\right), \tag{1}$$

and

$$\theta_{lp_i} = \arccos\left(\frac{\vec{R} \cdot \vec{r}_{lp_i}}{|R||r_{lp_i}|}\right), \tag{2}$$

with $i = 1, 2$ identify the orientation of the tetrahedron vertexes with respect to the direction identified by $\vec{R}$, while

$$\phi = \arccos\left(\frac{\vec{R}\cdot\vec{d}}{|R||d|}\right), \quad (3)$$

measures the orientation of the water dipole.

4.4 Joint and Conditional Probability

For each of the 2,775 residues, we computed the hydrogen joint probability, $P(R,\theta)$, which gives the probability of finding a water molecule with a given $\theta_{OH-lp} = \theta$ angle at distance R from the nearest heavy atom of the reside and dipole joint probability, $P(R,\phi)$, of finding a water molecule with a given ϕ angle at distance R. In both cases, the probabilities are computed discretizing the distance range 0–6 $\AA$ with steps of 0.05 Å, and the angular interval $0 - 180^\circ$ with a step of 1°. See **Figure 1** for an example.

From the joint probabilities, we obtained the distance marginal probability, $P(R)$ as

$$P(R) = \int d\theta P(R,\theta), \quad (4)$$

while we calculated the conditional probabilities as

$$P(\theta|R) = \frac{P(R,\theta)}{P(R)}. \quad (5)$$

Considering each residue as a reference, $P(R)$ encodes the overall probability of finding a water molecule at distance R from the reference. As one can see from **Figure 1**, the typical shape of probability is that of a damped sinusoidal function, showing a series of maxima (and minima) with decreasing amplitude. This behavior originates from the molecular interactions between the residue and the water molecules and those between water molecules. Water molecules tend to form a shell around the residue with a higher density of molecules in correspondence of the $P(R)$ maxima and a lower density in its minima.

Using the $P(R)$ profile, we identified the shells as follows:

- the first shell starts at R_0 the first non-null value of $P(R)$;
- the border between the first shell and the second (R_1) coincide with the minimum following the maximum in the first shell;
- the end of the second shell, R_2 coincides with the minimum following the maximum in the second shell.

When the $P(R)$ profile does not allow us to identify the minima, we add them according to their average values calculated on the respective residue.

Once the shells were identified, we calculated $P(\theta_1)$ and $P(\theta_2)$ as

$$P(\theta_i) = \int_{R_{i-1}}^{R_i} dR P(\theta|R), \quad (6)$$

where $i = 1, 2$ and θ can be either the hydrogen angle or the dipole one. The $P(\theta)$ were calculated on both the joint ($P_j(\theta_{1,2})$) and conditional ($P_j(\theta_{1,2})$) probability. Since some $P(R)$ exhibited an anomalous profile they were discarded from subsequent analyzes, reducing the dataset to 2,740 residuals. Ultimately, we obtained three descriptors for the conditional and joint probability histograms of the OH-lp and dipole angles: $P(R)$, $P_{j,c}(\theta_1)$ and $P_{j,c}(\theta_2)$. Analyses were performed using R standard libraries (R Core Team, 2020).

4.5 Principal Component Analysis and Clustering

Principal component analysis (PCA) was performed over 1) the vector obtained by concatenating the discretized (75 points) probability distribution $P(R)$ for each of the 2,740 residues; 2) on the vector obtained by concatenating the discretized (180 points) probability distribution $P_j(\theta_1)$ for each of the 2,740 residues; 3) on the vector obtained by concatenating $P_j(\theta_2)$ of all 2,740 residues and 4) using the vector obtained by concatenating together all the previous probabilities. We used the "prcomp" function of R software (R Core Team, 2020). The same procedure has been repeated also using $P(R)$ and the two conditional probability marginals, $P_c(\theta_1)$, and $P_c(\theta_2)$.

A clustering analysis was performed on the points on the first two components plane relating to $P_c(\theta_1)$ and $P_c(\theta_2)$ through a hierarchical clustering, using the Euclidean distance and the Ward method as linkage function (Ward, 1963) via the "hclust" function of the "Stats" package of R (R Core Team, 2020). Finally, we computed the Silhouette for the hierarchical cluster via the R package "cluster" (Maechler et al., 2019).

Finally, to measure the dispersion of the points regarding the various residues in the PCA plane, we calculated the amino acids gyration radius as

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i^2)}, \quad (7)$$

where r_i are the distances between each of the N points and the centroid.

4.6 Hydrophobicity Measure in Principal Component Plane

Starting from the plane shown in **Figure 3A**, we defined a measure of hydrophobicity. We take as reference the point C, the centroid of all the Ile points with coordinates PC1 = 0.75 and PC2 = −0.39. For a generic point in the plane, i, we calculated the distance d_i from C. Defining the angle variable α, like the one starting from the x-axis in an anticlockwise direction, we thus fixed a reference angle, $\alpha_{ref} = 2.8rad$. Now it is possible to define, for a generic point i on the plane with distance d_i and angle α_i, the Hydrophobicity index as follows:

$$H_r = d_i + k|\alpha_i - \alpha_{ref}|, \quad (8)$$

where $k = 2$.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

EM conceived research; LDR and MM designed and performed computational analysis. LB performed molecular dynamics simulations and statistical analysis. EM, DR, and GR supervised the research and performed the analysis. All authors analyzed results, wrote and revised the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.626837/full#supplementary-material.

REFERENCES

Aftabuddin, M., and Kundu, S. (2007). Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys. J.* 93, 225–231. doi:10.1529/biophysj.106.098004

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014). Scop2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42, D310–D314. doi:10.1093/nar/gkt1242

Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2020). The scop database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48, D376–D382. doi:10.1093/nar/gkz1064

Babiaczyk, W. I., Bonella, S., Guidoni, L., and Ciccotti, G. (2010). Hydration structure of the quaternary ammonium cations. *J. Phys. Chem. B* 114, 15018–15028. doi:10.1021/jp106282w

Ball, P. (2008). Water as an active constituent in cell biology. *Chem. Rev.* 108, 74–108. doi:10.1021/cr068037a

Bonella, S., Raimondo, D., Milanetti, E., Tramontano, A., and Ciccotti, G. (2014). Mapping the hydropathy of amino acids based on their local solvation structure. *J. Phys. Chem. B* 118, 6604–6613. doi:10.1021/jp500980x

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30, 1545–1614. doi:10.1002/jcc.21287

Brooks, C. L., Gruebele, M., Onuchic, J. N., and Wolynes, P. G. (1998). Chemical physics of protein folding. *Proc. Natl. Acad. Sci. United States* 95, 11037–11038. doi:10.1073/pnas.95.19.11037

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126, 014101. doi:10.1063/1.2408420

Cheatham, T. E. I., Miller, J. L., Fox, T., Darden, T. A., and Kollman, P. A. (1995). Molecular dynamics simulations on solvated biomolecular systems: the particle mesh ewald method leads to stable trajectories of DNA, RNA, and proteins. *J. Am. Chem. Soc* 117, 4193–4194. doi:10.1021/ja00119a045

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* 248, 338–339. doi:10.1038/248338a0

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105, 1–12. doi:10.1016/0022-2836(76)90191-1

Deber, C. M., Wang, C., Liu, L. P., Prior, A. S., Agrawal, S., Muskat, B. L., et al. (2001). Tm finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci.* 10, 212–219. doi:10.1110/ps.30301

Dorairaj, S., and Allen, T. W. (2007). On the thermodynamic stability of a charged arginine side chain in a transmembrane helix. *Proc. Natl. Acad. Sci. United States* 104, 4943–4948. doi:10.1073/pnas.0610470104

Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., and Grubmüller, H. (2012). Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLoS ONE* 7, e33931. doi:10.1371/journal.pone.0033931

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., et al. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433, 377–381. doi:10.1038/nature03216

Jones, D. D. (1975). Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch. *J. Theor. Biol.* 50, 167–183. doi:10.1016/0022-5193(75)90031-4

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi:10.1002/bip.360221211

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132. doi:10.1016/0022-2836(82)90515-0

Leopizzi, M., Cocchiola, R., Milanetti, E., Raimondo, D., Politi, L., Giordano, C., et al. (2017). IKKα inibition by a glucosamine derivative enhances Maspin expression in osteosarcoma cell line. *Chem. Biol. Interact.* 262, 19–28. doi:10.1016/j.cbi.2016.12.005

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104, 59–107. doi:10.1016/0022-2836(76)90004-8

Levy, Y., and Onuchic, J. N. (2006). Water mediation IN protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* 35, 389–415. doi:10.1146/annurev.biophys.35.040405.102134

MacCallum, J. L., Bennett, W., and Tieleman, D. P. (2007). Partitioning of amino acid side chains into lipid bilayers: results from computer simulations and comparison to experiment. *J. Gen. Physiol.* 129, 371–377. doi:10.1085/jgp.200709745

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). cluster: cluster Analysis Basics and Extensions. *R package version 2.1.0—for new features, see the "Changelog" file (in the package source)*. https://CRAN.R-project.org/package=cluster

Milanetti, E., Carlucci, G., Olimpieri, P. P., Palumbo, P., Carlucci, M., and Ferrone, V. (2019). Correlation analysis based on the hydropathy properties of non-steroidal anti-inflammatory drugs in solid-phase extraction (spe) and reversed-phase high performance liquid chromatography (hplc) with photodiode array detection and their applications to biological samples. *J. Chromatogr. A* 1605, 360351. doi:10.1016/j.chroma.2019.07.005

Milanetti, E., Raimondo, D., and Tramontano, A. (2016). Prediction of the permeability of neutral drugs inferred from their solvation properties. *Bioinformatics* 32, 1163–1169. doi:10.1093/bioinformatics/btv725

Miotto, M., Olimpieri, P. P., Di Rienzo, L., Ambrosetti, F., Corsi, P., Lepore, R., et al. (2018). Insights on protein thermal stability: a graph representation of molecular interactions. *Bioinformatics* 35, 2569–2577. doi:10.1093/bioinformatics/bty1011

Mondal, S., Mukherjee, S., and Bagchi, B. (2017). Origin of diverse time scales in the protein hydration layer solvation dynamics: a simulation study. *J. Chem. Phys.* 147, 154901. doi:10.1063/1.4995420

Moon, C. P., and Fleming, K. G. (2011). Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc. Natl. Acad. Sci. United States* 108, 10174–10177. doi:10.1073/pnas.1103979108

Moret, M., and Zebende, G. (2007). Amino acid hydrophobicity and accessible surface area. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 75, 011920. doi:10.1103/PhysRevE.75.011920

Nicolau, D. V., Paszek, E., Fulga, F., and Nicolau, D. V. (2014). Mapping hydrophobicity on the protein molecular surface at atom-level resolution. *PLoS One* 9, e114042. doi:10.1371/journal.pone.0114042

Oobatake, M., and Ooi, T. (1988). Characteristic thermodynamic properties of hydrated water for 20 amino acid residues in globular proteins. *J. Biochem.* 104, 433–439. doi:10.1093/oxfordjournals.jbchem.a122485

Ooi, T., Oobatake, M., Némethy, G., and Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. United States* 84, 3086–3090. doi:10.1073/pnas.84.10.3086

Parrinello, M., and Rahman, A. (1980). Crystal structure and pair potentials: a molecular-dynamics study. *Phys. Rev. Lett.* 45, 1196–1199. doi:10.1103/physrevlett.45.1196

Peters, C., and Elofsson, A. (2014). Why is the biological hydrophobicity scale more accurate than earlier experimental hydrophobicity scales? *Proteins* 82, 2190–2198. doi:10.1002/prot.24582

Qvist, J., Persson, E., Mattea, C., and Halle, B. (2009). Time scales of water dynamics at biological interfaces: peptides, proteins and cells. *Faraday Discuss* 141, 131–207. doi:10.1039/b806194g

R Core Team (2020). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raschke, T. M. (2006). Water structure and interactions with protein surfaces. *Curr. Opin. Struct. Biol.* 16, 152–159. doi:10.1016/j.sbi.2006.03.002

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* 229, 834–838. doi:10.1126/science.4023714

Rothschild, L. J., and Mancinelli, R. L. (2001). Life in extreme environments. *Nature* 409, 1092–1101. doi:10.1038/35059215

Schow, E. V., Freites, J. A., Cheng, P., Bernsel, A., Von Heijne, G., White, S. H., et al. (2011). Arginine in membranes: the connection between molecular dynamics simulations and translocon-mediated insertion experiments. *J. Membr. Biol.* 239, 35–48. doi:10.1007/s00232-010-9330-x

Sweet, R. M., and Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* 171, 479–488. doi:10.1016/0022-2836(83)90041-4

Tanford, C. (1972). Hydrophobic free energy, micelle formation and the association of proteins with amphiphiles. *J. Mol. Biol.* 67, 59–74. doi:10.1016/0022-2836(72)90386-5

Tarek, M., and Tobias, D. J. (2000). The dynamics of protein hydration water: a quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. *Biophys. J.* 79, 3244–3257. doi:10.1016/S0006-3495(00)76557-X

Tomobe, K., Yamamoto, E., Kojić, D., Sato, Y., Yasui, M., and Yasuoka, K. (2017). Origin of the blueshift of water molecules at interfaces of hydrophilic cyclic compounds. *Sci. Adv.* 3, e1701400. doi:10.1126/sciadv.1701400

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., et al. (2015). A series of pdb-related databanks for everyday needs. *Nucleic Acids Res.* 43, D364–D368. doi:10.1093/nar/gku1028

Vagenende, V., and Trout, B. L. (2012). Quantitative characterization of local protein solvation to predict solvent effects on protein structure. *Biophys. J.* 103, 1354–1362. doi:10.1016/j.bpj.2012.08.011

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi:10.1002/jcc.20291

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi:10.1080/01621459.1963.10500845

Wilce, M. C. J., Aguilar, M.-I., and Hearn, M. T. W. (1995). Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from rp-hplc of peptides. *Anal. Chem.* 67, 1210–1219. doi:10.1021/ac00103a012

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

frontiers in Molecular Biosciences

TECHNOLOGY AND CODE
published: 07 April 2021
doi: 10.3389/fmolb.2021.617176

BIO-GATS: A Tool for Automated GPCR Template Selection Through a Biophysical Approach for Homology Modeling

*Amara Jabeen[1], Ramya Vijayram[2] and Shoba Ranganathan[1]**

[1] Department of Molecular Sciences, Macquarie University, Sydney, NSW, Australia, [2] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India

G protein-coupled receptors (GPCRs) are the largest family of membrane proteins with more than 800 members. GPCRs are involved in numerous physiological functions within the human body and are the target of more than 30% of the United States Food and Drug Administration (FDA) approved drugs. At present, over 400 experimental GPCR structures are available in the Protein Data Bank (PDB) representing 76 unique receptors. The absence of an experimental structure for the majority of GPCRs demand homology models for structure-based drug discovery workflows. The generation of good homology models requires appropriate templates. The commonly used methods for template selection are based on sequence identity. However, there exists low sequence identity among the GPCRs. Sequences with similar patterns of hydrophobic residues are often structural homologs, even with low sequence identity. Extending this, we propose a biophysical approach for template selection based principally on hydrophobicity correspondence between the target and the template. Our approach takes into consideration other relevant parameters, including resolution, similarity within the orthosteric binding pocket of GPCRs, and structure completeness, for template selection. The proposed method was implemented in the form of a free tool called Bio-GATS, to provide the user with easy selection of the appropriate template for a query GPCR sequence. Bio-GATS was successfully validated with recent published benchmarking datasets. An application to an olfactory receptor to select an appropriate template has also been provided as a case study.

Keywords: biophysical approach, hydrophobicity correspondence, template selection, homology modeling, GPCR, olfactory receptor, automated tool

OPEN ACCESS

Edited by:
Alfredo Iacoangeli,
King's College London,
United Kingdom

Reviewed by:
Tomasz Stepniewski,
Institute of Metallurgy and Materials Science (PAN), Poland
Tommaso Biagini,
Casa Sollievo della Sofferenza (IRCCS), Italy

****Correspondence:***
Shoba Ranganathan
shoba.ranganathan@mq.edu.au;
shoba.ranganathan1@gmail.com

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *14 October 2020*
Accepted: *24 February 2021*
Published: *07 April 2021*

Citation:
Jabeen A, Vijayram R and Ranganathan S (2021) BIO-GATS: A Tool for Automated GPCR Template Selection Through a Biophysical Approach for Homology Modeling. Front. Mol. Biosci. 8:617176. doi: 10.3389/fmolb.2021.617176

INTRODUCTION

The three-dimensional (3-D) structure of the proteins is important for deciphering its biological function and gaining mechanistic insights of biological events. Analyzing the relationship between sequence, structure, and function between proteins might help in transferring functional annotation between proteins. Cyrus Chothia's contribution in incorporating computational approaches for a sequence-structure relationship, such as the development of Structural Classification of Proteins (SCOP) database (Lo Conte et al., 2000), has opened up

new avenues for structural bioinformatics. The hierarchical division of proteins into classes, folds, superfamilies, and families based on structural and functional similarities by SCOP has enabled linking of known protein structures with homologous sequences lacking a known structure. Distant homologies can also be tracked through the SCOP database (Redfern et al., 2008). The use of homolog structures for generating the structural model of a protein lacking experimental structure forms the basis of homology modeling. The success of the homology model is greatly determined by the selected template and the alignment generated between the target and the template (Wallner and Elofsson, 2005; Haddad et al., 2020). In this article, we have developed a graphical user interface for selecting suitable templates for GPCRs. Our biophysical method for GPCR template selection is based primarily on hydrophobic correspondence (HC) between the target and the template, inspired by the work of Cyrus Chothia on the conceptual methods for hydrophobicity determination (Chothia, 1976).

G protein-coupled receptors, also known as seven transmembrane (TM) domain receptors, constitute the largest family of cell surface receptors with above 800 members in humans. All GPCRs share a common architecture of seven TM helices connected through three extracellular (ECL 1–3) and three intracellular (ICL 1–3) loops with an extracellular amino (N-) terminus and intracellular carboxyl (C-) terminus (Miyagi et al., 2020). The most common classification system used for GPCRs is based on sequence and functional similarities. This schema classifies GPCRs into six classes, *viz.* class A (rhodopsin-like family), class B (secretin family), class C (metabotropic glutamate family), class D (fungal mating pheromone receptors), class E (cyclic adenosine monophosphate or cAMP receptors), and class F (frizzled/smoothened receptors). All classes of GPCRs govern myriad functionalities within the human body, ranging from sensory perception (smell, taste, vision) to neurotransmission, metabolism, immune response, blood pressure regulation, and cognition (Hu et al., 2017). GPCRs recognize diverse ligands including peptides, hormones, odorants, tastants, vitamins, photons, ions, and metabolites, among others (Wacker et al., 2017). The extracellular ligands bind to the inactive GPCRs and bring about a conformational change to the helical bundle, which in turn activates intracellular transducers such as G-proteins, or β-arrestins. The intracellular transducers are connected to the helical bundle through ICL3. Therefore, GPCRs exhibit multiple conformational states, with the active and inactive states being the predominant ones (Miyagi et al., 2020).

Dysfunction of GPCR signaling leads to pathological conditions within the human body, making GPCRs the largest druggable protein family. More than 34% of FDA approved drugs target GPCRs (Saikia et al., 2019). Currently, only ∼15% of the GPCRs are targeted. This under-representation is mainly due to the absence of known ligands for more than 30% of non-olfactory GPCRs (Insel et al., 2019). Virtual ligand screening coupled with experimentation has resulted in the discovery of novel ligands for numerous GPCRs (Congreve et al., 2020). Both ligand-based virtual screening (LBVS), as well as structure-based virtual screening (SBVs), have been used in finding novel ligands for GPCRs. LBVS can only be applied to the receptors having known ligands. Machine learning-based methods for LBVS are becoming popular for expanding the ligand set of the receptor with a large number of known ligands (Butkiewicz et al., 2019; Jabeen and Ranganathan, 2019). SBVS has also been used to find novel ligands for GPCRs (Congreve et al., 2020) but unfortunately, only 91 GPCRs have experimentally resolved structures to date, according to GPCRdb statistics (Munk et al., 2019) (as of 05.01.2021) with over 500 structures deposited in the Protein Data Bank (PDB) (Berman et al., 2000). This sequence to structure gap is mainly because of the challenges associated with structure determination of GPCRs (Baker et al., 2017; Jabeen et al., 2019a). Among the challenges are difficulties in heterologous expression, lower stability, maintaining the structural integrity by embedding into the membrane-like environment, and the existence of multiple conformations (Miyagi et al., 2020). The booming period for GPCR structural biology started in 2000 when the first GPCR structure (bovine rhodopsin) was resolved (Palczewski et al., 2000). Due to continuous improvement in structural biology methods, experimentally resolved GPCR structures are increasing but they are still under-represented compared to soluble, globular proteins. Experimental structures are now available for all classes except E (Munk et al., 2019). Most of the experimentally resolved structures belong to GPCR class A. Consequently, most of the available drugs in the market target class A receptors (Basith et al., 2018).

Homology modeling could be used for structure-based drug design (SBDD), in the absence of an experimental structure, as it is more reliable than *ab initio* modeling (Nikolaev et al., 2018). To assess the accuracy of GPCR structural model predictions, community-wide GPCR Dock competitions are conducted. Scientific research groups from all over the world are given the GPCR target sequences for blind structure prediction, with undisclosed 3D structures. The predicted models along with their atomistic interactions with pharmaceutically important small molecules, are then ranked based on the experimentally resolved structures (Kufareva et al., 2014). These competitions have shown that homology models are able to impart valuable insights into receptor-ligand interactions, especially when sequence identity between target and the template exceeds 35% (Alfonso-Prieto et al., 2019). In fact, ligand screening against dopamine D_3 receptor was conducted initially using a homology model and provided results comparable to the experimental receptor structure (Carlsson et al., 2011).

Homology modeling of GPCRs poses several challenges, with template selection being the most prominent one. This is due to the unavailability of a close structural template for many GPCRs and limited representation of structures in active and intermediate conformations. Active structures are available for 47 receptors from classes A, B1, C, D, and F, and the structures for 20 receptors (classes A, C, and B1) are present in intermediate conformation. Also, 63 receptors are present in inactive conformation (classes A, B1, C, and F).

The accuracy of homology models is largely dependent on the choice of the template structure (Rataj et al., 2014). There are

a number of servers designed specifically for GPCR homology modeling, such as GPCR-I-TASSER, GPCR Online MOdeling and DOcking server (GOMoDo) (Sandal et al., 2013), GPCR-Sequence-Structure-Feature-Extractor (SSFE) (Worth et al., 2017), GPCR-ModSim (Esguerra et al., 2016), and GPCRM (Miszta et al., 2018). The process of template selection varies among each server. GPCR-I-TASSER uses a local meta-threading server (LOMETS) (Zheng et al., 2019) to select templates for a particular GPCR. LOMETS uses eleven different threading programs (CEthreader, FFAS3D, HHpred, HHsearch, MUSTER, Neff-MUSTER, PPAS, PRC, PROSPECT2, SP3, and SparksX) to select templates for a GPCR target. GOMoDo uses the HHsearch protocol to select the template for a query GPCR sequence. The user can either use the server-generated alignment, supply their own alignment, or use a previously stored alignment for GPCR homology model building. GPCR-SSFE selects the template based on the sequence-structure profile generated by HMMER2. The webserver provides a TM-wise template suggestion. It uses 27 GPCR structures as templates. The server-generated alignment is used for model building within GPCR-SSFE. The GPCR-ModSim server uses a set of 33 structures (22 inactive, eight intermediate, and three active) and a GPCR query sequence to generate the profile alignment and then selects the suitable templates. The templates for a specific region can be also selected by the user. The server-generated alignment, as well as a manually edited alignment, can be used for model building. The GPCRM server uses sequence identity calculated by ClustalW2 for selecting the template structures. Single or multiple templates may be selected, depending upon the sequence identity between the query and the template. The server also provides the feature of selecting the template based on the user's choice. The user can also opt for inactive or active templates. The set of templates include 63 inactive and 31 active GPCR structures.

Numerous benchmarking studies have been conducted by incorporating global and local similarity measures to select the appropriate template for GPCRs. Models based on local similarity measures have produced better results in virtual screening experiments (Castleman et al., 2019; Szwabowski et al., 2020). Multiple studies have shown that sequence identity above 30% could result in good GPCR homology models (within 3 Å) (Shahaf et al., 2016; Loo et al., 2018; Jaiteh et al., 2020). But most of the GPCRs share low sequence identity with available templates. It is also known from the literature that models based on greater sequence identity are not always the best ones and models based on distant homologs have performed well in virtual screening experiments (Rataj et al., 2014; Perry et al., 2015). Therefore, additional measures other than sequence identity must be considered for appropriate template selection. Also, a detailed inspection of all available homolog structures is essential for finding an optimal template, rather than randomly selecting a template based on the closest homolog, to generate better homology models (Kosinski et al., 2013). Sequences with similar hydrophobic patterns are often homologs, resulting in hydrophobicity being used in determining even distant homologs (Lolkema and Slotboom, 1998; Silva, 2008). The consideration of hydrophobic information for GPCR model building enables the representation of functional aspects as well (Crasto, 2010).

We proposed a biophysical approach recently for GPCR template selection (Jabeen and Ranganathan, 2020), which was applied to an olfactory receptor (OR), based on hydrophobicity correspondence (HC), the resolution, completeness of structures (or query coverage), and similarity between the residues within the orthosteric binding pocket for GPCRs (hotspot residues). Bio-GATS presents a GUI for template selection of GPCRs, based on this biophysical approach (**Figure 1**). Ligand profiles for selected templates and the target can be compared to get an optimal template. Further incorporation of mutagenesis data while refining the binding pocket of the model might help in improving the overall model.

As a case study, we have selected OR1A1, a human OR, as a query sequence. ORs are the largest superfamily of GPCRs and have no known experimental structure. Only 30 of 405 human ORs are currently known as proteins, with the rest regarded as "missing" proteins on account of insubstantial proteomic evidence (Jabeen et al., 2019a). ORs share low sequence identity with available GPCR structures. Therefore, it is challenging to get a reliable homology model for any OR. OR1A1 is ectopically expressed in gut enterochromaffin cells and proposed to be involved in serotonin release (Braun et al., 2007). Also, OR1A1 is known to be ectopically expressed in HepG2 liver cells where it is responsible for hepatic triglyceride metabolism modulation (Wu et al., 2015).

MATERIALS AND METHODS

Bio-GATS is written in Python 3 programming language (Van Rossum and Drake, 2011). The interface was built using PyQt5. The computing was performed through pandas. The numPy library was utilized for mathematical tasks. Biopython (Cock et al., 2009) was used for running BLAST (Altschul et al., 1990) locally through the command line, and for aligning the query sequence with that of the template. The HC plots were visualized using matplotlib. The hydrophobicity moment was calculated and plots were visualized using modlAMP package (Müller et al., 2017). A downloadable result summary file, from which images and data can be easily extracted, is generated in Microsoft (MS) Word format, using the docx library.

Bio-GATS requires Python, Biopython and also local BLAST to be installed locally to align sequences and then calculate the sequence identity values. Bio-GATS is linked to the GPCR dataset stored in an MS Excel file, which can be updated locally, as new GPCR structures are solved. The template selection process is divided into three steps: TM splitting and alignment, HC calculation, and finally, sequence similarity calculation among hotspot residue positions within the target and the template.

Also, a scoring matrix has been defined to rank the templates. The final score of the template is calculated based on resolution, the HC score, and binding site (or hotspot) residue similarity (BRS) score.

GPCR Dataset

The dataset used by Bio-GATS comprises GPCR sequences, available GPCR structural templates, TM definition of each

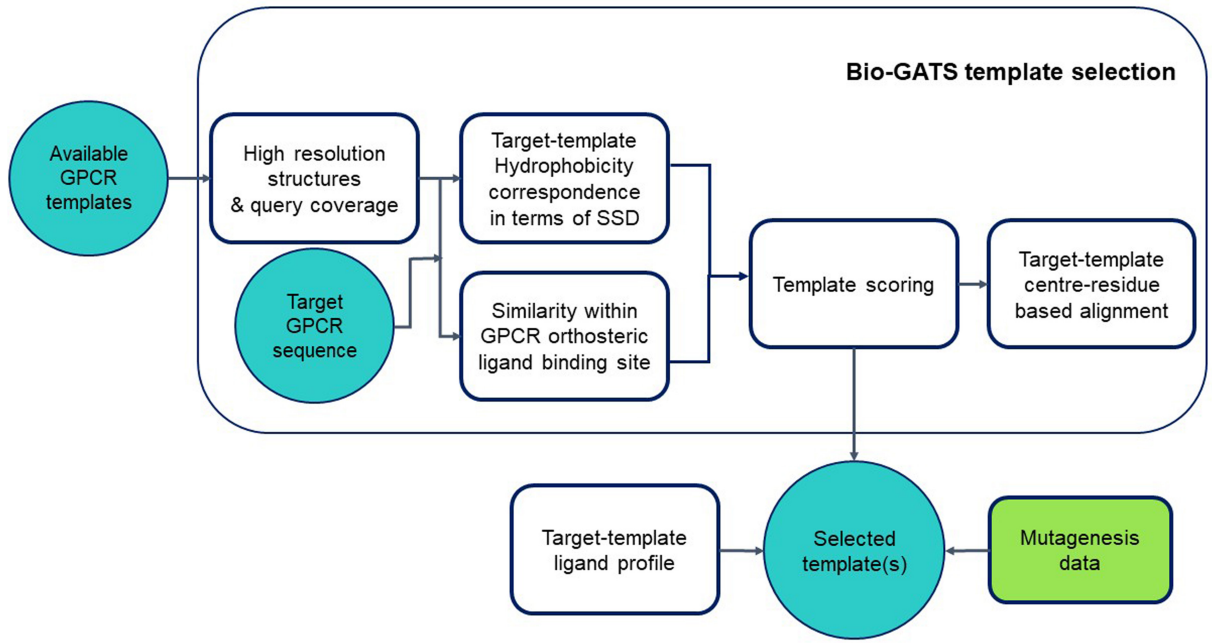


FIGURE 1 | Workflow for GPCR template selection through a biophysical approach, with details of how templates are selected and the alignment is generated.

entry and structure resolution, conformation, and positions having structural information for each of the available templates (query coverage). The data for available GPCR structures were downloaded from GPCRdb. It contains 76 unique receptors and over 400 PDB entries (as of 05.08.2020). The resolution of GPCR structures varies from 1.7 to 7.7 Å. Some GPCRs are over-represented, with 52 different structures of variable resolution available for bovine rhodopsin (UniProtKB OPSD_BOVIN) followed by 49 structures for human adenosine receptor A2a (AA2AR_ HUMAN). The data for 814 GPCR sequences and their TM definitions were taken from the published GpcR Sequence-Structure (GRoSS) alignment (Cvicek et al., 2016).

TM Splitting and Alignment

During the first step, the sequence of each TM was retrieved after splitting the sequence of both target and template according to the TM definitions taken from the GRoSS alignment. The corresponding TMs of target and template were then aligned together by tethering the center residues of each helix, as adopted by several groups (Wolf et al., 2017; Abaffy et al., 2018). The center residue for each helix is labeled as X.50 (X being the TM number), according to *Ballesteros–Weinstein* numbering scheme (Ballesteros and Weinstein, 1995).

Hydrophobicity Profile Generation

The hydrophobicity profile for each helix was generated using the Eisenberg scale (Eisenberg et al., 1984), as detailed in our recent publication (Jabeen and Ranganathan, 2020) and briefly outlined here. A moving window of size 11 was set up as suggested for the identification of putative transmembrane α-helices (Wallace et al., 2004). The average value over all the residues in a window was taken and ascribed to the center residue of the window. We then measured the HC between each aligned helix of the target and the template. The HC is represented as the sum of squared differences (SSD) (eq. 1 and eq. 2):

$$H_n = \sum_{i=n-5}^{n+5} h_i/_{11} \quad (1)$$

$$SSD = \sqrt{\sum_{n=1}^{N} (H_{template,n} - H_{target,n})^2} \quad (2)$$

where H_n is the calculated hydrophobicity for the aligned template-target residue in the nth position of the alignment and h_i is the hydrophobicity of the ith residue from the Eisenberg scale. The value, is normalized by dividing with the total number of residues in a particular helix, as the SSD value is length dependent and will only be relevant if a per-residue value is considered.

Calculating Sequence Similarity Between Hotspot Residues Known for GPCRs

We have taken the 24 traditional orthosteric ligand binding positions observed in most of the available GPCR structures. The positions are labeled according to *Ballesteros–Weinstein* numbering scheme and include 3.28, 3.29, 3.32, 3.33, 3.36, 3.37, 4.52, 5.39. 5.40, 5.43, 5.44, 5.47, 5.53, 6.44, 6.48, 6.51, 6.52, 6.55, 6.58, 7.31, 7.34, 7.38, 7.41, 7.42 (Chan et al., 2019). The similarities between these hotspot residues among the target-template pairs were computed using GPCRtm scoring matrix,

designed specifically for GPCRs considering the compositional bias of hydrophobic TM regions (Rios et al., 2015).

Target-Template Scoring

Each of the selected templates is scored based on two parameters: the HC-score and the BRS score (Munk et al., 2019). For each aligned helix, if the SSD per residue is between 0 and 0.1, 2 is added to the HC-score, while for SSD per residue >0.1, 1 is subtracted from the HC-score. This scheme is adapted from the BLAST match and mismatch scoring scheme and provides significant weighting for hydrophobicity. The overall HC-score is computed for each target-template pair using eq. 3,

$$HC\text{-}score = S_h = \sum_{i=1}^{7} s_i \quad (3)$$

where S_h is the overall hydrophobicity correspondence score ranging from helix 1 to 7, and s_i is the SSD per residue per helix. S_b is computed through GPCRtm matrix, S_r is the resolution score. If the resolution is $\leq$ 2.5 Å, the value for S_r is 1, otherwise it is 0. The total score S_t is computed by eq. 4.

$$S_t = S_h + S_b + S_r \quad (4)$$

S_h can attain a maximum value of 14 while S_b may exceed 70, depending upon the score computed by GPCRtm. To avoid biases, we normalized both S_h and S_b between 0 and 1 and computed the ranking score, S_{rank} for ranking the top three templates while searching for templates, using eq. 5,

$$S_{rank} = S_h^n + S_b^n + S_r \quad (5)$$

where S_{rank} is the total score between the target-template pair, S_h^n is the normalized HC-score, S_b^n is the normalized BRS score and S_r is the resolution score, retained from eq. 4.

Homology Modeling

Bio-GATS provides a complete alignment that was used to build a 3-D structural model for SBVS using Modeller 9.18 (Webb and Sali, 2017) by a previously established protocol for GPCR homology modeling (Jabeen et al., 2019b). The sequence alignment between the target and the template can be manually adjusted using MEGA7 (Kumar et al., 2016) by tethering center residues, class A GPCR conserved motifs, and cysteine residues forming a disulphide bridge. Bio-GATS uses predicted transmembrane regions from the GRoSS sequence alignment of all known GPCRs sequences (Cvicek et al., 2016). The ligand of each template was initially copied to the 3-D model and removed later to create an empty binding pocket within the query model structure for the OR1A1 case study.

Molecular Docking

For OR1A1, molecular docking of ligands was performed with ICM software (Abagyan et al., 1994). The binding pocket was predicted though ICMPocketFinder (An et al., 2005) and selected based on the available mutagenesis data for all ORs (Jabeen et al., 2019a).

RESULTS AND DISCUSSION

Bio-GATS has been tested on multiple computers, running on Linux as well as Windows platforms, and found to run successfully with the required dependencies installed. To validate our approach, we applied it to recent target-template datasets from published benchmarking studies and compared the results. We also considered representative receptors from each class (A, B, C, D, and F) with known experimental structure and built their models on the basis of templates selected by Bio-GATS. The models were then compared with the cognate experimental structures by calculating their root mean square deviation (RMSD) values. Further, we carried out a case study using an ectopically expressed olfactory receptor, OR1A1. We used the best templates from our approach, to build the models for OR1A1, which were validated by molecular docking with known ligands of the receptor, to check for retrieval of mutagenesis data important for ligand binding.

Performance of Bio-GATS on Published Benchmarking Datasets

To assess the performance of Bio-GATS, we collated the already published target-template pairs used in benchmarking studies and/or virtual ligand screening (VLS) runs. The best benchmarked modeling pair choices, as well as pairs which did not perform well, were considered for the analysis. The performance of the templates was ranked as good or bad, in published studies, on the basis of good ligand enrichment in VLS (Perry et al., 2015; Loo et al., 2018; Jaiteh et al., 2020), local and global (RMSD) from crystal structures (Castleman et al., 2019), and both ligand enrichment and RMSD from the crystal structure (Shahaf et al., 2016). Researchers have compared varied parameters in these studies among the target-template pairs, including global sequence identity, TM-wise sequence identity, local sequence identity (identity within the binding pocket), model refinement through molecular dynamics and/or induced-fit docking, and the ligand binding site plasticity. These parameters were applied to classify templates as good or bad in their publications.

We applied our approach to these selected target-template pairs and compared the results of published studies and our approach. A total of 28 target-template pairs for nine different GPCR targets belonging to class A and published within last 5 years were considered for comparison. We calculated S_t for each target-template pair. All target-template pairs rankings in the benchmarking studies corresponded to the numerical S_t values (**Table 1** and **Supplementary Table 1**). The top S_t scores for each target was ranked "good" in the benchmarking studies.

It was also evident from the collected dataset that high sequence identity does not always imply a good HC.

PAR2_HUMAN shows good HC with both PAR1_HUMAN and OPRX_HUMAN, in accord with the VLS results (Perry et al., 2015), although it is closer to PAR1_HUMAN (sequence identity: 41%) than to OPRX_HUMAN (sequence identity: 28%). There are many instances where good HC is observed among the target-template pairs even the sequence identity falls below 30% (**Supplementary Table 1**).

Also, sequence-structure correlation is not always implied according to the published studies, for instance, the model of $P2Y_{12}R$ based on $P2Y_{12}R$- PAR1_HUMAN pair (sequence identity: 23%) was closer to the $P2Y_{12}R$ crystal structure in comparison with the model based on the $P2Y_{12}R$-OPRK_HUMAN pair (sequence identity: 28%) (Castleman et al., 2019). We note that the *St* scores reported here correctly rank PAR1_HUMAN as the best template over the other three templates (**Table 1**), without model building and VLS.

In the case of PAR2- PAR1_HUMAN and PAR2-OPSD_BOVIN pairs, although both have good HC, the hotspot residues are dissimilar, with S_b(PAR2-OPSD_BOVIN) of −2, and S_b(PAR2- PAR1_HUMAN) of 51. Thus, BRS comparison is a useful parameter in selecting the appropriate template for GPCRs. Overall, the S_t score is able to identify the best template for each of the nine target receptors in **Table 1**.

TABLE 1 | Performance of Bio-GATS on recent published target-template pairs.

Target receptor	Template pairs	Published ranking	S_t
hPAR2	hPAR1 [36]	Good	52
	hOPRX [36]	Good	31
	bOPSD [36]	Bad	10
h5HT7	hOPRX [34]	Good	41
	hPAR1 [34]	Bad	30
hPAR1	hOPRK [33]	Good	42
	hOPRX [33]	Good	40
	hAA2AR [33]	Bad	19
hADRB2	hOPRK [33]	Good	31
	hAA2AR [33]	Good	17
	$hP2Y_{12}R$ [33]	Bad	9
$hP2Y_{12}R$	hPAR1 [32]	Good	26
	hOPRK [33]	Bad	15
	h5HT1B [32]	Bad	10
	hADRB2 [33]	Bad	9
hACM2	hDRD3 [32]	Good	44
	hOPRK [33]	Good	26
	$hP2Y_{12}R$ [33]	Bad	3
hFFAR1	hAT1R [32]	Good	24
	$hP2Y_{12}R$ [32]	Bad	22
h5-HT2AR	h5-HT2CR [35]	Good	71
	bOPSD [35]	Bad	20
	hAA2AR [35]	Bad	19
	hCXCR4 [35]	Bad	11
	hCNR1 [35]	Bad	9
hDRD2	hCXCR4 [35]	Good	26
	bOPSD [35]	Bad	11
	hCNR1 [35]	Bad	2

Validating Bio-GATS Template Selection Through Experimentally Resolved GPCR Structures

To further validate Bio-GATS, we selected 20 class A, 10 class B, four class C, one class D, and three class F receptors having experimentally solved structures. In all cases, the experimental structure was selected as the top ranked target template by Bio-GATS. Ignoring this top ranked structure, homology models for 38 receptors were build using Modeller (Webb and Sali, 2017) based on the second top template selected through Bio-GATS. The alignment was manually edited within loop regions through MEGA7 (Kumar et al., 2016). The generated models were compared with experimental structures through RMSD calculation for TM regions. For all models the RMSD of structurally aligned region was in the range 0.5–2.5 Å (**Supplementary Table 2**) as shown in **Figure 2** (mean = 1.38 ± 0.43 Å, median = 1.29 Å). The interquartile range (IQR) for all classes is 0.60 Å. For individual classes, class A is showing the IQR from 0.62 Å with sample size of 20. The IQR for class B and C is 0.36 and 0.16 with sample size of 10 and 4, respectively. To date, as only one structure is available for class D, this template was selected for this receptor, although it is phylogenetically distant and therefore showing a high RMSD value. The IQR for modeled class F receptors 0.1 with sample size 3 although two of three models were built on the basis of class B templates. The results of this study on 38 representative receptors from each class are showing the utility of hydrophobicity correspondence as a measure for template selection. The median for individual classes was under 1.5 Å except for classes D and F.

Subsequently, three receptors from classes A, B, C, F, and the single class D receptor was modeled through GPCR modeling servers such as GPCR-ModSim (Esguerra et al., 2016), GoMoDo (Sandal et al., 2013), GPCRM (Miszta et al., 2018), and GPCR-SSFE (Worth et al., 2017). The RMSDs for TM regions of automated models and models constructed using Bio-GATS

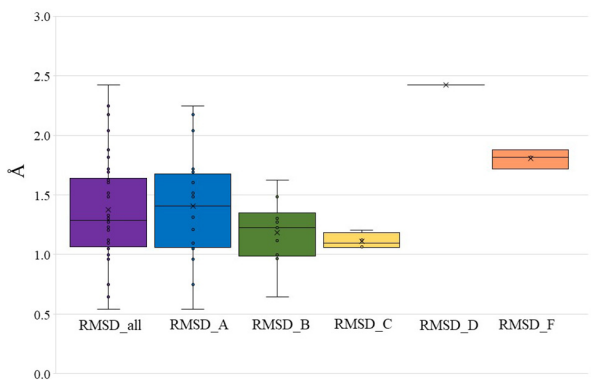


FIGURE 2 | RMSD between modeled structures and experimental structures for all GPCR classes, class A, B, C, D, and F (presented in **Supplementary Table 2**). The boundary of the box closest to zero indicates the 25th percentile, a black line within the box marks the median, and the boundary of the box farthest from zero indicates the 75th percentile.

TABLE 2 | The templates selected by Bio-GATS and the automated servers for representative GPCRs from each class along with RMSD values between the generated models and experimental structures for TM residues.

Receptor and PDBID	Bio-GATS		GPCRM		GPCR-SSFE		GPCR-ModSim		GoMoDo	
	Template and PDBID	RMSD (Å)	Template	RMSD (Å)	Template	RMSD (Å)	Template	RMSD (Å)	Template	RMSD (Å)
h5HT2A (6A94)	tADRB1 (4BVN)	**1.313**	tADRB1 (5F8U and 2VT4)	1.347	Many[1]	1.717	tADRB1 (2VT4)	1.427	None[6]	–
hTA2R (6IIU)	bOPSD (1U19)	2.248	hCNR1 (5TGZ) hAA2AR (5UIG)	**1.804**	Many[2]	1.975	hOPRX (4EA3)	1.948	hOPRX (4EA3)	1.911
hPE2R3 (6AK3)	hOPRM1 (5C1M)	*1.623*	h5HT2C (6BQG and 6BQH)	**1.484**	Many[3]	1.939	bOPSD (3PQR)	2.013	hOPRX (4EA3)	2.005
hCRFR1 (4K5Y)	hGLR (5EE7)	**1.626**	hCRFR1* (4Z9G)	0.966*	None[4]	–	hP2Y12 (4NTJ)	2.738	h5HT1B (4IAR)	1.853
hPACR (6P9Y)	hSCTR (6WZG)	**0.645**	hCALCR (5UZ7)	1.273	None[4]	–	bOPSD (3PQR)	2.192	None[6]	–
hSCTR (6WZG)	hCALRL (6UVA)	**1.304**	hCALCR (5UZ7)	1.466	None[4]	–	hACM2 (4MQS)	2.174	hCRFR1 (4K5Y)	1.773
hGRM1 (4OR2)	hGRM5 (6N52)	**1.122**	hGRM1* (4OR2)	0.108*	None[4]	–	None[5]	–	None[6]	–
hGRM5 (6N52)	hGRM1 (4OR2)	**1.207**	hGRM5* (5CGC, 5CGD)	0.875*	None[4]	–	None[5]	–	hOPRM1 (4DKL)	2.369
hGABR1 (6W2Y)	hGABR2 (7C7S)	**1.057**	hGRM1 (4OR2) hGRM5 (5CGC)	1.537	None[4]	–	None[5]	–	hGRM1 (4OR2)	1.492
ySTE2 (7AD3)	hGLP1R (6X19)	*2.425*	hOPRM1 (5C1M) h5HT2C (6BQH)	**2.298**	None[4]	–	hADRB2 (3SN6)	2.968	hP2Y12 (4PXZ)	2.721
hFZD4 (6BD4)	hPTH1R (6FJ3)	**1.878**	tADRB1 (5F8U, 2VTR)	2.916	None[4]	–	hPAR1 (3VW7)	–	hPAR1 (3VW7)	2.179
hFZD5 (6WW2)	hPTH1R (6FJ3)	1.817	hSMO (4O9R,4QIN)	**1.427**	None[4]	–	hADRB2 (2RH1)	–	hSMO (4JKV)	1.461
hSMO (5V56)	mSMO (6O3C)	**1.717**	hSMO (5L7I)*	0.745*	None[4]	–	None[5]	–	h5HT1B (4IAR)	1.944

The minimum RMSD values are in bold and second best values are in italics. The human GPCRs are prefixed by h, mouse by m, zebra fish by z, common turkey by t, yeast by y, and bovine by b.

**self template used; RMSD values were therefore not considered.*

[1]GPCR-SSFE templates: hACM4 (5DSG), hHRH1 (3RZE), hDRD3 (3PBL), hPAR2 (5NDD), h ADRB2 (2RH1), hP2Y12 (4NTJ), bOPSD (1U19), hACM3 (4U15).

[2]GPCR-SSFE templates: hPAR1 (3VW7), zLPA6 (5XSZ), hP2Y12 (4NTJ), hCXCR4 (3ODU), hAA2AR (4EIY), hPAR2 (5NDD).

[3]GPCR-SSFE templates: mOPRD1 (4EJ4), hP2Y12 (4NTJ), hCXCR4 (3ODU), sOPSD (2Z73), hCCR5 (4MBS), hHRH1 (3RZE), hPAR2 (5NDD).

[4]GPCR-SSFE does not work on non-Class A GPCRs.

[5]GPCR-ModSim does not work sequences greater than 600 residues such as hGRM1, hGRM5, hGABR1, and hSMO.

[6]GOMoDo does not work for h5HT2A, hPACR, and hGRM1.

suggested templates were compared (**Table 2**). We chose to compare RMSDs of TM regions only as loop modeling and refinement within servers is a time taking process. GPCR-SSFE was only able to generate models for class A GPCRs. While GPCR-ModSim cannot accept input sequence greater than 600 residues therefore, could not model selected class C GPCRs and one class F GPCR, i.e., SMO_human. Also, for all the receptors from class A to F considered for this study, GPCR-ModSim always selected the template from class A. Of 13 GPCRs, five models built on the basis of templates selected by Bio-GATS showed minimum RMSD with experimental structure of the receptor. The four models constructed by GPCRM (CRFR1_human, GRM1_human, GRM5_human, SMO_human) were based on receptor's own structure as a template therefore, showing the minimum RMSD (**Table 2**). The RMSD comparison shows the utility of our biophysical method to select the appropriate templates for all classes of GPCRs.

To further extend the application of Bio-GATS we built three models each for class A and C orphans through servers as well as on the basis of Bio-GATS suggested templates. The structural alignment of automated models and manual model (based on Bio-GATS template) for GPR35_human showed the differences in modeling TM1 by GPCRSSFE and TM6 by GPCRM. For P2RY10, the model built by GoMoDo was distorted with disoriented TM1 (**Supplementary Figure 1**). For class C orphans, there were significant differences among all the automated and manual models as shown by structural superposition (**Supplementary Figure 2**) and RMSD values (**Supplementary Table 3**).

Case Study on OR1A1

Currently, there exists no close homolog for ORs as evident from the phylogenetic tree between available GPCR templates and OR1A1 (**Figure 3**). We used Bio-GATS to search for

an optimal template for OR1A1. We selected OR1A1 as a case study because it contains the maximum mutagenesis data against six ligands among OR superfamily. The selection of templates was done on the basis of resolution (Insel et al., 2019), matching hydrophobicity profiles (S_h), and the BRS score (Munk et al., 2019). We considered inactive structures having $\leq$ 2.5 Å resolution, in accord with our earlier study on OR1A2 (Jabeen and Ranganathan, 2020). The top three templates selected by Bio-GATS for OR1A1 are human NK-1 or tachykinin receptor 1, NK1R_HUMAN (PDBID: 6HLP), bovine rhodopsin, OPSD_BOVIN (PDBID: 1U19) and the human thromboxane A2 receptor, TA2R_Human (PDBID: 6IIU). We also considered one template (CXCR4_HUMAN, PDBID: 3ODU) that was showing poor HC and low BRS score with OR1A1, for comparison, from the downloadable Bio-GATS result summary table (available from Bio-GATS Github page). All four structures belong to class A GPCRs. 6HLP and 6IIU show greater than 35% sequence identity with OR1A1 (**Table 3**).

Hydrophobic correspondence for each TM of the top two templates 6HLP and 1U19 compared to OR1A1 are shown in **Supplementary Figures 3**, **4**, with the other two templates to OR1A1 shown in **Supplementary Figures 5**, **6**. All OR1A1 TMs have good HC with 6HLP TMs, except TM6. OR1A1 shows good HC with 1U19 from TM1 to TM5 but not for TM6 and TM7, while it shares good HC with 6IIU in TM1, 2, 3, 5, and 6 but not in TM4 and TM7. The OR1A1 has poor HC throughout with 3ODU except within TM1, 3, and 5. The hydrophobic moment was calculated for both the target

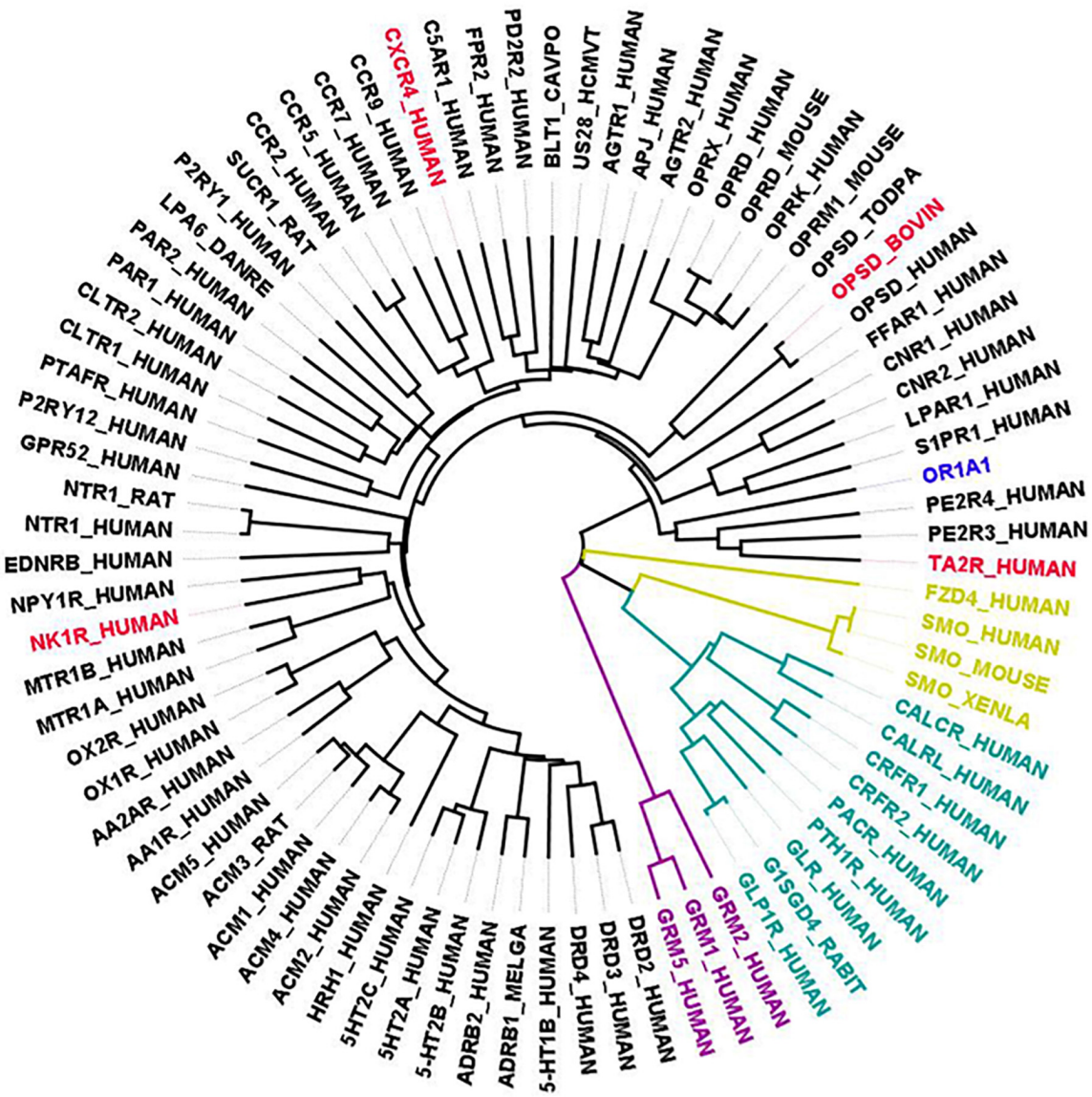


FIGURE 3 | Phylogenetic tree showing all available GPCR templates are distantly related to OR1A1. The selected templates for OR1A1 are shown in red color, members having known structures for class A are in black, Class B1 are in green color, Class C are in purple, and Class F are in gold color.

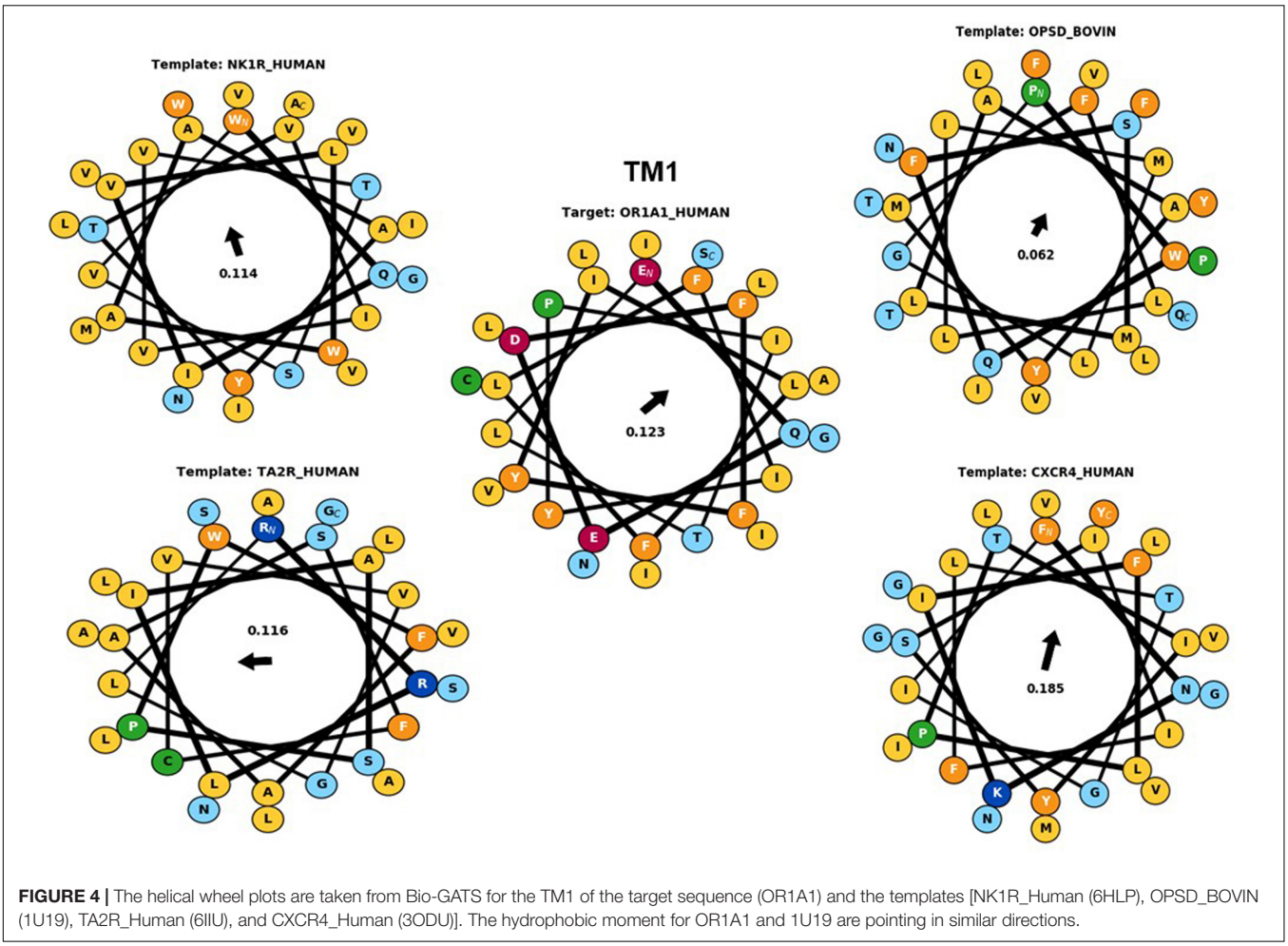


FIGURE 4 | The helical wheel plots are taken from Bio-GATS for the TM1 of the target sequence (OR1A1) and the templates [NK1R_Human (6HLP), OPSD_BOVIN (1U19), TA2R_Human (6lIU), and CXCR4_Human (3ODU)]. The hydrophobic moment for OR1A1 and 1U19 are pointing in similar directions.

sequence as well as the template sequences. The hydrophobic moment plots show the amphiphilic nature of the helices for the target as well as templates (TM1 in **Figure 4**, TM2–7 in **Supplementary Figures 7–9**). Amphiphilic helices are partly in the membrane and partly exposed to the aqueous phase. We used the Eisenberg scale and a window size of 11 as suitable for membrane proteins (Eisenberg et al., 1984) to calculate the hydrophobic moment of each helix. The hydrophobic moment points in the direction of maximum hydrophobicity (shown by an arrow within the hydrophobic moment plots) and it often faces the lipid surface (Liu et al., 2004). A large hydrophobic moment value shows the amphiphilicity of the helix perpendicular to its axis (Eisenberg et al., 1982). TMs 5, 6, and 7 for OR1A1 are more amphiphilic as compared to the rest of the helices. The hydrophobic moments for OR1A1 TMs 1, 2, 5, and 6 are pointing almost in the same direction as 1U19 (**Figure 4** and **Supplementary Figures 7–9**). The incorporation of hydrophobic moment information into the structural model building is essential in the proper positioning of helices within the model (Crasto, 2010).

An example of the downloadable Bio-GATS summary file, with details of helix-wise alignment, HC comparison and hydrophobic moment results, along with the overall GRoSS alignment, is provided for the OR1A1-1U19 target-template pair in **Supplementary Note 1**.

TABLE 3 | Parameters used by Bio-GATS to predict top templates for OR1A1.

Rank	Template	Sequence identity (%)	Resolution (Å)	S_h	S_b	S_r	S_t	S_{rank}
1	6HLP	37	2.2	11	6	1	18	2.91
2	1U19	20	2.2	8	8	1	17	2.75
3	6lIU	36	2.5	8	8	1	17	2.75
22	3ODU	25	2.5	2	−9	1	−6	1.54

Sequence identity is listed for comparison.

For most queries, there best scoring template can be selected for analysis, and the Bio-GATS alignment can be used directly for model building and SBVS. For OR1A1, the top three templates show very similar S_{rank} scores (**Table 3**), suggesting that they may all be suitable for the query sequence, due to the evolutionary distance of OR1A1 (and other ORs in general) from available templates (**Figure 2**). Further analysis such as ligand profiling is required from our previous study on OR1A2 (Jabeen and Ranganathan, 2020), to see if all three templates are equally suitable or one is better than the other two.

We calculated the Tanimoto score between the known OR1A1 ligands and the ligand bound to the template structures, based on PubChem fingerprints computed using Knime (Berthold et al., 2009). Retinal (PubChem CID: 638015), the ligand for 1U19 (**Figure 5** in blue) is more similar to the known ligands for OR1A1 followed by ramatroban (PubChem CID: 123879, **Figure 4** in green) in 6IIU and netupitant (PubChem CID: 6451149, **Figure 5** in gold) in 6HLP. We also compared the ligand profile for the lower scoring 3ODU and OR1A1. An isothiourea derivative, ITD (PubChem CID: 25147749, **Figure 5** in pink), the ligand for 3ODU, did not match with any OR1A1 ligand (**Figure 5**), listed in listed in **Supplementary Table 4** and is clearly not suitable for OR1A1.

The available structure for 6HLP is not complete, also the ligand profile for netupitant does not match with OR1A1 ligands. The 2nd best template 1U19 possesses a complete structure and contain a hydrophobic ligand that matches with OR1A1 ligand profile. It has the same resolution as 6HLP and S_b (8) is also better than that of 6HLP. Therefore, we selected 1U19 as a final template. To validate the Bio-GATS template selection, we built the homology model based on the suggested template (1U19) and performed molecular docking with known OR1A1 ligands having mutagenesis data and inspected whether we are able to recover the mutagenesis residues or not. For comparison, we also built a model with a template showing poor correspondence with OR1A1 in terms of S_h, S_b and ligand profile.

We built models for OR1A1 based on 1U19 and 3ODU (template showing low S_{rank}, and mismatched ligand profile), to differentiate between good and bad templates. We built 50 models using each template. The models with minimum Modeller objective function were selected for mutagenesis data analysis by molecular docking. Currently, OR1A1 has site-directed mutagenesis data for 13 sites for six ligands. Five positions 3.36, 3.37, 3.40, 4.56, and 5.46 are involved in (*S*)-(-)-citronellol (PubChem ID: 7793) and (*S*)-(-)-citronellal (PubChem ID: 443157) binding, 11 positions 3.34, 3.36, 3.37, 3.39, 4.53, 4.56, 5.46, 6.47, 6.48, 7.41, and 7.42 are important for (*S*)-(+)-carvone (PubChem ID: 16724) and (*R*)-(-)-carvone (PubChem ID: 439570) binding, and positions 6.48 and 6.55 are crucial for musk tibetene (PubChem ID: 67350) and musk xylene (PubChem ID: 62329) binding to OR1A1. Overall, seven positions 3.36, 3.37, 6.48, 6.55, 7.41, and 7.42 are part of the orthosteric binding site of GPCRs.

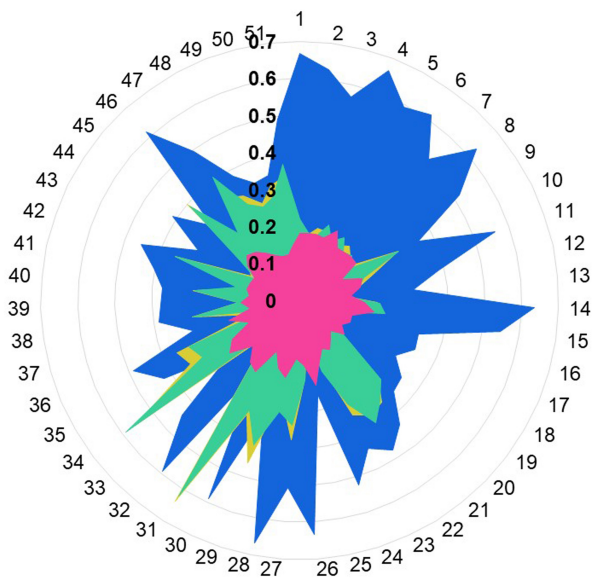


FIGURE 5 | Ligand profile for OR1A1 and selected templates. The similarity of OR1A1 ligands with: retinal (from 1U19) is represented in blue color, netupitant (from 6HLP) is in gold color, ramatroban (from 6IIU) is represented in green color, and ITD (from is represented in pink color). OR1A1 ligands from 1 to 51 are listed in **Supplementary Table 4**. Tanimoto scores between OR1A1 ligands and the template ligands range from 0.1 to 0.7 (in bold).

We downloaded the structures for these six ligands from PubChem and docked them to the predicted binding pocket of OR1A1, selected on available mutagenesis data. After docking (*S*)-(-)-citronellol and (*S*)-(-)-citronellal, we recovered 5/5 sites with the 1U19-based OR1A1 model but only 2/5 sites with the 3ODU-based OR1A1 model. Upon docking (*S*)-(+)-carvone and (*R*)-(-)-carvone, we were able to recover 6/11 sites with a 1U19-based model but only 3/11 sites with a 3ODU-based model. Docking musk xylene and musk tibetene into the binding pockets of OR1A1 models resulted in the recovery of both sites with a 1U19-based model and just one site using a 3ODU-model. In summary, we were able to recover maximum mutagenesis sites with the 1U19-based OR1A1 model (**Supplementary Table 5**). Thus, comparing the ligand profile of the target and candidate templates might be a useful measure in validating an appropriate template, in addition to the other measures. Mutagenesis data might also help in refining the predicted binding pocket of the model and has previously been incorporated to improve GPCR homology models in the literature (Ivanov et al., 2009; Perry et al., 2015).

We also used GPCR modeling servers to select the templates for OR1A1 and downloaded the generated alignment. Unfortunately, GOMoDo, and GPCR-ModSim servers did not permit submission of the query sequence therefore, results from these two servers are not included in the current study. GPCR-SSFE did not work for OR1A1 as the sequence did not match with the HMMER2 generated profile. Both GPCRM and GPCR-I-TASSER suggested AA2AR (PDBID: 3EML, resolution: 2.6 Å) as the top template. 3EML has resolution >2.5 Å and is not considered by Bio-GATS, although the high resolution AA2AR template, 5IU4 was identified as the 5th ranking template (in the result summary table, available from Bio-GATS Github page). The alignment generated by the two servers and Bio-GATS are shown in **Supplementary Figures 10–12**. The TM6 center residues were not aligned within the GPCRM and GPCR-I-TASSER server generated alignments but it was aligned properly by Bio-GATS. The Bio-GATS generated alignment needs manual adjustment within loop regions before proceeding to the model building step (**Supplementary Figure 12**).

Bio-GATS Features

Bio-GATS is connected to a local data file which contains manually curated 814 GPCR sequences, their TM definitions,

PDBIDs of currently available 443 GPCR structures, their conformation, resolution, and query coverage in terms of completeness of the structure. Bio-GATS provides three main features to the users. Firstly, the user can retrieve the top three templates for the queried sequence by clicking on the search button (**Figure 6**).

The top three templates are retrieved on the basis of three biophysical parameters, namely the resolution, hydrophobicity profile, and BRS score. The user can navigate among inactive, active, and intermediate conformational states as indicated in GPCRdb. The choice for selecting from a list of high resolution ($\leq$ 2.5 Å) structures is also provided (**Figure 7**). For some receptors, there exist multiple PDBs as in the case of OPSD_BOVIN, with 44 PDBs available. For such a scenario, only high-quality structures were shortlisted. The quality of the structure was determined on the basis of resolution and completeness of the structure (query coverage >75%). Hence, for the search template option, high-quality structures for 54 receptors in inactive, 34 receptors in active, and 19 receptors in intermediate conformations were considered. A detailed report (shown in **Supplementary Note 1**) with alignments and helix-wise HC and hydrophobicity moment of each target-template pair can be downloaded for comparison and data/figure extraction. A comprehensive data table with all scoring parameters for all templates considered is also available for further analysis (examples available from Bio-GATS Github page).

For consideration of options other than resolution, HC, and BRS score for template selection, the browse functionality is available, as an advanced feature in Bio-GATS. Within this feature, the expert user might browse for the best template among the complete list of 76 receptors with 443 available PDBs. In addition to the parameters considered earlier, the

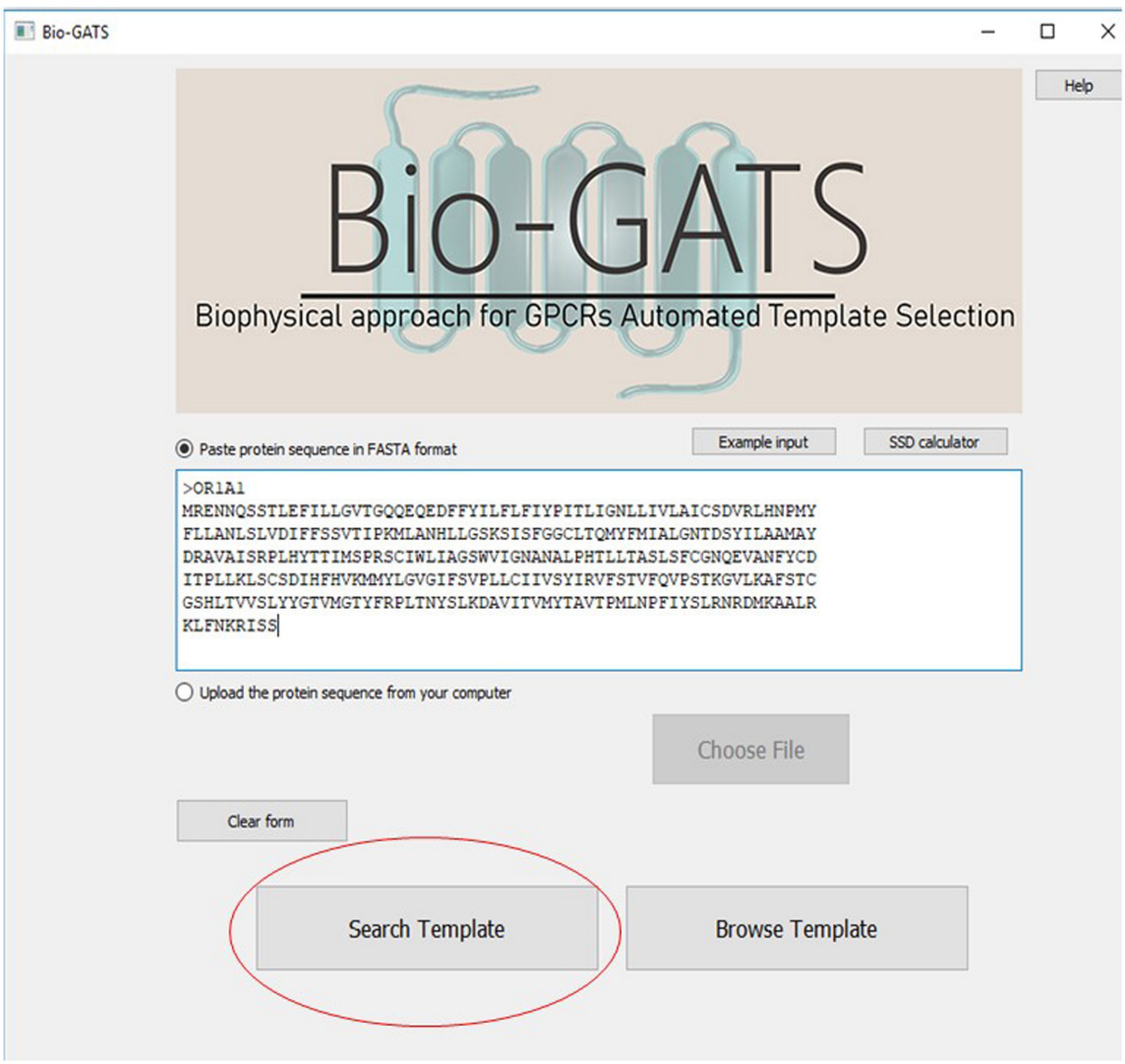


FIGURE 6 | The main interface of Bio-GATS. Automated selection of templates can be done by clicking on the search template button.

browse template page provides sequence identity and TM-wise sequence identity for each template (**Supplementary Figure 13**). The sequence identity is calculated through a locally installed BLAST alignment. Also, all the available PDB entries, their resolution, and query coverage for each receptor can be displayed for comparison purposes (**Supplementary Figure 14**). The *Browse template* feature thus lists comprehensive biophysical parameters comparing the query sequence to all available templates, which might also help the user in selecting multiple templates. HC between the target and the template within the search and browse template features are based on TM definitions derived from the GRoSS alignment (Cvicek et al., 2016). For customized TM definition, a third feature, the SSD calculator, has been added to Bio-GATS, where HC is calculated based on user-defined TM definitions for both the target and the template (**Supplementary Figure 15**). This feature is also useful for GPCR sequences that are not present within the curated data.

The hydrophobicity plots can be visualized and downloaded for each selected target-template pair (**Supplementary Figure 3**). The helical wheel plots can also be shown which might help the user in identifying the helical amphiphilicities (**Figure 3**). Also, the center residue-based TM alignment between the target and the template can be visualized and downloaded

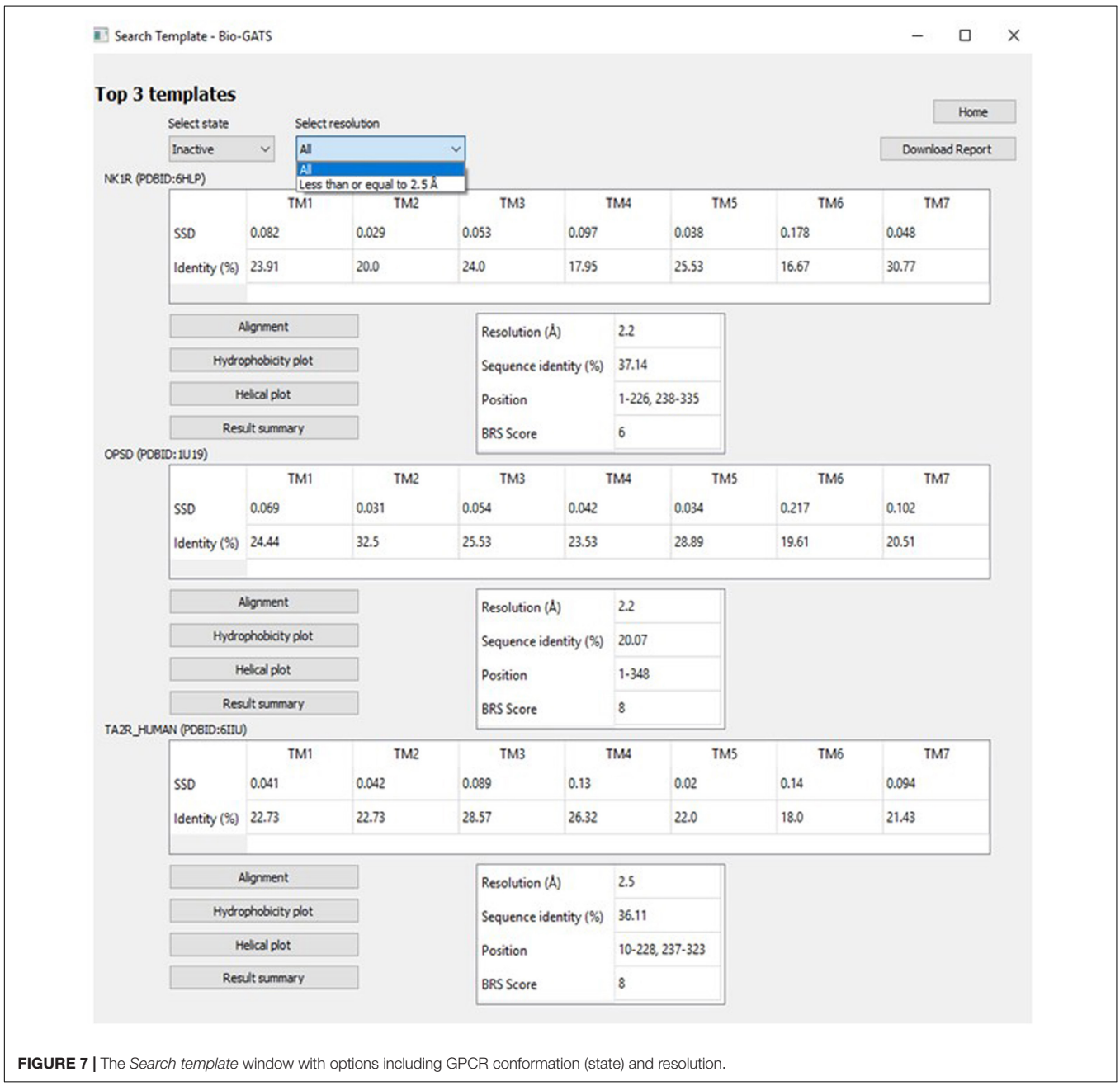


FIGURE 7 | The *Search template* window with options including GPCR conformation (state) and resolution.

(**Supplementary Figure 16**). The full-length alignment between the target and the selected template can also be downloaded in FASTA format for editing using available programs such as MEGA (Kumar et al., 2016), and AliView (Larsson, 2014), or directly building homology models through online servers such as GOMoDo (Sandal et al., 2013) or locally installed independent programs, for instance, Modeller (Webb and Sali, 2017). All these options are available from the different Bio-GATS windows. Further, a summary report (**Supplementary Note 1**) with the full-length alignment, TM-wise alignment, HC plots, and helical wheel plots of the target-template pair can be downloaded for detailed analysis and for use in reports and publications.

CONCLUSION

The existence of low sequence identity among available GPCR structures and sequences particularly OR sequences demands additional parameters for template selection. HC, similarities within the GPCR hotspot residues and matching the target-template ligand profile might serve as additional local parameters for GPCR template selection. Further, the incorporation of mutagenesis data might be helpful in refining GPCR homology models. Bio-GATS provides a convenient and user interactive way of selecting an appropriate template for a target GPCR, based on hydrophobicity profile and hotspot residue similarity while displaying global sequence identity as well as TM sequence identity for more advanced usage. The tool provides a comprehensive biophysical comparison between a target sequence and all the available templates which might assist in selecting more than one templates, commemorating Chothia's pioneering work in structural bioinformatics.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/amara86/Bio-GATS.

AUTHOR CONTRIBUTIONS

AJ and SR designed the study and wrote the manuscript. AJ acquired the data. AJ and RV implemented the interface. All authors read and approved the final manuscript.

FUNDING

This work was partially supported by the award of an Australian Research Council grant (DP180102727) to SR.

ACKNOWLEDGMENTS

AJ is grateful to Macquarie University for the award of an International Macquarie University Research Excellence Scholarship (iMQRES).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.617176/full#supplementary-material

REFERENCES

Abaffy, T., Bain, J. R., Muehlbauer, M. J., Spasojevic, I., Lodha, S., and Bruguera, E. (2018). A testosterone metabolite 19-hydroxyandrostenedione induces neuroendocrine trans-differentiation of prostate cancer cells via an ectopic olfactory receptor. *Front. Oncol.* 8:162.

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* 15, 488–506. doi: 10.1002/jcc.540150503

Alfonso-Prieto, M., Navarini, L., and Carloni, P. (2019). Understanding ligand binding to G-protein coupled receptors using multiscale simulations. *Front. Mol. Biosci.* 6:29.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

An, J., Totrov, M., and Abagyan, R. (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteom.* 4, 752–761. doi: 10.1074/mcp.m400159-mcp200

Baker, M. S., Ahn, S. B., Mohamedali, A., Islam, M. T., Cantor, D., et al. (2017). Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* 8:14271.

Ballesteros, J. A., and Weinstein, H. (1995). [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* 25, 366–428. doi: 10.1016/s1043-9471(05)80049-7

Basith, S., Cui, M., Macalino, S. J. Y., Park, J., Clavio, N. A. B., Kang, S., et al. (2018). Exploring G protein-coupled receptors (GPCRs) ligand space via cheminformatics approaches: impact on rational drug design. *Front. Pharmacol.* 9:128.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explor. Newslet.* 11, 26–31. doi: 10.1145/1656274.1656280

Braun, T., Voland, P., Kunz, L., Prinz, C., and Gratzl, M. (2007). Enterochromaffin cells of the human gut: sensors for spices and odorants. *Gastroenterology* 132, 1890–1901. doi: 10.1053/j.gastro.2007.02.036

Butkiewicz, M., Rodriguez, A. L., Rainey, S. E., Joshua, W., Luscombe, V. B., Stauffer, S. R., et al. (2019). Identification of novel allosteric modulators of metabotropic glutamate receptor subtype 5 Acting at site distinct from 2-Methyl-6-(phenylethynyl)-pyridine Binding. *ACS Chem. Neurosci.* 10, 3427–3436. doi: 10.1021/acschemneuro.8b00227

Carlsson, J., Coleman, R. G., Setola, V., Irwin, J. J., Fan, H., Schlessinger, A., et al. (2011). Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat. Chem. Biol.* 7, 769–778. doi: 10.1038/nchembio.662

Castleman, P. N., Sears, C. K., Cole, J. A., Baker, D. L., and Parrill, A. L. (2019). GPCR homology model template selection benchmarking: global versus local similarity measures. *J. Mol. Graph. Model* 86, 235–246. doi: 10.1016/j.jmgm.2018.10.016

Chan, H. C. S., Li, Y., Dahoun, T., Vogel, H., and Yuan, S. (2019). New Binding sites, new opportunities for GPCR drug discovery. *Trends Biochem. Sci.* 44, 312–330. doi: 10.1016/j.tibs.2018.11.011

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105, 1–12. doi: 10.1016/0022-2836(76)90191-1

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Cox, C. J., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Congreve, M., de Graaf, C., Swain, N. A., and Tate, C. G. (2020). Impact of GPCR structures on drug discovery. *Cell* 181, 81–91. doi: 10.1016/j.cell.2020.03.003

Crasto, C. J. (2010). Hydrophobicity profiles in G protein-coupled receptor transmembrane helical domains. *J. Receptor. Ligand. Channel Res.* 2010, 123–133. doi: 10.2147/jrlcr.s14437

Cvicek, V., Goddard, W. A. III, and Abrol, R. (2016). Structure-based sequence alignment of the transmembrane domains of all human GPCRs: phylogenetic, structural and functional implications. *PLoS Comput. Biol.* 12:e1004805. doi: 10.1371/journal.pcbi.1004805

Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 179, 125–142. doi: 10.1016/0022-2836(84)90309-7

Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* 299, 371–374. doi: 10.1038/299371a0

Esguerra, M., Siretskiy, A., Bello, X., Sallander, J., and Gutiérrez-de-Terán, H. (2016). GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled receptors. *Nucleic Acids Res.* 44, W455–W462.

Haddad, Y., Adam, V., and Heger, Z. (2020). Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLoS Comput. Biol.* 16:e1007449. doi: 10.1371/journal.pcbi.1007449

Hu, G. M., Mai, T. L., and Chen, C. M. (2017). Visualizing the GPCR network: classification and evolution. *Sci. Rep.* 7:15495.

Insel, P. A., Sriram, K., Gorr, M. W., Wiley, S. Z., Michkov, A., Salmerón, C., et al. (2019). GPCRomics: An approach to discover GPCR drug targets. *Trends Pharmacol. Sci.* 40, 378–387. doi: 10.1016/j.tips.2019.04.001

Ivanov, A. A., Barak, D., and Jacobson, K. A. (2009). Evaluation of homology modeling of G-protein-coupled receptors in light of the A(2A) adenosine receptor crystallographic structure. *J. Med. Chem.* 52, 3284–3292. doi: 10.1021/jm801533x

Jabeen, A. V., and Ranganathan, S. (2020). A two-stage computational approach to predict ligands for a chemosensory receptor. *Curr. Res. Struct. Biol.* 2, 213–221. doi: 10.1016/j.crstbi.2020.10.001

Jabeen, A., and Ranganathan, S. (2019). Applications of machine learning in GPCR bioactive ligand discovery. *Curr. Opin. Struct. Biol.* 55, 66–76. doi: 10.1016/j.sbi.2019.03.022

Jabeen, A., Mohamedali, A., and Ranganathan, S. (2019a). *Looking for Missing Proteins*, *Reference Module in Life Sciences*. Amsterdam: Elsevier.

Jabeen, A., Mohamedali, A., and Ranganathan, S. (2019b). "Protocol for protein structure modelling," in *Encyclopedia of Bioinformatics and Computational Biology*, ed. S. Ranganathan, et al. (Oxford: Academic Press), 252–272. doi: 10.1016/b978-0-12-809633-8.20477-9

Jaiteh, M., Rodríguez-Espigares, I., Selent, J., and Carlsson, J. (2020). Performance of virtual screening against GPCR homology models: Impact of template selection and treatment of binding site plasticity. *PLoS Comput. Biol.* 16:e1007680. doi: 10.1371/journal.pcbi.1007680

Kosinski, J., Barbato, A., and Tramontano, A. (2013). MODexplorer: an integrated tool for exploring protein sequence, structure and function relationships. *Bioinformatics* 29, 953–954. doi: 10.1093/bioinformatics/btt062

Kufareva, I., Katritch, V., Participants of Gpcr Dock 2013, Stevens, R. C., and Abagyan, R. (2014). Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure* 22, 1120–1139. doi: 10.1016/j.str.2014.06.012

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531

Liu, W., Eilers, M., Patel, A. B., and Smith, S. O. (2004). Helix packing moments reveal diversity and conservation in membrane protein structure. *J. Mol. Biol.* 337, 713–729. doi: 10.1016/j.jmb.2004.02.001

Lo Conte, L., Bart, A., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G., Chothia, C., et al. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 28, 257–259. doi: 10.1093/nar/28.1.257

Lolkema, J. S., and Slotboom, D. J. (1998). Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins. *FEMS Microbiol. Rev.* 22, 305–322. doi: 10.1111/j.1574-6976.1998.tb00372.x

Loo, J. S., Emtage, A. L., Ng, K. W., Yong, A. S. J., and Doughty, S. W. (2018). Assessing GPCR homology models constructed from templates of various transmembrane sequence identities: Binding mode prediction and docking enrichment. *J. Mol. Graph. Model.* 80, 38–47. doi: 10.1016/j.jmgm.2017.12.017

Miszta, P., Pasznik, P., Jakowiecki, J., Sztyler, A., Latek, D., and Filipek, S. (2018). GPCRM: a homology modeling web service with triple membrane-fitted quality assessment of GPCR models. *Nucleic Acids Res.* 46, W387–W395.

Miyagi, H., Asada, H., Suzuki, M., Takahashi, Y., Yasunaga, M., Suno, C., et al. (2020). The discovery of a new antibody for BRIL-fused GPCR structure determination. *Sci. Rep.* 10:11669.

Müller, A. T., Gabernet, G., Hiss, J. A., and Schneider, G. (2017). modlAMP: python for antimicrobial peptides. *Bioinformatics* 33, 2753–2755. doi: 10.1093/bioinformatics/btx285

Munk, C., Mutt, E., Isberg, V., Nikolajsen, L. F., Bibbe, J. M., Flock, T., et al. (2019). An online resource for GPCR structure determination and analysis. *Nat. Methods* 16, 151–162. doi: 10.1038/s41592-018-0302-x

Nikolaev, D. M., Shtyrov, A. A., Panov, M. S., Jamal, A., Chakchir, O. B., Kochemirovsky, V. A., et al. (2018). A comparative Study of modern homology modeling algorithms for rhodopsin structure prediction. *ACS Omega* 3, 7555–7566. doi: 10.1021/acsomega.8b00721

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., et al. (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* 289, 739–745. doi: 10.1126/science.289.5480.739

Perry, S. R., Xu, W., Wirija, A., Lim, J., Yau, M-K., Stoermer, M. J., et al. (2015). Three homology models of PAR2 derived from different templates: application to antagonist discovery. *J. Chem. Inf. Model* 55, 1181–1191. doi: 10.1021/acs.jcim.5b00087

Rataj, K., Witek, J., Mordalski, S., Kosciolek, T., and Bojarski, A. J. (2014). Impact of template choice on homology model efficiency in virtual screening. *J. Chem. Inf. Model* 54, 1661–1668. doi: 10.1021/ci500001f

Redfern, O. C., Dessailly, B., and Orengo, C. A. (2008). Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* 18, 394–402. doi: 10.1016/j.sbi.2008.05.007

Rios, S., Fernandez, M. F., Caltabiano, G., Campillo, M., Pardo, L., and Gonzalez, A. (2015). GPCRtm: an amino acid substitution matrix for the transmembrane region of class A G protein-coupled receptors. *BMC Bioinformatics* 16:206.

Saikia, S., Bordoloi, M., and Sarmah, R. (2019). Established and In-trial GPCR Families in clinical trials: a review for target selection. *Curr. Drug Targets* 20, 522–539. doi: 10.2174/1389450120666181105152439

Sandal, M., Tran, P. D., Cona, M., Zung, H., Carloni, P., Musiani, F., et al. (2013). GOMoDo: A GPCRs online modeling and docking webserver. *PLoS One* 8:e74092. doi: 10.1371/journal.pone.0074092

Shahaf, N., Pappalardo, M., Basile, L., Guccione, S., and Rayan, A. (2016). How to choose the suitable template for homology modelling of GPCRs: 5-HT7 receptor as a test case. *Mol. Inform.* 35, 414–423. doi: 10.1002/minf.201501029

Silva, P. J. (2008). Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis. *Proteins* 70, 1588–1594. doi: 10.1002/prot.21803

Szwabowski, G. L., Castleman, P. N., Sears, C. K., Wink, L. H., Cole, J. A., and Baker, D. L. (2020). Benchmarking GPCR homology model template selection in combination with de novo loop generation. *J. Comput. Aided Mol. Des.* 34, 1027–1044. doi: 10.1007/s10822-020-00325-x

Van Rossum, G., and Drake, F. L. (2011). *The Python Language Reference Manual*. Surrey, UK: Network Theory Ltd.

Wacker, D., Stevens, R. C., and Roth, B. L. (2017). How ligands illuminate GPCR molecular pharmacology. *Cell* 170, 414–427. doi: 10.1016/j.cell.2017.07.009

Wallace, J., Onkabetse, A. D., Harris, F., and Phoenix, D. A. (2004). Investigation of hydrophobic moment and hydrophobicity properties for transmembrane alpha-helices. *Theor. Biol. Med. Model* 1:5.

Wallner, B., and Elofsson, A. (2005). All are not equal: a benchmark of different homology modeling programs. *Protein Sci.* 14, 1315–1327. doi: 10.1110/ps.041253405

Webb, B., and Sali, A. (2017). Protein Structure Modeling with MODELLER. *Methods Mol. Biol.* 1654, 39–54. doi: 10.1007/978-1-4939-7231-9_4

Wolf, S., Nikolina, J., Gelis, L., Pietsch, S., Hatt, H., and Gerwert, K. (2017). Dynamical binding modes determine agonistic and antagonistic ligand effects in the prostate-specific G-protein coupled receptor (PSGR). *Sci. Rep.* 7:16007.

Worth, C. L., Kreuchwig, F., Tiemann, J. K. S., Kreuchwig, A., Ritschel, M., Kleinau, G., et al. (2017). GPCR-SSFE 2.0-a fragment-based molecular modeling web tool for Class A G-protein coupled receptors. *Nucleic Acids Res.* 45, W408–W415.

Wu, C., Jia, Y., Lee, J. H., Kim, Y., Sekharan, S., Batista, V. S., et al. (2015). Activation of OR1A1 suppresses PPAR-γ expression by inducing HES-1 in cultured hepatocytes. *Int. J. Biochem. Cell Biol.* 64, 75–80. doi: 10.1016/j.biocel.2015.03.008

Zheng, W., Chengxin, Z., Wuyun, Q., Pearce, R., Li, Y., and Zhang, Y. (2019). LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res.* 47, W429–W436.

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 30 April 2021
doi: 10.3389/fmolb.2021.626729

Abundance Imparts Evolutionary Constraints of Similar Magnitude on the Buried, Surface, and Disordered Regions of Proteins

*Benjamin Dubreuil and Emmanuel D. Levy**

Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel

An understanding of the forces shaping protein conservation is key, both for the fundamental knowledge it represents and to allow for optimal use of evolutionary information in practical applications. Sequence conservation is typically examined at one of two levels. The first is a residue-level, where intra-protein differences are analyzed and the second is a protein-level, where inter-protein differences are studied. At a residue level, we know that solvent-accessibility is a prime determinant of conservation. By inverting this logic, we inferred that disordered regions are slightly more solvent-accessible on average than the most exposed surface residues in domains. By integrating abundance information with evolutionary data within and across proteins, we confirmed a previously reported strong surface-core association in the evolution of structured regions, but we found a comparatively weak association between disordered and structured regions. The facts that disordered and structured regions experience different structural constraints and evolve independently provide a unique setup to examine an outstanding question: why is a protein's abundance the main determinant of its sequence conservation? Indeed, any structural or biophysical property linked to the abundance-conservation relationship should increase the relative conservation of regions concerned with that property (e.g., disordered residues with mis-interactions, domain residues with misfolding). Surprisingly, however, we found the conservation of disordered and structured regions to increase in equal proportion with abundance. This observation implies that either abundance-related constraints are structure-independent, or multiple constraints apply to different regions and perfectly balance each other.

Keywords: protein abundance, protein evolution, protein structure, misfolding, intrinsic disorder, contact number, misinteraction, yeast proteome

OPEN ACCESS

Edited by:
Paolo Marcatili, Technical University of Denmark, Denmark

Reviewed by:
Andrew James Doig, The University of Manchester, United Kingdom
Karen N. Allen, Boston University, United States

***Correspondence:**
Emmanuel D. Levy
emmanuel.levy@weizmann.ac.il

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *06 November 2020*
Accepted: *29 March 2021*
Published: *30 April 2021*

Citation:
Dubreuil B and Levy ED (2021) Abundance Imparts Evolutionary Constraints of Similar Magnitude on the Buried, Surface, and Disordered Regions of Proteins. Front. Mol. Biosci. 8:626729. doi: 10.3389/fmolb.2021.626729

INTRODUCTION

During the course of evolution, mutations arise throughout genomes and can impact every protein at every site. However, contemplating a multiple sequence alignment of orthologous sequences typically shows widely differing levels of conservation across sites. Additionally, comparing multiple sequence alignments of different orthogroups shows even larger differences: certain groups such as those of ribosomal genes can be well conserved despite hundreds of millions of years of divergence, while others accumulate mutations much faster.

Amino-acid residues within proteins are subject to functional, biophysical, and structural constraints that are interconnected. These constraints result in different degrees of purifying selection along the sequence (i.e., purging of deleterious mutations by natural selection), which yields different levels of positional conservation. We discuss here structural aspects related to these constraints while placing an emphasis on works of Cyrus Chothia, to whom this issue is dedicated, and refer the reader to several reviews for a comprehensive overview (Liberles et al., 2012; Sikosek and Chan, 2014; Echave et al., 2016; Echave and Wilke, 2017). Following the characterization of the first few structures of proteins, their comparative analysis made it clear that the burial of non-polar residues accompanied with Van der Waals interactions and hydrogen bonding were the main contributors to the folding free energy (Chothia, 1974, 1975, 1976; Miller et al., 1987). Confirming the "hydrophobic bonding" intuition of Kauzmann (Kauzmann, 1959) and relying on calculations of molecular surfaces based on the algorithm of Lee and Richards (1971), Chothia estimated that each square Ångstrom of accessible surface removed from contact with water provides a free energy gain of 25 cal. Mol^{-1} (Chothia, 1974, 1975). At the same time, he provided universal relationships governing protein folding, e.g., on the proportion of the total accessible surface of a polypeptide chain that becomes buried upon folding (Chothia, 1975). This simple relationship has a profound meaning with respect to surface-to-volume ratios in folded proteins, notably that longer proteins should fold following a beads-on-a-string model rather than by forming larger beads (Wetlaufer, 1973) – indeed it was soon realized that beads (domains) are fundamental units of protein evolution (Chothia, 1992; Murzin et al., 1995; Bateman et al., 2002; Gough and Chothia, 2002). On top of hydrophobic bonding energy, a high degree of steric complementarity creates a well-packed protein interior (Chothia, 1975), in which mutations are incrementally accommodated by small structural changes (Lesk and Chothia, 1980). Ultimately, as sequences diverge, structures do too, albeit more slowly (Chothia and Lesk, 1986, 1987). Considering that structures are globally maintained during the course of evolution, it is intuitive that buried residues, which contribute to folding and stability more than surface residues (Creighton and Chothia, 1989; Lim and Sauer, 1989; Tokuriki et al., 2007), are more conserved (Koshi and Goldstein, 1995; Goldman et al., 1998; Guo et al., 2004; Bloom et al., 2006; Sasidharan and Chothia, 2007; Goldstein, 2008; Conant and Stadler, 2009; Franzosa and Xia, 2009; Liberles et al., 2012; Yeh et al., 2014; Echave et al., 2015; Shahmoradi and Wilke, 2016; Spielman and Wilke, 2016; Echave and Wilke, 2017; Liu et al., 2017).

We saw that the structure of a protein could help explain why certain positions – notably those buried and in contact with a large number of neighboring residues, are more conserved than others. Protein structure can also help to rationalize why certain proteins, e.g., those with more designable folds, evolve faster than others (Shakhnovich et al., 2005; Bloom et al., 2006). Globally, however, structural information only explains a small fraction of the heterogeneity in evolutionary rates seen across different proteins. Several studies have singled out other protein-centric properties associated with this heterogeneity (Zhang and Yang, 2015), including function (Wall et al., 2005; Lopez-Bigas et al., 2008; Xia et al., 2009), essentiality (Hurst and Smith, 1999; Hirsh and Fraser, 2001; Jordan et al., 2002; Liao et al., 2006), the number of interaction partners (Fraser et al., 2002; Bloom and Adami, 2004; Fraser and Hirsh, 2004; Hahn and Kern, 2005; Kim et al., 2006; Xia et al., 2009), or cellular abundance (Pal et al., 2001; Krylov et al., 2003; Rocha and Danchin, 2004; Subramanian and Kumar, 2004; Drummond et al., 2005; Bloom et al., 2006; Liao et al., 2006; Popescu et al., 2006; Pál et al., 2006; Sällström et al., 2006; Drummond and Wilke, 2008; Xia et al., 2009; Zhang and Yang, 2015). The latter is, by far, the most significant, in particular among unicellular organisms where there is no complexity added by tissue-specific expression. Several mechanistic interpretations of this abundance-conservation association have been proposed (Drummond et al., 2005; Drummond and Wilke, 2008; Cherry, 2010; Gout et al., 2010; Plata et al., 2010; Levy et al., 2012; Yang et al., 2012; Park et al., 2013; Zhang and Yang, 2015) and remain a matter of active debate (Plata and Vitkup, 2018; Razban, 2019). We will scrutinize this relationship further in the results and discussion section, in the context of the results presented.

We have seen how protein structure helped to interpret and rationalize data on evolutionary conservation. Here, we invert this logic to characterize structural properties of disordered regions from data on their evolutionary conservation. First, we compared the evolutionary rate of disordered regions to that of surface residues in the same protein and found that disordered regions are equivalent to super-accessible surface residues. Second, we know that the divergence of surface and core residues is interdependent. In other words, a protein's surface can hardly diverge without mutations arising in its interior as well, and vice-versa. We confirmed this finding in showing that evolutionary rates of surface and interior regions are correlated within proteins ($R > 0.85$). In contrast, the evolutionary rates of disordered and domain regions were poorly coupled ($R \sim 0.25$), indicating that disordered regions are, for the most part, structurally independent from domains in the same sequence. Finally, the structural differences and the lack of interdependence between disordered and structured regions supports that they can be influenced differently by biophysical and structural constraints. For example, an increased purifying selection for protein stability is expected to impact buried residues more than disordered ones. This idea led us to examine whether abundance impacts the relative conservation between these regions. Surprisingly, however, the relative conservation between different regions appeared independent from abundance.

RESULTS AND DISCUSSION

Disordered Regions Are Equivalent to Super-Accessible Surface Residues in Terms of Their Conservation

Among proteins that need to fold into stable structures to function, amino-acid residues buried in the protein interior

contribute the most to stability. Consequently, these residues are under stronger purifying selection than surface amino-acid residues, and are, on average, more conserved in the sequence. Two measures of residue burial have been associated with the heterogeneity of conservation in sequences: (i) solvent accessible surface area or ASA (Lee and Richards, 1971; Shrake and Rupley, 1973; Goldman et al., 1998; Bloom et al., 2006; Lin et al., 2007; Conant and Stadler, 2009; Franzosa and Xia, 2009), which measures the surface or fractional surface of an amino-acid residue that is in contact with bulk water, and (ii) the packing density of an amino-acid residue, which measure the density of its neighbors. Different metrics capture this information, including the contact number and the weighted contact number, with the latter containing longer-range information (Franzosa and Xia, 2009; Yeh et al., 2014). While not strictly equivalent, both accessible surface area and packing density correlate strongly (Echave et al., 2016), and both measures show that the less buried is a residue, the less conserved it is within a protein sequence.

This conservation-structure relationship prompts us to infer structural properties of disordered regions from their pattern of conservation within proteins. We know that disordered regions are devoid of a hydrophobic core and therefore cannot autonomously adopt a stable three-dimensional structure. However, if we consider the spectrum of solvent accessibility and packing density found among folded domains, where would disordered regions position themselves on average? Would they appear much less conserved than even the most solvent-exposed regions? Some disordered regions serve purely as linkers or entropic springs and are expected to show very weak sequence conservation (Dyson and Wright, 2005; Van der Lee et al., 2014). At the same time, disordered regions can also form secondary structure elements and bind to partners (Tompa, 2005; Vacic et al., 2007; Uversky and Dunker, 2010; Wright and Dyson, 2015; Banani et al., 2017; Dignon et al., 2019), thereby burying residues and transiently increasing their packing density. For example, p27Kip1 can wrap around the structure of Cdk2 to regulate its function (Russo et al., 1996; Galea et al., 2008).

To position disordered regions on the solvent accessibility spectrum observed in structured regions, we compared the evolutionary rate of residues in both region types. Specifically, we selected 3,350 proteins from *Saccharomyces cerevisiae*, which contain at least 20 residues in both structured regions and disordered regions. We inferred residue-level conservation using Rate4Site (Pupko et al., 2002) on multiple sequence alignments of orthologs from 14 fungal species (see section "Materials and Methods"). Evolutionary and structural information were mapped along the reference sequence from the multiple alignment as illustrated for STI1, a conserved Hsp90 co-chaperone (**Figure 1A**). We calculated a ratio per protein *i*, corresponding to the mean evolutionary rate of residues in disordered regions (R_i^{diso}) divided by the mean rate of residues in a domain (R_i^{domain}). Overall, considering 2607 proteins with known orthologs, containing both types of regions, the median ratio (R_i^{diso}/R_i^{domain}) is equal to 2.2 (**Figure 1B**). If we now consider domains of known structure (i.e., present in PDB, currently ∼670) instead of those predicted, we find a similar median ratio equal to 2.0. For those proteins, we compared the conservation of disordered regions to that of buried and surface residues separately and found ratios equal to 3.1 and 1.4, respectively. Thus, in an average protein of this dataset, disordered regions evolve 3.1 and 1.4-fold faster than buried and surface residues, respectively (**Figure 1B**).

This result is based on a definition of surface that includes residues with >25% relative ASA. As higher ASA is associated with lower conservation, we asked whether increasing the cut-off progressively from >25 to >80% would yield a point where surface residues evolve faster than disordered ones (**Figure 1C**). We did not reach such a point as the ratio remained above 1 for all values. However, the ratio did converge to a value close to 1, highlighting that in an average protein, disordered residues are almost equivalent in their conservation to the most exposed residues at the surface of structured regions.

If we assume that the differential conservation of sites within protein sequences largely reflects different structural constraints, we can infer that disordered regions are, on average, highly solvent-exposed and under weak structural constraints. In sum, our results place disordered regions in the continuum of protein structure, at the end of the solvent-accessibility spectrum. It will be interesting to refine this relationship in the future. For example, by comparing additional properties such as hydrophobicity (Kyte and Doolittle, 1982) or stickiness (Levy, 2010), by considering where disordered segments fall in the sequence (e.g., N/C-ter and inside domains), or by breaking down disorder into different types (Bellay et al., 2011).

Conservation of Disorder Versus Domains Is Poorly Correlated Among Low Abundance Proteins and the Correlation Increases With Abundance

Individual residues within a structure contribute to stability together. As a result, we can expect the evolutionary conservation of residues within a structure to be uniform. To examine this idea, we compared the average evolutionary rate of surface and buried amino-acid residues within structures. Importantly, we know that protein abundance imposes global constraints on the conservation of proteins, which may also result in a uniform evolutionary pressure across the sequence, independently of the structure. Thus, we initially focused on low abundance proteins in which such global constraints are minimized. We observed the conservation of surface and buried regions to correlate strongly ($R > 0.83$, **Figure 2A**), which is reminiscent of the surface-core association described previously (Tóth-Petróczy and Tawfik, 2011).

We next compared the association in evolutionary conservation between disordered regions and domains found in the same protein. In this case, the correlation was reduced

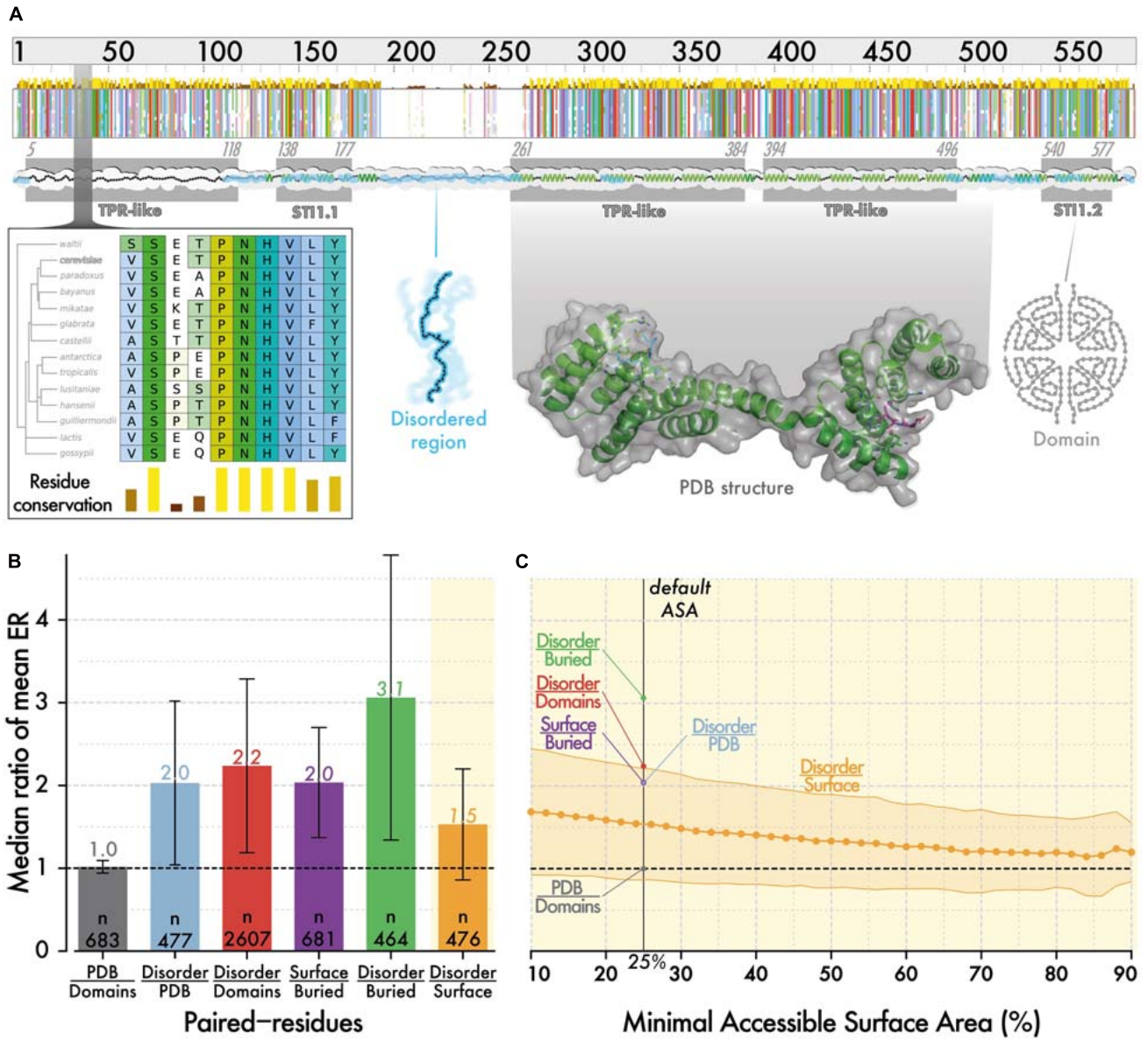


FIGURE 1 | The evolutionary rate of disordered regions is comparable to that of super-exposed regions in folded proteins. **(A)** Evolutionary information and structural features are mapped onto protein sequences from *S. cerevisiae*. The minimap represents the multiple sequence alignment of orthologous sequences to STI1. The amino acids are colored using CLUSTAL's color scale (Thompson et al., 1994) depending on residue type and conservation. The zoomed-in panel illustrates residue-level conservation, which we calculated with Rate4Site (Pupko et al., 2002). We mapped the positions of PFAM (Bateman et al., 2002) and SUPERFAMILY (Gough and Chothia, 2002) domains (gray box), and of disordered regions predicted by IUPRED (Dosztányi, 2018) (cyan ribbon). We also mapped structural information available from PDB (Rose et al., 2017; Armstrong et al., 2019) and 3DComplex (Levy et al., 2006) on sequences. For this particular sequence, structural information was partially available based on PDB code 3UQ3 (Schmid et al., 2012). **(B)** Within proteins, the evolutionary rate of residues in different regions are averaged, and we compare the ratio of these averages. We show the median of ratios with error bars corresponding to the median absolute deviation. Surface and buried residues are defined based on relative ASA of >25 and ≤25%, respectively (Levy, 2010). **(C)** We calculate the same ratio as in panel **(B)**, between disordered regions and surface regions, using an increasingly stringent relative ASA cut-off to define surface residues. As we increase the cutoff, the median ratio tends toward 1, which highlights that disordered residues evolve only slightly faster than the most exposed residues at protein surfaces.

greatly (R = 0.25), indicating that the structural connectivity and interdependence between disordered regions and domains are globally weak. These results are consistent with those of the previous section, which depict disordered regions as being highly solvent-accessible and structurally independent from domains. However, proteins expressed at higher levels show increasing correlation, from R = 0.40 among medium abundance proteins, to R = 0.63 in the class of proteins with the highest abundance (**Figure 2B**, lower row). This apparent coupling in evolutionary rates is unlikely to have a structural

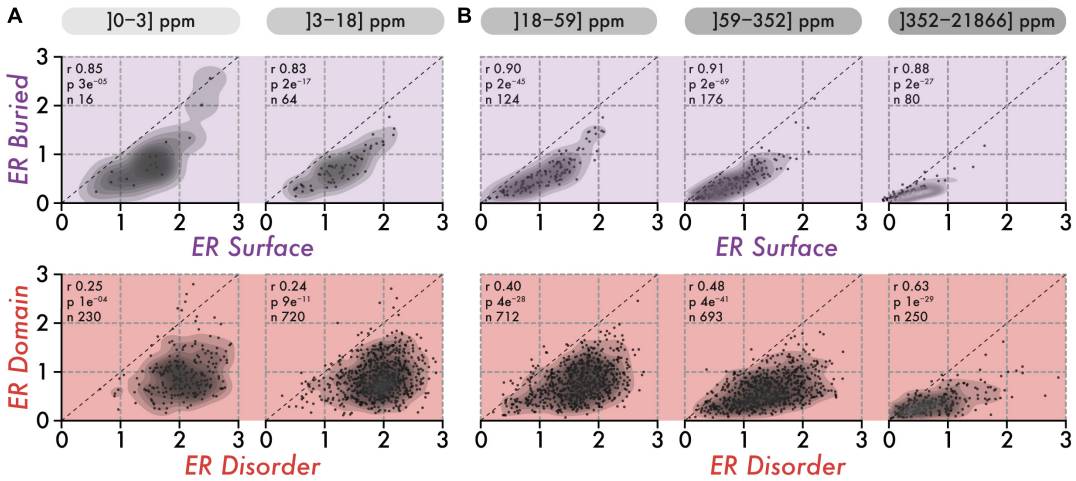


FIGURE 2 | The correlation in the conservation of disorder vs domain regions is poor among low abundance proteins and increases with abundance. **(A)** The top row shows the average evolutionary rate (ER) of surface residues (*x*-axis) vs buried residues (*y*-axis) per protein, for two classes of abundance (0–3 and 3–18 ppm or parts per millions). The lower row shows the average ER of disordered residues (*x*-axis) vs residues in domains (*y*-axis) per protein, for the same two classes of abundance. A protein falling on the diagonal (dashed line) means that residues in the two regions being compared have equal evolutionary rates (i.e., a ratio of 1). The Spearman rank correlation coefficient (r), the associated *p*-value (*p*, two-sided Spearman's rank correlation test), and the number of proteins (n) within each class of abundance are given above each scatterplot. **(B)** Same as in panel **(A)**, for three classes of abundance (18–59, 59–352, and 352–21,866 ppm or parts per million).

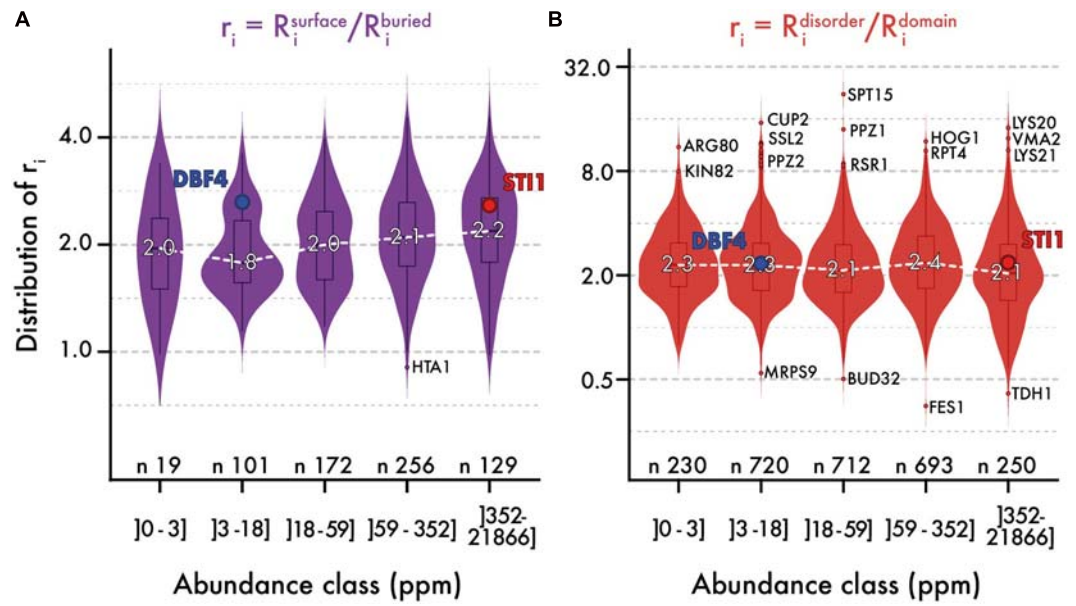


FIGURE 3 | The relative evolutionary rates of different protein regions are steady with abundance. Distribution of evolutionary rates ratio between different regions in the sequence (*y*-axis), across five classes of protein abundance (*x*-axis). A ratio is calculated by dividing the average evolutionary rate of residues found in two regions panel **(A)** surface vs. buried, panel **(B)** disorder vs. domain. The white dashed line highlights the median ratio across bins of abundance. Overlaid box plots show the interquartile range (IQR = 25 to 75% quantiles) with their whiskers extending to 1.58 × IQR. Beyond this interval, the three most extreme outlier values are annotated. The number of proteins contributing to each distribution is given. We also highlight the relative rates for a pair of proteins, one with low and one with high abundance (STI1 and DBF4). These two proteins show comparable structural features, different evolutionary rates (respectively, 0.575 and 1.34 for their full sequence), and similar ratios.

origin. Rather, it probably results from global constraints linked to abundance and exerted on the whole protein sequence. This apparent coupling also implies that different regions in a sequence all experience increasingly strong purifying selection with increasing abundance. This observation led us to quantify whether such negative selection increases equally in all regions, or whether some regions become more constrained than others.

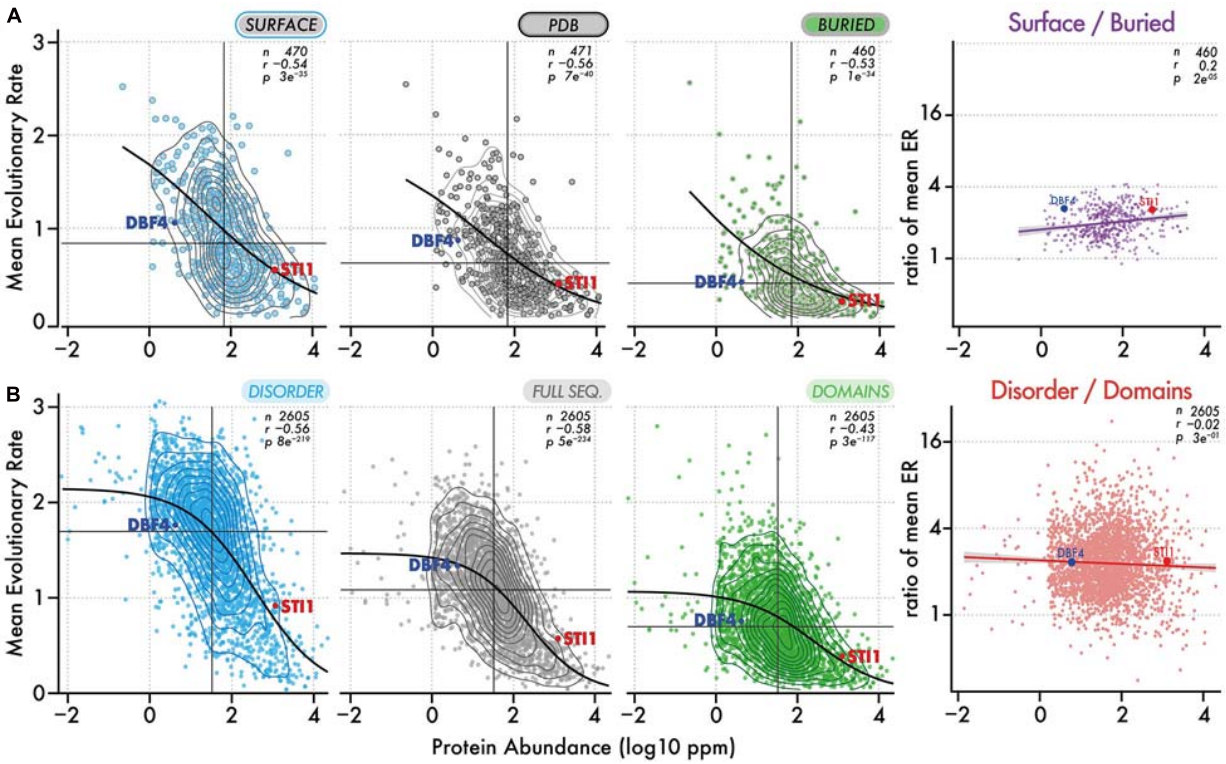


FIGURE 4 | Evolutionary rates of different regions and their ratio as a function of abundance. **(A)** Evolutionary rates (*y*-axis) as a function of protein abundance (*x*-axis) for surface regions, full-length structures, and buried regions. The ratio of evolutionary rate for surface vs buried regions is also shown as a function of abundance. Contour lines show the density of points. The median evolutionary rate and median protein abundance are shown by a vertical and horizontal line, respectively. The Spearman rank correlation coefficient and *p*-value are given with the number of proteins in each dataset. A black line shows the fitted sigmoidal regression for each plot. We highlight two proteins, one with a low and one with a high abundance (DBF4 and STI1). Both have comparable structural features but different evolutionary rates. **(B)** Same representation as in panel **(A)**, now considering disordered versus domain regions.

Evolutionary Constraints Imparted by Protein Abundance Scale Similarly Among Surface, Buried, and Disordered Regions

We saw that surface residues in a protein evolve twice as fast as buried residues on average. This difference, which has long been recognized, is mainly explained by solvent-accessibility/packing density and reflects that protein structures are more likely to be destabilized by mutations at buried positions than by mutations at the surface (Koshi and Goldstein, 1995; Goldman et al., 1998; Guo et al., 2004; Bloom et al., 2006; Sasidharan and Chothia, 2007; Goldstein, 2008; Conant and Stadler, 2009; Franzosa and Xia, 2009; Liberles et al., 2012; Yeh et al., 2014; Echave et al., 2015; Shahmoradi and Wilke, 2016; Spielman and Wilke, 2016; Echave and Wilke, 2017; Liu et al., 2017). Similarly, residues in disordered regions evolve faster than those in domains. Interestingly, this reflects that surface, buried, and disordered residues experience different structural and biophysical constraints. Thus, we propose to examine whether the ratio of their conservation is changing as a function of abundance. For example, observing that buried residues are twice more conserved than surface residues among low abundance proteins, and become four-times more conserved among high abundance proteins would suggest that stability is increasingly constrained with higher abundance.

We analyzed the ratio of conservation (**Figures 3A**, **4A**) of surface and buried residues as a function of abundance. The distribution of these ratios showed comparable median values of about ∼2. In the highest abundance class, this ratio reached ∼2.2 (**Figure 3A**) creating a significant albeit weak ($R = 0.2$) correlation (**Figure 4A**). Overall, the ratio is relatively stable, implying that both regions are constrained to a similar extent with increasing abundance. Alternatively, a relatively constant ratio could be favored by the coupling we observed between interior and surface regions (**Figure 2**, top row). Accordingly, constraints placed on the protein surface could percolate to interior regions and vice versa (Tóth-Petróczy and Tawfik, 2011). To control for this effect, we next compared disordered and domain regions, which show minimal structural coupling. We also observed a stable ratio of ∼2 across the five same abundance classes (**Figure 3B**), and we observed no dependence of the ratio with abundance even at the highest levels ($R = -0.02$, **Figure 4B**). Additionally, focusing on disorder and domain regions increased the size of the dataset as we were not limited by the availability of atomic-resolution structures, so this observation applies to the yeast proteome.

By definition, disordered regions and domains should experience distinct structural and biophysical constraints. Thus, the fact that these two regions appear equally constrained with increasing abundance is puzzling and can be interpreted in different ways. One possible explanation is that constraints associated with abundance apply to entire sequences independently of structure. Such constraints could include translational selection (Akashi, 2003), although region-specific codon-bias constraints may exist as well (Tuller et al., 2010; Pechmann and Frydman, 2013), cost of expression (Dekel and Alon, 2005; Wagner, 2005; Cherry, 2010; Gout et al., 2010; Plata et al., 2010), as well as other functional elements and sequence properties that may impact transcription or translation (Stergachis et al., 2013; Zhou et al., 2016). Alternatively or in addition, region-specific structural and biophysical constraints associated with protein concentration could increase in similar proportions with abundance, resulting in a stable ratio. In this case, two primary constraints have been characterized: a first on protein stability (Serohijos et al., 2012, 2013) leading to selection against misfolding (Drummond et al., 2005; Drummond and Wilke, 2008), would dominate among interior residues. A second, on protein solubility (Knowles et al., 2014; Garcia-Seisdedos et al., 2017, 2018; Dubreuil et al., 2019; Foy et al., 2019; Macossay-Castillo et al., 2019; Vecchi et al., 2020), with selection against promiscuous interactions (Deeds et al., 2007; Levy et al., 2009, 2012; Liberles et al., 2011; Yang et al., 2012), would dominate among solvent-exposed residues. However, the fact that constraints on different regions scale proportionally with abundance may appear surprising and will need to be explored in future works.

CONCLUSION

We analyzed the evolutionary conservation of sites within proteins, and of proteins within proteomes. We found that disordered regions evolve about three-fold faster than buried regions, and 1.4-fold faster than surface regions. Additionally, disordered regions evolve about as fast as the most solvent-exposed surface regions, highlighting that they extend the continuum of protein structure as a "super-accessible" surface. Unlike regular surface residues, however, disordered regions evolve more independently from domains in the same sequence. This independence allowed us to examine how abundance constrains different regions that are not structurally connected in sequences. Notably, the evolution of disordered regions and domains changed in a similar proportion with abundance: on average, disordered regions evolved twice as fast as domains across the entire range of abundance. Since different regions are subject to different structural and biophysical constraints, we foresee that such comparative analyses of conservation-ratios as a function of abundance will help identify mechanisms underlying the abundance-conservation relationship. It is likely that multiple mechanisms are at play (Mehlhoff et al., 2020) and may be captured by targeted analyses of specific regions and protein subsets.

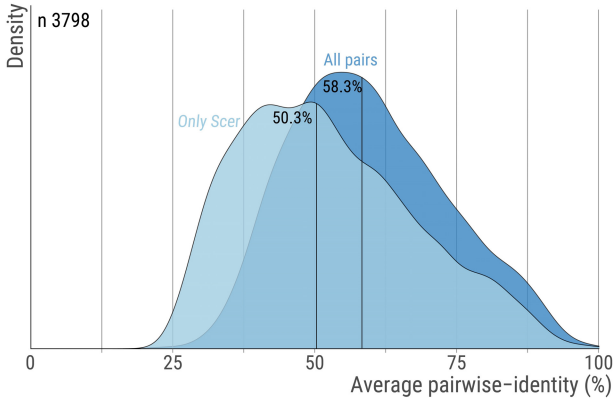


FIGURE 5 | Pairwise sequence identity across orthologs pairs. For each orthogroup we calculate the average percent sequence-identity using all ortholog pairs or only pairs that include the *S. cerevisiae* protein. The distribution for these two measures are shown with dark and light blue, respectively. Vertical lines highlight the median. The number of orthogroups is 3,798.

MATERIALS AND METHODS

Reference Proteome Sequences

The sequences were taken from the reference *S. cerevisiae* proteome maintained by SGD (Cherry et al., 2012). To facilitate data integration, we also mapped those reference sequences against the UniprotKB complete proteome for *S. cerevisiae* (Stutz et al., 2006; UniProt Consortium, 2019).

Crystallographic Structures

We relied on the 3DComplex database (Levy et al., 2006) to map UNIPROT sequences onto atomic coordinates of protein structures. For each yeast protein, the structures matching the UNIPROT sequence with the largest sequence overlap (minimum 20%) and identity above 90% were retained. Only experimentally determined crystallographic structures with resolutions below 3.0 Ångtrsoms were considered.

Cellular Abundance

Protein abundances were obtained from Pax-Db (v4.0, May 2015) (Wang et al., 2012, 2015), which provides relative abundances for unicellular and multicellular organisms including tissue-specific data. We use overall abundance inferred from all available data sets (integrated data set).

Orthologs Alignment and Position-Specific Evolutionary Rate

The orthologs' alignments were obtained from the original work by Wapinski et al. (2007). Briefly, genes sharing significant sequence similarity were clustered into putative orthogroups and their phylogeny was constructed by a modified neighbor-joining procedure based on pre-computed residues similarities and shared synteny scores. This process was repeated and optimized until each orthogroup consisted

of genes that shared a single common ancestor. Here, we used 3798 groups of orthologous proteins along with their multiple sequence alignment encompassing 14 fungal species (*S.cerevisiae, Saccharomyces paradoxus, Saccharomyces mikatae, Saccharomyces bayanus, Naumovozyma castellii (Saccharomyces castellii), Candida glabrata, Kluyveromyces lactis, Debaryomyces hansenii, Yarrowia lipolytica, Eremothecium gossypii (Ashbya gossypii), Lachancea waltii (Kluyveromyces waltii), Candida albicans, Aspergillus nidulans, Fusarium graminearum, Magnaporthe grisea, Neurospora crassa, Cryptococcus neoformans, Schizosaccharomyces pombe)* were used. Only 6 orthogroups had one sequence missing and these were replaced by indels. The median pairwise sequence identity within these 3,798 orthogroups is 58.3% (**Figure 5**).

All alignments were computed using MUSCLE (Edgar, 2004) and then concatenated to estimate residue-level evolutionary rate using the software Rate4Site (Pupko et al., 2002). Additional details on how evolutionary rates were estimated are available in Landry et al. (2009).

Intrinsic Disorder Predictions

We predicted disordered regions in the yeast proteome by combining short and long disorder segments predicted by IUPred (Mészáros et al., 2009; Dosztányi, 2018). We considered the 20% amino-acid residues with the highest disorder probabilities among all proteins. In all analyses, we required a minimum number of 20 residues in a particular region to calculate an average evolutionary rate. When fewer residues were available, the average rate of the region was considered undefined.

Domains Assignment

To assign domains, we aligned profiles from Pfam-A (v27.0, May 2013) (Bateman et al., 2002; Finn et al., 2014) and SUPERFAMILY (v1.75, March 2013) (Gough, 2002; Oates et al., 2015) to reference proteome sequences, filtering the hits with an *E*-value score above 10^{-3}. Finally, domain residues are those that were identified as part of a hit from either Pfam, SUPERFAMILY, or both.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s. Data used in this work are available on Figshare in a tabulated format: https://doi.org/10.6084/m9.figshare.13738657.

AUTHOR CONTRIBUTIONS

BD and EL designed the analyses and experiments, analyzed the data, and wrote the manuscript. BD carried out the analyses. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Israel Science Foundation (grant No. 1452/18), by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 819318), by a research grant from A.-M. Boucher, by research grants from the Estelle Funk Foundation, the Estate of Fannie Sherr, the Estate of Albert Delighter, the Merle S. Cahn Foundation, Mildred S. Gosden, the Estate of Elizabeth Wachsman, and the Arnold Bortman Family Foundation.

ACKNOWLEDGMENTS

We thank H. Greenblatt for helping with the computer infrastructure and Tal Pupko for his advice.

REFERENCES

Akashi, H. (2003). Translational selection and yeast proteome evolution. *Genetics* 164, 1291–1303.

Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Gutmanas, A., Anyango, S., Choudhary, P., et al. (2019). PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* 48, D335–D343.

Banani, S. F., Lee, H. O., Hyman, A. A., and Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18, 285–298.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., et al. (2002). The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280.

Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B. J., et al. (2011). Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* 12:R14.

Bloom, J. D., and Adami, C. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. *BMC Evol. Biol.* 4:14. doi: 10.1186/1471-2148-4-14

Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. (2006). Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* 23, 1751–1761.

Cherry, J. L. (2010). Expression level, evolutionary rate, and the cost of expression. *Genome Biol. Evol.* 2, 757–769.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., et al. (2012). *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, D700–D705.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature* 248, 338–339.

Chothia, C. (1975). Structural invariants in protein folding. *Nature* 254, 304–308.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105, 1–12.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5, 823–826.

Chothia, C., and Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harb. Symp. Quant. Biol.* 52, 399–405.

Conant, G. C., and Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol. Biol. Evol.* 26, 1155–1161.

Creighton, T. E., and Chothia, C. (1989). Protein structure. Selecting buried residues. *Nature* 339, 14–15.

Deeds, E. J., Ashenberg, O., Gerardin, J., and Shakhnovich, E. I. (2007). Robust protein protein interactions in crowded cellular environments. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14952–14957.

Dekel, E., and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436, 588–592.

Dignon, G. L., Zheng, W., and Mittal, J. (2019). Simulation methods for liquid–liquid phase separation of disordered proteins. *Curr. Opin. Chem. Eng.* 23, 92–98.

Dosztányi, Z. (2018). Prediction of protein disorder based on IUPred. *Protein Sci.* 27, 331–340.

Drummond, D. A., and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134, 341–352.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14338–14343.

Dubreuil, B., Matalon, O., and Levy, E. D. (2019). Protein abundance biases the amino acid composition of disordered regions to minimize non-functional interactions. *J. Mol. Biol.* 431, 4978–4992.

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208.

Echave, J., and Wilke, C. O. (2017). Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Annu. Rev. Biophys.* 46, 85–103.

Echave, J., Jackson, E. L., and Wilke, C. O. (2015). Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys. Biol.* 12:025002.

Echave, J., Spielman, S. J., and Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* 17, 109–121.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230.

Foy, S. G., Wilson, B. A., Bertram, J., Cordes, M. H. J., and Masel, J. (2019). A shift in aggregation avoidance strategy marks a long-term direction to protein evolution. *Genetics* 211, 1345–1355.

Franzosa, E. A., and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* 26, 2387–2395.

Fraser, H. B., and Hirsh, A. E. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol. Biol.* 4:13. doi: 10.1186/1471-2148-4-13

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science* 296, 750–752.

Galea, C. A., Nourse, A., Wang, Y., Sivakolundu, S. G., Heller, W. T., and Kriwacki, R. W. (2008). Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. *J. Mol. Biol.* 376, 827–838.

Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N., and Levy, E. D. (2017). Proteins evolve on the edge of supramolecular self-assembly. *Nature* 548, 244–247.

Garcia-Seisdedos, H., Villegas, J. A., and Levy, E. D. (2018). Infinite assembly of folded proteins in evolution, disease, and engineering. *Angew. Chem. Int. Ed. Engl.* 58, 5514–5531. doi: 10.1002/anie.201806092

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149, 445–458.

Goldstein, R. A. (2008). The structure of protein evolution and the evolution of protein structure. *Curr. Opin. Struct. Biol.* 18, 170–177.

Gough, J. (2002). The SUPERFAMILY database in structural genomics. *Acta Crystallogr. D Biol. Crystallogr.* 58, 1897–1900.

Gough, J., and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30, 268–272.

Gout, J.-F., Kahn, D., Duret, L., and Paramecium Post-Genomics Consortium (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* 6:e1000944. doi: 10.1371/journal.pgen.1000944

Guo, H. H., Choe, J., and Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9205–9210.

Hahn, M. W., and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22, 803–806.

Hirsh, A. E., and Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature* 411, 1046–1049.

Hurst, L. D., and Smith, N. G. (1999). Do essential genes evolve slowly? *Curr. Biol.* 9, 747–750.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12, 962–968.

Kauzmann, W. (1959). "Some factors in the interpretation of protein denaturation11the preparation of this article has been assisted by a grant from the national science foundation," in *Advances in Protein Chemistry*, eds C. B. Anfinsen, M. L. Anson, K. Bailey, and J. T. Edsall (Cambridge, MA: Academic Press), 1–63.

Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314, 1938–1941. doi: 10.1126/science.1136174

Knowles, T. P., Vendruscolo, M., and Dobson, C. M. (2014). The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* 15, 384–396.

Koshi, J. M., and Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein Eng. Des. Sel.* 8, 641–645.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13, 2229–2235.

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132.

Landry, C. R., Levy, E. D., and Michnick, S. W. (2009). Weak functional constraints on phosphoproteomes. *Trends Genet.* 25, 193–197. doi: 10.1016/j.tig.2009.03.003

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400.

Lesk, A. M., and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136, 225–270.

Levy, E. D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* 403, 660–670.

Levy, E. D., De, S., and Teichmann, S. A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20461–20466.

Levy, E. D., Landry, C. R., and Michnick, S. W. (2009). How perfect can protein interactomes be? *Sci. Signal.* 2:e11.

Levy, E. D., Pereira-Leal, J. B., Chothia, C., and Teichmann, S. A. (2006). 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* 2:e155. doi: 10.1371/journal.pcbi.0020155

Liao, B.-Y., Scott, N. M., and Zhang, J. (2006). Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol. Biol. Evol.* 23, 2072–2080.

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., et al. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* 21, 769–785.

Liberles, D. A., Tisdell, M. D. M., and Grahnen, J. A. (2011). Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. *Proc. Biol. Sci.* 278, 1930–1935.

Lim, W. A., and Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* 339, 31–36.

Lin, Y.-S., Hsu, W.-L., Hwang, J.-K., and Li, W.-H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol. Biol. Evol.* 24, 1005–1011.

Liu, J.-W., Lin, J.-J., Cheng, C.-W., Lin, Y.-F., Hwang, J.-K., and Huang, T.-T. (2017). On the relationship between residue structural environment and sequence conservation in proteins. *Proteins* 85, 1713–1723.

Lopez-Bigas, N., De, S., and Teichmann, S. A. (2008). Functional protein divergence in the evolution of Homo sapiens. *Genome Biol.* 9:R33.

Macossay-Castillo, M., Marvelli, G., Guharoy, M., Jain, A., Kihara, D., Tompa, P., et al. (2019). The balancing act of intrinsically disordered proteins: enabling

functional diversity while minimizing promiscuity. *J. Mol. Biol.* 431, 1650–1670. doi: 10.1016/j.jmb.2019.03.008

Mehlhoff, J. D., Stearns, F. W., Rohm, D., Wang, B., Tsou, E.-Y., Dutta, N., et al. (2020). Collateral fitness effects of mutations. *Proc. Natl. Acad. Sci. U.S.A.* 117, 11597–11607.

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* 5:e1000376. doi: 10.1371/journal.pcbi.1000376

Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* 196, 641–656.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.

Oates, M. E., Stahlhacke, J., and Vavoulis, D. V. (2015). The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res* 43, D227–D233.

Pal, C., Papp, B., and Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931.

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nat. Rev. Genet.* 7, 337–348.

Park, C., Chen, X., Yang, J.-R., and Zhang, J. (2013). Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 110, E678–E686.

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20, 237–243.

Plata, G., and Vitkup, D. (2018). Protein stability and avoidance of toxic misfolding do not explain the sequence constraints of highly expressed proteins. *Mol. Biol. Evol.* 35, 700–703.

Plata, G., Gottesman, M. E., and Vitkup, D. (2010). The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biol.* 11:R98.

Popescu, C. E., Borza, T., Bielawski, J. P., and Lee, R. W. (2006). Evolutionary rates and expression level in *Chlamydomonas*. *Genetics* 172, 1567–1576.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl. 1), S71–S77.

Razban, R. M. (2019). Protein melting temperature cannot fully assess whether protein folding free energy underlies the universal abundance–evolutionary rate correlation seen in proteins. *Mol. Biol. Evol.* 36, 1955–1963.

Rocha, E. P., and Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* 21, 108–116.

Rose, P. W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* 45, D271–D281.

Russo, A. A., Jeffrey, P. D., Patten, A. K., Massagué, J., and Pavletich, N. P. (1996). Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* 382, 325–331.

Sällström, B., Arnaout, R. A., Davids, W., Bjelkmar, P., and Andersson, S. G. E. (2006). Protein evolutionary rates correlate with expression independently of synonymous substitutions in *Helicobacter* pylori. *J. Mol. Evol.* 62, 600–614.

Sasidharan, R., and Chothia, C. (2007). The selection of acceptable protein mutations. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10080–10085.

Schmid, A. B., Lagleder, S., Gräwert, M. A., Röhl, A., Hagn, F., Wandinger, S. K., et al. (2012). The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *EMBO J.* 31, 1506–1517.

Serohijos, A. W. R., Lee, S. Y. R., and Shakhnovich, E. I. (2013). Highly abundant proteins favor more stable 3D structures in yeast. *Biophys. J.* 104, L1–L3.

Serohijos, A. W. R., Rimas, Z., and Shakhnovich, E. I. (2012). Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* 2, 249–256.

Shahmoradi, A., and Wilke, C. O. (2016). Dissecting the roles of local packing density and longer-range effects in protein sequence evolution. *Proteins Struct. Funct. Bioinf.* 84, 841–854.

Shakhnovich, B. E., Deeds, E., Delisi, C., and Shakhnovich, E. (2005). Protein structure and evolutionary history determine sequence space topology. *Genome Res.* 15, 385–392.

Shrake, A., and Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79, 351–371.

Sikosek, T., and Chan, H. S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface* 11:20140419.

Spielman, S. J., and Wilke, C. O. (2016). Extensively parameterized mutation–selection models reliably capture site-specific selective constraint. *Mol. Biol. Evol.* 33, 2990–3002.

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., et al. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 342, 1367–1372.

Stutz, A., Bairoch, A., and Estreicher, A. (2006). UniProtKB/Swiss-Prot: the protein sequence knowledgebase. *FEBS J.* 273, 62–62.

Subramanian, S., and Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168, 373–381.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D. S. (2007). The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* 369, 1318–1332.

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579, 3346–3354.

Tóth-Petróczy, A., and Tawfik, D. S. (2011). Slow protein evolutionary rates are dictated by surface-core association. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11151–11156.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., et al. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141, 344–354.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515.

Uversky, V. N., and Dunker, A. K. (2010). Understanding protein non-folding. *Biochim. Biophys. Acta* 1804, 1231–1264.

Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., et al. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* 6, 2351–2366.

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.

Vecchi, G., Sormanni, P., Mannini, B., Vandelli, A., Tartaglia, G. G., Dobson, C. M., et al. (2020). Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proc. Natl. Acad. Sci. U.S.A.* 117, 1015–1020.

Wagner, A. (2005). Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* 22, 1365–1374.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., et al. (2005). Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5483–5488.

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15, 3163–3168. doi: 10.1002/pmic.201400441

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., et al. (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* 11, 492–500.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449, 54–61.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 70, 697–701. doi: 10.1073/pnas.70.3.697

Wright, P. E., and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18–29.

Xia, Y., Franzosa, E. A., and Gerstein, M. B. (2009). Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput. Biol.* 5:e1000413. doi: 10.1371/journal.pcbi.1000413

Yang, J. R., Liao, B. Y., Zhuang, S. M., and Zhang, J. Z. (2012). Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 109, E831–E840.

Yeh, S.-W., Huang, T.-T., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., et al. (2014). Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed. Res. Int.* 2014: 572409.

Zhang, J., and Yang, J. R. (2015). Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* 16, 409–420.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.-H., Fu, J., et al. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci. U.S.A.* 113, E6117–E6125.

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 30 April 2021
doi: 10.3389/fmolb.2020.612920

Universal Architectural Concepts Underlying Protein Folding Patterns

Arun S. Konagurthu[1], Ramanan Subramanian[1], Lloyd Allison[1], David Abramson[2], Peter J. Stuckey[1,3], Maria Garcia de la Banda[1] and Arthur M. Lesk[4,5]**

[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Clayton, VIC, Australia, [2]Research Computing Center, University of Queensland, Brisbane, QLD, Australia, [3]School of Computing and Information Systems, University of Melbourne, Melbourne, VIC, Australia, [4]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, United States, [5]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

What is the architectural "basis set" of the observed universe of protein structures? Using information-theoretic inference, we answer this question with a dictionary of 1,493 substructures—called *concepts*—typically at a subdomain level, based on an unbiased subset of known protein structures. Each *concept* represents a topologically conserved assembly of helices and strands that make contact. Any protein structure can be dissected into instances of concepts from this dictionary. We dissected the Protein Data Bank and completely inventoried all the concept instances. This yields many insights, including correlations between concepts and catalytic activities or binding sites, useful for rational drug design; local amino-acid sequence–structure correlations, useful for *ab initio* structure prediction methods; and information supporting the recognition and exploration of evolutionary relationships, useful for structural studies. An interactive site, PROÇODIC, at http://lcb.infotech.monash.edu.au/prosodic (click), provides access to and navigation of the entire dictionary of concepts and their usages, and all associated information. This report is part of a continuing programme with the goal of elucidating fundamental principles of protein architecture, in the spirit of the work of Cyrus Chothia.

Keywords: architectural concepts, protein-building blocks, structural motifs, lossless compression, information theory, folding pattern

OPEN ACCESS

Edited by:
Alfredo Iacoangeli, King's College London, United Kingdom

Reviewed by:
Nick Grishin, Quantitative Biomedical Research Center, United States
Patrick Senet, Université de Bourgogne, France

****Correspondence:***
Arun S. Konagurthu
arun.konagurthu@monash.edu
Arthur M. Lesk
aml25@psu.edu

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *01 October 2020*
Accepted: *16 December 2020*
Published: *30 April 2021*

Citation:
Konagurthu AS, Subramanian R, Allison L, Abramson D, Stuckey PJ, Garcia de la Banda M and Lesk AM (2021) Universal Architectural Concepts Underlying Protein Folding Patterns. Front. Mol. Biosci. 7:612920. doi: 10.3389/fmolb.2020.612920

1 INTRODUCTION

The polypeptide chains of amino acids (primary structure) contain, in most proteins, regions that fold into helices and strands of sheets (secondary structure), which in turn assemble to give proteins their intricate three-dimensional shapes and folding patterns (tertiary and quaternary structures). As of April 2021, experimental methods have already provided more than 167,000 entries in the Protein Data Bank (PDB) (Berman et al., 2003), containing the three-dimensional coordinates of proteins and protein–nucleic acid complexes from a wide range of species.

Unraveling protein architecture and discovering the relationship among these major levels of structural description provide the key to understanding how proteins function, how their 3D folding patterns form, and how they evolve (Lesk, 2016). Investigations of protein folding patterns have revealed recurrent themes (Pauling and Corey, 1951; Pauling et al., 1951; Levitt and Chothia, 1976; Lesk and Rose, 1981; Chothia and Lesk, 1986; Richards and Kundrot, 1988), which form the basis for widely used hierarchical classifications of protein structures (Murzin et al., 1995; Orengo et al., 1997; Andreeva et al., 2013; Schaeffer et al., 2016). Nevertheless, many aspects of the relationships across structural levels remain unresolved. Further, François Jacob observed that proteins evolve by "*bricolage*," that is, through evolutionary tinkering by reusing "pieces" from other proteins

(Jacob, 1977; Duboule and Wilkins, 1998). Despite much previous work to unravel these "pieces," the problem of precisely characterizing them has remained open.

Chothia and Lesk (1986) introduced the idea of a *core* of the folding patterns of homologous proteins. This core comprises a maximal set of secondary structural elements (SSEs) that assemble in a common 3D topology, while withstanding a certain amount of distortion. The parts outside the core are structurally more variable.

Many related proteins share some but not all of the substructures that form their cores. Therefore, it is of great interest to discover the nature of the substructures that contribute to the cores of protein families. Some of these are *supersecondary structures*—small recurrent combinations of *successive* elements of secondary structure, such as the β-α-β subunit. Supersecondary structures recur within many protein folds and can be shared even by unrelated proteins. For example, the β-α-β subunit appears in NAD-binding domains, in TIM barrels, and in many other proteins.

Early definitions of supersecondary structures relied strongly on experts' spotting and naming them (Rao and Rossmann, 1973; Kister, 2013). With the steady growth of the PDB, several methods have been developed to identify automatically, with varying operational definitions, a *library* of substructures that form what can be considered as the 3D building blocks of protein structures (Unger et al., 1989; Rooman et al., 1990; Unger and Sussman, 1993; Camproux et al., 1999; Micheletti et al., 2000; Kolodny et al., 2002; Friedberg and Godzik, 2005; Joseph et al., 2010; Chitturi et al., 2016; Dybas and Fiser, 2016; Mackenzie et al., 2016; Nepomnyachiy et al., 2017; de Oliveira et al., 2018; Joshi, 2018). However, these approaches have yielded limited libraries containing mostly short oligopeptide fragments, or assemblies of typically 2–4 secondary structural elements. It has been a challenge so far to go further than that and dissect protein structures into a more complete set that includes *larger* conserved substructures. (A more detailed exploration of key prior work on this topic is provided under "Comparison with previous work" within the "Results" section.) Apart from the enormous computational challenge this problem poses, the attempts made so far have lacked a rigorous framework in which to describe, compute, identify, and resolve a dictionary of conserved assemblies of secondary structures.

Thus, the key focus of this work is to go beyond definitions of recurring substructural patterns that are identified using *ad hoc* formulations and adjustments. This work utilizes new statistical models to describe all observed protein folding patterns in terms of their substructural constituents. It provides an attempt toward a systematic description of recurrent substructures of protein folding patterns using methodological devices never previously explored in the literature on this topic. Finally, this work is broadly analogous (in scope and application) to finding a formalized description of "syntactic structures" that now underpins linguistic analyses of natural languages (Chomsky, 1957).

Specifically, this work unravels observed protein folding patterns into a dictionary of architectural building blocks (*concepts*) containing topologically conserved assemblies of helices and strands that make contact. We note that several databases such as SCOP (Murzin et al., 1995; Andreeva et al., 2013; Chandonia et al., 2017), CATH (Orengo et al., 1997), and ECOD (Schaeffer et al., 2016) classify protein structures at the level of domains, and include multiple instances of domains with very similar structures. *Concepts*, in contrast, provide a dictionary of independent structural patterns, into which full domains can be dissected.

We distinguish concepts both from motifs and from domains as follows:

- We understand the term motifs to mean recurrent structural patterns in proteins that can—in their entirety or partially—be superposed with low root-mean-square deviation of the backbone (or at least of the C_α) atoms. The idea of a concept focuses instead on conservation of the *topology* of secondary structure assembly, but instances of the same concept in different proteins can less-rigidly preserve structure and have varying lengths.
- Domains in proteins are individual compact units. Although some concepts do correspond to domains, some are not in themselves entirely compact, some are subdomains, and others comprise portions or even all of multiple domains.

The determination of the dictionary was completely automatic (i.e., *unsupervised*), and unbiased by any previously known sequence or structural patterns. Our framework to infer this dictionary can be best understood as an imaginary communication between a transmitter and receiver pair over a communication (Shannon) channel. The transmitter has a collection of protein shapes she wants to share with the receiver. The transmitter has two possible methods of communication. The first involves communicating the collection *as is*—this constitutes the null or baseline model. But another approach is to communicate the whole collection more efficiently using a dictionary of concepts, followed by the details of the collection specified with the aid of that stated dictionary. Here, the role of the dictionary is to illuminate common patterns observed in the collection and is stated one-off over all shapes in the collection. It is intuitive to observe that the better a dictionary, in terms of its ability to describe (i.e., fit) the shapes in the collection, the more economical will be the description of the source collection. An optimal dictionary in this framework is the one that yields the most economical one-off statement of the dictionary and the collection using that dictionary.

Our approach relies on an information-theoretic framework that allows the inference of a dictionary that a) avoids overfitting (i.e., avoiding inferring a dictionary that is more complex than necessary to explain the observed folding patterns) and b) achieves an objective trade-off between the descriptive complexity of concepts in the dictionary and their fidelity (i.e., the amount of compression) gained when explaining the observed protein folding patterns. This dictionary of concepts advances the current knowledge of conserved subdomain structural patterns significantly beyond the classical supersecondary structures and other known patterns. Thus, this work presents a "basis set" of substructural concepts underlying all observed protein folding patterns, and allows

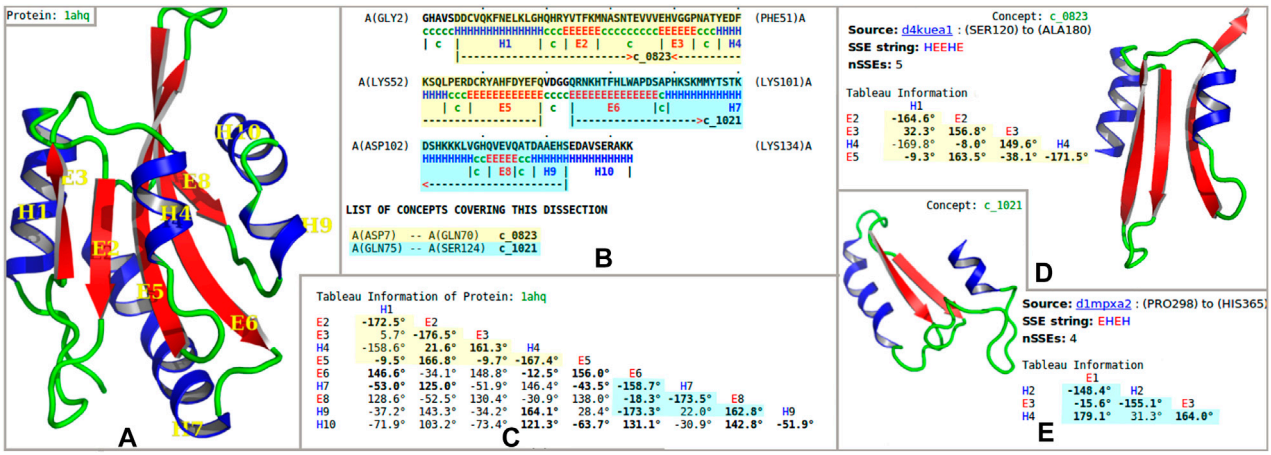


FIGURE 1 | Example dissection of an actin-binding protein into "concepts" from the inferred dictionary. **(A)** Secondary-structural cartoon representation of the crystal structure of the actin-binding protein actophorin from *Acanthamoeba* (1AHQ) (Leonard et al., 1997). **(B)** Secondary structural assignment [using SST (Konagurthu et al., 2012); H = helix, E = strand of sheet] and the optimal dissection of the protein chain into nonoverlapping regions, using the inferred concept dictionary. This information is shown with reference to the amino-acid sequence information in a marked-up format: the dissection of 1AHQ uses *concepts* (see the text) c_0823 (highlighted in yellow) and c_1021 (highlighted in blue). **(C)** Tableau representation of the folding pattern of 1AHQ. The highlighted subtableaux corresponds to concepts c_0823 and c_1021. Here, only the lower-triangle part of the tableau information is shown because the full tableau is a symmetric matrix. The rows and columns are indexed by secondary structure elements in order of appearance in the polypeptide chain. Off-diagonal elements record the angles between the pairs of secondary structural elements; boldface indicates that there is a contact between the corresponding pair of secondary structural elements. **(D, E)** The concepts c_0823 and c_1021 are shown, together with their *archetypal* tableaux and corresponding secondary structural representation.

any protein chain to be decomposed optimally into parts corresponding to substructures from this set. It thereby contributes a plethora of useful biological insights, such as the following:

1. Understanding the fundamental components of protein folding patterns. Our dictionary of concepts will support innovative projects aimed at the analysis of protein structures.
2. Correlation, in many cases, of concepts with functions directly, or indirectly *via* ligand-binding sites. This provides useful predictions in the case of proteins with known structure but unknown function.
3. Many concepts show amino-acid sequence correlation; that is, some conservation of sequence patterns. These results are applicable to protein structure prediction by suggesting conformations of local regions.

The results of dissecting all the structures in the current PDB, or of dissecting a user-supplied set of protein coordinates, are accessible from the PROÇODIC website: http://lcb.infotech.monash.edu.au/prosodic (click). This site supports the interactive exploration of protein structures and their relationships.

2 RESULTS

2.1 Automatic Inference of a Dictionary of Substructural Concepts

This work uses the concise *tableau* representation of protein folding patterns introduced by Lesk (1995), which is based on the idea that the essence of a protein folding topology is captured by the order, patterns of contacts, and geometry of the assembly of secondary structural elements along the amino-acid chain. A tableau corresponds to the 3D structure of a single-protein domain (or sometimes chain), and has the form of a symmetric matrix (**Figures 1A,C**). Importantly, in this representation supersecondary structures can be defined in a compact and computable way as subtableaux containing two or more *successive* secondary structure elements in contact (**Figures 1D,E**).

We have constructed the dictionary reported here using our recently developed method to infer, automatically, conserved assemblies of secondary structural elements within *any* given source collection of tableaux (Subramanian et al., 2017). We call these substructures *concepts*. This idea of a concept is constrained by the requirement that every secondary structural element in the concept must be in contact with at least one other secondary-structure element in that concept. Our concept inference approach (Subramanian et al., 2017) is based on the minimum message length criterion for statistical inference (Wallace and Boulton, 1968; Wallace, 2005; Allison, 2018) and lossless data compression. We have applied this method to compress the source collection of tableaux corresponding to ASTRAL SCOP domains (Murzin et al., 1995; Andreeva et al., 2013; Chandonia et al., 2017). This has allowed us to infer a dictionary of 1,493 substructural concepts that *most concisely* and *losslessly* describes the entire source collection, and does so without any prior knowledge or preconceived notions regarding these recurrent substructures.

The total computational effort required to identify this dictionary is equivalent to about 7 years of runtime on a

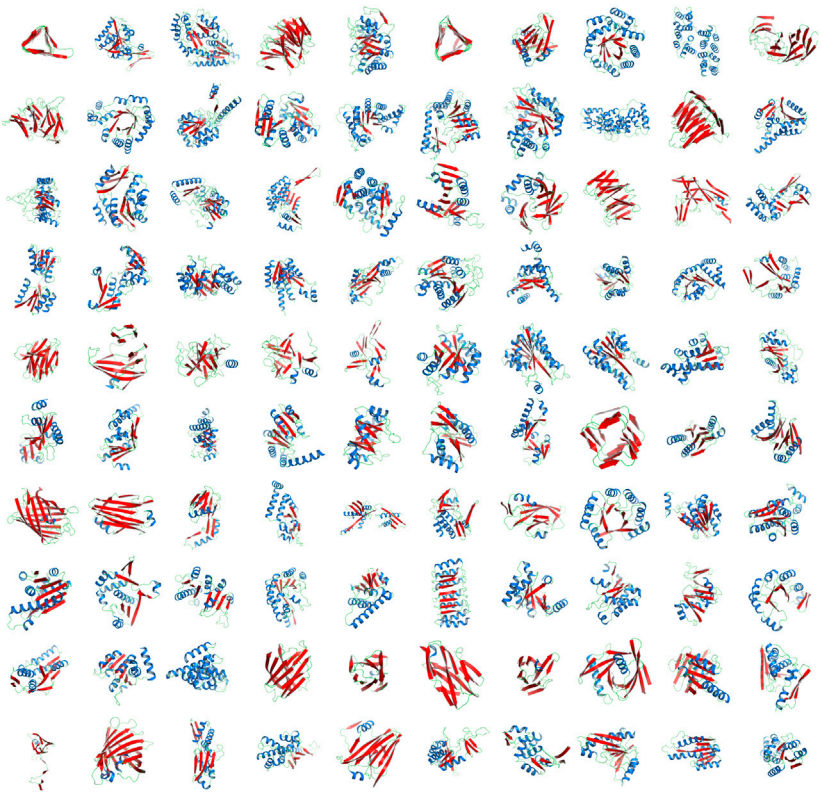


FIGURE 2 | Top 100 concepts from the inferred dictionary. The representative structural cartoons of the top 100 concepts from the inferred dictionary containing 1,493 concepts, ranked in a decreasing order of number of secondary-structure elements (row-wise top-left to bottom-right: c_0001 to c_0100). Strands of sheet are shown in Red; helices in Blue. (See the website for the full interactive listing.) The inference of the whole dictionary was automatic without any prior knowledge or preconceived notions of these recurrent themes. The inferred concepts subsume known patterns; for example, shown in the figure are: "α-β Barrel" (c_0005), "Armadillo repeat" (c_0083), "β Barrel" (c_0061), "β Propeller" (c_0004), "Icosahedral virus coat protein" (c_0067), Immunoglobulin (c_0062), "Jellyroll architecture" (c_0084), "Left-handed β-Helix" (c_0001), "Leucine-rich repeat" (c_0076), "Right-handed quadrilateral β-Helix" (c_0058) "NAD-binding domain" (c_0002), "TIM barrel" (c_0008), etc. Other classical supersecondary structures not shown in this figure such as β-hairpin (c_1442), α-hairpin (c_1484), β-α-β unit (c_1240) appear lower down in the dictionary of concepts, ordered from largest to smallest.

modern computer. Therefore, we parallelized our method and ran it on a high-performance computing cluster using 240 cores to identify the PROÇODIC dictionary in 14 days (see **Section 4**).

2.2 PROÇODIC: The Dictionary of Inferred Concepts

Each of the 1,493 concepts in the dictionary is designated by an identifier of the form "c_" followed by 4 digits: c_0001—c_1493. This order follows 1) the decreasing length in the number of secondary structural elements (nSSEs) defining each concept, and 2) for concepts containing the same number of SSEs, the lexicographic order of their secondary structural strings, where we represent any helix by "H" and any strand by "E."

Figure 2 shows the top 100 concepts in the dictionary, ordered by number of SSEs included. The largest concept (c_0001) contains 28 secondary structural elements. The smallest concepts (c_1441—c_1493)—not shown in **Figure 2**—contain only two elements. (Note that a single-helix or a single-strand/extended region is not considered here as a concept.) The distribution of inferred concept sizes is shown in **Figure 3A**: 9 concepts (c_0001—c_0009) are composed of an assembly of ≥ 20 secondary structural elements, 48 concepts (c_0010—c_0057) have between 15 and 19 SSEs, 217 concepts (c_0058—c_0274) contain between 10 and 14 SSEs, 217 concepts (c_0058—c_0274) contain between 10 and 14 SSEs, and 368 concepts (c_0275—c_0642) contain between 9 and 6 SSEs. The remaining concepts contain between 5 and 2 SSEs. The median concept size is 5 SSEs.

On average, a concept archetype is significantly smaller (with 47.6% of the number of SSEs) than its source protein domain. Yet, there are several concepts inferred in our dictionary that describe conserved folding patterns at the level of domains. These include: NAD-binding domain (c_0173), β-grasp fold (e.g., c_729), β-propeller (c_0382), Swiss/Jelly roll fold (c_0406), Ferredoxin (plait) fold (c_0581), TIM barrel (c_0008), Immunoglobulin fold (c_0118, c_0121), Ubiquitin roll (c_0737), and large β-barrel (c_0061). This shows that our dictionary encompasses a broader set of substructural invariants than previous studies (see **Section 2.5**). This advantage is due to our use of tableaux to capture concisely the essence of protein folding patterns, together with the information-theoretic criterion of minimum message length

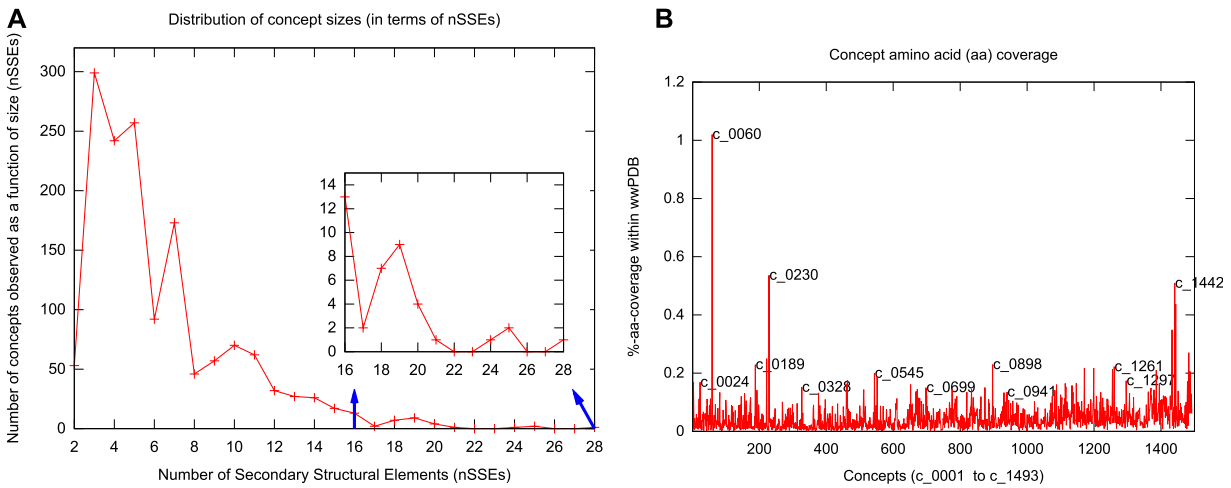


FIGURE 3 | Distributions of concept length and amino-acid coverage. **(A)** Distribution of concept lengths in terms of the number of secondary structural elements (nSSEs) they contain. The smallest concepts have 2 secondary structural elements; the largest has 28. **(B)** Individual concept amino-acid coverage (as a percentage of the total 74,246,836 residues) in the serial order of concept identifiers (with some concepts highlighted).

to yield an objective dictionary complexity-versus-fidelity trade-off.

The null model encoding length of our source collection is 33,352,380 bits. The encoding length after compressing the same collection using the inferred dictionary is 31,927,340 bits. The resultant compression is 1,425,040 bits (or 4.3%) over the null model. We emphasize that this compression gain is over the null model encodings of the tableaux representations which are themselves compact 2D representations of 3D structural information.

The complete inferred dictionary is available *via* the interactive website PROÇODIC (for Protein Concept dictionary—the cedilla allows the pronunciation as "prosodic") at http://lcb.infotech.monash.edu.au/prosodic. As discussed later, this site allows the exploration of any structure that the user provides as input, or of specific concepts that are of motivating focus for the user, including: the usages of concepts in other structures, both homologous and nonhomologous; or the inspection of frequently occurring keywords within the "KEYWDS" records and the ligand-binding information from the "HETATM" records extracted from the source PDB coordinate files (see **Section 3**).

2.3 Our Dictionary Subsumes Known Supersecondary Structural Motifs

Our dictionary includes many concepts that match the known repertoire of supersecondary structural motifs (Efimov, 2013). Matched motifs involving assemblies of a small number of helices and strands include: antiparallel (c_1442) and parallel (c_1443) β-β assemblies, α-α hairpin (c_1484) α-β/β-α assembly, (c_1459/c_1472), basic helix–loop–helix (c_1351), β-α-β motif (c_1240), EF-Hand (c_1342, c_1491), ϕ-motif (c_1178), helix–turn–helix motifs (c_0826 – winged type I, c_0870 – winged type II, c_1373 – plain), four-helix bundle (c_1101 – type I, c_1117 – type II), β-meander (c_1187), Greek key (c_0964), Zinc finger (c_1230), helix–hairpin–helix motif (c_1068), β-sandwich (c_0390), and $\alpha\beta$-sandwich (c_0603), among others.

Our dictionary also includes larger assemblies of helices and strands that match known *repeating* structural motifs. These include three-sided left-handed β-helix (c_0001, c_0380), three-sided right-handed β-helix (c_0388), right-handed quadrilateral β-helix (c_0058), ankyrin repeat (c_0370, c_0632), armadillo repeat (c_0083, c_0888), kelch repeat (c_0395), α-solenoid (c_0270, c_0271), and leucine-rich repeat (c_0076), among others.

PROÇODIC yields a flat (nonhierarchical) dictionary of 1,493 concepts. The inference of these concepts is unsupervised, driven by information-theoretic trade-off between the dictionary complexity and its fidelity to explain the source collection of tableaux. Visual inspection reveals shared topological relationships between certain subsets of concepts (e.g., c_0001 and c_0006; see **Figure 2**). Therefore, to explore the topological relationships between the inferred concepts, we undertake an agglomerative clustering exercise to construct a hierarchy from that otherwise flat dictionary of concepts. We emphasize that this exercise is *not* meant to suggest any structural pathways [cf. Efimov structural trees (Efimov, 2013)] or evolutionary relationships between concepts, but merely provides a device to explore their topological relationships. (We also emphasize that a systematic approach to finding hierarchical relationships and structural pathways requires the unsupervised Bayesian inference of a hierarchical dictionary of concepts, which is beyond the scope of the current work.)

To undertake this agglomerative clustering, since each concept archetype defines a (sub)tableau derived from a tableau of the domain in the source collection, we can infer the dictionary of *meta-concepts* (i.e., concepts of "concepts") that best explains all the PROÇODIC concept tableaux. This is achieved by using exactly the same unsupervised (flat dictionary) inference methodology

that was used to infer PROÇODIC concepts. That is, we now treat the tableaux representing 1,493 archetypes from our inferred PROÇODIC concept dictionary as the source collection, and rerun our inference method (see **Section 4**). This in turn yielded 34 meta-concepts that dissect (i.e., best explain) the inferred 1,493 concepts. The text file containing these meta-concepts, along with the corresponding list of PROÇODIC concepts that use each meta-concept within their dissections, is available in the supporting data file: metaConceptsAndUsageList.txt (click).

This permits the decomposition of each PROÇODIC concept in terms of these 34 meta-concepts. Thus, each PROÇODIC concept can be represented as a 34-dimensional feature vector in the meta-concept space, where each vector component denotes the number of times the corresponding meta-concept is used in that concept dissection. We note that this representation is similar to the bag-of-words model (Harris, 1954) used in information retrieval and natural language processing. Using this feature vector representation, the 1,493 PROÇODIC concepts are clustered hierarchically using the following method:

1. A $1,493 \times 1,493$ similarity matrix between PROÇODIC concepts is constructed using the cosine similarity measure (Singhal, 2001) between all the pairs of these 34-dimensional vectors.
2. Using the resultant similarity matrix, we cluster all the PROÇODIC concepts hierarchically, based on the unweighted pair-groups method using arithmetic averages (Sokal, 1958).

This procedure yields a hierarchical tree of concept relationships, available in an interactive format from: prosodicConceptClustering.html (click). This tree reveals similarities that are also detectable by comparing the concept archetypes, their usages, and keywords. For example, c_0009 and c_0018 are both helical bundles related to the architecture of Annexin proteins, with c_0009 having one extra helix compared to c_0018. Another example is the cluster containing c_0001, c_0006, c_0113, and c_0380, where all represent left-handed β-helical motifs composed of 28, 20, 12, and 7 β-strands, respectively.

2.4 Dissection of PDB and Coverage of Concepts Across the Protein Folding Space

The methods used for this work also permit the optimal *dissection*, within seconds (on a single processor), of any protein chain into nonoverlapping regions that are explained (compressed) using the concepts from the inferred dictionary. **Figure 1** shows an example of the dissection of the crystal structure of the actin-binding protein actophorin from *Acanthamoeba* (1AHQ) (see the PROÇODIC website to dissect any protein structure of interest; either a PDB entry or a user-supplied coordinate set). We note that regions not assigned to any dictionary concept (notionally designated to the *null* concept, c_0000) remain uncompressed. These include the small set of proteins that have no secondary structure, for instance wheat-germ agglutinin (9WGA).

We have dissected the entire PDB, which at the time of calculation resulted in tableaux corresponding to 275,014 protein chains containing 74,246,839 amino-acid residues overall. (Note that the dictionary was constructed using an *unbiased* set of domains from ASTRAL, but the subsequent dissection of the entire PDB reflects the biases in the distribution of protein folding patterns in the full database.) The usages of the resulting concepts cover regions within proteins that account for 66.35% (49,262,577) of the total (74,246,839) amino acids in the PDB protein chains we dissected (**Supplementary Figure S3A**). The remaining 33.65% is dominated by single secondary structural elements, plus loops between successive concept assignments along a dissected chain. **Figure 3B** shows the distributions of amino-acid coverage of concept usages within the PDB. Concept c_0060 has the largest coverage in terms of the number of amino acids its usages cover. This concept is composed of 14 secondary structural elements (SSE string: EEEEHHEEEEHHEE) assembling into a four-layer architecture, with its core containing two layers of closely packed five-stranded β-sheets (Chothia et al., 1977) that are sandwiched between two outer layers, containing two α-helices each (see **Figure 2**, the rightmost structure on the sixth row). In total, this concept was used within 3,892 protein chains, with a median value of amino-acid coverage equal to 194 residues (**Supplementary Figures S3A,B**). Examination of these usages reveals that they all come from the protein chains of 285 proteasome complexes. At the other extreme is concept c_0568, which has the smallest amino-acid coverage: 561 residues over 13 protein chains related to plant and bacterial Ferredoxins (Tagawa and Arnon, 1962). This concept is composed of 6 secondary structural elements (SSE string: EEHEEE).

Insights about the concepts can be gained from their usage information. For example, consider the concepts c_0060 and c_0568 mentioned earlier: the concept c_0060 covers the β5 subunit of a recently solved structure of the native human 20S proteasome at 1.8 Å resolution (5LE5) (Schrader et al., 2016). This landmark study revealed a number of functionally important differences with respect to what was known from the previously published 20S proteasome structures. In particular, it identified chloride ions within all active sites, thus significantly revising the description of the proteasome active site, and providing new insights into peptide hydrolysis that underpin the "development of next-generation proteasome-based cancer therapeutics" (Schrader et al., 2016). The examination of the usages of c_0060 within the dissection of 5LE5 (chain Y – β5 subunit) reveals that this concept is directly linked to proteolytic active sites (**Figure 4A**). Analyses of the human-annotated keywords used in the PDB coordinate files from these usages showed among its top 10 frequently used phrases terms such as "Cancer (therapy)," "Drug resistance," and "Bortezomib"—an anticancer drug and the first therapeutic proteasome inhibitor to be used in humans. This is strong evidence of the concept's link to a proteolytic active site. A similar examination of the usage instances of the concept c_0568 directly links it to the Fe_2S_2-cluster binding ferredoxins (see **Figure 4B**), which mediate electron transfer (Nechushtai et al., 2011).

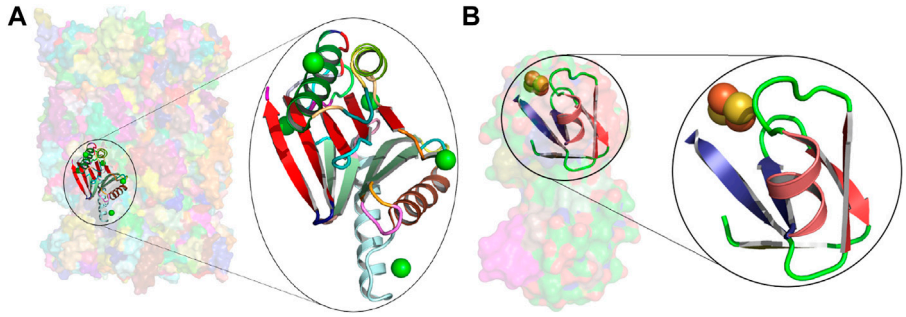


FIGURE 4 | Usages of concepts c_0060 and c_0568. **(A)** Transparent surface rendering of the native human 20S proteasome at 1.8 Å (5LE5), with the usage of concept c_0060 in the β5 subunit (chain Y in the amino-acid region THR1 to ASN191) shown in cartoon. The closeup of this region reveals a chloride ion in all active sites. Chloride ions are known to facilitate a proton shuttle catalytic mechanism (Schrader et al., 2016). **(B)** Similar rendering as above for the usage of concept c_0568 in the 2.3 Å Ferredoxin structure from *Mastigocladus laminosus* (3P63 chain A in the amino acid region THR48 to GLU90). The closeup shows the region linked to the Fe_2S_2-cluster binding.

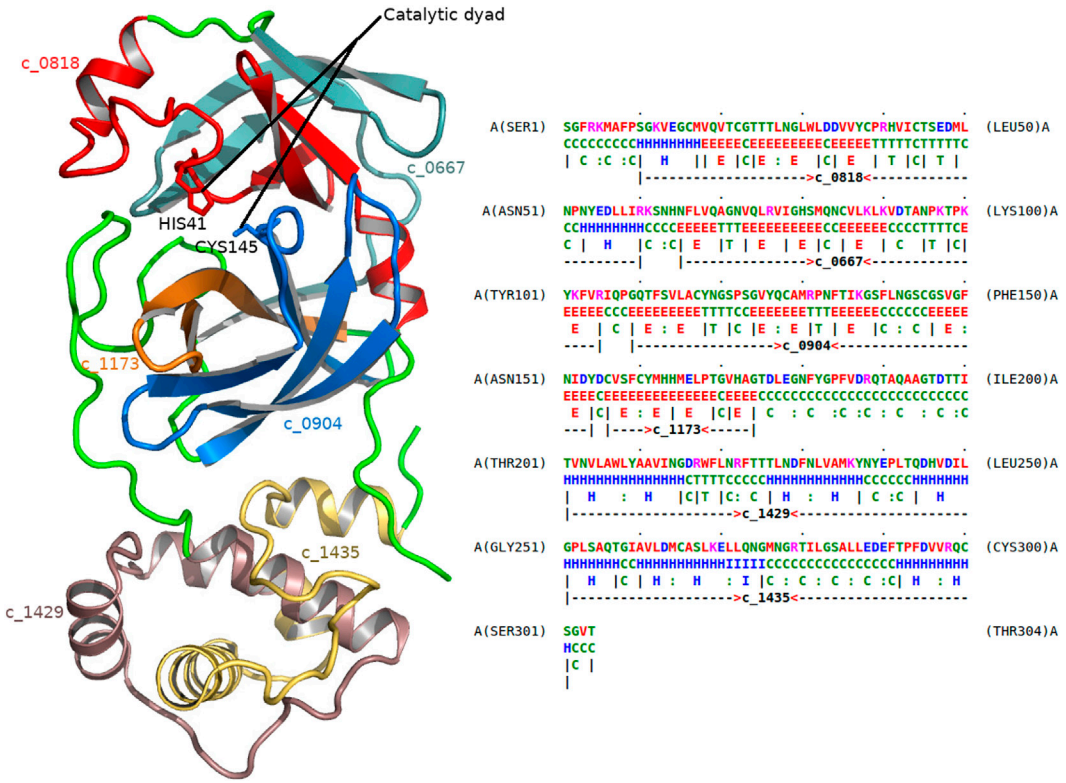


FIGURE 5 | Dissection of the main protease of SARS-CoV-2 virus. The left frame shows the 1.8 Å crystal structure of the main protease of SARS-CoV-2 (5R84). The right frame gives the dissection of this protein as markup under 5R84's amino-acid sequence (chain A). The successive regions of 5R84 chain A are explained using the following concepts (in that order): c_0818, c_0667, c_0904, c_1173, c_1429, and c_1435, respectively. Their corresponding substructural regions of the protease are shown with varying colors (left frame). Cysteine 145 (CYS145) and Histidine 41 (HIS41) residues form the catalytic dyad of this protease, and are associated with concepts c_0904 and c_0818, respectively.

As another example, consider the dissection of the main protease 5R84 (**Figure 5**) of the SARS-CoV-2 virus. This virus is the cause of the coronavirus disease (COVID-19). 5R84 is a cysteine protease that is responsible for cleaving the SARS-CoV-2 polyprotein chain that prepares the molecular machinery responsible for viral replication and infection. The dissection involves, among others, the following two concepts: c_0818 and c_0904. (For a full list of concepts in the dissection, see **Figure 5**.) Studying the usages of these concepts, it becomes clear that they are composed of highly conserved substructures that are specific

to viral proteases, mainly coronaviruses (SARS and MERS). Concept c_0904 explains the region of 5R84 containing the catalytic cysteine-145 residue (CYS145) of this main protease, whereas c_0814 explains the other residue in the catalytic-dyad, histidine-41 (HIS41). Therefore, these concepts are directly linked to the catalytic function of the protease.

2.5 Comparison With Previous Work

Many previous studies have attempted to identify a canonical set of recurrent patterns that encompass the structures of proteins.

Dissection of protein folding patterns into substructures began with the recognition of recurrent patterns. The first of these were the canonical secondary structures (α-helix and β-sheet) followed by descriptions of supersecondary structures (α-hairpin, β-hairpin, and β-α-β unit). At this point the approach was observation and intuition-based rather than systematic, and the field lacked attempts to determine a set of substructures from which *complete* domain structures could be assembled. The earliest attempts to generate a roster of supersecondary structures automatically, with varying motivations, include those of Lesk and Rose (1981), Jones and Thirup (1986), and Richards and Kundrot (1988).

To identify a set of building blocks that *cover* protein structures, Unger et al. (1989) analyzed protein main chain conformations in terms of hexamers (oligopeptides of six amino-acid residues). Their analysis involved a refined set of 82 proteins in the (then) known structures, which contributed to a total of 12,973 hexamers. Using a normalized root-mean-square-deviation (RMSD)–based membership function (with an RMSD threshold of 1 Å) and a variant of *K*-nearest-neighbor clustering, they demonstrated that most hexamers grouped into 55 disjoint clusters.

Much subsequent work followed along similar lines of clustering short oligopeptide fragments using variations of clustering heuristics and membership-deciding thresholds to produce different local fragment libraries (Tramontano et al. 1989; Rooman et al., 1990; Hutchinson and Thornton, 1996; Micheletti et al., 2000; Kolodny et al., 2002; Kihara and Skolnick, 2003; Friedberg and Godzik, 2005; Joseph et al., 2010). For instance, Micheletti et al. (2000) sought a minimal set of "oligons" that can represent protein structures, by clustering oligopeptide conformations extracted from known structures. They considered oligopeptide lengths from 4 to 7 and created libraries containing 8, 202, 932, and 2561 elements—within which they recognized redundancies. They were able to fit a set of test structures to within an RMSD of approximately 1 Å.

The main limitations of these approaches are at least two-fold: 1) The nature of the covering substructures is *imposed*—in these cases, short oligopeptide fragments—rather than allowing their method to identify more general possibilities, and 2) the definition of cluster membership of various oligopeptide fragments remains extremely sensitive to the chosen RMSD threshold values and clustering heuristic.

Complementing the above strategies that rely on clustering local 3D fragments, Bystroff et al. (1996) and Bystroff and Baker (1998) proposed a fully automated method to cluster short 1D sequence segments into a library (I-sites) of amino-acid patterns that correlate strongly with their 3D (local) structure. These sequence segments were clustered using a weighted amino-acid frequency profile (Vingron and Argos, 1989) over a *K*-means clustering approach. Subsequently, over an iterative procedure, pairs of peptide segments within each cluster are evaluated based on their structural characteristics (C_α–C_α distance profiles and backbone torsion angles) to select a "paradigm" local structure for their sequence cluster. Latest I-sites library (v5.3) reports 128 clusters containing motifs of length ranging from 3 to 15 amino acids. This popular library, together with the inferred local sequence–structure relationships, now underpins successful and popular *ab initio* structure prediction methods (Rohl et al., 2004). Despite being a noteworthy milestone in the literature, this library is not geared toward identifying topologically conserved assemblies of SSEs, which is the main focus of the work presented here.

Camproux et al. (1999) used an *a priori* method based on hidden Markov models (HMM) to identify a recurrent 3D structural alphabet. In their work, proteins are described using a sequence of overlapping tetrapeptide states on which a HMM is used to infer libraries of fragments together with their local conformational dependencies. This work mainly yielded 12 distinct tetrapeptide states derived from a data set of about 100 proteins. These states correspond predominantly to conformations of classical helices, strands, and turns, plus a few others. Further extension of this work (Camproux et al., 2004) gathered 27 tetrapeptide states. This work also examined the restrictions on the sequences of such states that appear in proteins. The inferred 27 tetrapeptide states correspond to α, 3_{10}, and π helices, extended strands, turns of various descriptions and coil, respectively. Using different models, Pandini et al. (2010) also clustered tetrapeptide fragments (using the internal angles between the C_α coordinates) from known proteins to determine another structural alphabet. Nevertheless, similar to the other libraries, these structural alphabets remain extremely short and limited in scope.

Going beyond the clustering of oligopeptide fragments, some key studies have iteratively assembled SSEs under specific rules to explore structural "pathways" of observed protein folds. Specifically, Efimov (1997) used a constructive approach to introduce the notion of "structural trees." These trees reveal how folding patterns can be constructed from root structural motifs *via* addition of helices and strands in a stepwise fashion, subject to a restricted set of growth-rules. Efimov examined five types of structural trees corresponding to five protein superfamilies. The key outcome of this work was the demonstration that the structural trees give pathways of growth that lead to known protein folding patterns. Murzin and Finkelstein (1988) presented a model for the possible arrangements of α-helices in globular proteins. Subsequently, Taylor (2002) also explored a similar idea. Taylor's work constructed idealized topologies of protein structures by applying SSE packing rules that build on a set of basic "forms." These forms are represented using stick models of SSEs in different layered arrangements, where the spacing between idealized helices (of arbitrary lengths) within a layer

is fixed to 10 Å, whereas that between idealized strands is set to 5 Å. To match any protein to the sets of idealized forms, a protein structure is converted to a stick representation and then a fast filtering step is applied to find potential matches (using a bipartite matching algorithm), followed by a more exhaustive pairwise comparison between the filtered stick forms and the proteins (based on a double-dynamic programming algorithm and RMSD threshold for match set to 6 Å.)

By demonstrating the limitation on the number of realizable folding patterns, arising due to the restrictions imposed by the growth rules on feasible spatial assemblies of SSEs, the studies by Efimov (1997) and Taylor (2002) confirm the observations of Finkelstein and Ptitsyn (1987) and Chothia (1992). Moreover, these works inform new schemes to classify the observed protein folds (Gordeev et al., 2010).

Grishin and colleagues (Chitturi et al., 2016) recently proposed a method to enumerate constructively all idealized *parallel/antiparallel* arrangements of up to 5 SSEs. This work proposed a systematic enumeration of all possible parallel/antiparallel arrangements using a 3D lattice model. This allowed them to model theoretical arrangements of SSEs and use them to search for observed occurrences of each arrangement within the PDB. However, their idealized models are limited to parallel/antiparallel orientations, which poses a severe restriction in exploring the full set of SSE arrangements observed in the PDB.

Alva et al. (2015) sought regions of proteins that might comprise a set of ancestral fragments, conceivably vestiges of a pre-cellular "RNA-peptide world." They identify 40 fragments, typically containing few secondary structure elements, that recur in many protein structures, including in sets of proteins not recognized as homologous. Some of these are similar to certain of our concepts; for instance, their set includes the standard supersecondary structures α-α hairpin, β-hairpin, and β–α–β unit. However, comprehensive coverage of observed protein folding patterns was not a goal of that work.

Other motif libraries have also been recently proposed: the Smotif library of Dybas and Fiser (2016) and the TERMs library of Mackenzie et al. (2016). An Smotif is designated by the arrangement of a *pair* of SSEs (of one of the following types: EE, EH, HE, and HH). A library of Smotifs is a collection of such SSE-pairs with different geometries. Their work utilizes an RMSD threshold of 2.5Å to cluster 11,068 observed pairs of SSEs from a collection of 1,200 protein structures (i.e., one randomly chosen protein domain per SCOP fold). These fragments serve in their work as the representatives of the protein structural space. Thus, any consecutive pair of secondary structures within a protein chain is assigned to the closest (based on RMSD) representative Smotif.

The tertiary motif (TERMs) library (Mackenzie et al., 2016) was able to find bigger assemblies of short oligopeptide fragments using the following approach. For each amino-acid residue i in the nonredundant collection of 29,000 residues, a candidate TERM is defined using one or more oligopeptide fragments formed by the union of the residues $i-2, \ldots, i+2$ together with all penta-peptide regions around residues that form a "potential contact" with the residue i. For each candidate TERM, the method finds matching tertiary fragments using an RMSD-based search method. A subset of candidate TERMs is realized by posing it as the classical set cover problem and realizing the minimal cover using a greedy approximation method that iteratively identifies the TERMs (based on their coverage) that match proteins in the considered set. This iterative procedure yields about *half a million* (458,251) TERMs. The minimum TERM has 1 oligopeptide fragment containing 5 amino acids, whereas the maximum TERM has 10 fragments with 52 amino acids. Importantly, an average TERM in their library is composed of 3 oligopeptide fragments covering 19 amino acids (i.e., 6 amino acids per fragment). Furthermore, inspecting the TERMs that cover 50% of their proteins in their considered collection of 29,000 protein structures, we find that each TERM averages 2 fragments with 12 amino acids. Moreover, inspecting the top 24 TERMs [see **Figure 2A** of Mackenzie et al. (2016)], we find many repetitions of short helices and antiparallel strands.

Nepomnyachiy et al. (2017) recently proposed a pipeline to explore "reuse" of regions in proteins based on their amino acid sequence relationships. This work reported repeated occurrences of sequence segments between 35 and 200 amino acids in length. However, relying on amino-acid sequence relationships is rather limiting because sequences diverge more drastically than structures in evolution.

In comparison, our work results in only 1,493 architectural concepts (two orders of magnitude more concise than TERMs), where our smallest concepts contain 2 SSEs covering, on average, 19 amino acids—this is the median length of the regions where concepts with 2 SSEs are used, in the dissections of the structures from the PDB. The biggest concept is composed of 28 SSEs covering 171 amino acids. An average PROÇODIC concept in our dictionary is composed of 6 SSEs covering 75 amino acids. Considering the PROÇODIC concepts that cover 50% of the PDB, an average concept has 5 SSEs covering 66 amino acids. Thus, using this framework, our dictionary yields concepts that are a substantially larger than TERMs, and define a significantly more economical dictionary that explains the entire PDB. Moreover, the methodology we use defines a direct and efficient (dynamic-programming based) way to dissect any given protein structure using the inferred PROÇODIC dictionary.

These results are achieved due to the expressive power of tableaux to represent compactly the essence of protein folding patterns. This tableau representation, together with the minimum message length inference methodology, provides a reliable framework to compress without loss and identify relationships in the protein folding space.

3 DISCUSSION

3.1 Many Concepts Are Linked to Ligand-Binding Sites

The molecular function of proteins is often mediated *via* interactions with chemical components such as metal ions, coenzymes, metabolic substrates, and nucleic acids, amongst others. Knowledge of such interactions is central to annotate protein function (Whisstock and Lesk, 2003; Goldstein, 2008), engineer new proteins (Gutteridge and Thornton, 2005), and

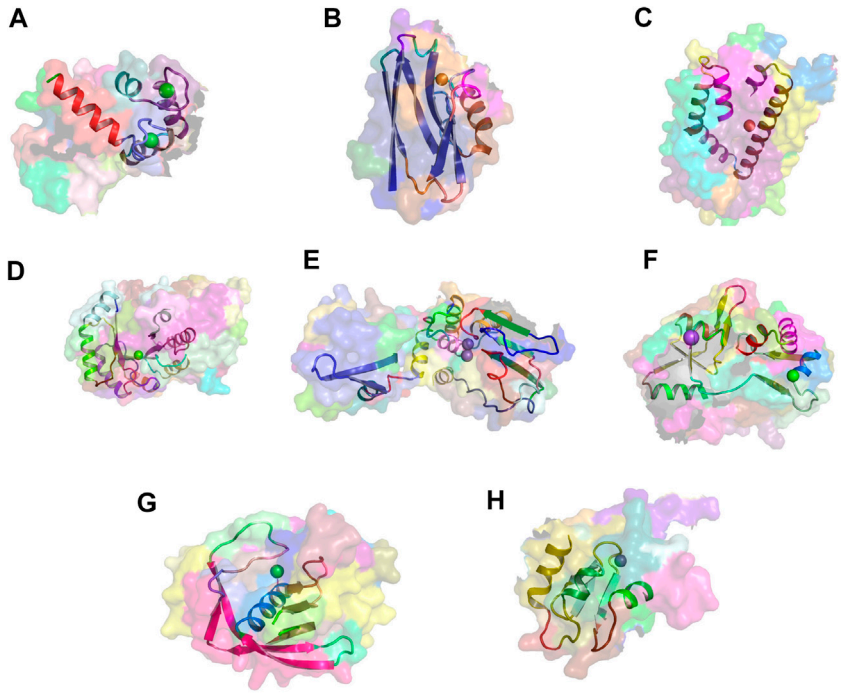


FIGURE 6 | Metal-binding activity examples. Exemplars of usages of eight concepts linked to metal-binding activity. The region of concept usage is shown in cartoon in the context of the surface rendering of the source protein chain. **(A)** Usage of concept c_1099 within the calcium-bound calmodulin [1CDL (Meador et al., 1992)]. **(B)** Usage of concept c_432 within the copper-bound electron transfer protein [1A4B (Messerschmidt et al., 1998)]. **(C)** Usage of concept c_885 within the iron-bound oxidoreductase (2VUX). **(D)** Usage of concept c_139 within the magnesium-bound lyase (3TTE). **(E)** Usage of concept c_186 within the manganese-bound hydrolase [1K23 (Ahn et al., 2001)]. **(F)** Usage of concept c_133 within the sodium-bound Kainate and AMPA receptors [3G3G (Chaudhry et al., 2009)]. **(G)** Usage of concept c_280 within the nickel-bound peptide deformylase (2AIA). **(H)** Usage of concept c_624 within the zinc-bound melanoma-inhibiting anti-apoptotic protein [1OY7 (Franklin et al., 2003)].

design drugs (Rognan, 2007; Kinjo and Nakamura, 2009). These functionally critical interactions impose structural constraints on protein structures, as their domains evolve from a common ancestor. As noted by Lesk and Chothia (1980), in many cases active sites are the best-conserved regions within a family of protein structures (as seen in **Figures 4**, **5**).

We have analyzed our dictionary and systematically identified concepts directly related to protein–ligand interactions. To achieve this, we mined and catalogued frequent ligand information (from "HETATM" records) derived from the source PDB entries of each concept usage (i.e., each instance in the PDB where the concept appears in the dissection of that protein's tableau). Our definition of a *ligand* comes from the inventory of 23,258 chemical components specified by the LigandExpo (Feng et al., 2004) database. We note that this inventory does *not* exclude simple monovalent ions (such as Na^+, K^+, and Cl^-) or those that are often not biologically functional (such as sulfate SO_4^{2-} ions). To complement this information, we also mined and cataloged keywords (from "KEYWDS" records) derived again from the source PDB entries of these concept usages. We used the observed frequencies of the bound ligands within the regions of concept usages, to narrow the initial set down to the 463 (31%) concepts that stand out in terms of recurrent patterns of interactions with the same set of ligand(s). These encompass interactions with monovalent ions, di-/tri-/tetra-valent ionic species, small molecules (including nucleotides), and macromolecular compounds, among others.

The fully annotated list of concepts with observed interactions with ligands/chemical components is available in the supporting data file: conceptsWithLigandInteractions.txt (click).

Figures 6A–G show examples of concept usages for a random selection of 8 concepts associated with metal-binding activity. **Table 1** shows a partial list of concepts for which all (100%) of their usages show binding to the specified ligand/chemical components. Also shown are the extracted high-frequency keywords associated with usages of that concept, providing useful insights to impute functional roles. Among the shortlisted set of 463 concepts are those that demonstrably show binding specificity linked with target recognition, reception, and signaling (see **Table 2**).

The full list of inferred concepts putatively linked to molecular reception, recognition, and signaling is available in the supporting data file: receptorConcepts.pdf (click).

3.2 Inferring Biological Function From Concept Usage Information

Many proteins are deposited into the PDB with unspecified functional annotation, especially those coming from structural

TABLE 1 | A partial list of concepts for which all (100% of) their usages show interactions with ligands or chemical components. This is derived by inspecting the ligand ("HETATM") records within the source coordinate files of each concept usage. The bound ligands are shown (in the second column) using their standardized abbreviations, along with their observed frequency within the usages in parentheses. Also shown (in the third column) are the top keyword terms (from "KEYWDS" records specified by the structures' authors) recurring within the usage coordinate files with their associated frequencies. (Note: CA = calcium ion).

Concept ID	Ligand/chemical component (freq)	Keyword (freq)
c_0011	PQQ (100%), CA (100%)	OXIDOREDUCTASE (90%), QUINOPROTEIN (27%)
c_0036	ZN (100%)	HYDROLASE (85%), EXOPEPTIDASE (46%), CARBOXYPEPTIDASE B (46%)
c_0065	FES (100%)	OXIDOREDUCTASE (96%), XANTHINE OXIDASE (32%), IRON SULFUR (30%)
c_0096	FMN (100%)	OXIDOREDUCTASE (100%), ROSSMANN FOLD (55%)
c_0108	HEM (100%), CA (100%)	OXIDOREDUCTASE (85%), PEROXIDASE (63%)
c_0110	HEM (100%)	OXIDOREDUCTASE (82%), MONOOXYGENASE (43%), CYTOCHROME P450 (34%)
c_0124	SF4 (100%), MG (100%)	OXIDOREDUCTASE (91%), [NIFE]HYDROGENASE (26%)
c_0144	CA (100%)	TRANSFERASE (81%), CGTASE (36%), ACARBOSE (33%)
c_0156	ZN (100%)	TRANSFERASE (90%), SET DOMAIN (39%), EPIGENETICS (28%)
c_0159	SF4 (100%)	OXIDOREDUCTASE (96%), NIFE HYDROGENASE (17%)
c_0208	CU (100%)	OXIDOREDUCTASE (97%), BETA BARREL (34%), LACCASE (32%)
c_0374	HEM (100%)	OXYGEN TRANSPORT (56%), HEMOGLOBIN (26%)
c_0397	ZN (100%)	OXIDOREDUCTASE (94%), SUPEROXIDE DISMUTASE (27%)
c_0424	PCA (100%)	HYDROLASE (95%), GLYCOSIDASE (35%), CELLULOSE DEGRADATION (32%)
c_0546	ZN (100%)	HYDROLASE (88%), PHOSPHODIESTERASE (32%), PDE (28%)
c_0568	FES (100%)	ELECTRON TRANSPORT (77%), FERREDOXIN (38%)
c_0604	HEM (100%)	ELECTRON TRANSPORT (100%), HEME (57%), CYTOCHROME B5 (40%)
c_0624	ZN (100%)	APOPTOSIS (47%), ZINC FINGER (44%), METAL BINDING (30%)
c_0714	NAG (100%)	VIRAL PROTEIN (84%), HEMAGGLUTININ (39%), GLYCOPROTEIN (22%)

TABLE 2 | A partial list of concepts putatively linked to molecular reception, recognition, and signaling.

Concept ID	Ligand/chemical component (freq)	Frequent keywords (freq)
c_0062	NAG (96%), BMA (70%)	IMMUNE RECOGNITION (21%)
c_0133	ZN (35%)	AMPA RECEPTOR (26%), NEUROTRANSMITTER RECEPTOR (20%)
c_0205	GAL (36%)	CARBOHYDRATE RECOGNITION (11%)
c_0252	MYR (40%)	RHINOVIRUS COAT PROTEIN (20%), RECEPTOR (17%), ANTIVIRAL COMPOUND (10%)
c_0304	CA (60%)	ANTIBODY RECEPTOR (18%), CARBOHYDRATE RECOGNITION DOMAIN (15%)
c_0335	NAG (34%)	CELL ADHESION (29%), RECEPTOR (16%), GLYCOPROTEIN (11%)
c_0352	GOL (67%)	PEPTIDOGLYCAN RECOGNITION PROTEIN (10%)
c_0423	NAG (63%)	IMMUNE SYSTEM (87%), ANTIGEN PRESENTATION (26%), T CELL RECEPTOR (12%)
c_0572	FMN (67%)	PHOTORECEPTOR (36%), LIGHT-INDUCED SIGNAL TRANSDUCTION (13%)
c_0819	ZN (58%)	SIGNALING PROTEIN (19%), PHOTORECEPTOR (13%)

genomic initiatives. Functional characterization of such proteins is of crucial importance to the structural biology community. Its importance can be evidenced by the community-wide Critical Assessment of protein Function Annotation program (CAFA, biofunctionprediction.org/cafa/), which assesses methods dedicated to predicting protein function from an amino-acid sequence.

As previously shown (**Figure 4A**), the rich source of information within this concept dictionary is useful to investigate and impute biological function. More evidence of this is shown by another case study involving the haze-forming thaumatin-like protein in white wines made from *Vitis vinifera* (4JRU containing 201 residues). **Figure 7** gives the dissection of 4JRU composed of two concepts c_0111 and c_1442.

Concept c_1442 is of less functional interest as it defines a common β-hairpin unit consisting of two antiparallel β-strands. On the other hand, c_0111 contains 12 strands that assemble to form mainly two face-to-face packed antiparallel β-sheets with an extended β-ribbon connected by an Ω-loop (Leszczynski and Rose, 1986). This multistranded motif is characteristic of thaumatin-like proteins (Ogata et al., 1992). Examining the usages of this concept within the PDB *via* our PROÇODIC web site, we find it is used at 15 other loci, most of them thaumatin/osmatin-like proteins, with their top two keywords displaying "antifungal protein (53.3%)" and "plant protein (46.7%)," respectively. **Figures 7B,C** show the structural alignment of 4JRU with the usage in the pathogenesis-related PR-5D protein of tobacco (*Nicotiana tabacum*; 1AUN with 208 residues) that results in a superposition with 1.47 Å root-mean-square deviation over 201 amino-acid residues between the C_α coordinates of the two structures. This specific PR-5D protein is classified functionally as an antifungal protein, and, in general, proteins of this class have known pathogenesis-related antifungal activity. This suggests that the haze-forming protein might exhibit the same biological function.

In some cases, the information provided by this dictionary can lead to a reliable but less-specific functional classification prediction, for example putatively identifying a general type of

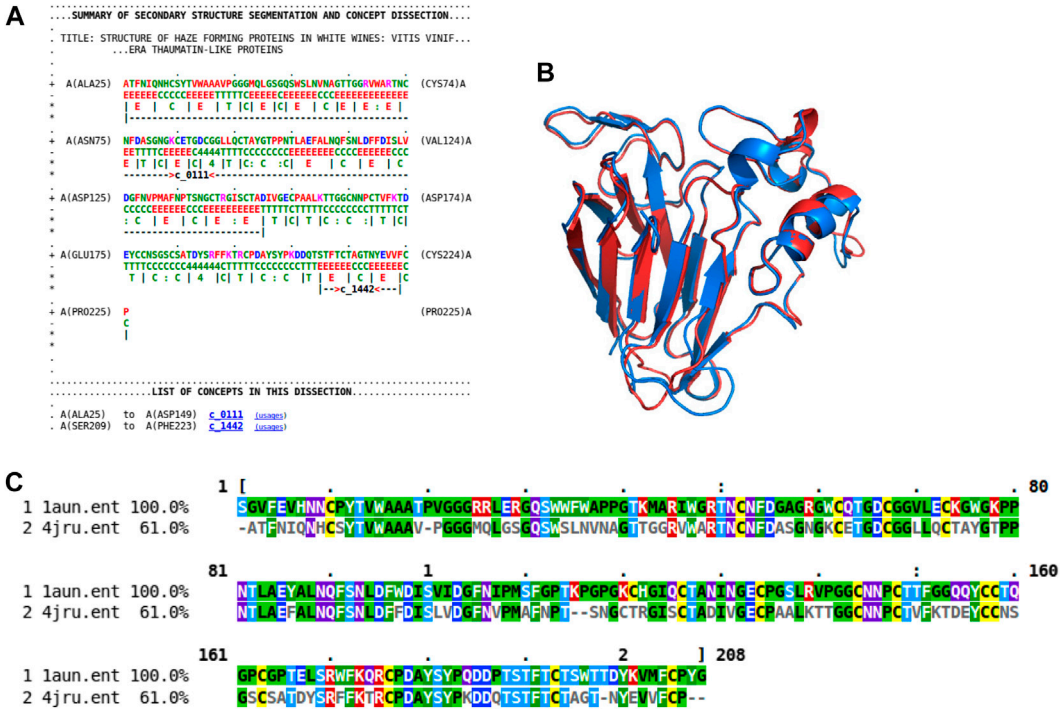


FIGURE 7 | Dissection case-study on the haze-forming thaumatin-like protein, 4JRU. **(A)** Dissection output from PROÇODIC of the haze-forming thaumatin-like protein in white wines from *Vitis vinifera* (4JRU). **(B)** Superposition of this haze-forming protein and the pathogenesis-related PR-5d protein of tobacco (*Nicotiana tabacum*; 1AUN). 4JRU is shown in blue; 1AUN is shown in red. This superposition was based on the structural alignment produced by MMLigner (Collier et al., 2017), which is shown in **(C)**.

function such as "oxidoreductase" or "lyase." Such generic functional classification can be useful, as it may provide guidance for laboratory experiments aimed at defining the function more precisely, especially if clues about a ligand-binding site are available. For example, consider the crystal structure of dihydrodipicolinate synthase (DapA) from *Agrobacterium tumefaciens* (2HMC). The dissection of this DapA structure shows the usage of concept c_0008 covering its entire chain A. About 90% of c_0008's 118 usages show the functional classification as "lyase." DapA belong to the family of amine-lyases that catalyze the cleaving of carbon–nitrogen bonds, playing an important role in lysine biosynthesis in prokaryotes, phycomycetes, and plants (Mirwaldt et al., 1995).

3.3 Local Sequence–Structure Correlation Within Concept Usages

The identification of structural features that have strong amino-acid sequence preferences is central to structure prediction (Bystroff and Baker, 1998). Therefore, we studied the concept usages within the PDB to explore the conformational preferences of local sequences. To achieve this, for each concept, the amino-acid sequences in the regions of concept usages within the PDB were extracted, and the sequences in each set were aligned and clustered (Sievers and Higgins, 2014).

Almost 20% of the concepts in our dictionary (288 out of 1,493) have associated amino-acid sequence patterns that cluster into a single group (**Supplementary Figure S4A**). When considering the (normalized) ratio of clusters over the number of *nonidentical* amino-acid sequences of concept usages (**Supplementary Figure S4B**), almost 30% of the concepts (441) have a ratio smaller than 0.05, whereas almost 50% (738) have a ratio between 0.05 and 0.1. Together, this indicates that for almost 80% of the concepts in our dictionary, their usages of amino-acid sequences cluster into a small number of groups (<10% of their total unique amino-acid sequences).

This strong sequence dependence is expected, particularly for concepts linked to ligand binding or other functional units. For example, **Figure 8** shows the sequence logo obtained from the multiple sequence alignment of the usage sequences of the concept c_0397. This concept is related to the Cu-Zn type I (SODI) superoxide dismutase, which has a β–barrel-like subunit with copper and zinc ions bound at the active site. This is common in many Gram-negative bacterial pathogens (amongst others) to counteract a burst of toxic superoxide radicals under oxidative stress (Forest et al., 2000).

There is a potential application of the observed sequence–structure correlations to structure prediction. We downloaded the coordinate files of 33 PDB structures specified in the description field of the CASP12 target list available at http://predictioncenter.org/casp12/targetlist.cgi. Each chain from these 33 structures was independently dissected using the PROÇODIC dictionary of concepts. The dissection of protein chains defines nonoverlapping regions assigned either to one

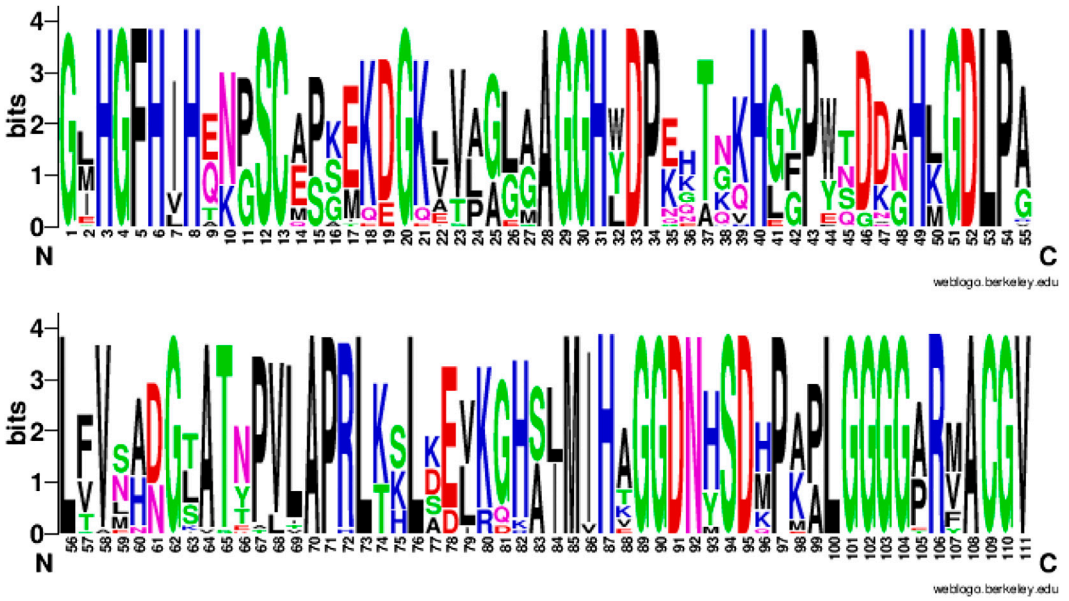


FIGURE 8 | Sequence consensus across the usages of concept c_0397. amino-acid sequence logo (in two parts: columns 1-55 and 56-111) showing the sequence consensus across the usages of a randomly chosen concept c_0397 directly related to the Cu-Zn binding superoxide dismutase. Of the 111 columns in the multiple sequence alignment (of the c_0397's 33 usage sequences) corresponding to this logo, 46 aligned columns show a consensus of 100%.

of the dictionary concepts (c_0001 – c_1493), or a *null* concept (c_0000). For each region assigned to a dictionary concept, we extracted the associated target amino-acid sequence and performed a pairwise *sequence* alignment with each of the local amino-acid sequences defined by the concept usages. This exercise identified a subset of concept usages in the PDB whose local amino-acid sequences have a detectable similarity with the target. **Table 3** quantifies the extent of coverage of these regions for each of the 33 CASP12 targets. This table shows that in 26 of 33 cases, more than 50% of the target amino-acid chain has detectable sequence similarity that can be derived from the usage information.

It should be noted that we used the structural information of CASP12 targets to dissect the protein chains, before identifying the sequence relationships of the target sequence and those within the concept usages. However, for the proper application to structure prediction, the identification of sequence hits with concept usages should be carried out using only the target sequence. In principle, this can be done by sliding along the target sequence with varying window sizes, and exploring the sequence similarity with the sequences across all usages of every concept in the dictionary. Nevertheless, this preliminary analysis can be used to hypothesize reasonably that these local sequence–structure relationships provide a strong potential to support structure prediction efforts, especially since an average concept usage spans significantly longer stretches along the protein chain than the currently considered oligopeptide-fragment libraries used by fragment-based *ab initio* protein modeling approaches. Thus, this information can be potentially utilized to model several nonoverlapping regions in the target protein chains by the structure-prediction servers (Kim et al., 2004; Källberg et al., 2012; Waterhouse et al., 2018; Zheng et al., 2019; Senior et al., 2020).

As a note on the latest breakthrough in the field of structure prediction, convolutional neural network-based prediction architectures [especially AlphaFold (Senior et al., 2020)] have seen groundbreaking success in the CASP13 and CASP14 rounds. These neural network methods train on multiple sequence alignments as inputs, involving either whole or part of the target sequence whose structure is being predicted. At the time of writing this article, the technical details of the AlphaFold system used in CASP14 remain unpublished. When these details become open, it would be useful to explore whether sequence–structure correlations at the level of concepts can be incorporated into training the neural networks more efficiently—as per the information disclosed by Google Deepmind the current architecture requires in the order of 128 Tensor Processing Units and over a few weeks to predict structure from sequence, but with groundbreaking accuracy.

The amino-acid subsequences of nonoverlapping regions dissected using the PROÇODIC dictionary of concepts are available at: casp12_prosodic_dissections.tgz (click). The information of dissected target region followed by other subsequences in the usages of the corresponding (assigned) concept with demonstrable sequence similarity (under pairwise sequence alignment with the target subsequence) is available at: casp12_concept_usage_hits.tgz (click). The multiple sequence alignments [using MUSCLE (Edgar, 2004) with default options] of the identified sequence hits are available at: casp12_concept_usage_hits_msa.tgz (click).

TABLE 3 | Statistics showing the extent of detectable sequence similarity on each of the 33 CASP12 targets with their PDBIDs specified at http://predictioncenter.org/casp12/targetlist.cgi. First column: PDBID of the 3D experimental structure of each CASP12 target. Second column: The coverage statistics in terms of the total number of amino acids (#a.a.) within the amino acid (sub-)sequences defined by the dissected regions of the target protein with detectable sequence similarity with amino acid (sub-)sequences of their corresponding concept usage instances (see the main text). Third column: The total number of amino acids in the target protein, cumulative over all chains. Fourth column: Percentage coverage = Second column*100/Third column.

Target's PDBID	#a.a.'s in regions with usage seq hits	Total #a.a.'s in all chains	%Covered
3JB5	1046	2076	50.4
4YMP	202	215	94.0
5A7D	4468	5065	88.2
5AOT	91	102	89.2
5AOZ	125	141	88.7
5D9G	154	502	30.7
5ERE	417	540	77.2
5FHY	155	458	33.8
5FJL	88	136	64.7
5G3Q	100	168	59.5
5G5N	580	1022	56.8
5HKQ	160	263	60.8
5IDJ	63	242	26.0
5J4A	339	440	77.0
5J5V	815	1065	76.5
5JMB	103	182	56.6
5JMU	172	219	78.5
5JO9	215	239	90.0
5JZR	203	262	77.5
5KKP	166	509	32.6
5KO9	73	253	28.9
5LEV	323	375	86.1
5M2O	171	211	81.0
5MQP	2674	4801	55.7
5NSJ	150	284	52.8
5NV4	713	1377	51.8
5SY1	786	1458	53.9
5T87	444	745	59.6
5TF2	331	338	97.9
5TJ4	2640	5462	48.3
5UNB	378	681	55.5
5UVN	954	2496	38.2
5UW2	211	332	63.6

3.4 Exploration of Substructures and Structural Relationships

In addition to the applications explored above, the dictionary can be used to complement standard protein structural studies. Researchers can approach the dictionary with a particular structure or family of structures in mind. For example, dissecting the human hemoglobin (1HHO, chain A) at the PROÇODIC website identifies the concepts c_0375, c_0894, and c_1410. Choosing one of the concepts, for example, c_0894, its archetype is found in d1x9fd, a globin from the annelid *Lumbricus terrestris*. Note that related proteins can present dissections into different concepts. However, these concepts may still be related (see **Section 3** on hierarchical clustering of concepts). Our dictionary subsumes known supersecondary structural motifs. For example, c_0375 and c_0894 are related concepts linked to globins, with the former being more elaborate (with three extra helices) than the latter. Examining the corresponding concept "usages" link on the PROÇODIC website reveals that many usages of these related concepts appear in other globins. **Supplementary Material S1** contains several examples of use of PROÇODIC to explore protein substructural similarities.

4 MATERIALS AND METHODS

4.1 Tableau Representation

Tableaux are concise two-dimensional representations of protein folding patterns (Lesk, 1995). A tableau represents a protein folding pattern in terms of a) the order of secondary structural elements that appear along the polypeptide chain, and b) the geometry of interactions of pairs of SSEs in contact. This provides a computable definition for protein folding patterns, useful to study many aspects of protein architecture (Kamat and Lesk, 2007; Konagurthu et al., 2008; Konagurthu and Lesk, 2010).

For a protein of known 3D structure, the construction of a tableau involves first assigning the SSEs from the set of coordinates. In this work, we assign the secondary structure using the program SST (Konagurthu et al., 2012). This identifies the order in which helices (H) and strands of sheet (E) appear along the polypeptide chain. The succession of SSEs in any protein therefore appears as a string of characters H or E. The relative orientation of each pair of SSEs is computed as a dihedral angle between two planes formed by the least-squares vectors fitting the Cα coordinates of each SSE (directed from N- to C-terminus) and their mutual perpendicular.

The geometry of pairs of SSEs is represented as a square-symmetric matrix of orientation angles, with rows and columns indexed by successive SSEs. A corresponding contact matrix stores the contact patterns between pairs of SSEs. Two SSEs are said to be in contact if there exists at least a pair of residues (one from each SSE) that are in contact. Two residues are in contact if there is at least one pair of atoms (one from each residue) the distance between which is less than the sum of their Van der Waals radii plus a small constant (1 Å).

The idea is that the essence of a protein folding pattern is contained in the SSEs, their contact patterns, and the relative orientations of pairs of SSEs in contact.

4.2 Source Collection Used for the Inference of PROÇODIC Dictionary of Concepts

A *source collection* is a collection of (source) tableaux $\mathcal{T}$. Since the full PDB has redundancy and bias in terms of entries with similar structures, to infer the dictionary of concepts we use the ASTRAL SCOP-95 (Murzin et al., 1995; Andreeva et al., 2013; Chandonia et al., 2017) (v2.05) dataset that has been produced to remove bias due to over-represented structures, while explicitly incorporating structure quality at each step of the domain selection (Brenner et al., 2000). This data set is composed of 26,949 domains, representing only 12% of the full SCOPe (v2.05) domain dataset. Of these, 13,365 domains have < 40% sequence similarity to its closest neighbor. Although the maximum

sequence similarity two proteins can share is 95%, the average sequence similarity is significantly lower (< 53%). The full list of ASTRAL SCOP-95 domains used to infer the reported dictionary is available in the supporting data file: prosodicInferenceList.txt (click). Further, the inferred PROÇODIC dictionary was used to dissect the PDB (Berman et al., 2003). Analysis presented here includes the dissections of 113,724 protein coordinates files: prosodicDissectedWWPDBList.txt (click). In addition to these dissections, the PROÇODIC website allows users interactively to dissect any protein structure on demand.

4.3 Definitions of a Concept and a Dictionary of Concepts

Any subtableau comprising ≥ 2 consecutive rows and columns is potentially a concept, provided that the graph defined by the corresponding contact matrix is connected. (An undirected graph is said to be connected if there exists a path between any pair of vertices.) The rationale for this definition is supported by the analysis by Kamat and Lesk (2007) who demonstrated that almost all the information required to identify a folding pattern is inherent in the local structure, which can be captured using successive diagonals of a tableau. We also note that relaxing the concept definition to general subtableaux with nonconsecutive SSEs would render the problem of finding an optimal set of concepts computationally intractable.

A candidate *dictionary* $\mathcal{C}$ is a set of concepts. Any possible dictionary is a set of substructures, each satisfying the definition of a concept, that appear in the source collection. Our goal is to determine the optimal concept dictionary to *explain* the entire source collection as efficiently as possible. Technically, this is the dictionary that gives the most (lossless) compression of the source collection.

Associated with each concept $c \in \mathcal{C}$ is a concentration parameter, κ, corresponding to a von Mises circular (angular) probability distribution (Mardia and Jupp, 2009). This parameter controls the assignment of probabilities used to estimate the encoding length of entries in Ω when compressing regions of the source tableaux. That is, κ controls the *flexibility* of an inferred concept. A smaller/larger κ_y yields greater/lesser flexibility of the concept's usages for compressing source tableaux regions. These values are inferred as a part of the dictionary search (see **Algorithm 1**).

4.4 Inference of Dictionary of Concepts

We recently described a lossless compression-based methodology to infer recurrent subtableaux on any source collection of tableaux using the Minimum Message Length (MML) criterion (Subramanian et al., 2017). The dictionary we report here has been subsequently inferred using the methodology described in that work. For convenience of the reader, the overview of our methodology is summarized later, and we refer the reader to our published methodology (Subramanian et al., 2017) for formal details.

The main goal here is to learn a flat (nonhierarchical) dictionary of concepts $\mathcal{C}$ that yields the best lossless compression of the source collection of tableaux $\mathcal{T}$. The inference of $\mathcal{C}$ was undertaken using the Bayesian criterion of minimum message length (MML) (Wallace, 2005). MML provides a statistical inference framework to learn propositions from any observed data set. A proposition can be made as a hypothesis, model, explanation, or theory (Allison, 2018). The MML framework combines ideas from the field of information theory developed by Shannon (Shannon, 1948) and Bayesian inference. Using MML, the descriptive complexity of any stated hypothesis (model, theory, etc.) and its fidelity to explain the observed data can be accurately quantified in terms of Shannon information content (measured in *bits*). This allows the MML framework to provide a reliable complexity-versus-fidelity trade-off, and overcome the well-known problem of over-fitting that is observed in many statistical inference problems. Thus, the best hypothesis is chosen to be the one that yields the most succinct two-part encoding, where the first part encodes the hypothesis, whereas the second encodes the observed data *given* the stated hypothesis. From the Bayesian standpoint (Bayes and Price, 1763), this translates to finding the hypothesis on the data that maximizes their joint probability. Applying MML to this work, the best concept-dictionary $\mathcal{C}$ that explains a source collection of tableaux $\mathcal{T}$ is the one that minimizes the length of the two-part encoding of the form: $\mathcal{I}(\mathcal{C}\&\mathcal{T}) = \mathcal{I}(\mathcal{C}) + \mathcal{I}(\mathcal{T}|\mathcal{C})$, where $\mathcal{I}(\cdot) = -\log_2(\Pr(\cdot))$ measures the Shannon information content (Shannon, 1948) of each of the two parts.

The MML framework provides a natural null-hypothesis test. A dictionary $\mathcal{C}$ explaining $\mathcal{T}$ is accepted if and only if its two-part lossless encoding length is *shorter* than the encoding length of the observed data communicated independently (without the support of any dictionary). The latter is termed the null model message length and denoted as $\mathcal{I}_{\text{null}}(\cdot)$. Further, the quality of an inferred dictionary can be measured by the amount of lossless compression gained with respect to the null model: $\mathcal{I}_{\text{null}}(\mathcal{T}) - \mathcal{I}(\mathcal{C}\&\mathcal{T})$. Thus, using MML, the best dictionary can be equivalently chosen using the following objective: $\arg\max \mathcal{I}_{null}(\mathcal{T}) - \mathcal{I}(\mathcal{C}\&\mathcal{T})$. Formal details of the search methodology for the best dictionary and MML methods to estimate the lossless encoding lengths given by the terms $\mathcal{I}_{\text{null}}(\mathcal{T})$ and $\mathcal{I}(\mathcal{C}\&\mathcal{T})$ appear in (Subramanian et al., 2017).

Broadly speaking, central to the inference of the dictionary is the procedure to generate the optimal encoding of any single tableau using a given concept-dictionary $\mathcal{C}$. This involves: 1) the optimal partitioning of the tableau into subtableaux, 2) the optimal assignment of those regions to concepts in the dictionary (or to a null concept), and 3) the encoding of information within the whole tableau using the assignment of regions to their respective concepts. This optimal partitioning and encoding is chosen as the one that yields the minimum encoding length, and can be derived using an efficient dynamic programming algorithm. Therefore, using MML, the best dictionary for any source collection of tableaux is defined as the one that yields the shortest overall encoding of stating the dictionary, plus the optimal encodings for each tableau in the collection given that dictionary.

Finally, the search for the best dictionary is carried out using simulated annealing (see **Algorithm 1**). Starting from the initial

Algorithm 1 | Simulated annealing algorithm to find the optimal dictionary.

Input: Collection $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_{|\mathcal{T}|}\}$ of tableaux.
Output: Dictionary $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ of concepts.
Begin
Initialise dictionary $\mathcal{C} \leftarrow \emptyset$. ▷ Start with an empty dictionary
Initialise temperature (i.e. control parameter) $t \leftarrow 5000$. ▷ Anneal from a high temperature
Initialise $\alpha \leftarrow 0.88$. ▷ Rate of annealing constant
while $t > 0.1$ **do** ▷ Repeat until stopping criterion is met
  if $t > 10$ **then** ▷ Set length of Markov chain (MC)
    $\texttt{numIter} \leftarrow 50,000$.
  else ▷ Increase MC length for low temperatures
    $\texttt{numIter} \leftarrow 500,000$.

  for $1 \leq i \leq \texttt{numIter}$ **do**
    $I_{current} \leftarrow \mathcal{I}(\mathcal{C}\&\mathcal{T})$. ▷ Compute message length using $\mathcal{C}$

    ▷ Choose a random perturbation primitive
    $prmv \leftarrow random(\{\texttt{addConcept}, \texttt{removeConcept}, \texttt{perturbLength}, \texttt{perturbKappa}, \texttt{swapWithUsage}\})$.
    $\mathcal{C}' \leftarrow$ perturbation$(\mathcal{C}, prmv)$. ▷ Get perturbed dictionary
    $I_{perturbed} \leftarrow \mathcal{I}(\mathcal{C}'\&\mathcal{T})$. ▷ Compute message length using $\mathcal{C}'$

    ▷ Apply Metropolis Criterion for acceptance/rejection of $\mathcal{C}'$
    $\Delta I \leftarrow I_{perturbed} - I_{current}$. ▷ Compute difference in message lengths
    if $\Delta I < 0$ **then** ▷ Accept perturbation with a probability of 1
      $\mathcal{C} \leftarrow \mathcal{C}'$
    else ▷ Accept perturbation with a probability based on ΔI
      Compute $\Pr \leftarrow 2^{-\Delta I/T}$.

      Generate uniform random $u \in [0, 1)$.
      if $u \leq \Pr$ **then**
        $\mathcal{C} \leftarrow \mathcal{C}'$. ▷ Accept perturbed dictionary $\mathcal{C}'$

  $t \leftarrow t * \alpha$. ▷ Decrement control parameter and continue

End

state of an *empty* dictionary, the search involves iterative and stochastic explorations of local neighborhood of the solution-space using the following perturbation primitives: 1) Add concept: randomly choose a subtableau (candidate concept) from the source collection and add it to the current dictionary. 2) Remove concept: randomly choose and delete an existing concept from the current dictionary. 3) Perturb concept length: randomly choose an existing concept from the dictionary and extend/shorten it by one SSE, with reference to the concept's original source (in the source collection). 4) Perturb kappa: randomly choose a concept and perturb its statistical parameter (κ) that controls its flexibility. 5) Swap concept with usage: randomly choose a concept from the current dictionary, and swap it with a region in the collection that is currently encoded by it.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Conceptualization: AK and AL; methodology: AK, LA, DA, MG, PS, and AL; software: RS and AK; validation: AK, LA, and AL; analysis: AK, RS, LA, and AL; investigation: AK, LA, PS, MG, and AL; resources: DA and AL; data curation: AK; writing—original draft: AK and AL; writing—review and editing: RS, LA, DA, PS, and MG; visualization: AK and AL; supervision: AK; project administration: AK; and funding acquisition: AK, PS, MG, and AL.

FUNDING

This research is funded by an Australian Research Council (ARC) Discovery Project grant (DP150100894).

ACKNOWLEDGMENTS

We thank Research Computing Centre, University of Queensland, for the High-Performance Cluster Infrastructure that supported this project over the last 3 years. AL thanks the Medical Research Council Laboratory of Molecular Biology for their hospitality during his sabbatical year. We thank Sureshkumar Balasubramanian for proofreading this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.612920/full#supplementary-material

REFERENCES

Ahn, S., Milner, A. J., Fütterer, K., Konopka, M., Ilias, M., Young, T. W., et al. (2001). The "open" and "closed" structures of the type-C inorganic pyrophosphatases from *Bacillus subtilis* and *Streptococcus gordonii*. *J. Mol. Biol.* 313 (4), 797–811. doi:10.1006/jmbi.2001.5070

Allison, L. (2018). *Coding Ockham's Razor*. Berlin, Germany: Springer.

Alva, V., Söding, J., and Lupas, A. N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. *eLife* 4, e09410. doi:10.7554/elife.09410

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2013). SCOP2 prototype: a new approach to protein structure mining. *Nucl. Acids Res.* 42 (D1), D310–D314. doi:10.1093/nar/gkt1242

Bayes, T., and Price, R. (1763). An essay towards solving a problem in the doctrine of chance. *Philos. Trans. R. Soc.* 53, 370–418. doi:10.1098/rstl.1763.0053

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.* 10 (12), 980. doi:10.1038/nsb1203-980

Brenner, S. E., Koehl, P., and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucl. Acids Res.* 28 (1), 254–256. doi:10.1093/nar/28.1.254

Bystroff, C., and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281 (3), 565–577. doi:10.1006/jmbi.1998.1943

Bystroff, C., Simons, K. T., Han, K. F., and Baker, D. (1996). Local sequence-structure correlations in proteins. *Curr. Opin. Biotechnol.* 7 (4), 417–421. doi:10.1016/s0958-1669(96)80117-0

Camproux, A. C., Tuffery, P., Chevrolat, J. P., Boisvieux, J. F., and Hazout, S. (1999). Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng.* 12 (12), 1063–1073. doi:10.1093/protein/12.12.1063

Camproux, A. C., Gautier, R., and Tufféry, P. (2004). A hidden Markov model derived structural alphabet for proteins. *J. Mol. Biol.* 339 (3), 591–605. doi:10.1016/j.jmb.2004.04.005

Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2017). SCOPe: manual curation and artifact removal in the structural classification of proteins - extended database. *J. Mol. Biol.* 429 (3), 348–355. doi:10.1016/j.jmb.2016.11.023

Chaudhry, C., Weston, M. C., Schuck, P., Rosenmund, C., and Mayer, M. L. (2009). Stability of ligand-binding domain dimer assembly controls kainate receptor desensitization. *EMBO J.* 28 (10), 1518–1530. doi:10.1038/emboj.2009.86

Chitturi, B., Shi, S., Kinch, L. N., and Grishin, N. V. (2016). Compact structure patterns in proteins. *J. Mol. Biol.* 428 (21), 4392–4412. doi:10.1016/j.jmb.2016.07.022

Chomsky, N. (1957). *Syntactic structures*. 2nd Edn. New York, NY: Walter de Gruyter.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5 (4), 823. doi:10.1002/j.1460-2075.1986.tb04288.x

Chothia, C., Levitt, M., and Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. *Proc. Natl. Acad. Sci.* 74 (10), 4130–4134. doi:10.1073/pnas.74.10.4130

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature* 357, 543–544. doi:10.1038/357543a0

Collier, J. H., Allison, L., Lesk, A. M., Stuckey, P. J., Garcia de la Banda, M., and Konagurthu, A. S. (2017). Statistical inference of protein structural alignments using information and compression. *Bioinformatics* 33 (7), 1005–1013. doi:10.1093/bioinformatics/btw757

de Oliveira, S. H. P., Deane, C. M., and Valencia, A. (2018). Combining co-evolution and secondary structure prediction to improve fragment library generation. *Bioinformatics* 34 (9), 2219–2227. doi:10.1093/bioinformatics/bty084

Duboule, D., and Wilkins, A. S. (1998). The evolution of 'bricolage'. *Trends Genet.* 14 (2), 54–59. doi:10.1016/s0168-9525(97)01358-9

Dybas, J. M., and Fiser, A. (2016). Development of a motif-based topology-independent structure comparison method to identify evolutionarily related folds. *Proteins* 84 (12), 1859–1874. doi:10.1002/prot.25169

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32 (5), 1792–1797. doi:10.1093/nar/gkh340

Efimov, A. V. (1997). Structural trees for protein superfamilies. *Proteins* 28 (2), 241–260. doi:10.1002/(SICI)1097-0134(199706)28:2%3C241::AID-PROT12%3E3.0.CO;2-I

Efimov, A. V. (2013). "Super-secondary structures and modeling of protein folds," in *Protein Supersecondary Structures*. 2nd Edn (Berlin, Germany: Springer), Vol. 932, 177–189. doi:10.1007/978-1-62703-065-6_11

Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., et al. (2004). Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 20 (13), 2153–2155. doi:10.1093/bioinformatics/bth214

Finkelstein, A. V., and Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* 50 (3), 171–190. doi:10.1016/0079-6107(87)90013-7

Forest, K. T., Langford, P. R., Kroll, J. S., and Getzoff, E. D. (2000). Cu, Zn superoxide dismutase structure from a microbial pathogen establishes a class with a conserved dimer interface. *J. Mol. Biol.* 296 (1), 145–153. doi:10.1006/jmbi.1999.3448

Franklin, M. C., Kadkhodayan, S., Ackerly, H., Alexandru, D., Distefano, M. D., Elliott, L. O., et al. (2003). Structure and function analysis of peptide antagonists of melanoma inhibitor of apoptosis (ML-IAP). *Biochemistry* 42 (27), 8223–8231. doi:10.1021/bi034227t

Friedberg, I., and Godzik, A. (2005). Connecting the protein structure universe by using sparse recurring fragments. *Structure* 13 (8), 1213–1224. doi:10.1016/j.str.2005.05.009

Goldstein, R. A. (2008). The structure of protein evolution and the evolution of protein structure. *Curr. Opin. Struct. Biol.* 18 (2), 170–177. doi:10.1016/j.sbi.2008.01.006

Gordeev, A. B., Kargatov, A. M., and Efimov, A. V. (2010). PCBOST: protein classification based on structural trees. *Biochem. Biophys. Res. Commun.* 397 (3), 470–471. doi:10.1016/j.bbrc.2010.05.136

Gutteridge, A., and Thornton, J. M. (2005). Understanding nature's catalytic toolkit. *Trends Biochem. Sci.* 30 (11), 622–629. doi:10.1016/j.tibs.2005.09.006

Harris, Z. S. (1954). Distributional structure. *Word* 10 (2-3), 146–162. doi:10.1080/00437956.1954.11659520

Hutchinson, E. G., and Thornton, J. M. (1996). PROMOTIF-a program to identify and analyze structural motifs in proteins. *Protein Sci.* 5 (2), 212–220. doi:10.1002/pro.5560050204

Jacob, F. (1977). Evolution and tinkering. *Science* 196 (4295), 1161–1166. doi:10.1126/science.860134

Jones, T. A., and Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* 5 (4), 819–822. doi:10.1002/j.1460-2075.1986.tb04287.x

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J.-C., Swapna, L. S., Offmann, B., et al. (2010). A short survey on protein blocks. *Biophys. Rev.* 2 (3), 137–145. doi:10.1007/s12551-010-0036-1

Joshi, R. R. (2018). Diversity and motif conservation in protein 3D structural landscape: exploration by a new multivariate simulation method. *J. Mol. Model.* 24 (4), 76. doi:10.1007/s00894-018-3614-y

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7 (8), 1511. doi:10.1038/nprot.2012.085

Kamat, A. P., and Lesk, A. M. (2007). Contact patterns between helices and strands of sheet define protein folding patterns. *Proteins* 66 (4), 869–876. doi:10.1002/prot.21241

Kihara, D., and Skolnick, J. (2003). The PDB is a covering set of small protein structures. *J. Mol. Biol.* 334 (4), 793–802. doi:10.1016/j.jmb.2003.10.027

Kim, D. E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucl. Acids Res.* 32 (suppl. 2), W526–W531. doi:10.1093/nar/gkh468

Kinjo, A. R., and Nakamura, H. (2009). Comprehensive structural classification of ligand-binding motifs in proteins. *Structure* 17 (2), 234–246. doi:10.1016/j.str.2008.11.009

Kister, A. E. (Editor) (2013). *Protein supersecondary structures*. New York, NY: Springer-Humana Press.

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.* 323 (2), 297–307. doi:10.1016/s0022-2836(02)00942-7

Konagurthu, A. S., and Lesk, A. M. (2010). Cataloging topologies of protein folding patterns. *J. Mol. Recognit.* 23 (2), 253–257. doi:10.1002/jmr.1006

Konagurthu, A. S., Stuckey, P. J., and Lesk, A. M. (2008). Structural search and retrieval using a tableau representation of protein folding patterns. *Bioinformatics* 24 (5), 645–651. doi:10.1093/bioinformatics/btm641

Konagurthu, A. S., Lesk, A. M., and Allison, L. (2012). Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* 28 (12), i97–i105. doi:10.1093/bioinformatics/bts223

Leonard, S. A., Gittis, A. G., Petrella, E. C., Pollard, T. D., and Lattman, E. E. (1997). Crystal structure of the actin-binding protein actophorin from *Acanthamoeba*. *Nat. Struct. Mol. Biol.* 4 (5), 369–373. doi:10.1038/nsb0597-369

Lesk, A. M., and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136 (3), 225–270. doi:10.1016/0022-2836(80)90373-3

Lesk, A. M., and Rose, G. D. (1981). Folding units in globular proteins. *Proc. Natl. Acad. Sci.* 78 (7), 4304–4308. doi:10.1073/pnas.78.7.4304

Lesk, A. M. (1995). Systematic representation of protein folding patterns. *J. Mol. Graph.* 13 (3), 159–164. doi:10.1016/0263-7855(95)00037-7

Lesk, A. M. (2016). *Introduction to protein science: architecture, function, and genomics*. 3rd Edn. Oxford, United Kingdom: Oxford University Press.

Leszczynski, J., and Rose, G. D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science* 234 (4778), 849–855. doi:10.1126/science.3775366

Levitt, M., and Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261 (5561), 552. doi:10.1038/261552a0

Mackenzie, C. O., Zhou, J., and Grigoryan, G. (2016). Tertiary alphabet for the observable protein structural universe. *Proc. Natl. Acad. Sci. USA* 113 (47), E7438–E7447. doi:10.1073/pnas.1607178113

Mardia, K. V., and Jupp, P. E. (2009). *Directional statistics*. New York, NY: John Wiley & Sons, Vol. 494.

Meador, W., Means, A., and Quiocho, F. (1992). Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. *Science* 257 (5074), 1251–1255. doi:10.1126/science.1519061

Messerschmidt, A., Prade, L., Kroes, S. J., Sanders-Loehr, J., Huber, R., and Canters, G. W. (1998). Rack-induced metal binding vs. flexibility: Met121His azurin crystal structures at different pH. *Proc. Natl. Acad. Sci.* 95 (7), 3443–3448. doi:10.1073/pnas.95.7.3443

Micheletti, C., Seno, F., and Maritan, A. (2000). Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* 40 (4), 662–674. doi:10.1002/1097-0134(20000901)40:4<662::aid-prot90>3.0.co;2-f

Mirwaldt, C., Korndorfer, I., and Huber, R. (1995). The crystal structure of dihydrodipicolinate synthase from *Escherichia coli* at 2.5 Å resolution. *J. Mol. Biol.* 246 (1), 227–239. doi:10.1006/jmbi.1994.0078

Murzin, A. G., and Finkelstein, A. V. (1988). General architecture of the α-helical globule. *J. Mol. Biol.* 204 (3), 749–769. doi:10.1016/0022-2836(88)90366-x

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247 (4), 536–540. doi:10.1016/s0022-2836(05)80134-2

Nechushtai, R., Lammert, H., Michaeli, D., Eisenberg-Domovich, Y., Zuris, J. A., Luca, M. A., et al. (2011). Allostery in the ferredoxin protein motif does not involve a conformational switch. *Proc. Natl. Acad. Sci.* 108 (6), 2240–2245. doi:10.1073/pnas.1019502108

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc. Natl. Acad. Sci. USA* 114 (44), 11703–11708. doi:10.1073/pnas.1707642114

Ogata, C.M., Gordon, P. F., de Vos, A. M., and Kim, S. H. (1992). Crystal structure of a sweet tasting protein thaumatin I, at 1.65 A resolution. *J. Mol. Biol.* 228 (3), 893–908. doi:10.1016/0022-2836(92)90873-i

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J. (1997). CATH - a hierarchic classification of protein domain structures. *Structure* 5 (8), 1093–1109. doi:10.1016/s0969-2126(97)00260-8

Pandini, A., Fornili, A., and Kleinjung, J. (2010). Structural alphabets derived from attractors in conformational space. *BMC Bioinform.* 11 (1), 97. doi:10.1186/1471-2105-11-97

Pauling, L., and Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci.* 37 (5), 251. doi:10.1073/pnas.37.5.251

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci.* 37 (4), 205–211. doi:10.1073/pnas.37.4.205

Rao, S. T., and Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* 76 (2), 241–256. doi:10.1016/0022-2836(73)90388-4

Richards, F. M., and Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* 3 (2), 71–84. doi:10.1002/prot.340030202

Rognan, D. (2007). Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* 152 (1), 38–52. doi:10.1038/sj.bjp.0707307

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). "Protein structure prediction using Rosetta," in *Methods in Enzymology* (Amsterdam, Netherlands: Elsevier), Vol. 383, 66–93. doi:10.1016/S0076-6879(04)83004-0

Rooman, M. J., Rodriguez, J., and Wodak, S. J. (1990). Automatic definition of recurrent local structure motifs in proteins. *J. Mol. Biol.* 213 (2), 327–336. doi:10.1016/s0022-2836(05)80194-9

Schaeffer, R. D., Liao, Y., Cheng, H., and Grishin, N. V. (2016). ECOD: new developments in the evolutionary classification of domains. *Nuc. Acids Res.* 45 (D1), D296–D302. doi:10.1093/nar/gkw1137

Schrader, J., Henneberg, F., Mata, R. A., Tittmann, K., Schneider, T. R., Stark, H., et al. (2016). The inhibition mechanism of human 20S proteasomes enables

next-generation inhibitor design. *Science* 353 (6299), 594–598. doi:10.1126/science.aaf8993

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577 (7792), 706–710. doi:10.1038/s41586-019-1923-7

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Sievers, F., and Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* 1079, 105–116. doi:10.1007/978-1-62703-646-7_6

Singhal, A. (2001). Modern information retrieval: a brief overview. *IEEE Data Eng. Bull.* 24 (4), 35–43.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationship. *Univ. Kans. Sci. Bull* 28, 1409–1438.

Subramanian, R., Allison, L., Stuckey, P. J., Garcia de la Banda, M., Abramson, D., Lesk, A. M., et al. (2017). "Statistical compression of protein folding patterns for inference of recurrent substructural themes," in Data Compression Conference (DCC), Snowbird, UT, April 4–7, 2017 (New York, NY: IEEE), 340–349.

Tagawa, K., and Arnon, D. I. (1962). Ferredoxins as electron carriers in photosynthesis and in the biological production and consumption of hydrogen gas. *Nature* 195 (4841), 537–543. doi:10.1038/195537a0

Taylor, W. R. (2002). A 'periodic table' for protein structures. *Nature* 416 (6881), 657. doi:10.1038/416657a

Tramontano, A., Chothia, C., and Lesk, A. M. (1989). Structural determinants of the conformations of medium-sized loops in proteins. *Proteins* 6 (4), 382–394. doi:10.1002/prot.340060405

Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 5 (4), 355–373. doi:10.1002/prot.340050410

Unger, R., and Sussman, J. L. (1993). The importance of short structural motifs in protein structure analysis. *J. Comput. Aided Mol. Des.* 7 (4), 457–472. doi:10.1007/bf02337561

Vingron, M., and Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *Bioinformatics* 5 (2), 115–121. doi:10.1093/bioinformatics/5.2.115

Wallace, C. S., and Boulton, D. M. (1968). An information measure for classification. *J. Comput.* 11 (2), 185–194. doi:10.1093/comjnl/11.2.185

Wallace, C. S. (2005). *Statistical and inductive inference by minimum message length*. Berlin, Germany: Springer.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucl. Acids Res.* 46 (W1), W296–W303. doi:10.1093/nar/gky427

Whisstock, J. C., and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quart. Rev. Biophys.* 36 (3), 307–340. doi:10.1017/s0033583503003901

Zheng, W., Zhang, C., Bell, E. W., and Zhang, Y. (2019). I-TASSER gateway: a protein structure and function prediction server powered by XSEDE. *Future Gener. Comput. Syst.* 99, 73–85. doi:10.1016/j.future.2019.04.011

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 03 May 2021
doi: 10.3389/fmolb.2021.626600

Structural Profiling of Bacterial Effectors Reveals Enrichment of Host-Interacting Domains and Motifs

*Yangchun Frank Chen and Yu Xia**

Department of Bioengineering, McGill University, Montreal, QC, Canada

Effector proteins are bacterial virulence factors secreted directly into host cells and, through extensive interactions with host proteins, rewire host signaling pathways to the advantage of the pathogen. Despite the crucial role of globular domains as mediators of protein-protein interactions (PPIs), previous structural studies of bacterial effectors are primarily focused on individual domains, rather than domain-mediated PPIs, which limits their ability to uncover systems-level molecular recognition principles governing host-bacteria interactions. Here, we took an interaction-centric approach and systematically examined the potential of structural components within bacterial proteins to engage in or target eukaryote-specific domain-domain interactions (DDIs). Our results indicate that: 1) effectors are about six times as likely as non-effectors to contain host-like domains that mediate DDIs exclusively in eukaryotes; 2) the average domain in effectors is about seven times as likely as that in non-effectors to co-occur with DDI partners in eukaryotes rather than in bacteria; and 3) effectors are about nine times as likely as non-effectors to contain bacteria-exclusive domains that target host domains mediating DDIs exclusively in eukaryotes. Moreover, in the absence of host-like domains or among pathogen proteins without domain assignment, effectors harbor a higher variety and density of short linear motifs targeting host domains that mediate DDIs exclusively in eukaryotes. Our study lends novel quantitative insight into the structural basis of effector-induced perturbation of host-endogenous PPIs and may aid in the design of selective inhibitors of host-pathogen interactions.

Keywords: structural biology, bacterial effector, host-pathogen interaction, protein-protein interaction, globular domains, short linear motifs, structural homology, convergent evolution

OPEN ACCESS

Edited by:
Charlotte Deane,
University of Oxford, United Kingdom

Reviewed by:
Julien Bergeron,
King's College London,
United Kingdom
Ester Boix,
Universitat Autònoma de Barcelona,
Spain

****Correspondence:***
Yu Xia
brandon.xia@mcgill.ca

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *06 November 2020*
Accepted: *21 April 2021*
Published: *03 May 2021*

Citation:
Chen YF and Xia Y (2021) Structural Profiling of Bacterial Effectors Reveals Enrichment of Host-Interacting Domains and Motifs. Front. Mol. Biosci. 8:626600. doi: 10.3389/fmolb.2021.626600

INTRODUCTION

An important goal of systems microbiology is to understand how host-pathogen protein-protein interactions (PPIs) impact host-endogenous signaling networks. Effector proteins are virulence factors secreted by pathogenic bacteria and injected directly into the host cytoplasm via specialized secretion systems (Galan, 2009). Effectors are key mediators of host-pathogen interactions throughout the infection cycle, from initial host attachment and pathogen internalization, to migration and proliferation in the host. Among the diverse biochemical activities of effectors discovered so far are guanine nucleotide exchange factors and dissociation inhibitors, GTPase-activating proteins, kinases and phosphatases, ubiquitin ligases, and so on (Fu and Galan, 1998; Steele-Mortimer et al., 2000; Janjusevic et al., 2006). A common virulence mechanism of effectors is functional mimicry of host activities, whereby effectors compete with host proteins for control of

host signaling pathways. This functional mimicry can be achieved in one of two ways: horizontal acquisition of eukaryotic globular domains, or convergent evolution of domains and short linear motifs in bacteria that bear little sequence or structural similarity to eukaryotic proteins (Stebbins and Galan, 2001; Popa et al., 2016; Scott and Hartland, 2017). These structural modules allow effectors to interact seamlessly with host-endogenous factors involved in actin remodelling, protein degradation and cell cycle regulation, helping the pathogen to survive and thrive in the host while bypassing immune surveillance. Previous studies have uncovered a large repertoire of bacterial effectors that are structural homologs of eukaryotic proteins, giving rise to models for predicting effectors based on the premise that whereas most domains are uniformly distributed among all species of bacteria, eukaryotic-like domains are overrepresented in the genomes of pathogenic and symbiotic species (Jehl et al., 2011; Marchesini et al., 2011). Although useful for identifying candidate effectors in metagenomic analyses, a main caveat of these models is their treatment of domains as individual, rather than interacting, entities that contribute to protein-protein interactions. Eukaryotic domains and their domain-domain interaction (DDI) partners are of special interest to the study of host-pathogen interactions, as they are often mimicked or targeted by pathogens to subvert host signaling pathways (Arnold et al., 2012). Despite studies pointing to the presence of many eukaryotic-like domains in bacterial effectors, there has yet to be a comprehensive, quantitative analysis of the relevance of such domains to host-endogenous PPIs, which is crucial to understanding systems-level changes in the host upon infection with pathogens.

Past studies on host-bacteria protein-protein interactions (PPIs) have examined either individual interactions at the domain level (Cazalet et al., 2004), or interactome networks at the whole protein level (Schweppe et al., 2015), but never both at the domain level and on an interactome scale. In this work, we ask the following new question: how do host-bacteria PPIs mimic and modulate host-endogenous PPIs at the protein domain level on an interactome scale? To answer this question, we carried out two analyses of host-interacting bacterial proteins: the first on mimicry of host-endogenous binding sites by bacterial effectors, and the second on enrichment of host-interacting domains and short linear motifs in bacterial effectors. In the first analysis, we examined the mechanism of host binding site mimicry by bacterial proteins where, rather than creating new binding sites, bacteria recruit existing binding sites involved in host-endogenous PPIs for host-bacteria PPIs (Cazalet et al., 2004). Previous studies on host-virus interactions found that while two human proteins sharing binding sites on a common target tend to be structurally similar, a viral protein and a human protein sharing binding sites on a common target tend to be structurally distinct (Franzosa and Xia, 2011; Garamszegi et al., 2013). In other words, binding site sharing among human proteins is largely attributable to divergent evolution through gene duplication, whereas binding site mimicry by viral proteins tends to involve convergent evolution of unique host-interacting modules in viruses. To our knowledge, similar analyses have yet to be performed for host-bacteria interactions. In the second analysis, we tested the hypothesis that compared to non-effector proteins, bacterial effectors are enriched for domains that either mimic or target host domains involved in eukaryote-specific domain-domain interactions (DDIs). In addition to domains, we also tested whether effectors tend to contain a higher variety and density of short linear motifs that interact with host domains mediating DDIs exclusively in eukaryotes.

RESULTS

Mechanism of Binding Site Sharing in Host-Endogenous *vs.* Host-Bacteria Protein-Protein Interaction Network

Previous studies have established binding site mimicry via convergent evolution as a key feature of human-virus PPIs where, rather than securing new binding sites, viruses have evolved unique, non-host-like structural domains and short linear motifs to compete with host proteins for the same binding sites on a common host target (Franzosa and Xia, 2011; Garamszegi et al., 2013). As bacteria and viruses are both known to hijack host molecular machinery through interacting with host proteins, we performed similar analyses on a domain-resolved host-bacteria PPI network with regard to binding site mimicry and its evolutionary mechanism. To this end, we acquired eukaryote-endogenous (within animals/plants/fungi), bacteria-endogenous and host-bacteria (between animals/plants and pathogenic bacteria) PPI data, and resolved each PPI into domain-domain interactions (DDIs) between interacting proteins, based on DDI templates previously derived from 3D structures of protein complexes (Materials and Methods). The resulting domain-resolved, eukaryote-bacteria PPI network consists of: 1) 57,019 PPIs among 22,110 eukaryotic proteins, resolved into 4,953 DDIs among 2,859 eukaryotic domains; 2) 3,362 PPIs among 3,000 bacterial proteins, resolved into 1,434 DDIs among 1,120 bacterial domains; and 3) 173 PPIs between 107 host proteins and 103 bacterial proteins, resolved into 87 DDIs between 53 host domains and 63 bacterial domains. The entire list of domain-resolved host-bacteria, bacteria-endogenous, and eukaryote-endogenous PPIs can be found in **Supplementary File S1**.

We found that of the 103 host-targeting bacterial proteins, 95 (92%) bind to the same domains on their host target that are otherwise bound by host-endogenous proteins, suggesting that like viruses, bacteria also tend to recruit domains involved in host-endogenous PPIs for host-pathogen PPIs (Cazalet et al., 2004). We then determined whether bacterial and host proteins binding to the same domain on another host protein are structurally similar. We found that of 18,331 cases where two host proteins A and B bind to the same domain on a common target, 13,139 (72%) cases involve domains which are conserved between A and B, while in the remaining 5,192 (28%) cases there is no domain conserved between A and B. Conversely, among 95 cases where host protein X and bacterial protein Y bind to the same domain on another host protein, only 8 cases (8%) involve domains which are conserved between X and Y, while in the

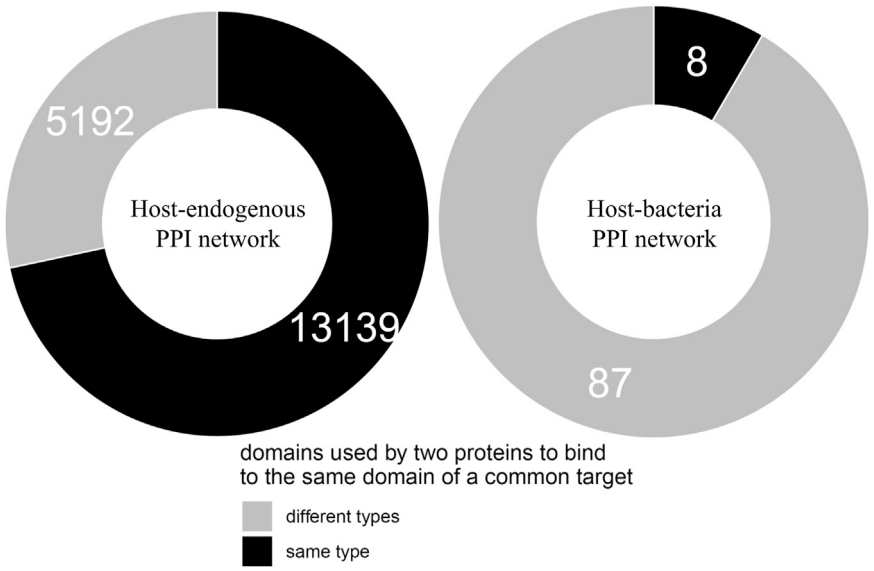


FIGURE 1 | Different evolutionary mechanisms for binding site sharing in the host-endogenous vs. host-bacteria PPI network. Two host proteins are more likely to use the same type of domain to bind to the same domain of a common target, suggesting divergent evolution followed by gene duplication. A host protein and a bacterial protein are more likely to use different types of domain to bind to the same domain of a common host target, suggesting convergent evolution (or extreme divergent evolution) of non-host-like domains in bacteria. The difference in dominant evolutionary mechanisms is statistically significant (Fisher's exact test, two-tailed $p < 2.2 * 10^{-16}$).

remaining 87 (92%) cases there is no domain conserved between X and Y. In other words, compared to binding site sharing among host proteins, binding site mimicry by bacterial proteins appears significantly more likely to involve convergent evolution (or extreme divergent evolution) of bacteria-exclusive domains, rather than horizontal acquisition of host domains (Fisher's exact test, two-tailed $p < 2.2 * 10^{-16}$). **Figure 1** shows the contrast in dominant evolutionary mechanisms behind binding site sharing in the host-endogenous *vs.* host-bacteria PPI network.

Effectors Structurally Mimic Host Domains Involved in Eukaryote-Specific Protein-Protein Interactions

Having examined the mechanism of binding site mimicry in the host-bacteria PPI network, we then asked whether bacterial effectors tend to mimic or target host domains that mediate DDIs predominantly in eukaryotes, as opposed to domains that mediate DDIs in eukaryotes and bacteria with similar likelihood - the rationale being that the former are involved in eukaryote-specific processes such as protein ubiquitination (Grau-Bove et al., 2015), which are prime targets for pathogens to manipulate, whereas the latter are involved in highly conserved, essential cellular processes (Walhout et al., 2000; Matthews et al., 2001), which are unlikely to be perturbed in host-pathogen interactions. We found that of the 63 host-binding bacterial domains in our dataset, 12 have homologs in eukaryotes, among which 7 mediate DDIs exclusively in eukaryotes (PF12796, PF00092, PF12799, PF02205, PF04564, PF00646, PF13676), 3 mediate DDIs primarily in eukaryotes (PF00069, PF00583, PF00183), and 2 have similar numbers of DDI partners in eukaryotes and bacteria (PF13472, PF00085) (**Supplementary File S2**). Meanwhile, of the 31 host domains targeted by bacteria-exclusive domains, 28 otherwise mediate DDIs exclusively in eukaryotes, 2 mediate DDIs primarily in eukaryotes, and 1 has similar numbers of DDI partners in eukaryotes and bacteria (**Supplementary File S3**). In summary, effectors tend to mimic or target host domains that mediate DDIs predominantly in eukaryotes. Given that effectors comprise nearly half (43/103) of the host-targeting bacterial proteins in our PPI dataset, we hypothesized that compared to the rest of the pathogen proteome, effectors are generally enriched for: 1) eukaryotic-like domains that mediate DDIs predominantly in eukaryotes; and 2) bacteria-exclusive domains that target host domains which, when not involved in host-pathogen DDIs, mediate DDIs exclusively in eukaryotes. To test this hypothesis, we systematically compared 238 effectors and 3,921 non-effectors with unique domain signatures, encoded by 84 bacterial species of verified pathogenicity (Urban et al., 2020). Effectors encoded by the 84 pathogenic species are selected from PHI-base, if a pathogen gene's "Gene Function" or "Mutant Phenotype" column contains the keyword "effector", as well as from the UniProt database if the gene name or cellular location contains keywords such as "type * effector", "t*ss effector", or "secreted effector". To ensure the same selection criteria are applied to all proteins, non-effectors encoded by the 84 pathogenic species include proteins not already in the effector set, whose cellular location is any one of cytoplasmic, membrane or secreted (Materials and Methods).

We first tested the hypothesis that effectors are enriched for domains that mediate DDIs exclusively in eukaryotes. We found

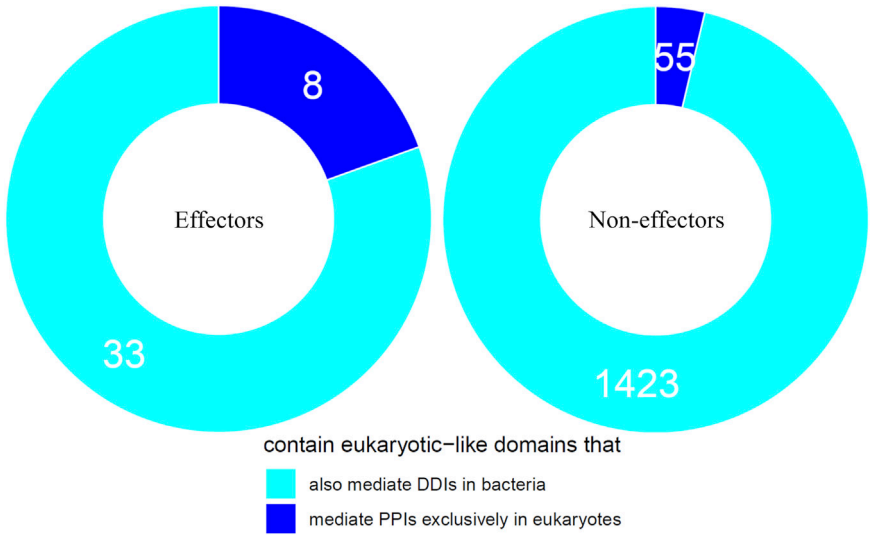


FIGURE 2 | Effectors are enriched for eukaryotic-like domains that mediate PPIs exclusively in eukaryotes. Among pathogen proteins containing domains that mediate experimentally verified PPIs in eukaryotes, 20% effectors and 4% non-effectors contain domains that mediate PPIs exclusively in eukaryotes, suggesting that effectors are six times as likely as non-effectors to repurpose eukaryote-specific processes via eukaryotic-like domains (Fisher's exact test, two-tailed $p = 2 * 10^{-4}$).

TABLE 1 | Effectors containing domains that mediate PPIs exclusively in eukaryotes.

UniProt accession	Representative species	Domain(s)	# Pathogenic spp. encoding proteins with domain(s)
B0RMF9	*Xanthomonas campestris*	**PF00560**	14
A0A0A8VF40	*Yersinia ruckeri*	**PF00560**; PF13855	12
A0A199P7E1	*Xanthomonas translucens*	**PF00646**	11
A0A0S4VGA6	*Ralstonia solanacearum*	**PF00646**; PF13516	4
F6G106	*Ralstonia solanacearum*	**PF00646**; PF13516; PF13855	4
A0A1Y0FB05	*Ralstonia solanacearum*	**PF00665**; PF13276	74
A0A286NT26	*Vibrio parahaemolyticus*	**PF02205**	3
D8NFZ7	*Ralstonia solanacearum*	PF01535; PF12854; **PF13812**	1

Domains mediating PPIs exclusively in eukaryotes are marked in bold.

TABLE 2 | Weighted average host-interacting potential of a multi-domain bacterial protein.

DDI	Eukaryotic species encoding both interacting domains	Eukaryotic species encoding either one or both interacting domains	Bacterial species encoding both interacting domains	Bacterial species encoding either one or both interacting domains	Odds ratio of domain mediating DDIs in eukaryotes *vs.* in bacteria	Weight
A_B	m	H_1	n	B_1	$OR_1 = \frac{m*(B_1-n)}{n*(H_1-m)}$	$w_1 = \frac{m*(B_1-n)}{H_1+B_1}$
A_C	p	H_2	q	B_2	$OR_2 = \frac{p*(B_2-q)}{q*(H_2-p)}$	$w_2 = \frac{p*(B_2-q)}{H_2+B_2}$
D_E	x	H_3	y	B_3	$OR_3 = \frac{x*(B_3-y)}{y*(H_3-x)}$	$w_3 = \frac{x*(B_3-y)}{H_3+B_3}$

$$\text{Host-interacting potential} = \log\left(\frac{\sum_{i=1}^{3} OR_i * w_i}{\sum_{i=1}^{3} w_i}\right)$$

The host-interacting potential of a bacterial protein containing domains A and D, where A and D have DDI partners (domains B, C, E) in both eukaryotes and bacteria, is computed as the Mantel-Haenszel weighted average log odds ratio of domains A and D co-occurring with interacting domains in eukaryotes vs. in bacteria. The odds of domain co-occurring with DDI partners are the number of species encoding both interacting domains (i.e. DDI is possible) divided by the number of species encoding either one, but not both, of the interacting domains (i.e. DDI is not possible).

that among 41 effectors and 1,478 non-effectors containing domains that mediate experimentally verified PPIs in eukaryotes, 8 effectors (20%) and 55 non-effectors (4%) contain domains that mediate PPIs exclusively in eukaryotes, suggesting that effectors are six times as likely as non-effectors to repurpose eukaryote-specific processes via eukaryotic-like domains (Fisher's exact test, two-tailed $p = 2 * 10^{-4}$) (**Figure 2**). **Table 1** is a list of effectors containing domains involved in interprotein DDIs in eukaryotes, but neither interprotein nor intraprotein DDIs in bacteria. Next, we tested the hypothesis that effectors are enriched for domains that mediate DDIs primarily in eukaryotes. For domains having

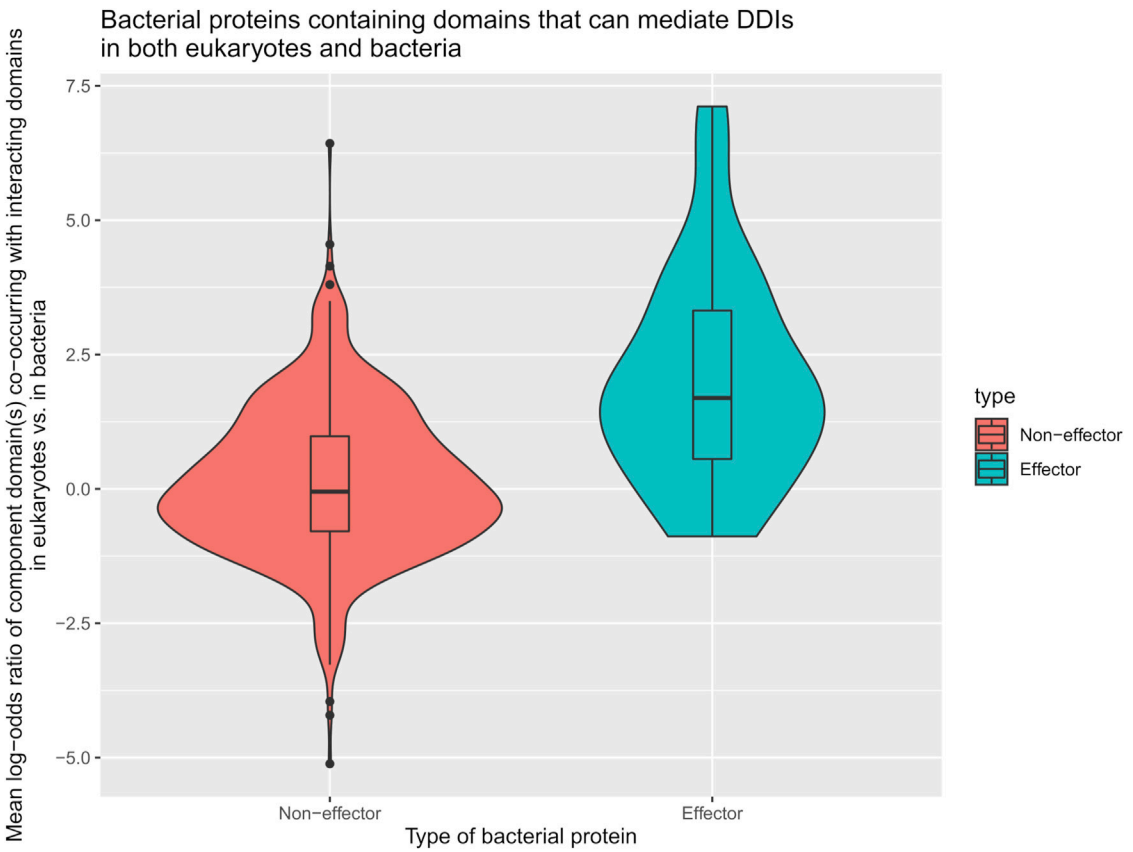


FIGURE 3 | Effectors are enriched for eukaryotic-like domains that mediate PPIs primarily in eukaryotes. Among pathogen proteins containing domains that have DDI partners in both eukaryotes and bacteria, the average domain in effectors is seven times as likely as that in non-effectors to co-occur with DDI partners in eukaryotes rather than in bacteria (Wilcoxon test, two-tailed $p = 4 * 10^{-7}$).

TABLE 3 | Effectors containing domains that mediate PPIs primarily in eukaryotes.

UniProt accession	Species	Domain(s)	# Pathogenic spp. encoding proteins with domain(s)	Log odds ratio of domains co-occurring with DDI partners in eukaryotes *vs.* in bacteria
Q5ZRQ0	*Legionella pneumophila*	**PF04564**	1	7.1
O84875	*Chlamydia trachomatis*	**PF02902**	13	6.2
P74873	*Salmonella enterica*	**PF00102**; PF03545; PF09119	2	4.4
Q9KS43	*Vibrio cholerae*	**PF01764**	40	3.9
Q3BQY9	*Xanthomonas euvesicatoria*	**PF13202**; **PF13499**	19	3.8
D8P6Z5	*Ralstonia solanacearum*	**PF13516**	14	3.7
Q8XT98	*Ralstonia solanacearum*	**PF00069**	69	3.5
D2TI55	*Citrobacter rodentium*	**PF00557**	84	2.8
Q8XZN9	*Ralstonia solanacearum*	**PF13516**; **PF13855**	13	2.2
A0A6C9X110	*Escherichia coli*	PF00805; **PF01391**; PF13599	27	2

Domains with DDI partners in both eukaryotes and bacteria are marked in bold.

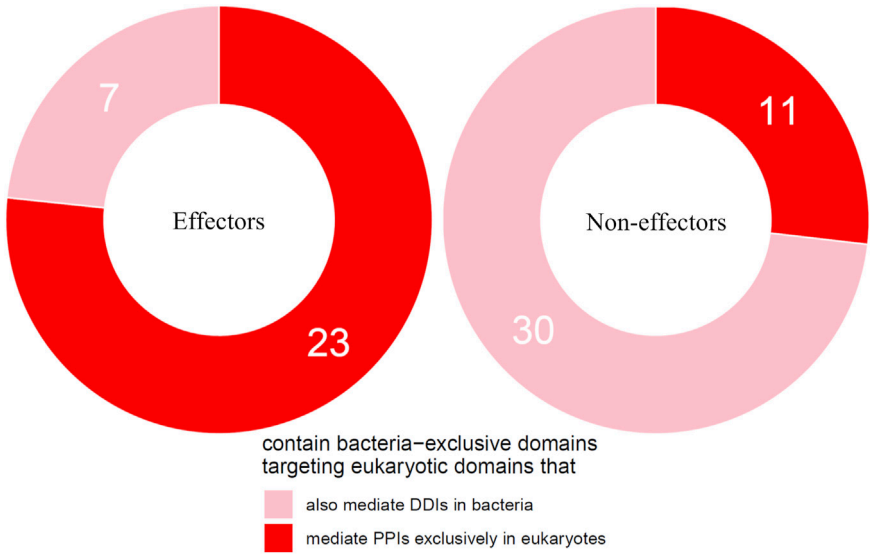


FIGURE 4 | Effectors are enriched for bacteria-exclusive domains targeting host domains that otherwise mediate DDIs exclusively in eukaryotes. Among pathogen proteins containing bacteria-exclusive domains with the potential to target host domains that mediate DDIs in eukaryotes, 77% effectors and 27% non-effectors target host domains that mediate DDIs exclusively in eukaryotes, suggesting that effectors are nine times as likely as non-effectors to disrupt eukaryote-specific processes via bacteria-exclusive domains (Fisher's exact test, two-tailed $p = 4 * 10^{-5}$).

DDI partners in both eukaryotes and bacteria, we estimated their propensity for mediating eukaryote-specific DDIs by computing the odds ratio of the domain's co-occurrence with DDI partners in eukaryotes *vs.* in bacteria. If a bacterial protein contains multiple such domains, we computed a weighted average odds ratio (**Table 2**). We found that among 26 effectors and 635 non-effectors containing domains that have DDI partners in both eukaryotes and bacteria, the average domain in effectors is seven times as likely as that in non-effectors to co-occur with DDI partners in eukaryotes rather than in bacteria (Wilcoxon test, two-tailed $p = 4 * 10^{-7}$) (**Figure 3**). **Table 3** lists effectors with the top 10 highest odds ratios of their component domains co-occurring with DDI partners in eukaryotes rather than in bacteria.

Effectors Convergently Target Host Domains Involved in Eukaryote-Specific Protein-Protein Interactions

We then tested the hypothesis that effectors are enriched for bacteria-exclusive domains that target host domains which, when not involved in host-pathogen DDIs, mediate DDIs exclusively in eukaryotes. Given that experimental PPI data often suffer from limitations such as false negatives and investigator bias in pathogen selection, we supplemented host-interacting bacteria-exclusive domains supported by PPI data with host-interacting bacteria-exclusive domains supported by interprotein DDI templates (Mosca et al., 2014). In this manner, we identified a total of 207 bacteria-exclusive domains with the potential to target host domains that mediate DDIs in eukaryotes, 52 of which target host domains that mediate DDIs exclusively in eukaryotes. We found that among 30 effectors and 41 non-effectors containing bacteria-exclusive domains with the potential to target host domains that mediate DDIs in eukaryotes, 23 effectors (77%) and 11 non-effectors (27%) target host domains that mediate DDIs exclusively in eukaryotes, suggesting that effectors are nine times as likely as non-effectors to disrupt eukaryote-specific processes via bacteria-exclusive domains (Fisher's exact test, two-tailed $p = 4 * 10^{-5}$) (**Figure 4**). **Supplementary File S4** is a list of effectors with bacteria-exclusive domains targeting host domains that otherwise mediate DDIs exclusively in eukaryotes.

In addition to encoding globular domains that either mimic or target host domains, effectors also encode short linear motifs that bind to host domains with similar specificities as host-endogenous proteins, albeit sharing little homology with the latter (Samano-Sanchez and Gibson, 2020). These short linear motifs follow particular sequence patterns and are predominantly located in intrinsically disordered regions of proteins that are accessible to interacting partners (Davey et al., 2012). To determine whether effectors are enriched for host-interacting motifs, we counted the number of unique classes and instances of eukaryotic linear motifs (ELMs) (Kumar et al., 2020) in long disordered regions of bacterial proteins as annotated by the MobiDB database (Piovesan et al., 2018). When comparing 162 effectors and 8,414 non-effectors with unique ELM compositions and containing eukaryotic-like domains, we found that effectors and non-effectors encode 9 and 8 ELM classes per protein (**Figure 5**, Left), along with 0.30 and 0.28 ELM instances per disordered residue (**Figure 6**, Left), respectively. In other words, in the presence of eukaryotic-like domains, effectors encode a slightly higher variety (Wilcoxon test, two-tailed $p = 3 * 10^{-3}$), but similar density of ELMs (Wilcoxon test, two-

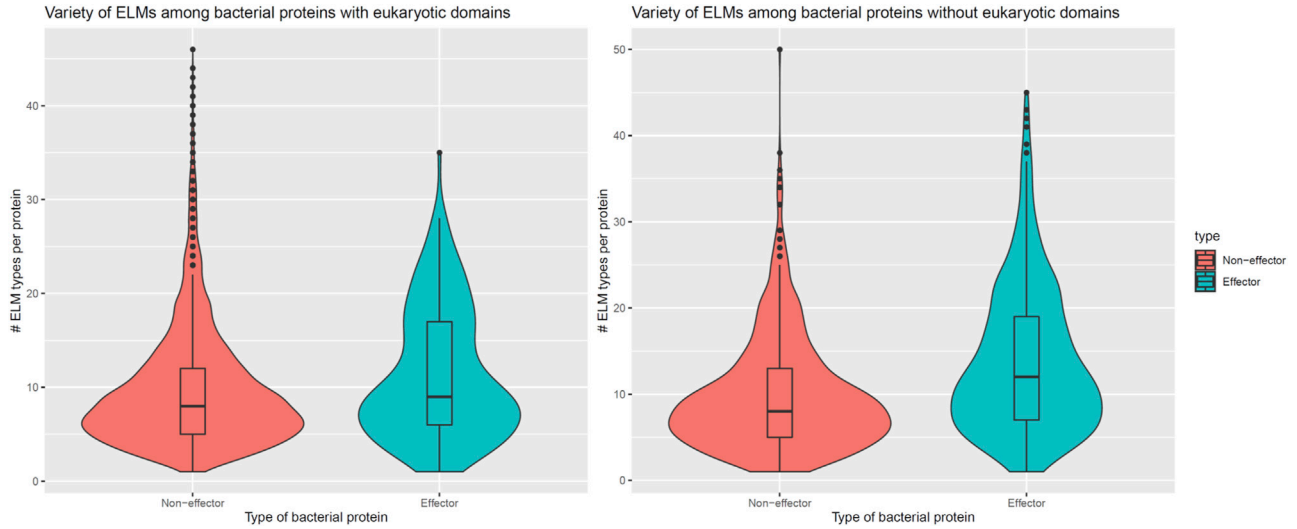


FIGURE 5 | In the absence of eukaryotic-like domains or domains in general, effectors contain a higher variety of eukaryotic linear motifs compared to non-effectors. **Left:** In the presence of eukaryotic-like domains, effectors encode 9 ELM classes per protein, a slightly higher variety (Wilcoxon test, two-tailed $p = 3 * 10^{-3}$) compared to 8 ELM classes per non-effector protein. **Right:** In the absence of eukaryotic-like domains or among proteins without Pfam domains, effectors encode 12 ELM classes per protein, a much higher variety (Wilcoxon test, two-tailed $p < 2.2 * 10^{-16}$) compared to 8 ELM classes per non-effector protein. ELM = eukaryotic linear motif.

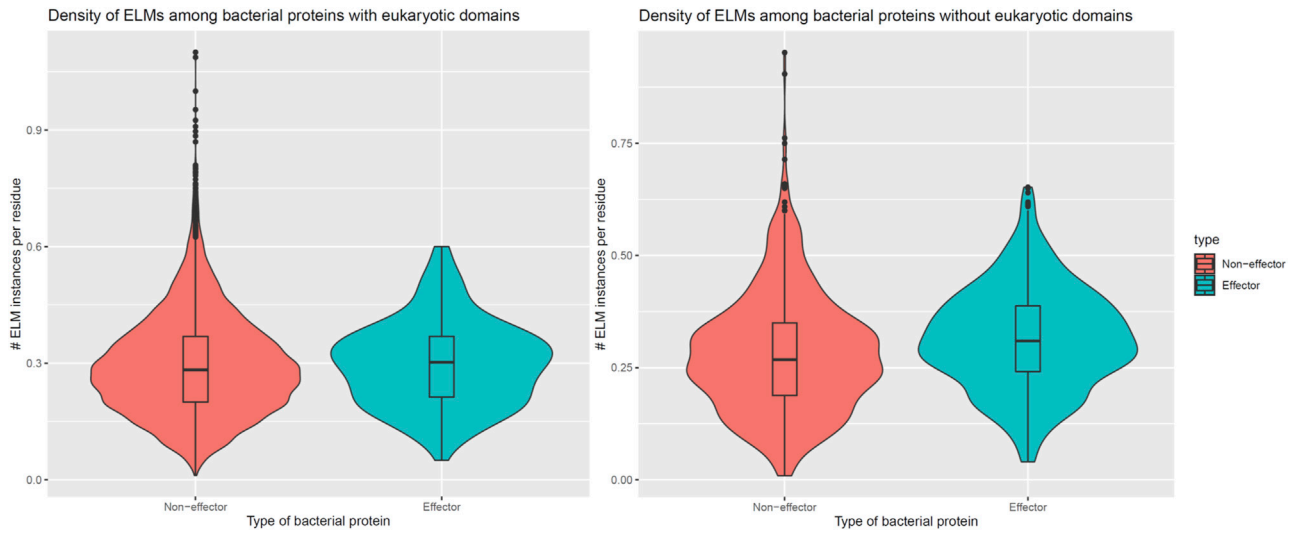


FIGURE 6 | In the absence of eukaryotic-like domains or domains in general, effectors contain a higher density of eukaryotic linear motifs compared to non-effectors. **Left:** In the presence of eukaryotic-like domains, effectors encode 0.30 ELM instances per disordered residue, a slightly higher density that is not statistically significant (Wilcoxon test, two-tailed $p = 0.3$) compared to 0.28 ELM instances per disordered residue in non-effectors. **Right:** In the absence of eukaryotic-like domains or among proteins without Pfam domains, effectors encode 0.31 ELM instances per disordered residue, a higher density that is statistically significant (Wilcoxon test, two-tailed $p = 2 * 10^{-8}$) compared to 0.27 ELM instances per disordered residue in non-effectors. ELM = eukaryotic linear motif.

tailed $p = 0.3$) compared to non-effectors. When comparing 521 effectors and 794 non-effectors with unique ELM compositions and not containing eukaryotic-like domains or any Pfam domains, however, we found that effectors and non-effectors encode 12 and 8 ELM classes per protein (**Figure 5**, Right), along with 0.31 and 0.27 ELM instances per disordered residue (**Figure 6**, Right), respectively. In other words, in the absence of eukaryotic-like domains or among pathogen proteins without Pfam domains, effectors encode a higher variety (Wilcoxon test, two-tailed $p < 2.2 * 10^{-16}$) as well as higher density of ELMs (Wilcoxon test, two-tailed $p = 2 * 10^{-8}$) compared to non-effectors.

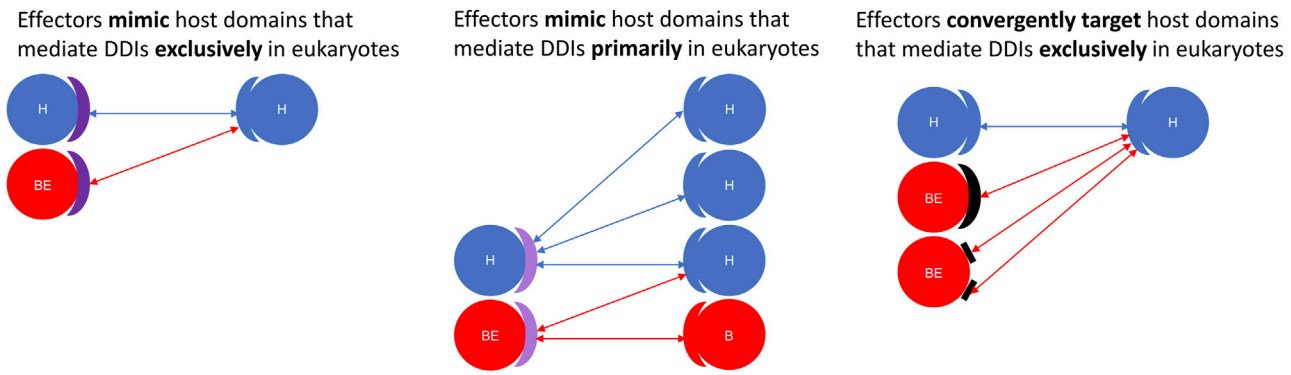


FIGURE 7 | Effectors are enriched for structural modules that either mimic or target host domains mediating eukaryote-specific domain-domain interactions. Using a domain-resolved eukaryote-bacteria protein-protein interaction network, we identified domains that mediate PPIs exclusively in eukaryotes (Left: deep purple arc), as well as domains that are more likely to co-occur with domain-domain interaction partners in eukaryotes rather than in bacteria (Middle: light purple arc). We found that compared to generic proteins encoded by pathogenic bacteria, effector proteins are significantly enriched for domains and short linear motifs that either mimic or target host domains mediating DDIs either exclusively or primarily in eukaryotes. **Left:** Effectors mimic host domains that mediate DDIs exclusively in eukaryotes; **Middle:** Effectors mimic host domains that mediate DDIs primarily in eukaryotes; **Right:** Effectors convergently evolve domains and short linear motifs to target host domains that mediate DDIs exclusively in eukaryotes. H = host protein; B = generic bacterial protein; BE = bacterial effector protein. Domains are represented as blue (eukaryotic), red (bacterial), purple (eukaryotic-like and host-targeting), and black (convergently evolved and host-targeting) arcs. Short linear motifs are represented as rectangles.

DISCUSSION

Pathogenic bacteria have evolved a plethora of strategies to survive and thrive in eukaryotic hosts. A key strategy is functional mimicry of host activities, which is achieved through one of two orthogonal evolutionary mechanisms: horizontal acquisition of eukaryotic domains or convergent evolution of bacteria-exclusive domains (Stebbins and Galan, 2001; Popa et al., 2016; Scott and Hartland, 2017). Current literature contains many case studies of bacterial effectors targeting host domains involved in host-endogenous PPIs via eukaryotic-like domains or bacteria-exclusive domains. For instance, *Ralstonia solanacearum* have acquired a host-like F-box domain (PF00646) that competes with host-endogenous F-box protein for binding to SKP1, thus hijacking the ubiquitin-proteasome pathway in *Arabidopsis thaliana* (Angot et al., 2006), while *Shigella flexneri* have convergently evolved a GEF domain (PF03278) that competes with host Rho GEF, thus activating the Rho GTPase signaling pathway in humans (Huang et al., 2009). In addition to mechanistic studies on individual host-targeting domains in bacteria, databases of eukaryotic-like domains and short linear motifs provide a snapshot of the extent to which bacterial pathogens mimic host structural modules. For instance, EffectiveDB, a database for predicting bacterial effectors based on several criteria including the presence of eukaryotic-like domains, currently reports 2,636 eukaryotic-like domains as being significantly enriched (Z-score $\geq$ 4) in the genomes of pathogenic *vs.* non-pathogenic bacteria (Eichinger et al., 2016). Meanwhile, the Eukaryotic Linear Motif Resource currently contains ~100 instances of bacteria-mimicked eukaryotic short linear motifs from a small number of extensively studied pathogenic species (Samano-Sanchez and Gibson, 2020).

Here, we constructed a domain-resolved network consisting of eukaryote-endogenous, bacteria-endogenous and host-bacteria protein-protein interactions (PPIs), based on which we studied the mechanism of host binding site mimicry by bacterial proteins, and systematically probed the proteomes of pathogenic bacteria for domains that mimic or target host domains engaging in domain-domain interactions (DDIs) that are specific to eukaryotes, as opposed to DDIs that are conserved between eukaryotes and bacteria. Our comprehensive and quantitative profiling of bacterial proteomes reveals statistically significant enrichment of domains and short linear motifs in bacterial effectors that interact with host domains engaged in eukaryote-specific DDIs, which allows host-bacteria PPIs to mimic host-endogenous PPIs on an interactome scale (**Figure 7**). We found that consistent with previous results for host-virus interactions, binding site sharing among host proteins largely results from gene duplication followed by divergent evolution, whereas binding site mimicry by bacterial proteins seems to largely result from convergent evolution (or extreme divergent evolution) of structural modules in bacteria that bear little resemblance to those in host. Our results indicate that: 1) effectors are six times as likely as non-effectors to contain host-like domains that mediate DDIs exclusively in eukaryotes (**Figure 2**); 2) the average domain in effectors is seven times as likely as that in non-effectors to co-occur with DDI partners in eukaryotes rather than in bacteria (**Figure 3**); and 3) effectors are nine times as likely as non-effectors to contain bacteria-exclusive domains that target host domains mediating DDIs exclusively in eukaryotes (**Figure 4**). Moreover, in the absence of host-like domains or among pathogen proteins without domain assignment, effectors harbor a higher variety and density of short linear motifs targeting host domains that mediate DDIs exclusively in eukaryotes.

While our dataset does contain more host-endogenous than bacteria-endogenous or host-bacteria PPIs and DDIs, this imbalance should not confound our results, as domain assignment and DDI templates are not taxonomy-specific, but rather are used to resolve all PPIs, regardless of the species

involved. In fact, our estimation of domain's relevance to eukaryote-specific DDIs anticipates and accounts for DDIs that are exclusive to host species, by giving more weight to domains engaging in such DDIs. In **Tables 1**,**3** showing examples of effectors containing domains mediating DDIs either exclusively or primarily in eukaryotes, the fact that many domains can be traced to a few species is a technical consequence of proteins containing the same domains being merged into UniRef50 clusters, and only the species of the representative member of each cluster being retained. It is also a testament to extensive domain sharing among diverse pathogenic species. Taxonomic information may be useful when comparing effectors that are indistinguishable at the domain level but exhibit more variations at the residue level. Pooled analysis of proteins with identical domain compositions across different species can reveal general patterns in the host-bacteria PPI network that may not be obvious on a species-by-species or protein-by-protein basis. On the one hand, host domains targeted by multiple effector domains can reveal convergent evolution of common virulence mechanisms among different pathogenic species, which may prove useful in developing broad spectrum antibiotics. For instance, the human Ras domain (PF00071) is targeted by structurally distinct domains in *Legionella* (PF14860, PF18172, PF18641), *Pseudomonas* (PF03496), *Salmonella* (PF03545, PF05925, PF07487), *Shigella* (PF03278) and *Yersinia* (PF00069, PF09632) effectors. On the other hand, effector domains targeting multiple host domains and thus potentially perturbing multiple host pathways represent targets for multipronged therapeutic intervention. Of the 103 host-targeting bacterial proteins in our PPI dataset, 71 interact with a single host protein, while 32 interact with multiple host proteins. For instance, the *Pseudomonas* effector ExoS contains the ADP ribosyltransferase domain (PF03496), which it uses to target host proteins containing either a 14-3-3 domain (PF00244) or Ras small GTPase domain (PF00071). These host domains participate in a wide array of signaling pathways (Xiao et al., 1995; Stenmark and Olkkonen, 2001). While experimentally determined protein-protein interactions (PPIs) may be biased towards well-studied species, and domain-domain interaction templates may be biased towards well-studied protein structures, our survey of the proteomes of 84 pathogenic bacterial species is nonetheless more comprehensive than case studies of bacterial effectors in uncovering general molecular recognition principles underlying the host-bacteria PPI network. To increase coverage of the structurally-resolved host-bacteria PPI network, future efforts should focus on emerging pathogenic strains of bacteria, more systematic mapping of host-bacteria interactomes, as well as new molecular modelling methods to predict structures of proteins and protein-protein interactions which do not have homologs with known structure (Wu and Zhang, 2008).

To identify domains in bacteria that are most likely involved in mimicking host-endogenous protein-protein interactions (PPIs), we excluded eukaryotic-like domains which engage in either interchain or intrachain domain-domain interactions (DDIs) in bacteria. Previous studies suggest that intrachain DDIs often occur between adjacent domains within the same protein (Littler and Hubbard, 2005); however, PPIs are much less likely attributable to DDIs derived solely from intrachain interactions, compared to DDIs derived from interchain interactions (Itzhaki et al., 2006). Although distinguishing biologically relevant interfaces from artifactual crystal contacts is beyond the scope of this work, several interface classification algorithms have been developed to address this specific issue, based on various criteria such as contact size and evolutionary conservation of interface residues (Valdar and Thornton, 2001; Duarte et al., 2012), thermodynamic prediction of interface stability (Krissinel and Henrick, 2007), and interface conservation across multiple crystal forms of a protein (Xu et al., 2008). Here, we take a conservative approach and exclude all eukaryotic-like domains engaging in intraprotein DDIs in bacteria, because while conserved eukaryotic-like DDIs may not contribute to homologous PPIs in bacteria, they may function in maintaining protein stability or metabolic processes in bacteria (Tsoka and Ouzounis, 2000), rather than mediate host-bacteria PPIs. One example of a domain mediating PPIs in eukaryotes but serving a structural function in bacteria is the Fibronectin type III domain (PF00041), which in animals is involved in cell adhesion, migration and differentiation, and whose interaction with 44 domains leads to 1,156 interactions among 738 proteins in eukaryotes. While PF00041 does not mediate PPIs between bacterial proteins, it forms crystal contacts with the domain PF00704 within the *Bacillus thuringiensis* chitinase protein (PDB: 6BT9), and likely acts as a linker in the multi-domain chitinase (Juarez-Hernandez et al., 2019), rather than mediating host-pathogen PPIs.

In conclusion, our demonstration of binding site mimicry and its mechanisms at the domain level in the host-bacteria PPI network provides novel insight into the evolution of host-interacting domains in bacterial effectors. In particular, we showed that convergent evolution (or extreme divergent evolution) appears to be the more dominant mechanism behind binding site mimicry in host-bacteria interactions. To date, similar analysis has only been done for viral proteins. In addition, our estimation of domain's relevance to eukaryote-specific DDIs provides quantitative, interaction-based criteria for identifying novel effectors, based on: 1) domains that exclusively or primarily mediate DDIs in eukaryotes; and 2) variety and density of short linear motifs targeting host domains that exclusively mediate DDIs in eukaryotes. Although predicting new effector-host interactions is beyond the scope of this paper, our study presents a first step toward resolving PPI interfaces in effector-host interactions: once the domains involved in effector-host PPIs are identified by our method, interface residues inside such domains can be predicted using machine learning methods (Meyer et al., 2018). By mapping the interface residues involved in host-effector PPIs, it may be possible to develop antibiotics that precisely inhibit host-pathogen PPIs, with minimal disruption to host-endogenous PPIs (Voter and Keck, 2018). Given the scarcity of host-bacteria PPI data and the rapidly increasing number of completely sequenced pathogen genomes, our framework for assessing the functional impact of structural modules within pathogen proteins, without needing direct experimental evidence of their interaction with host proteins, may help accelerate the discovery and mechanistic study of novel virulence factors, as well as the development of selective inhibitors of pathogen-subverted host signaling pathways.

MATERIALS AND METHODS

Domain-Resolved Eukaryote-Bacteria Protein-Protein Interaction Network

Eukaryote-endogenous, bacteria-endogenous, and host-bacteria protein-protein interaction (PPI) data were obtained from IntAct and HPIDB 3.0 (Orchard et al., 2014; Ammari et al., 2016). Domain-domain interaction (DDI) templates were obtained from 3did and Pfam (Mosca et al., 2014; El-Gebali et al., 2019). IntAct is one of the largest and most cited databases of literature-curated, high quality molecular interactions in multiple organisms (615,015 unique binary protein-protein interactions in 1,572 organisms). 3did is a comprehensive, regularly maintained resource for domain-domain and domain-motif interaction templates derived from PDB structures (14,278 domain-domain and 920 domain-motif interaction templates).

To resolve protein-protein interactions into domain-domain interactions, we first predicted the occurrence of domains in proteins with InterProScan, using Pfam's gathering threshold (Jones et al., 2014). We then considered all 14,278 unique types of DDI templates involving 8,048 interacting Pfam domains in the 3did database, which corresponds to ~1.77 DDIs per domain. The resulting integrated eukaryote-bacteria DDI network consists of 5,950 unique types of DDIs involving 3,558 interacting domains, which corresponds to ~1.67 DDIs per domain. In other words, the DDI-to-domain ratio is roughly preserved in the construction of domain-resolved eukaryote-bacteria interactome. Possible reasons for the absence of certain domains and DDIs in our interactome are: 1) they only occur in organisms not considered in our study, such as archaea and viruses; and 2) the DDI only occurs between domains within the same protein, rather than mediating PPIs between different proteins. When a PPI can be attributed to several possible DDIs, priority was given to interchain DDIs (derived from PDB structures consisting of at least two distinct protein entities), followed by intrachain DDIs.

Selection Criteria for Effector and Non-Effector Proteins

We included proteins encoded by pathogenic bacterial species catalogued in PHI-base (Urban et al., 2020). PHI-base contains expert curated, regularly updated data on pathogen genes with experimentally verified impact on host-pathogen interactions (216 host species, 274 pathogenic species, 7,681 pathogen genes). Effector protein IDs were retrieved from the PHI-base annotation file "phi-base_current.csv", by searching for genes whose "Gene Function" or "Mutant Phenotype" column contains the keyword "effector". In addition, effector protein IDs were also retrieved from UniProt (UniProt, 2019) using two sets of keywords. **By gene name:** taxonomy:"Bacteria [2]" name:effector (name:"type 1" OR name:"type 2" OR name:"type 3" OR name:"type 4" OR name:"type 5" OR name:"type 6" OR name:"type 7" OR name:"type 8" OR name:"type 9" OR name:t*ss OR name:"secretion system") **By cellular location:** taxonomy:"Bacteria [2]" (annotation(type:function effector) OR locations: (note:"type 1") OR locations(note:"type 2") OR locations: (note:"type 3") OR locations: (note:"type 4") OR locations: (note:"type 5") OR locations(note:"type 6") OR locations: (note:"type 7") OR locations: (note:"type 8") OR locations: (note:"type 9") OR locations: (note:t*ss) OR locations: (note:"secretion system")) (locations: (location:"Secreted [SL-0243]") OR locations: (location:"Host [SL-0431]")) Non-effectors consist of cytoplasmic, membrane as well as other secreted proteins encoded by the same pathogen species considered for effector proteins. **Cytoplasmic:** taxonomy:"Bacteria [2]" locations: (location:"Cytoplasm [SL-0086]") **Membrane:** taxonomy:"Bacteria [2]" (locations: (location:"Cell envelope [SL-0036]") OR locations: (location:"Membrane [SL-0162]")) **Secreted:** taxonomy:"Bacteria [2]" (locations: (location:"Secreted [SL-0243]") OR locations: (location:"Host [SL-0431]"))

Merging Bacterial Proteins With Identical Domain Compositions

Taxonomy and Pfam domain annotations of proteins were obtained from UniProt and InterPro (Mitchell et al., 2019). For each domain, we counted the number of eukaryotic and bacterial species encoding at least one protein containing that domain. To minimize the impact of spurious domains, such as arising from contaminated genomes or misannotated proteins, we required that each domain be found in at least three eukaryotic or bacterial proteomes–at least one of which must be a reference proteome or belong to a pan proteome. Bacterial proteins with identical domain compositions were merged into a single entry, as they are indistinguishable from one another at the domain resolution. To further reduce redundancy among highly related protein sequences (*e.g.* orthologs or fragments of the same protein) while also maintaining sufficient resolution, sequences belonging to the same UniRef50 cluster were ranked based on whether they are: 1) representative for the cluster, as assigned by UniRef50; 2) manually reviewed; 3) assigned high annotation score by UniProt; 4) from UniProt reference proteomes; and 5) longest. Only the top-ranking sequence was retained for each UniRef50 cluster. For domain compositions that are common to effectors and non-effectors, we assessed their relative frequency in effectors *vs.* non-effectors. Domain compositions that are significantly enriched (q-value < 0.1) in effectors were assigned to effectors, and domain compositions that are significantly depleted (q-value < 0.1) in effectors were assigned to non-effectors. Our final dataset thus contains 238 effectors and 3,921 non-effectors with unique domain signatures.

Statistical Tests Performed

Fisher's exact test was used for analyses based on odds ratios, and Wilcoxon test was used for analyses based on difference in means. To control the false discovery rate in multiple hypothesis testing, we calculated the positive false discovery rate, or q-values (Storey, 2003). All statistical analyses were conducted in R (Team, 2018).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: IntAct (ftp://ftp.ebi.ac.uk/pub/

databases/intact/current/psimitab/intact.zip) HPIDB (http://hpidb.igbb.msstate.edu/downloads/hpidb2.mitab.zip) InterPro (ftp://ftp.ebi.ac.uk/pub/databases/interpro/uniparc_match.tar.gz) Eukaryotic Linear Motif (http://elm.eu.org/downloads.html).

AUTHOR CONTRIBUTIONS

YC was responsible for conceptualization, data curation, formal analysis, and writing of the manuscript. YX was responsible for conceptualization, funding acquisition, supervision, and review of the manuscript.

FUNDING

This work was supported by Natural Sciences and Engineering Research Council of Canada grants RGPIN-2019-05952 and RGPAS-2019-00012, Canada Foundation for Innovation grants JELF-33732 and IF-33122, and Canada Research Chairs program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.626600/full#supplementary-material

REFERENCES

Ammari, M. G., Gresham, C. R., McCarthy, F. M., and Nanduri, B. (2016). *HPIDB 2.0: A Curated Database for Host-Pathogen Interactions*. Oxford: Database.

Angot, A., Peeters, N., Lechner, E., Vailleau, F., Baud, C., Gentzbittel, L., et al. (2006). Ralstonia Solanacearum Requires F-box-like Domain-Containing Type III Effectors to Promote Disease on Several Host Plants. *Proc. Natl. Acad. Sci.* 103 (39), 14620–14625. doi:10.1073/pnas.0509393103

Arnold, R., Boonen, K., Sun, M. G. F., and Kim, P. M. (2012). Computational Analysis of Interactomes: Current and Future Perspectives for Bioinformatics Approaches to Model the Host-Pathogen Interaction Space. *Methods*. 57 (4), 508–518. doi:10.1016/j.ymeth.2012.06.011

Cazalet, C., Rusniok, C., Brüggemann, H., Zidane, N., Magnier, A., Ma, L., et al. (2004). Evidence in the Legionella pneumophila Genome for Exploitation of Host Cell Functions and High Genome Plasticity. *Nat. Genet.* 36 (11), 1165–1173. doi:10.1038/ng1447

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., et al. (2012). Attributes of Short Linear Motifs. *Mol. Biosyst.* 8 (1), 268–281. doi:10.1039/c1mb05231d

Duarte, J. M., Srebniak, A., Schärer, M. A., and Capitani, G. (2012). Protein Interface Classification by Evolutionary Analysis. *BMC Bioinformatics*. 13, 334. doi:10.1186/1471-2105-13-334

Eichinger, V., Nussbaumer, T., Platzer, A., Jehl, M.-A., Arnold, R., and Rattei, T. (2016). EffectiveDB-updates and Novel Features for a Better Annotation of Bacterial Secreted Proteins and Type III, IV, VI Secretion Systems. *Nucleic Acids Res.* 44 (D1), D669–D674. doi:10.1093/nar/gkv1269

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam Protein Families Database in 2019. *Nucleic Acids Res.* 47 (D1), D427–D432. doi:10.1093/nar/gky995

Franzosa, E. A., and Xia, Y. (2011). Structural Principles within the Human-Virus Protein-Protein Interaction Network. *Proc. Natl. Acad. Sci.* 108 (26), 10538–10543. doi:10.1073/pnas.1101440108

Fu, Y., and Galán, J. E. (1998). Identification of a Specific Chaperone for SptP, a Substrate of the Centisome 63 Type III Secretion System ofSalmonella Typhimurium. *J. Bacteriol.* 180 (13), 3393–3399. doi:10.1128/jb.180.13.3393-3399.1998

Galán, J. E. (2009). Common Themes in the Design and Function of Bacterial Effectors. *Cell Host Microbe.* 5 (6), 571–579. doi:10.1016/j.chom.2009.04.008

Garamszegi, S., Franzosa, E. A., and Xia, Y. (2013). Signatures of Pleiotropy, Economy and Convergent Evolution in a Domain-Resolved Map of Human-Virus Protein-Protein Interaction Networks. *Plos Pathog.* 9 (12), e1003778. doi:10.1371/journal.ppat.1003778

Grau-Bové, X., Sebé-Pedrós, A., and Ruiz-Trillo, I. (2015). The Eukaryotic Ancestor Had a Complex Ubiquitin Signaling System of Archaeal Origin. *Mol. Biol. Evol.* 32 (3), 726–739. doi:10.1093/molbev/msu334

Huang, Z., Sutton, S. E., Wallenfang, A. J., Orchard, R. C., Wu, X., Feng, Y., et al. (2009). Structural Insights into Host GTPase Isoform Selection by a Family of Bacterial GEF Mimics. *Nat. Struct. Mol. Biol.* 16 (8), 853–860. doi:10.1038/nsmb.1647

Itzhaki, Z., Akiva, E., Altuvia, Y., and Margalit, H. (2006). Evolutionary Conservation of Domain-Domain Interactions. *Genome Biol.* 7 (12), R125. doi:10.1186/gb-2006-7-12-r125

Janjusevic, R., Abramovitch, R. B., Martin, G. B., and Stebbins, C. E. (2006). A Bacterial Inhibitor of Host Programmed Cell Death Defenses Is an E3 Ubiquitin Ligase. *Science*. 311 (5758), 222–226. doi:10.1126/science.1120131

Jehl, M.-A., Arnold, R., and Rattei, T. (2011). Effective--a Database of Predicted Secreted Bacterial Proteins. *Nucleic Acids Res.* 39 (Database issue), D591–D595. doi:10.1093/nar/gkq1154

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics*. 30 (9), 1236–1240. doi:10.1093/bioinformatics/btu031

Juarez-Hernandez, E. O., Casados-Vazquez, L. E., Brieba, L. G., Torres-Larios, A., Jimenez-Sandoval, P., and Barboza-Corona, J. E. (2019). The Crystal Structure of the Chitinase ChiA74 of Bacillus Thuringiensis Has a Multidomain Assembly. *Sci. Rep.* 9 (1), 2591. doi:10.1038/s41598-019-39464-z

Krissinel, E., and Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* 372 (3), 774–797. doi:10.1016/j.jmb.2007.05.022

Kumar, M., Gouw, M., Michael, S., Samano-Sanchez, H., Pancsa, R., Glavina, J., et al. (2020). ELM-the Eukaryotic Linear Motif Resource in 2020. *Nucleic Acids Res.* 48 (D1), D296–D306. doi:10.1093/nar/gkz1030

Littler, S. J., and Hubbard, S. J. (2005). Conservation of Orientation and Sequence in Protein Domain-Domain Interactions. *J. Mol. Biol.* 345 (5), 1265–1279. doi:10.1016/j.jmb.2004.11.011

Marchesini, M. I., Herrmann, C. K., Salcedo, S. P., Gorvel, J.-P., and Comerci, D. J. (2011). In Search of Brucella Abortus Type IV Secretion Substrates: Screening and Identification of Four Proteins Translocated into Host Cells through VirB System. *Cell Microbiol.* 13 (8), 1261–1274. doi:10.1111/j.1462-5822.2011.01618.x

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., et al. (2001). Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs". *Genome Res.* 11 (12), 2120–2126. doi:10.1101/gr.205301

Meyer, M. J., Beltrán, J. F., Liang, S., Fragoza, R., Rumack, A., Liang, J., et al. (2018). Interactome INSIDER: A Structural Interactome Browser for Genomic Studies. *Nat. Methods*. 15 (2), 107–114. doi:10.1038/nmeth.4540

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: Improving Coverage, Classification and Access to Protein Sequence Annotations. *Nucleic Acids Res.* 47 (D1), D351–D360. doi:10.1093/nar/gky1100

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a Catalog of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucl. Acids Res.* 42 (Database issue), D374–D379. doi:10.1093/nar/gkt887

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

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Mičetić, I., Camilloni, C., et al. (2018). MobiDB 3.0: More Annotations for Intrinsic Disorder, Conformational Diversity and Interactions in Proteins. *Nucleic Acids Res.* 46 (D1), D471–D476. doi:10.1093/nar/gkx1071

Popa, C. M., Tabuchi, M., and Valls, M. (2016). Modification of Bacterial Effector Proteins inside Eukaryotic Host Cells. *Front Cel. Infect Microbiol.* 6, 73. doi:10.3389/fcimb.2016.00073

Sámano-Sánchez, H., and Gibson, T. J. (2020). Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity. *Trends Biochem. Sci.* 45 (6), 526–544. doi:10.1016/j.tibs.2020.03.003

Schweppe, D. K., Harding, C., Chavez, J. D., Wu, X., Ramage, E., Singh, P. K., et al. (2015). Host-Microbe Protein Interactions during Bacterial Infection. *Chem. Biol.* 22 (11), 1521–1530. doi:10.1016/j.chembiol.2015.09.015

Scott, N. E., and Hartland, E. L. (2017). Post-translational Mechanisms of Host Subversion by Bacterial Effectors. *Trends Mol. Med.* 23 (12), 1088–1102. doi:10.1016/j.molmed.2017.10.003

Stebbins, C. E., and Galán, J. E. (2001). Structural Mimicry in Bacterial Virulence. *Nature*. 412 (6848), 701–705. doi:10.1038/35089000

Steele-Mortimer, O., Knodler, L. A., Marcus, S. L., Goh, B., Pfeifer, C. G., Finlay, B. B., et al. (2000). Activation of Akt/Protein Kinase B in Epithelial Cells by theSalmonella Typhimurium Effector SigD. *J. Biol. Chem.* 275 (48), 37718–37724. doi:10.1074/jbc.m008187200

Stenmark, H., and Olkkonen, V. M. (2001). The Rab GTPase Family. *Genome Biol.* 2 (5), reviews3007. doi:10.1186/gb-2001-2-5-reviews3007

Storey, J. D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value. *Ann. Stat.* 31 (6), 2013–2035. doi:10.1214/aos/1074290335

Team, R. C. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tsoka, S., and Ouzounis, C. A. (2000). Prediction of Protein Interactions: Metabolic Enzymes Are Frequently Involved in Gene Fusion. *Nat. Genet.* 26 (2), 141–142. doi:10.1038/79847

UniProt, C. (2019). UniProt: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47 (D1), D506–D515. doi:10.1093/nar/gky1049

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., et al. (2020). PHI-base: the Pathogen-Host Interactions Database. *Nucleic Acids Res.* 48 (D1), D613–D620. doi:10.1093/nar/gkz904

Valdar, W. S. J., and Thornton, J. M. (2001). Conservation Helps to Identify Biologically Relevant Crystal Contacts. *J. Mol. Biol.* 313 (2), 399–416. doi:10.1006/jmbi.2001.5034

Voter, A. F., and Keck, J. L. (2018). Development of Protein-Protein Interaction Inhibitors for the Treatment of Infectious Diseases. *Adv. Protein Chem. Struct. Biol.* 111, 197–222. doi:10.1016/bs.apcsb.2017.07.005

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., et al. (2000). Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development. *Science*. 287 (5450), 116–122. doi:10.1126/science.287.5450.116

Wu, S., and Zhang, Y. (2008). MUSTER: Improving Protein Sequence Profile-Profile Alignments by Using Multiple Sources of Structure Information. *Proteins*. 72 (2), 547–556. doi:10.1002/prot.21945

Xiao, B., Smerdon, S. J., Jones, D. H., Dodson, G. G., Soneji, Y., Aitken, A., et al. (1995). Structure of a 14-3-3 Protein and Implications for Coordination of Multiple Signalling Pathways. *Nature*. 376 (6536), 188–191. doi:10.1038/376188a0

Xu, Q., Canutescu, A. A., Wang, G., Shapovalov, M., Obradovic, Z., and Dunbrack, R. L. (2008). Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins. *J. Mol. Biol.* 381 (2), 487–507. doi:10.1016/j.jmb.2008.06.002

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

frontiers in Molecular Biosciences

REVIEW
published: 10 May 2021
doi: 10.3389/fmolb.2021.668184

Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds

Nicola Bordin[1], *Ian Sillitoe*[1], *Jonathan G. Lees*[2] *and Christine Orengo*[1]*

[1]*Institute of Structural and Molecular Biology, University College London, London, United Kingdom,* [2]*Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, United Kingdom*

This article is dedicated to the memory of Cyrus Chothia, who was a leading light in the world of protein structure evolution. His elegant analyses of protein families and their mechanisms of structural and functional evolution provided important evolutionary and biological insights and firmly established the value of structural perspectives. He was a mentor and supervisor to many other leading scientists who continued his quest to characterise structure and function space. He was also a generous and supportive colleague to those applying different approaches. In this article we review some of his accomplishments and the history of protein structure classifications, particularly SCOP and CATH. We also highlight some of the evolutionary insights these two classifications have brought. Finally, we discuss how the expansion and integration of protein sequence data into these structural families helps reveal the dark matter of function space and can inform the emergence of novel functions in Metazoa. Since we cover 25 years of structural classification, it has not been feasible to review all structure based evolutionary studies and hence we focus mainly on those undertaken by the SCOP and CATH groups and their collaborators.

Keywords: bioinformatics and computational biology, protein structural and functional analysis, structural bioinformatics, protein evolution, protein structure classification

OPEN ACCESS

Edited by:
Sarah Teichmann,
Wellcome Sanger Institute (WT),
United Kingdom

Reviewed by:
Patrick Senet,
Université de Bourgogne, France
Joost Schymkowitz,
VIB and KU Leuven Center for Brain and disease Research, Belgium

***Correspondence:**
Christine Orengo
c.orengo@ucl.ac.uk

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *15 February 2021*
Accepted: *27 April 2021*
Published: *10 May 2021*

Citation:
Bordin N, Sillitoe I, Lees JG and Orengo C (2021) Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds. Front. Mol. Biosci. 8:668184. doi: 10.3389/fmolb.2021.668184

THE EARLY DAYS–CHOTHIA THE PIONEER

Protein structures have helped us see more clearly into the evolutionary past. Cyrus Chothia, to whom this special issue is dedicated, was an early pioneer on these journeys and remained a leading figure throughout his life. As structures accumulated in the Protein Data Bank (PDB) from the early 1970s onwards, he was one of the first to realise the value of comparing them to capture their differences and thereby understand the mechanisms by which proteins evolve. In a similar timeframe i.e. the late 70s and early 80s, another early pioneer in the protein world, Margaret Dayhoff, was also cataloging evolutionary changes by considering the substitutions, insertions and deletions in the amino acid residues that can occur in the protein's polypeptide chain. By linking these data, we can see how genetic variations translate to structural and ultimately functional impacts. Over the last two decades the explosion in sequence data arising from increasingly sophisticated sequencing technologies, including sequences from thousands of completed genomes, have sharpened these insights. In parallel, structure prediction has seen some quantum leaps over the last decade including from exploitation of AI and deep learning strategies that may bring structural annotations to many mysterious regions of sequence space currently uncharacterised. In this review we highlight some of the major shifts in technology and data that have enabled better exploration of protein structure space and brought functional insights.

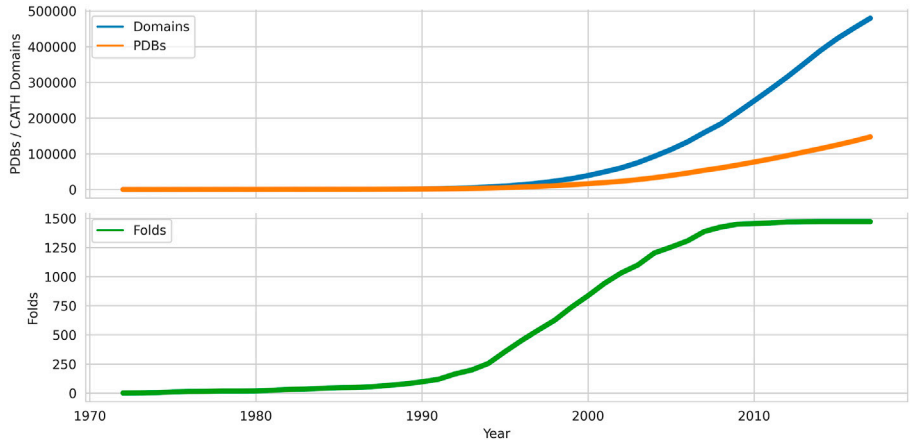


FIGURE 1 | Growth of domains, folds and chains deposited in the Protein Data Bank from 1972 onwards. Data sources: PDB, CATH.

Early Identification of Protein Families

The technical challenges of determining 3D structures of proteins has meant that the sequence data has always outstripped structural data–currently more than 300-fold. There are approximately 170,000 protein structures in the PDB (Armstrong et al., 2019) but more than 200 million sequences in UniProt (The UniProt Consortium, 2019), and metagenomic data adds billions more sequences (Mitchell et al., 2019). In the late 70s and early 80s, Dayhoff pioneered the comparison of protein sequences, designing residue substitution matrices which enabled the alignment of even relatively distant relatives diverged from a common ancestor. Many other approaches have been explored since then (e.g. BLOSUM (Henikoff and Henikoff, 1992)), see review for others (Jones et al., 1992)). These approaches and the dynamic programming algorithms (e.g. developed by Needleman and Wunsch (Needleman and Wunsch, 1970), Smith and Waterman (Smith and Waterman, 1981)) developed to align protein sequences started the identification of protein evolutionary families by Dayhoff and others.

How Constrained Are Protein Structures?

Adding structural data can help probe functional mechanisms more deeply and as the Protein Databank grew from the 1970s onwards (see **Figure 1**), algorithms for comparing structures emerged e.g. the still widely used rigid body approaches developed by Rossman and Argos (Rossmann and Argos, 1976) amongst others 9). As the PDB data grew it became clear that in some evolutionary superfamilies considerable divergence outside the structural core could occur.

One of the earliest and most important insights into structural divergence was captured by Cyrus Chothia and Arthur Lesk in their comparison of more than 32 pairs of protein homologues (Chothia and Lesk, 1986). This analysis showed the exponential relationship between sequence change and structural change and many of the characteristics captured in that study still hold when much larger datasets are examined. **Figure 2** shows the relationship detected for current data using the SSAP structure comparison algorithm (see below and (Orengo and Taylor, 1996)). For relatives having similar functional properties, the structure is highly conserved even at low sequence similarity. Extreme divergence occurs for relatives with different functional properties, likely to be paralogues, having different structural constraints imposed by these functions.

To expand on these insights, Chothia and Lesk published some detailed and beautifully described expositions of the sequence structure relationships for two important protein families the globins (Lesk and Chothia, 1980) and the immunoglobulins (Chothia and Lesk, 1982; Lesk and Chothia, 1982).

To capture structural properties between very diverse homologues, many new methods emerged to better cope with the extensive residue mutations, insertions and deletions occurring between them. These methods have continued to evolve since the late 1980s. Many built on the dynamic programming strategies successfully exploited in sequence comparison. In some, dynamic programming was applied at two levels to fully exploit the 3D data. First at a low level (i.e. residue views) and then to an upper summary level to obtain the final alignment (e.g. see SSAP (Orengo and Taylor, 1996)). Other approaches combined rigid body superposition with dynamic programming (see for example early approaches STAMP (Russell and Barton, 1992), STRUCTAL (Subbiah et al., 1993), CE (Shindyalov and Bourne, 1998)). One of the most popular algorithms with crystallographers and other structural biologists, DALI (Holm and Sander, 1993), effectively "chopped" the structures into hexapeptide fragments and used Monte Carlo optimization to determine the optimal order for concatenating matched fragments between the structures. Other approaches commonly used by structural biologists include MAMMOTH (Ortiz et al., 2002) and GESAMT (Krissinel, 2012). Fast approaches (e.g. CATHedral (Redfern et al., 2007)) were also developed that explicitly compared secondary structure elements between proteins giving up to 1000-fold speedups in the alignments but at the cost of accurate residue alignments. These approaches were driven by the exponential increase in the number of structures in the PDB and the need for rapid scans

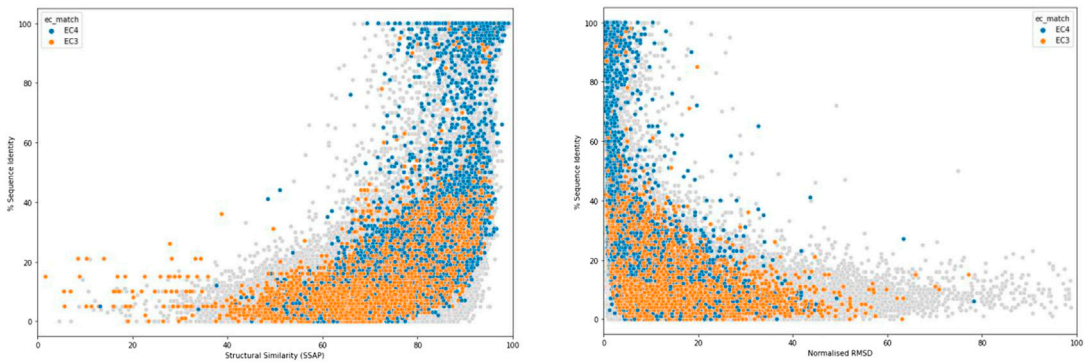


FIGURE 2 | Structural similarity measured by SSAP score (left) or normalised RMSD (right) vs % of sequence identity.

with newly solved structures to identify novel folds. More recent approaches (e.g. FATCAT (Ye and Godzik, 2003)) have been explicitly designed to optimize the alignments between loops, typically the most diverse regions, but often containing key functional residues.

Domain Based Structural Families

Chothia's examination of the globins and immunoglobulins was the first step toward a more comprehensive analysis of structure space and analyses performed in the following decade culminated in the establishment of one of the most widely used resources capturing protein domain structure superfamilies–SCOP (Murzin et al., 1995) in 1994. SCOP was co-founded by Alexey Murzin, who joined Chothia's team at the LMB and has remained a leading structure based evolutionary resource. Its first release contained 366 superfamilies, 866 non-redundant domain structures and 1182 protein domains from different species. As well as classifying domains by their superfamily, the superfamilies were also organized by class (determined by secondary structure composition) and fold group (determined by the order and orientation of those secondary structure elements in 3D space) in a hierarchical manner. Superfamilies in which relatives adopted regular arrangements in 3D were also annotated with architecture descriptions e.g. barrel, sandwich. Significant manual curation ensured very high quality in the assignments and annotations. SCOP has been expanded recently by inclusion of additional resources in SCOPe, managed by Steven Brenner and co-workers (Fox et al., 2014).

Continued expansion of the PDB has led to nearly a 10-fold increase in the number of superfamilies but the growth in new folds has been much slower (see **Figure 1** for numbers from a related resource). In parallel, Janet Thornton's group used a more automated approach by applying the SSAP structure comparison method (Orengo and Taylor, 1996), developed by Orengo and Taylor, to recognise homologues, including very distant homologues, and structures with similar folds. For extremely distant relatives, manual curation was also required but overall was not applied to the same extent as in SCOP. The CATH resource, set up by Orengo and Thornton, included a more formal architecture level within the hierarchy (see **Figure 3**).

As a result of the comparative ease of acquiring experimental data, the sequence databases (e.g. UniProt) expanded even more rapidly than the structure databank (PDB) and the increase in this information and more powerful profile based sequence comparison strategies to harness it e.g. PSI-BLAST (Altschul et al., 1997), HMMer (Eddy, 1998), HHsearch (Söding, 2005) aided in the confirmation of homologues in which structures had diverged considerably (see **Figure 4**). By capturing these extremely remote homologies, it became clear that sometimes only the structural core was conserved (see also **Figure 5**) (Dessailly et al., 2010). The variation in size across some superfamilies suggested a structural continuum and was also referred to as the "Russian Doll Effect" (Swindells et al., 1998). Furthermore, it was clear that some folding arrangements consisted of multiple repeat motifs e. g alpha-beta, beta-beta, alpha-alpha. Andrei Lupas and other groups highlighted primitive motifs appearing in early life that seeded the emergence of more complex folds through duplication and gene fusion (Lupas et al., 2001). In fact, a large scale application of the DALI algorithm on all known structures in the PDB, by Liisa Holm, identified a small set of very highly populated so-called "attractor" motifs (e.g. αβ, β–β, αβ) that link structural superfamilies (Holm and Sander, 1996).

More detailed SCOP and CATH-based analyses have suggested the need for a less rigid hierarchy and recent structural classifications, such as the ECOD resource developed by Nick Grishin (Cheng et al., 2014), have adopted this approach. In ECOD, domain structures are grouped into superfamilies annotated by class and architecture information but relatives within the superfamilies can be described as adopting different folds.

Both SCOP and CATH have also changed since their inception to reflect these phenomena. In 2014, SCOP2 was released (Andreeva et al., 2014, 2) providing many valuable links between superfamilies sharing common structural motifs. Rare structural motifs are also identified, and biochemical features highlighted. CATH now describes the topology annotations (fold or T-level) for each superfamily as reflecting "core structural motifs" since large scale comparisons of relatives show that for the majority of homologous pairs at least 50% of the structure is conserved and the core topological motif is a helpful structural fossil revealing even the most distant relationships.

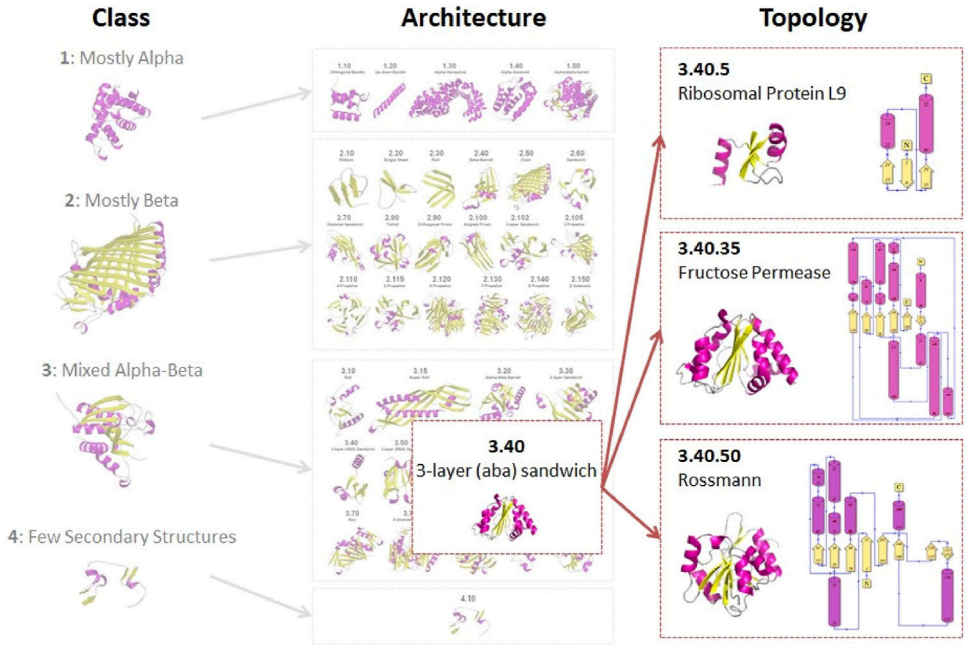


FIGURE 3 | Overview of the CATH classification scheme for protein domains.

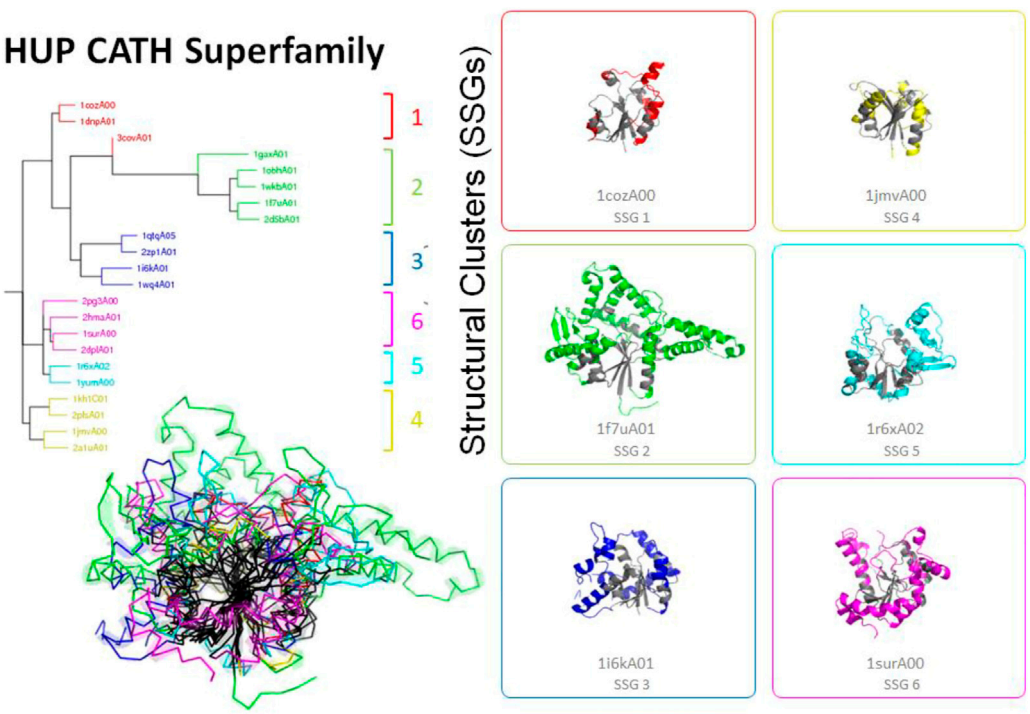


FIGURE 4 | Highly divergent structural homologues within the HUPs SuperFamily (CATH ID 3.40.50.620). Six diverse structural clusters (also called structurally similar groups, SSGs) are identified using SSAP to compare structures all against all (see tree top left and figures on the right). However, representatives from each SSG can be superposed to reveal the highly conserved structural core common to all (see central black region in the bottom left figure).

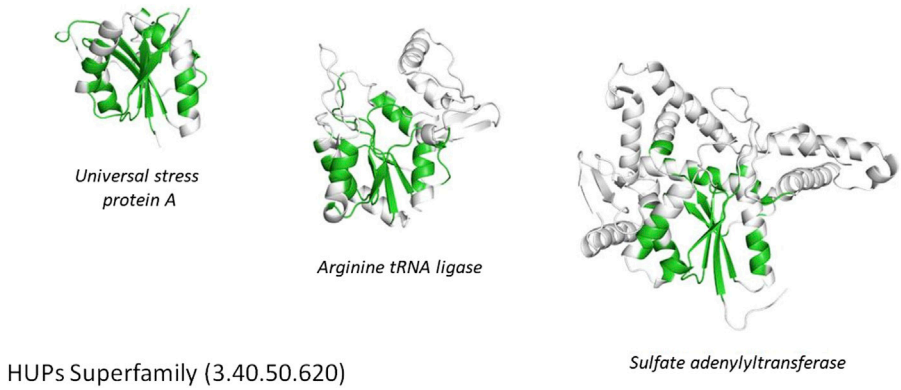


FIGURE 5 | Conservation of the structural core (highlighted in green) within the HUPs superfamily.

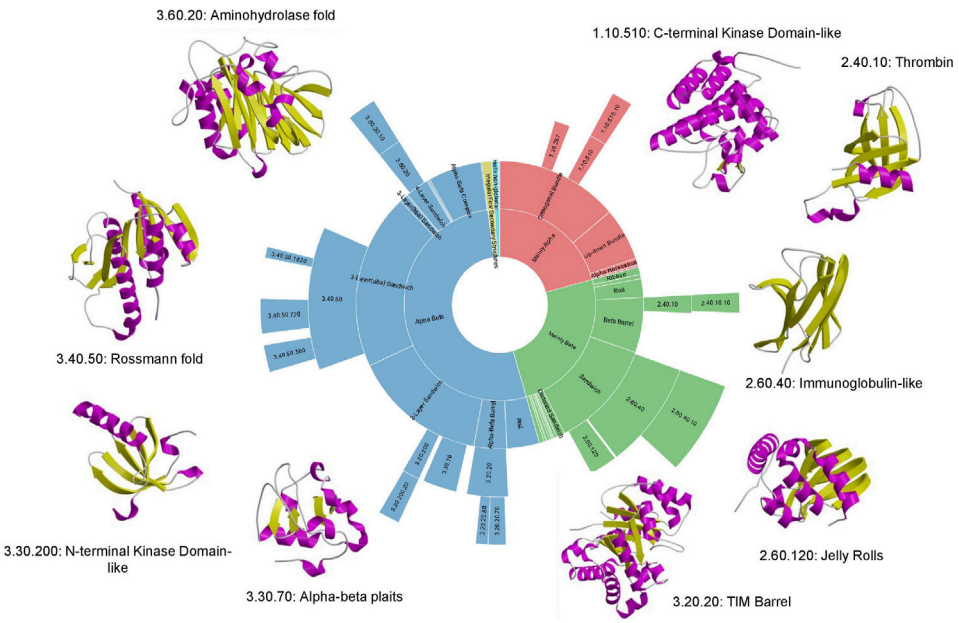
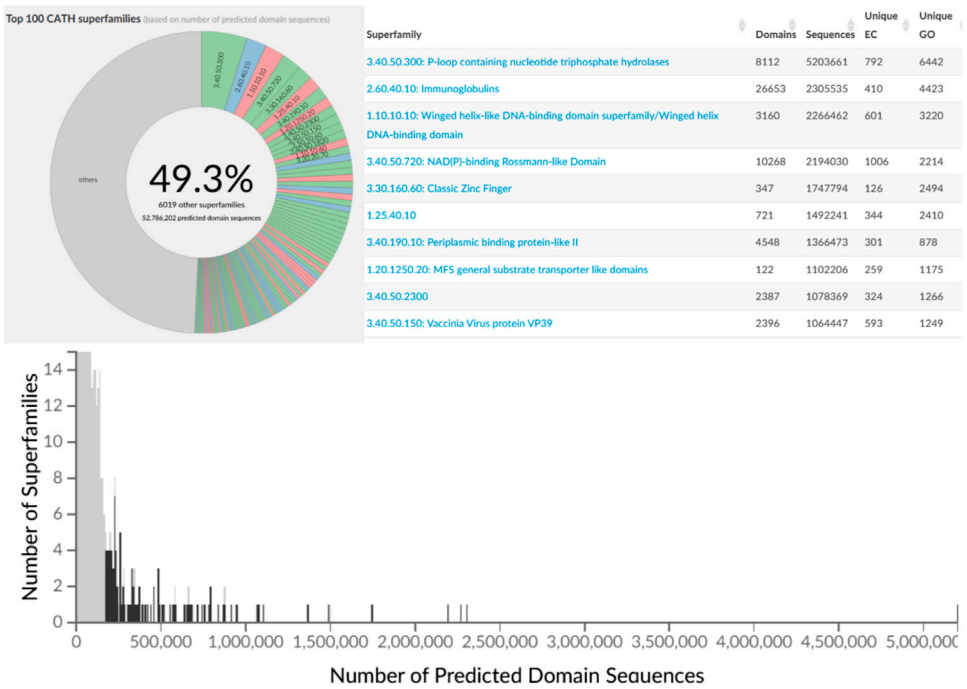


FIGURE 6 | Top 9 "super-folds" in CATH v4.3. The inner wheel shows the proportion of structures that fall into each class, architecture, fold group and superfamily respectively.

Unique Fold and Superfolds

Although structural classifications were clearly a valuable means of organising proteins and capturing evolutionary changes, a key question was the extent to which they reflected Nature or reflected bias in the Protein Data Bank. By using the more powerful profile-based sequence search strategies (e.g. PSI-BLAST) to map proteins with 3D-structures to all sequence relatives in UniProt, Chothia was able to show that even with the sparse structural data available at that time, a large proportion of sequences could be mapped to the SCOP families suggesting that these families were reasonably representative, though clearly they lacked many membrane associated proteins and disordered proteins (Chothia, 1992). That deficit still holds to some extent, although the PSI structural genomics initiatives which focused on membrane proteins helped to increase their representation in the structural classifications (Chandonia and Brenner, 2006). Current mapping done for some selected model organisms annotated in the integrated Genome3D resource ((Sillitoe et al., 2020) described below), shows that structural predictions based on SCOP or CATH superfamilies can be made for nearly 80% of proteins in many of these organisms, suggesting that a significant proportion of protein superfamilies in Nature are now represented in the protein structure classifications. In 1994, Cyrus Chothia made a prediction of fewer than 1000 folds in Nature (Chothia, 1992), an amazingly prescient estimation as 25 years later we possibly have as few as 1300–although there is still some controversy around the



Superfamily	Domains	Sequences	Unique EC	Unique GO
3.40.50.300: P-loop containing nucleotide triphosphate hydrolases	8112	5203661	792	6442
2.60.40.10: Immunoglobulins	26653	2305535	410	4423
1.10.10.10: Winged helix-like DNA-binding domain superfamily/Winged helix DNA-binding domain	3160	2266462	601	3220
3.40.50.720: NAD(P)-binding Rossmann-like Domain	10268	2194030	1006	2214
3.30.160.60: Classic Zinc Finger	347	1747794	126	2494
1.25.40.10	721	1492241	344	2410
3.40.190.10: Periplasmic binding protein-like II	4548	1366473	301	878
1.20.1250.20: MFS general substrate transporter like domains	122	1102206	259	1175
3.40.50.2300	2387	1078369	324	1266
3.40.50.150: Vaccinia Virus protein VP39	2396	1064447	593	1249

FIGURE 7 | Top 100 most populated CATH SuperFamilies (CATH v4.3) with additional details regarding sequence counts and unique EC and GO terms for the top 10 most populated SuperFamilies.

definition of fold! Furthermore, the dominance of some folding arrangements in Nature is still clear, with the top nine "superfolds" still accounting for more than 30% of all classified domain structures (see **Figure 6**). Of the current superfolds, five were detected in 1994 using CATH data (Orengo et al., 1994) and the remaining four superfolds (1.20.120, 1.10.490, 2.80.10, 3.10.20) were superseded by others (3.60.20, 2.40.10, 3.30.200, 1.10.510) that were less well populated in the original CATH release.

MAPPING SEQUENCE SPACE TO THE STRUCTURAL FAMILIES

While sequence to structure mapping has demonstrated that we have fold representatives for a large proportion of protein superfamilies in Nature, large parts of superfamily space are not yet covered by detailed structural and functional characterisation. This becomes even more apparent when metagenome sequence data is added e.g. from MGnify (Mitchell et al., 2019). Most structure classification resources make use of powerful tools like HMMer, developed by Sean Eddy and co-workers (Eddy, 1998) and HH-suite, developed by Johannes Soding to identify sequence relatives (Söding, 2005).

MGnify is 10-fold bigger than UniProt, currently comprising mostly prokaryotic data but the Earth Biome and Tree of Life sequencing projects (Lewin et al., 2018) will expand the data for eukaryotes too. The alpha-beta hydrolase superfamily, the 9th most populated superfamily in CATH (by number of non-redundant representatives at 90% sequence identity) is massively expanded (10-fold) by metagenome sequences extracted from a range of bacterial environments. Some of these e.g. from wastewater environments and oceans have changed in response to recent selection pressure leading to divergence in the binding site to accommodate PET and other plastics, which these enzymes can now degrade.

The second phase of the PSI structural genomics in the States (2005–2010) explicitly targeted structurally uncharacterised protein sequences mapping to SCOP, CATH or Pfam superfamilies to extend structural knowledge of these dark regions of sequence space (Norvell and Berg, 2007). These analyses further confirmed early expositions of the power law in structure-sequence space whereby some superfamilies had been massively expanded through extensive gene duplication throughout evolution. Many of these very highly populated superfamilies (described as "Megafamilies" by the structural genomic initiatives), are universal to all kingdoms of life and contain domains performing essential generic functions, like the many Rossman superfamilies which bind nucleotide cofactors e.g. NAD or NADP, in a common cleft in the structure formed by a crossover in the polypeptide chain. **Figure 7** shows that currently the top 100 superfamilies account for nearly 50% of all protein domain sequences mapped to CATH structure superfamilies.

PROTEIN DOMAINS ARE COMBINED IN MILLIONS OF DIFFERENT WAYS IN NATURE

Analyses of the SCOP and CATH superfamilies confirmed the generic functional role of many domain relatives (see further

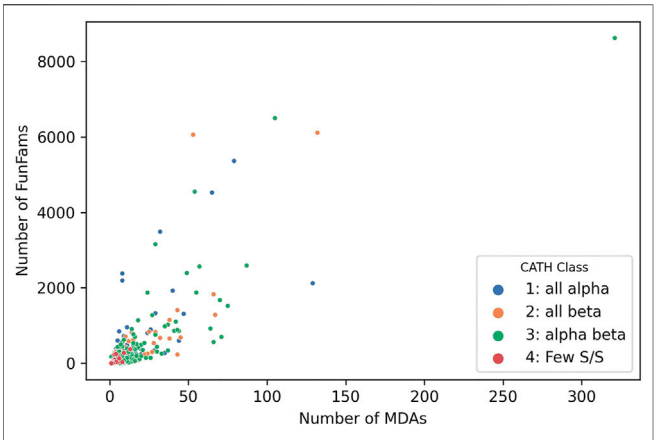


FIGURE 8 | Number of MDAs vs Number of sequence subfamilies (FunFams) for each SuperFamily in CATH v4.3.

discussion below) and the commonly used description of domains as independently folding functional units in evolution. The incredible enhancements in sequencing technologies at the turn of the millenium, allowing sequencing of whole genomes starting with human, meant that comparative genomics studies became possible exploring the different distribution of domain families and domain combinations within and between different kingdoms of life. There are now more than 300 complete and nearly complete genomes in ENSEMBL (Yates et al., 2020). This genomic data showed the extent of gene duplications, gene fusions and fissions occurring during evolution, with the former being more common (Björklund et al., 2005). Changes in these multidomain combinations or multidomain architectures (MDAs) result in expansions and divergence in the functional repertoires between species in response to selective pressures imposed by novel environmental contexts.

Studies inspired by Chothia's vision of domain units taken forward by various researchers he mentored, notably Sarah Teichmann and Mark Gerstein, characterised the "mosaic" nature of proteins and confirmed domains as the fundamental building blocks of life (Teichmann et al., 1999; Teichmann et al., 2001). Analyses of CATH-Gene3D which contains domain sequences from UniProt mapped to CATH and Pfam superfamilies using HMMer-based protocols currently reveal 311,575 different domain combinations. This is probably an underestimate since many proteins have regions of sequence that are still uncharacterized and may correspond to novel families that are unlikely to be common to multiple species. Unsurprisingly the more sequence sub-families found within a superfamily the more multidomain architectures identified (see **Figure 8**, below) and the power law is apparent again with the top 100 superfamilies occurring in the most MDA contexts occurring in a very large number of different MDA contexts (51% of all). Changes in domain context can modify the active site or binding pockets (discussed more below) and inevitably alter the surface features of the protein enabling diversity in protein interactions for paralogs expressed in different tissues. In addition, Teichmann and co-workers showed that some combinations of domains, described as supradomains, are particularly prevalent, probably corresponding to useful functional units (Vogel et al., 2004).

Comparative genome studies enabled by this vast sequence data could probe deep evolutionary relationships by using the structural families identified for different species. For example, CATH-based studies showed essential pathways populated by universal superfamilies that can be traced back to the Last universal Common Ancestor (LUCA) (Ranea et al., 2006).

STRUCTURE FAMILIES BRING DETAILED INSIGHTS INTO PROTEIN FUNCTION EVOLUTION

Chothia's eloquent reviews of domain structure families and evolutionary changes in protein structures were inspiring and played a key role in framing the questions around protein function evolution. In particular, he sought to elaborate on a "domain grammar of function" that would allow translation of a multi-domain "sentence" based on the functional roles of the constituent domains.

Other complementary studies added to the emerging picture. For example, Thornton's group analysed 31 highly populated and well structurally characterised superfamilies in CATH revealing the extent to which functions could diverge in particular in the megafamilies (Todd et al., 2001). A number of phenomena can drive this. Clearly the existence of multiple relatives in a genome means that extra copies (i.e. paralogues) will be more tolerant of mutations and these can drive functional shifts if they occur on or near key sites. In addition, as mentioned already, domain fusions can reshape functional sites or surfaces. Furthermore, relatives can oligomerise in different ways again driving structural modifications in the active site or functional surfaces and the creation of new surfaces capable of evolving functional roles.

However, dramatic changes in functional class or in the chemistry performed by an enzyme, for example, appear to be rare (Todd et al., 2001; Bashton and Chothia, 2007). It's hard to engineer the geometry and exquisite stereospecificity needed to perform an enzyme reaction and perhaps not surprising that these analyses revealed a significant tendency for chemical intermediates to be conserved along the reaction pathways of different relatives in the superfamily. More frequently, evolutionary changes (particularly residue insertions) cause changes in the geometry of the active site and binding pocket enabling relatives to perform the same or similar chemistry on a different substrate (Todd et al., 2001).

These evolutionary changes, which can sometimes be quite subtle involving just a handful of residues, combined with the expansion of paralogs through gene duplication give an effective mechanism for expanding the functional repertoire of an organism. For example, the kinase superfamily has been significantly expanded in eukaryotes where relatives perform essential functions in cell-cell communication and intracellular signaling. Most paralogs are involved in phosphorylation of protein targets, but these targets can vary and relatives may be

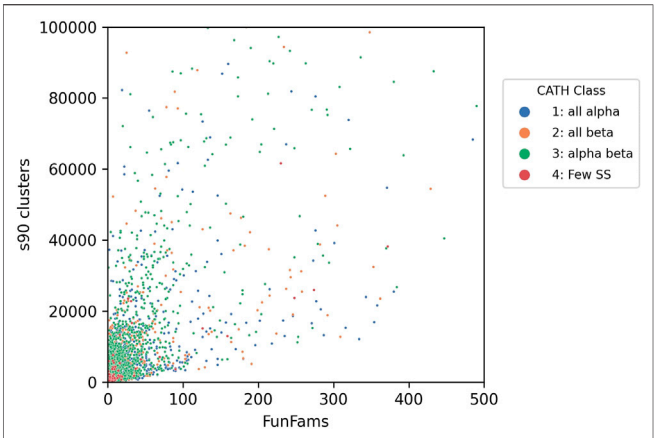


FIGURE 9 | Functional diversity (captured by number of functional families - FunFams) vs. sequence diversity (number of Gene3D s90 clusters i.e. in which relatives share 90% or more sequence identity) for CATH superfamilies. Each dot represents an individual superfamily.

expressed in different tissues having diverse interaction opportunities.

Bashton and Chothia (Bashton and Chothia, 2007) undertook a very detailed analysis of the extent to which key functional roles were conserved across domain superfamilies allowing domains to be used as "words" within a protein "functional sentence". This is a challenging task, and the challenges increasingly apparent as more experimentally characterised sequence relatives are classified within SCOP and CATH. In SCOP these predicted structural relatives are classified in the sister resource, SUPERFAMILY, managed by Julian Gough (Wilson et al., 2009). In CATH, sequences are directly integrated in superfamilies as well as being captured in the Gene3D sister resource (Lewis et al., 2018). Currently, the sequence data from UniProt expands the structural superfamilies 500-fold on average (up to 49 thousand-fold), depending on the superfamily allowing a deeper analysis of functional diversity. The correlation between sequence diversity and the number of sequence subfamilies and functional diversity can be seen for all types of superfamilies in **Figure 9**.

Chothia's analyses supported earlier hypotheses of conservation of function within a broad functional class (Bashton and Chothia, 2007). For example, the amino-acyl tRNA synthase superfamily is amongst the top 2% largest superfamiles and relatives perform multiple functions covering at least 31 EC3 categories (i.e. having different EC classifications at the third EC level associated with change in chemistry). Nevertheless, many relatives exploit the same co-factor pyridoxal 5-pyrophosphate binding to the same site and substrates tend to share a similar chemical moiety.

Another functionally diverse superfamily, the HUP superfamily, currently contains more than 640 thousand sequences from UniProt and 39,505 sequence subfamilies (at 50% sequence identity). This threshold is used because various studies have suggested 50% or 60% sequence identity for inferring functional similarity between homologues, provided there is reasonable overlap in sequence length (60% or more) (Rost, 2002; Rentzsch and Orengo, 2009). CATH identifies at least 55 EC terms and 594 diverse Gene Ontology (GO) terms for experimentally characterized relatives within this superfamily and like many other megafamilies less than 6% of relatives have experimental characterization. In addition, it's possible to characterize the structural diversity across this superfamily by clustering relatives according to structural similarity (e.g. < 5 A RMSD). There are currently 31 such structural clusters. Despite this structural and functional diversity the structural core is highly conserved (see **Figure 5** above), as also observed in other megafamilies. However, as seen in **Figure 4**, there can be considerable structural decorations or embellishments outside this core.

Phylogenetic Insights

The vast sequence data available for many species has allowed phylogenetic forays into protein superfamilies. For example, by combining both structural and sequence data as in CATH-Gene3D we can trace further back and explore the order of functional shifts within these superfamilies. The FunTree classification studies of Thornton and co-workers allowed tracing of shifts in enzyme chemistry (changes in the third number of the EC classification code) between homologues in all highly populated superfamilies in CATH (Furnham et al., 2012).

Similarly expansion in the structural data available for the superfamilies, thanks partly to targeted activities of the PSI structural genomics initiatives, provided insights into shifts in catalytic residues within enzyme superfamilies (Todd et al., 2005), confirming trends detected by early studies of Thornton and co-workers using much sparser data. As in the previous analysis interesting cases of convergence of catalytic machinery within superfamilies or "residue hopping" were detected (Todd et al., 2002). This was caused by divergence of functionally distinct homologues which then converged to the same chemistry via different mutational routes giving catalytic residues in different places in the active site pocket, but with the same chemical properties and necessary orientation to perform the chemistry.

FUNCTIONAL SUB-CLASSIFICATION OF PROTEIN REVEALS THE DARK MATTER OF FUNCTION SPACE

With <1% of protein sequences in UniProt having experimental characterisation, interest has grown in understanding the likely functional divergence across superfamilies, especially those with industrial value. Organising the sequence data to reveal highly conserved residues between putative functional relatives can give clues to possible changes in substrate specificity or enzyme chemistry. Because the structural data is so sparse, our approach to identifying functional families (FunFams) in CATH superfamilies has been to use sequence data and cluster relatives using an entropy-based method that segregates sets of relatives with differentially conserved residues (Das et al.,

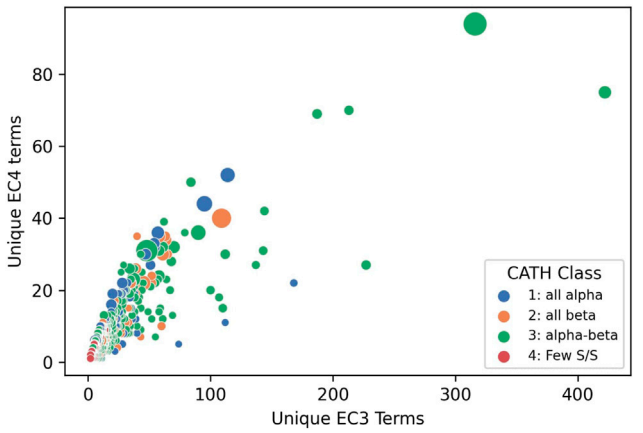


FIGURE 10 | Enzyme Commission terms distributions for each CATH v4.3 SuperFamilies, showing that 65 superfamilies have more than 20 different chemistries (i.e. EC3s).

2015b). Residues that are conserved across all relatives in a superfamily are likely to be important for folding or stability but residues that are conserved in different ways e.g. residues with different chemical properties, between different sets of relatives, are likely to be associated with the functional roles of the proteins. Some endorsement of this functional clustering is given by performance of CATH functional families in the independent CAFA Critical Assessment of Functional Annotations (Jiang et al., 2016; Zhou et al., 2019). Furthermore, residue sites conserved in FunFams are significantly enriched in known functional residues e.g. catalytic residues, protein interface residues, ligand binding residues etc (Das et al., 2015b).

Structural data, whether known or predicted, can then be exploited to determine where these putative functional determinants co-locate on the protein surface to glean further insights into functional properties. This clustering into functional families reveals the most promiscuous, highly diverse superfamilies. **Figure 10** shows that the top 65 most functionally diverse enzyme superfamilies have more than 20 different chemistries exhibited by relatives.

FunFams are only identified for sets of sequences where at least one relative has been experimentally characterized and has a GO functional annotation. On that basis, only about 36% of the 150 million domain sequences classified in CATH can be assigned to a functional family suggesting that there is still a large proportion of functional space to characterize. However, some superfamilies, particularly those containing important eukaryotic organisms (e.g. human, model organisms) tend to have a higher proportion of functional characterization. It's also important to remember that this is a domain based functional classification, but function is generally annotated at the protein level. However, analyses of selected superfamilies, namely the enolases, TPPs and HUPs suggest that by segregating on functional discriminants domain relatives occurring in different multidomain contexts are indeed clustered into separate functional groups (Das et al., 2015a).

FUNCTIONAL FAMILIES GIVE FINER INSIGHTS INTO THE EMERGENCE OF NOVEL FUNCTIONS IN METAZOA

As mentioned already above, globular domains are one of the key functional units of proteins, often with a specific functional role and with the ability to fold independently. Some globular domains have catalytic functions, facilitating enzymatic reactions, providing much of the complex chemistry that cells need to function. Other domains are responsible for detecting signals, by interacting with other protein domains and ligands, as part of signaling processes.

During the history of life on Earth there have been a number of major evolutionary events each requiring their own unique functional innovations. For example, early life-forms needed to establish much of the initial basic chemistry, energy production, metabolism etc. A number of domain superfamilies date back to the last universal common ancestor (LUCA) of cellular life, and these domain superfamilies provide much of the catalytic processes required by cells, such as the TIM-barrel domain superfamily, which provides the basic structural core for hundreds of different catalytic functions. Another major transition was the emergence of animals (Metazoans), which appeared several hundred million years ago, from single-celled ancestors (**Figure 11**). The emergence of Metazoans required many different functional innovations relating to cell communication, differentiation and migration. To support cellular complexity, coordinated regulation of gene expression was needed together with many other protein innovations such as the establishment of various signal transduction pathways that connect extracellular signals to transcriptional regulation.

As already mentioned, gene duplication and fusion can give proteins with novel domain combinations leading to new functions. For example, changing the multi-domain architecture, can give a novel protein that operates in a new cellular micro-environment. However, a change in multi-domain architecture is not a prerequisite for domain-based innovations and domains may gain novel functions with no change in domain partners. We can use a change in CATH FunFams, as a proxy for a change in domain function, allowing a preliminary exploration of various aspects of Metazoan evolution from a FunFam domain perspective.

For example, by examining the expansion in the number of FunFams within a domain superfamily we can track the expansion of functional diversity across that superfamily at different stages in evolution. By counting the number of FunFams for a given superfamily and clustering organisms using TreeFam (Ruan et al., 2007) we can show the FunFam expansions in CATH superfamilies at different evolutionary stages.

A large number (of domain superfamilies) show strong expansions (in the number of their FunFams) specifically at the emergence of Metazoans. Many of these expanded superfamilies are associated with signaling and regulatory processes, such as the SH2 domain family which undergoes significant expansion in Metazoans corresponding to its newly acquired role of phosphotyrosine binding domain in cell

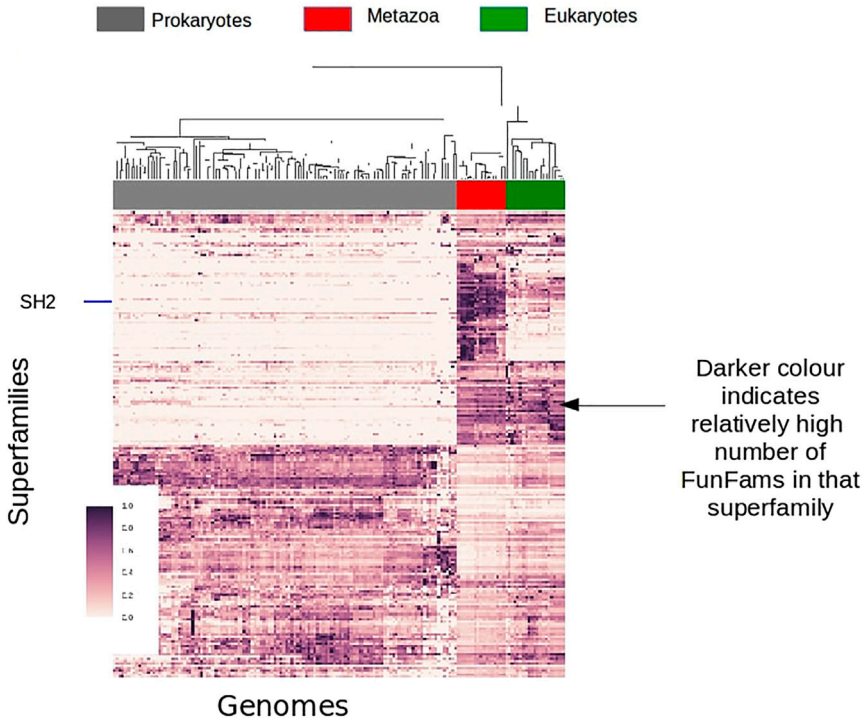


FIGURE 11 | Survey of FunFam expansions across the tree of life, darker colors show a higher number of FunFams in that superfamily.

signaling processes. Transcription factors are known to have a key role in Metazoan evolution/development. Many Transcription factor FunFams appear early in Metazoan evolution, prior to the separation of extant metazoan phyla but after the divergence of Choanoflagellates and Metazoans. There is also further lineage specific expansions in transcription factors, for example along the vertebrate lineage.

CONCLUSION

The pioneering work of Cyrus Chothia in characterising the relationship between sequence and structure and his subsequent analyses of specific families, namely the globins and immunoglobulins, together with structural and functional analyses by Janet Thornton amongst others, inspired algorithms and analytic protocols for detecting evolutionary relationships and the mechanisms by which genetic variations translate into structural and functional changes during evolution. These frameworks provided impetus for the establishment of comprehensive structural classifications which have been exploited in many analyses shedding light on divergence, particularly for enzyme superfamilies, but which also established general principles regarding functional shifts in all protein classes. To some extent the SCOP and CATH classifications have provided complementary perspectives as the former involved detailed manual curation and explicitly recognised domains found in diverse multi-domain contexts. In contrast, CATH aimed to exploit computational strategies that searched for globular domains and then classified them based on structural similarities in the core. Unlike many fields of science where competition often clouds judgment, Cyrus was a man of huge intellect and integrity who valued competition and the opportunities that diverse perspectives give in maximising the exploration and understanding of complex phenomena. He was one of the most supportive scientists in the Genome3D consortium which established formal collaborations between SCOP and CATH and which is currently enhancing the structural coverage of genome sequences in human, model organisms and Pfam families (Sillitoe et al., 2020). This collaboration is being continued in the new 3D-SCAfold initiative, being led by PDBe, which will ensure closer integration and disseminate the family data more widely to enable deeper studies of evolution.

AUTHOR CONTRIBUTIONS

NB, IS, JL generated and analyzed data. All authors contributed to the manuscript and figures. CO wrote the manuscript, with contributions and edits from NB, IS and JL.

FUNDING

IS and NB acknowledge funding from the Biotechnology and Biological Sciences Research Council (grant BB/R009597/1 to NB and BB/R014892/1 to I.S.)

REFERENCES

Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014). SCOP2 Prototype: a New Approach to Protein Structure Mining. *Nucl. Acids Res.* 42, D310–D314. doi:10.1093/nar/gkt1242

Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Gutmanas, A., Anyango, S., Choudhary, P., et al. (2019). PDBe: Improved Findability of Macromolecular Structure Data in the PDB. *Nucleic Acids Res.* 48, D335–D343. doi:10.1093/nar/gkz990

Bashton, M., and Chothia, C. (2007). The Generation of New Protein Functions by the Combination of Domains. *Structure*. 15, 85–99. doi:10.1016/j.str.2006.11.009

Björklund, Å. K., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A. (2005). Domain Rearrangements in Protein Evolution. *J. Mol. Biol.* 353, 911–923. doi:10.1016/j.jmb.2005.08.067

Chandonia, J.-M., and Brenner, S. E. (2006). The Impact of Structural Genomics: Expectations and Outcomes. *Science*. 311, 347–351. doi:10.1126/science.1121018

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., et al. (2014). ECOD: an Evolutionary Classification of Protein Domains. *Plos Comput. Biol.* 10, e1003926. doi:10.1371/journal.pcbi.1003926

Chothia, C., and Lesk, A. M. (1982). Evolution of Proteins Formed by β-sheets. *J. Mol. Biol.* 160, 309–323. doi:10.1016/0022-2836(82)90178-4

Chothia, C., and Lesk, A. M. (1986). The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* 5, 823–826. doi:10.1002/j.1460-2075.1986.tb04288.x

Chothia, C. (1992). One Thousand Families for the Molecular Biologist. *Nature*. 357, 543–544. doi:10.1038/357543a0

Das, S., Dawson, N. L., and Orengo, C. A. (2015a). Diversity in Protein Domain Superfamilies. *Curr. Opin. Genet. Develop.* 35, 40–49. doi:10.1016/j.gde.2015.09.005

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., and Orengo, C. A. (2015b). Functional Classification of CATH Superfamilies: a Domain-Based Approach for Protein Function Annotation. *Bioinformatics*. 31, 3460–3467. doi:10.1093/bioinformatics/btv398

Dessailly, B. H., Redfern, O. C., Cuff, A. L., and Orengo, C. A. (2010). Detailed Analysis of Function Divergence in a Large and Diverse Domain Superfamily: toward a Refined Protocol of Function Classification. *Structure* 18, 1522–1535. doi:10.1016/j.str.2010.08.017

Eddy, S. R. (1998). Profile Hidden Markov Models. *Bioinformatics* 14, 755–763. doi:10.1093/bioinformatics/14.9.755

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins-Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures. *Nucl. Acids Res.* 42, D304–D309. doi:10.1093/nar/gkt1240

Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Rahman, S. A., Laskowski, R. A., et al. (2012). FunTree: a Resource for Exploring the Functional Evolution of Structurally Defined Enzyme Superfamilies. *Nucleic Acids Res.* 40, D776–D782. doi:10.1093/nar/gkr852

Henikoff, S., and Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci.* 89, 10915–10919. doi:10.1073/pnas.89.22.10915

Holm, L., and Sander, C. (1993). Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* 233, 123–138. doi:10.1006/jmbi.1993.1489

Holm, L., and Sander, C. (1996). Mapping the Protein Universe. *Science*. 273, 595–602. doi:10.1126/science.273.5275.595

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., et al. (2016). An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy. *Genome Biol.* 17, 184. doi:10.1186/s13059-016-1037-6

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Bioinformatics*. 8, 275–282. doi:10.1093/bioinformatics/8.3.275

Krissinel, E. (2012). Enhanced Fold Recognition Using Efficient Short Fragment Clustering. *J. Mol. Biochem.* 1, 76–85.

Lesk, A. M., and Chothia, C. (1980). How Different Amino Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins. *J. Mol. Biol.* 136, 225–270. doi:10.1016/0022-2836(80)90373-3

Lesk, A. M., and Chothia, C. (1982). Evolution of Proteins Formed by β-sheets. *J. Mol. Biol.* 160, 325–342. doi:10.1016/0022-2836(82)90179-6

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: Sequencing Life for the Future of Life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333. doi:10.1073/pnas.1720115115

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., et al. (2018). Gene3D: Extensive Prediction of Globular Domains in Proteins. *Nucleic Acids Res.* 46, D435–D439. doi:10.1093/nar/gkx1069

Lupas, A. N., Ponting, C. P., and Russell, R. B. (2001). On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *J. Struct. Biol.* 134, 191–203. doi:10.1006/jsbi.2001.4393

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., et al. (2019). MGnify: the Microbiome Analysis Resource in 2020. *Nucleic Acids Res.* 48, D570–D578. doi:10.1093/nar/gkz1035

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* 247, 536–540. doi:10.1006/jmbi.1995.0159

Needleman, S. B., and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* 48, 443–453. doi:10.1016/0022-2836(70)90057-4

Norvell, J. C., and Berg, J. M. (2007). Update on the Protein Structure Initiative. *Structure* 15, 1519–1522. doi:10.1016/j.str.2007.11.004

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein Superfamilles and Domain Superfolds. *Nature* 372, 631–634. doi:10.1038/372631a0

Orengo, C. A., and Taylor, W. R. (1996). [36] SSAP: Sequential Structure Alignment Program for Protein Structure Comparison. *Methods in Enzymology* 266, 617–635. doi:10.1016/S0076-6879(96)66038-8

Ortiz, A. R., Strauss, C. E. M., and Olmea, O. (2002). MAMMOTH (Matching Molecular Models Obtained from Theory): an Automated Method for Model Comparison. *Protein Sci.* 11, 2606–2621. doi:10.1110/ps.0215902

Ranea, J. A. G., Sillero, A., Thornton, J. M., and Orengo, C. A. (2006). Protein Superfamily Evolution and the Last Universal Common Ancestor (LUCA). *J. Mol. Evol.* 63, 513–525. doi:10.1007/s00239-005-0289-7

Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M. G., and Orengo, C. A. (2007). CATHEDRAL: a Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures. *Plos Comput. Biol.* 3, e232. doi:10.1371/journal.pcbi.0030232

Rentzsch, R., and Orengo, C. A. (2009). Protein Function Prediction - the Power of Multiplicity. *Trends Biotechnol.* 27, 210–219. doi:10.1016/j.tibtech.2009.01.002

Rossmann, M. G., and Argos, P. (1976). Exploring Structural Homology of Proteins. *J. Mol. Biol.* 105, 75–95. doi:10.1016/0022-2836(76)90195-9

Rost, B. (2002). Enzyme Function Less Conserved Than Anticipated. *J. Mol. Biol.* 318, 595–608. doi:10.1016/S0022-2836(02)00016-5

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., et al. (2007). TreeFam: 2008 Update. *Nucleic Acids Res.* 36, D735–D740. doi:10.1093/nar/gkm1005

Russell, R. B., and Barton, G. J. (1992). Multiple Protein Sequence Alignment from Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels. *Proteins*. 14, 309–323. doi:10.1002/prot.340140216

Shindyalov, I. N., and Bourne, P. E. (1998). Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. *Protein Eng. Des. Selection*. 11, 739–747. doi:10.1093/protein/11.9.739

Sillitoe, I., Andreeva, A., Blundell, T. L., Buchan, D. W. A., Finn, R. D., Gough, J., et al. (2020). Genome3D: Integrating a Collaborative Data Pipeline to Expand the Depth and Breadth of Consensus Protein Structure Annotation. *Nucleic Acids Res.* 48, D314–D319. doi:10.1093/nar/gkz967

Smith, T. F., and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147, 195–197. doi:10.1016/0022-2836(81)90087-5

Söding, J. (2005). Protein Homology Detection by HMM-HMM Comparison. *Bioinformatics*. 21, 951–960. doi:10.1093/bioinformatics/bti125

Subbiah, S., Laurents, D. V., and Levitt, M. (1993). Structural Similarity of DNA-Binding Domains of Bacteriophage Repressors and the Globin Core. *Curr. Biol.* 3, 141–148. doi:10.1016/0960-9822(93)90255-M

Swindells, M. B., Orengo, C. A., Jones, D. T., Hutchinson, E. G., and Thornton, J. M. (1998). Contemporary Approaches to Protein Structure Classification. *Bioessays*. 20, 884–891. doi:10.1002/(sici)1521-1878(199811)20:11<884::aid-bies3>3.0.co;2-h

Teichmann, S. A., Chothia, C., and Gerstein, M. (1999). Advances in Structural Genomics. *Curr. Opin. Struct. Biol.* 9, 390–399. doi:10.1016/S0959-440X(99)80053-0

Teichmann, S., Rison, S. C., Thornton, J. M., Riley, M., Gough, J., and Chothia, C. (2001). Small-molecule Metabolism: an Enzyme Mosaic. *Trends Biotechnol.* 19, 482–486. doi:10.1016/s0167-7799(01)01813-3

The UniProt Consortium (2019). UniProt: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Todd, A. E., Marsden, R. L., Thornton, J. M., and Orengo, C. A. (2005). Progress of Structural Genomics Initiatives: an Analysis of Solved Target Structures. *J. Mol. Biol.* 348, 1235–1260. doi:10.1016/j.jmb.2005.03.037

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of Function in Protein Superfamilies, from a Structural Perspective. *J. Mol. Biol.* 307, 1113–1143. doi:10.1006/jmbi.2001.4513

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002). Plasticity of Enzyme Active Sites. *Trends Biochem. Sci.* 27, 419–426. doi:10.1016/s0968-0004(02)02158-8

Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S. A. (2004). Supra-domains: Evolutionary Units Larger Than Single Protein Domains. *J. Mol. Biol.* 336, 809–823. doi:10.1016/j.jmb.2003.12.026

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., et al. (2009). SUPERFAMILY-sophisticated Comparative Genomics, Data Mining, Visualization and Phylogeny. *Nucleic Acids Res.* 37, D380–D386. doi:10.1093/nar/gkn762

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688. doi:10.1093/nar/gkz966

Ye, Y., and Godzik, A. (2003). Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists. *Bioinformatics*. 19, ii246–ii255. doi:10.1093/bioinformatics/btg1086

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., et al. (2019). The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes through Experimental Screens. *Genome Biol.* 20, 244. doi:10.1186/s13059-019-1835-8

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

frontiers in Molecular Biosciences

MINI REVIEW
published: 11 May 2021
doi: 10.3389/fmolb.2021.643752

Recent Advances in Protein Homology Detection Propelled by Inter-Residue Interaction Map Threading

Sutanu Bhattacharya[1], *Rahmatullah Roche*[1], *Md Hossain Shuvo*[1] *and Debswapna Bhattacharya*[1,2]*

[1]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, United States, [2]Department of Biological Sciences, Auburn University, Auburn, AL, United States

Sequence-based protein homology detection has emerged as one of the most sensitive and accurate approaches to protein structure prediction. Despite the success, homology detection remains very challenging for weakly homologous proteins with divergent evolutionary profile. Very recently, deep neural network architectures have shown promising progress in mining the coevolutionary signal encoded in multiple sequence alignments, leading to reasonably accurate estimation of inter-residue interaction maps, which serve as a rich source of additional information for improved homology detection. Here, we summarize the latest developments in protein homology detection driven by inter-residue interaction map threading. We highlight the emerging trends in distant-homology protein threading through the alignment of predicted interaction maps at various granularities ranging from binary contact maps to finer-grained distance and orientation maps as well as their combination. We also discuss some of the current limitations and possible future avenues to further enhance the sensitivity of protein homology detection.

Keywords: protein homology, inter-residue interaction map, protein threading, homology modeling, protein structure prediction

OPEN ACCESS

Edited by:
Paolo Marcatili, Technical University of Denmark, Denmark

Reviewed by:
Dimitrios P. Vlachakis, Agricultural University of Athens, Greece
Kresten Lindorff-Larsen, University of Copenhagen, Denmark

***Correspondence:**
Debswapna Bhattacharya
bhattacharyad@auburn.edu

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *18 December 2020*
Accepted: *21 April 2021*
Published: *11 May 2021*

Citation:
Bhattacharya S, Roche R, Shuvo MH and Bhattacharya D (2021) Recent Advances in Protein Homology Detection Propelled by Inter-Residue Interaction Map Threading. Front. Mol. Biosci. 8:643752. doi: 10.3389/fmolb.2021.643752

INTRODUCTION

The development of computational approaches for accurately predicting the protein three-dimensional (3D) structure directly from the sequence information is of central importance in structural biology (Jones et al., 1992; Baker and Sali, 2001; Dill and MacCallum, 2012). While *ab initio* modeling aims to predict the 3D structure purely from the sequence information (Marks et al., 2011; Adhikari et al., 2015; Wang et al., 2016; Adhikari and Cheng, 2018; Greener et al., 2019; Senior et al., 2019; Xu, 2019; Yang et al., 2020; Roche et al., 2021), many protein targets have evolutionary-related (homologous) structures, also known as homologous templates, already available in the Protein Data Bank (PDB) (Berman et al., 2000). Correctly identifying these templates given the sequence of a query protein and building 3D models by performing query–template alignment, a technique broadly known as homology modeling (Altschul et al., 1997; Xu et al., 2003; Wu and Zhang, 2008; Lobley et al., 2009; Wu and Zhang, 2010; Källberg et al., 2012; Ma et al., 2014) often results in highly accurate predicted structural models (Abeln et al., 2017). As such, the success of homology modeling critically depends on the ability to identify the closely homologous template on the basis of sequence similarity and generate accurate query–template alignment. Intuitively, the performance of these methods sharply deteriorates when the direct evolutionary relationship between the query and templates becomes very low, typically when the sequence similarity falls

below 30%, the so-called distant-homology modeling scenarios (Bowie et al., 1991; Petrey and Honig, 2005). Protein threading, the most widely used distant-homology modeling technique, aims to address the challenge by leveraging multiple sources of information by mining the evolutionary profile of the query and templates to reveal potential distant homology and perform distant-homology modeling to predict the 3D structure of the query protein.

Existing threading methods exploit a wide range of techniques ranging from dynamic programming to profile-based comparison to machine learning (Jones, 1999; Rychlewski et al., 2000; Xu and Xu, 2000; Skolnick and Kihara, 2001; Ginalski et al., 2003; Marti et al., 2004; Jaroszewski et al., 2005; Söding, 2005; Zhou and Zhou, 2005; Cheng and Baldi, 2006; Peng and Xu, 2009; Lee and Skolnick, 2010; Peng and Xu, 2010; Yang et al., 2011; Ma et al., 2012; Ma et al., 2013; Gniewek et al., 2014). The recent advancement in predicting the inter-residue interaction maps using sequence coevolution and deep learning (Morcos et al., 2011; He et al., 2017; Wang et al., 2017; Adhikari et al., 2018; Hanson et al., 2018; Kandathil et al., 2019; Yang et al., 2020) has opened new possibilities to further improve the sensitivity of distant-homology protein threading by incorporating the predicted inter-residue interaction information. Fueled by this, several efforts have been made in the recent past to integrate interaction maps into threading. For instance, EigenTHREADER (Buchan and Jones, 2017), map_align (Ovchinnikov et al., 2017), CEthreader (Zheng et al., 2019a), CATHER (Du et al., 2020), and ThreaderAI (Zhang and Shen, 2020) have utilized predicted contact maps in protein threading. DeepThreader (Zhu et al., 2018) has exploited finer-grained distance maps for query proteins instead of using binary contacts to improve threading template selection and alignment. DisCovER (Bhattacharya et al., 2020) goes one step further by incorporating inter-residue orientation along with distance information together with topological network neighborhood (Chen et al., 2019) of query–template alignment to further improve threading performance. Here, we provide an overview of the latest advances in protein homology detection propelled by inter-residue interaction map threading.

GRANULARITIES OF PROTEIN INTER-RESIDUE INTERACTION MAPS

Protein inter-residue interaction maps are predicted at various resolutions ranging from binary contact maps to finer-grained distance and orientation maps as well as their combination. A low-resolution version of inter-residue interaction is a contact map, which is a square, symmetric matrix with binary entries, where a contact indicates the spatial proximity of a residue pair at a given cutoff distance, typically set to 8Å between the C_α or C_β carbons of the interacting residue pairs. Inter-residue distance map is finer-grained in that it captures the distribution of real-valued inter-residue spatial proximity information rather than the binary contacts at a fixed cutoff distance. Recent studies (Xu and Wang, 2019; Xu, 2019) have demonstrated the advantage of using distance maps in protein structure prediction over binary contacts as distances carry more physical constraint information of protein structures than contacts. The granularities of predicted distance maps vary from distance histograms to real-valued distances (Greener et al., 2019; Adhikari, 2020; Ding and Gong, 2020; Li and Xu, 2020; Wu et al., 2021; Yang et al., 2020). Very recently, trRosetta (Yang et al., 2020) has introduced inter-residue orientations in addition to distances to capture not only the spatial proximity information of the interacting pairs but also their relative angles and dihedrals. Collectively, inter-residue distances and orientations encapsulate the spatial positioning of the interacting pairs much better than only distances let alone binary contacts.

INTER-RESIDUE INTERACTION MAP THREADING

Figure 1 shows an overview of an interaction map threading of a query protein. Generally, threading has four components: (1) an effective scoring function to evaluate the fitness of query–template alignment; (2) efficient template searching or homology detection strategy; (3) optimal query–template alignments; and (4) building 3D models of query proteins based on alignments. One of the most important components of threading approaches is the scoring function, which is composed of standard threading features ranging from sequential features such as secondary structures, solvent accessibility, and sequence profiles to nonlinear features such as pairwise potentials (Bienkowska and Lathrop, 2005; Brylinski and Skolnick, 2010). Weights control the relative importance of different terms. An efficient scoring function should reliably differentiate a homologous template from the alternatives because the accuracy of the predicted model significantly depends on the evolutionary relatedness of the identified template. The inter-residue interaction map helps to improve the sensitivity of the threading scoring function by augmenting the standard scoring terms with additional contributions from the predicted interactions. Specifically, the score to align the i th residue of the query protein to the j th residue of the template can be defined as:

$$E(i,j) = w_1 E_{map}^{interaction}(i,j) + \sum_{\substack{k \in standard \\ threading\ features}} w_k E_k^{feature}(i,j)$$

where the first term accounts for the contribution of the interaction map and the second term accounts for the standard threading features with w_i being their relative weights. Typically, the similarity between the predicted inter-residue interaction map of the query protein and that derived from the template structure informs the interaction map term in the threading scoring function. It is worth noting here that the raw alignment score is biased to protein length (Xu et al., 2003). As such, most threading methods use a normalized alignment score in standard deviation units relative to the mean score of all

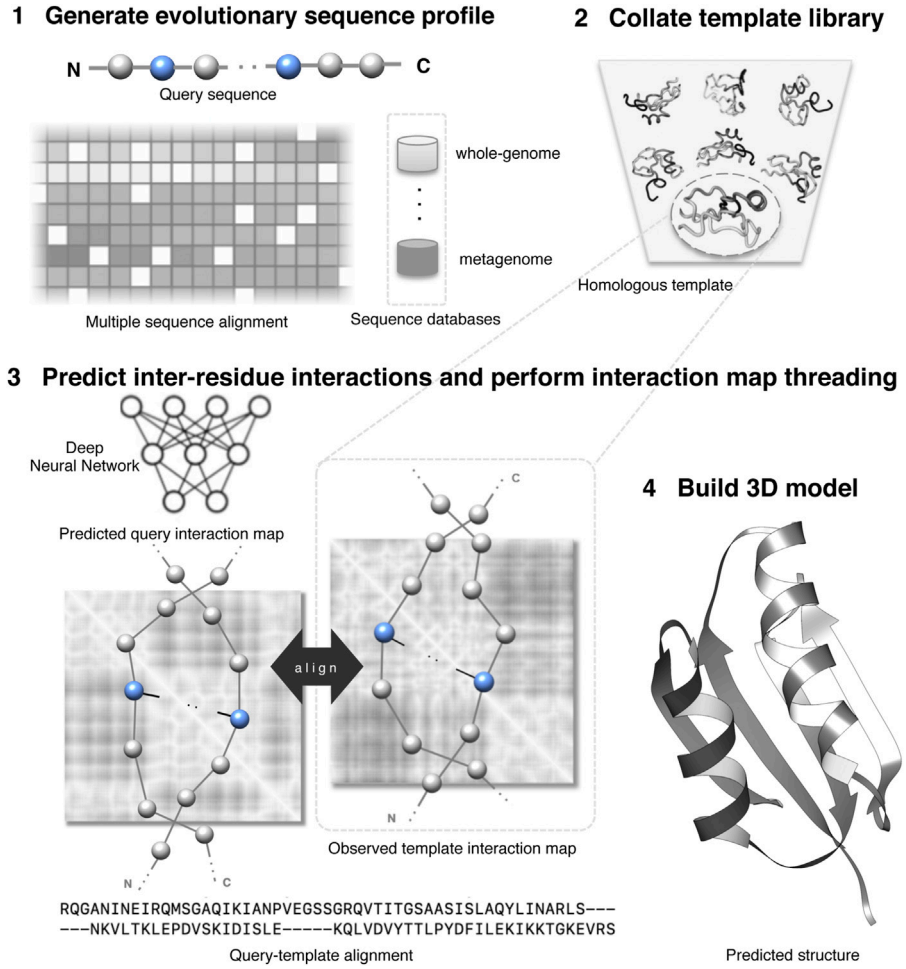


FIGURE 1 | Illustration of protein interaction map threading.

templates in the template library for homology detection—detecting best-fit templates from the PDB.

EMERGING TRENDS IN PROTEIN HOMOLOGY DETECTION BY INTERACTION MAP THREADING

With the recent advancement in contact prediction mediated by sequence coevolution and deep learning, significant research efforts have been made in the recent past to incorporate contact information as an additional scoring term into the threading scoring function for protein homology detection. For instance, Jones and coworkers developed EigenTHREADER (Buchan and Jones, 2017) that uses eigen-decomposition (Di Lena et al., 2010) of contact maps predicted using classical neural network–based predictor MetaPSICOV (Jones et al., 2015) to search a library of template contact maps for contact map threading. Baker and coworkers developed map_align (Ovchinnikov et al., 2017) that employs an iterative double dynamic programming framework (Taylor, 1999) for homology detection. map_align takes advantage of metagenomics sequence databases of microbial DNA (Söding, 2017) and uses contact maps predicted by coevolutionary contact predictor GREMLIN (Balakrishnan et al., 2011; Kamisetty et al., 2013) to perform contact map threading by maximizing the number of overlapping contacts and minimizing the number of gaps. Recently, Zhang and coworkers developed CEthreader (Zheng et al., 2019a) using contact maps predicted by deep learning–based contact map predictor ResPRE (Li et al., 2019). CEthreader also relies on eigen-decomposition and performs contact map threading through dynamic programming using a dot-product scoring function by integrating contacts as well as secondary structures and sequence profiles. Alongside, we developed a contact-assisted threading method (Bhattacharya and Bhattacharya, 2019) that incorporates contact information, predicted by deep learning–based predictor RaptorX (Wang et al., 2017), into threading using a two-stage approach. After selecting a subset of top templates from the template library using a standard profile-based threading technique in the first stage,

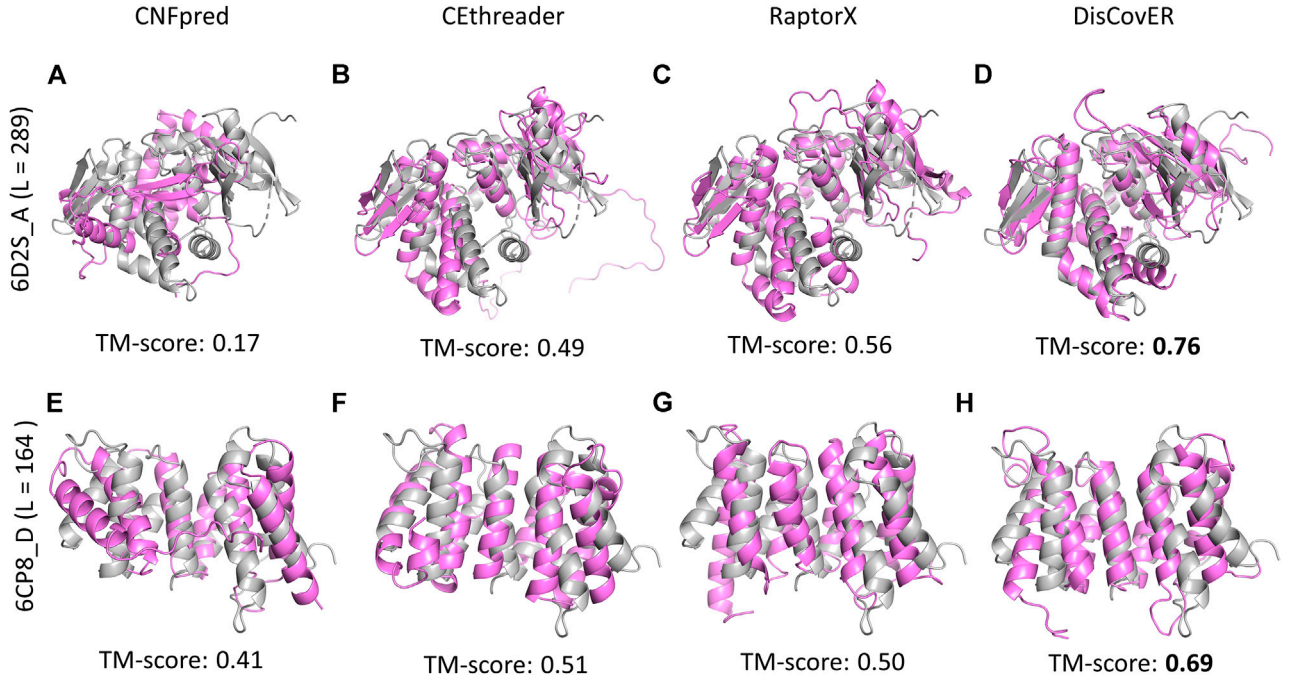


FIGURE 2 | Structural superposition between predicted models using various threading methods (in violet) and the corresponding experimental structures (in gray) for representative CAMEO targets 6D2S_A of length 289 residues and 6CP8_D of length 164 residues.

our method subsequently uses eigen-decomposition of the contact information along with the profile-based alignment score to select the best-fit template. We further analyze the impact of contact map quality on threading performance (Bhattacharya and Bhattacharya, 2020), which reveals that incorporating high-quality contact maps having the Matthews correlation coefficient (MCC) ≥ 0.5 improves the threading performance for ~ 30% cases in comparison to a baseline contact-free threading used as a control, while incorporating low-quality contacts with MCC <0.35 deteriorates the performance for 50% cases. Yang and coworkers developed CATHER (Du et al., 2020) by incorporating contact maps predicted by deep learning–based predictor MapPred (Wu et al., 2020) along with standard sequential information in the threading scoring function. Very recently, Shen and coworkers have developed ThreaderAI (Zhang and Shen, 2020) that implements a neural network for predicting alignments by incorporating deep learning–based contact information with conventional sequential and structural features into the scoring function.

Building on the successes of contact-assisted threading methods, Xu and coworkers developed a distance-based threading method called DeepThreader (Zhu et al., 2018). The method predicts distance maps by employing deep learning and then incorporates the predicted inter-residue distance information along with sequential features into threading through alternating direction method of multipliers (ADMM) algorithm. The inter-residue distance is binned into 12 bins: <5Å, 5–6Å, .., 14–15Å, and >15Å. Based on their reported results as well as the performance evaluation in the 13th Critical Assessment of protein Structure Prediction (CASP13), incorporating distance information boosts threading performance, particularly for distant-homology targets, outperforming contact-assisted threading methods by a large margin (Xu and Wang, 2019, 13). Zhang and coworkers have recently extended CEthreader to develop a distance-assisted threading method DEthreader introduced during the recently concluded CASP14 experiment by incorporating a distance-based scoring term into the scoring function. The method uses the C_α–C_α and C_β–C_β distance distribution, both are binned into 38 bins: 1 bin of <2Å, 36 bins of 2–20Å with a width of 0.5Å, and 1 bin of ≥20Å. Similarly, Yang and coworkers have extended CATHER into a distance-based threading approach by replacing contacts with distances in CASP14.

Powered by the development of the recent deep learning–based trRosetta method (Yang et al., 2020) for the prediction of inter-residue orientations and distances, our recent method DisCovER (Bhattacharya et al., 2020) goes one step further by incorporating predicted inter-residue orientations in addition to distances together with the neighborhood effect of the query–template alignment using an iterative double dynamic programming framework. The predicted distances are binned into 9 bins with a bin size of 1Å: <6Å to <14Å by summing up the likelihoods for distance bins below a distance threshold. The two orientation dihedrals (ω, θ) are binned into 24 bins with a width of 15°, and the orientation angle (ϕ) is binned into 12 bins with a width of 15°. Experimental results demonstrate the improved threading performance of DisCovER over the other

state-of-the-art threading approaches on multiple benchmark datasets across various target categories, especially for distantly homologous proteins. Representative examples on CAMEO targets 6D2S_A and 6CP8_D provide some insights into the origin of the improved performance. **Figure 2** shows our recent method DisCovER predicts correct folds (TM-score > 0.5) for both the targets 6D2S_A and 6CP8_D with a TM-score of 0.76 and 0.69, respectively, significantly better than the others. While the pure profile-based threading method CNFpred (Ma et al., 2012; Ma et al., 2013) and the recent contact-assisted threading method CEthreader fail to predict the correct fold for the target 6D2S_A, DisCovER and the CAMEO server RaptorX (Källberg et al., 2012; Zhu et al., 2018), employing the distance-based threading method DeepThreader (Haas et al., 2019), effectively predict the correct fold, with noticeably better performance by DisCovER (an improvement of 0.2 TM-score points) than the next best RaptorX. We also notice the superior performance of DisCovER for the target 6CP8_D where DisCovER significantly outperforms the other competing methods including the next best CEthreader by 0.18 TM-score points. It is worth mentioning both the targets are officially classified as "hard" by CAMEO (Haas et al., 2019), which warrants a distantly homologous nature in which current threading methods have limitations. Overall, the results show that the integration of the orientation information and the neighborhood effect in DisCovER results in improved threading, attaining state-of-the-art performance in (distant) homology detection.

THE ROLE OF SEQUENCE DATABASES IN INTERACTION MAP THREADING

The prediction of inter-residue interaction maps depends heavily on the availability of homologous sequences. As such, the role of the sequence databases is becoming increasingly important in protein homology detection via interaction map threading. In addition to the well-established whole-genome sequence databases such as the nr database from the National Center for Biotechnology Information (NCBI), UniRef (Suzek et al., 2015), UniProt (The UniProt Consortium, 2019), and Uniclust (Mirdita et al., 2017); emerging metagenome sequence databases from the European Bioinformatics Institute (EBI) Metagenomics (Markowitz et al., 2014; Mitchell et al., 2018) and Metaclust (Steinegger and Söding, 2018) are playing a prominent role. For example, Wang et al. (2019) have demonstrated the applications of marine metagenomics for improved protein structure prediction. map_align uses the Integrated Microbial Genomes (IMG) database (Markowitz et al., 2014), containing around 4 million unique protein sequences, to reliably predict high-quality models for distant-homology Pfam families of unknown structures. Another recent method for generating protein multiple sequence alignments, DeepMSA (Zhang et al., 2020), combines whole-genome and metagenome sequence databases and reports improved threading performance, particularly for distant-homology proteins. Newer sequence databases are getting larger and diverse. For example, BFD (Steinegger et al., 2019), a recent sequence database, is one of the largest sequence databases containing 2 billion protein sequences from soil samples and 292 million sequences of marine samples. Another very recent sequence database MGnify (Mitchell et al., 2020) contains around 1 billion nonredundant protein sequences. As such, the availability of evolutionary information of distant-homology proteins is getting enriched, likely leading to improved prediction accuracy of inter-residue interaction maps and hence more accurate interaction map threading for distant-homology protein modeling.

DISCUSSION

While the use of interaction maps is the main driving force behind the improved threading performance, the optimal granularity and information content of the predicted interaction maps remain elusive. Existing works consider various distance bins (Zhu et al., 2018; Bhattacharya et al., 2020) and subsets of predicted interactions either based on top predicted pairs sorted based on their likelihood values or using arbitrary likelihood cutoffs (Bhattacharya and Bhattacharya, 2019; Zheng et al., 2019a). A robust mechanism for defining and selecting interacting residue pairs can be beneficial to existing threading methods. Another challenge is how to integrate heterogeneous sources of available information from multiple interaction map predictors and/or sequence databases in a singular framework for unified interaction map threading. Finally, the use of multiple templates (Cheng, 2008; Peng and Xu, 2011; Meier and Söding, 2015) and meta-approaches (Wu and Zhang, 2007; Zheng et al., 2019b) possibly coupled with model quality assessment methods (Ray et al., 2012; Uziela et al., 2016; Uziela et al., 2017; 3; Alapati and Bhattacharya, 2018; Karasikov et al., 2019; Baldassarre et al., 2020; Eismann et al., 2020; Shuvo et al., 2020) and potentially aided by structure refinement (Bhattacharya and Cheng, 2013a; Bhattacharya and Cheng, 2013b; Bhattacharya and Cheng, 2013c; Bhattacharya et al., 2016; Bhattacharya, 2019; Wang et al., 2020; Heo and Feig, 2020) can collectively improve the accuracy of distant-homology protein modeling.

Recent CASP experiments have witnessed dramatic recent advances by DeepMind's AlphaFold series (Senior et al., 2019; Senior et al., 2020) in *ab initio* protein structure prediction, significantly outperforming the other groups. The success of AlphaFold series is primarily attributed to the successful application of deep neural networks for accurately predicting inter-residue spatial proximity information coupled with end-to-end training, significantly improving the accuracy of protein structure prediction (Pearce and Zhang, 2021). The integration of deep learning into various stages of protein modeling represents an exciting future direction that shall have a transformative impact on distant-homology protein modeling via interaction map threading, complementing and supplementing *ab initio* protein structure prediction methods developed by DeepMind.

AUTHOR CONTRIBUTIONS

All authors contributed in writing and revising the manuscript under the supervision of DB.

FUNDING

This work was partially supported by the National Science Foundation CAREER Award DBI-1942692 to DB, the National Science Foundation grant IIS-2030722 to DB, and the National Institute of General Medical Sciences Maximizing Investigators' Research Award (MIRA) R35GM138146 to DB.

ACKNOWLEDGMENTS

This work was made possible in part by Auburn University Early Career Development grant to DB.

REFERENCES

Abeln, S., Heringa, J., and Anton Feenstra, K. (2017). *Introduction to protein structure prediction. arXiv [arXiv:1712.00407]*. Available at: https://arxiv.org/abs/1712.00407v1.

Adhikari, B. (2020). A Fully Open-Source Framework for Deep Learning Protein Real-Valued Distances. *Scientific Rep.* 10 (1), 13374. doi:10.1038/s41598-020-70181-0

Adhikari, B., and Cheng, J. (2018). CONFOLD2: Improved Contact-Driven Ab Initio Protein Structure Modeling. *BMC Bioinformatics* 19 (1), 22. doi:10.1186/s12859-018-2032-6

Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). CONFOLD: Residue-Residue Contact-Guidedab Initioprotein Folding. *Proteins* 83 (8), 1436–1449. doi:10.1002/prot.24829

Adhikari, B., Hou, J., and Cheng, J. (2018). DNCON2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *Bioinformatics* 34 (9), 1466–1472. doi:10.1093/bioinformatics/btx781

Alapati, R., and Bhattacharya, D. (2018). "ClustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons," In Proceedings Of the 2018 ACM International Conference On Bioinformatics, Computational Biology, and Health Informatics. New York, NY, USA: Association for Computing Machinery. BCB '18. doi:10.1145/3233547.3233570

Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Baker, D., and Sali, A. (2001). Protein Structure Prediction and Structural Genomics. *Science* 294 (5540), 93–96. doi:10.1126/science.1065659

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. (2011). Carbonell, Su-In Lee, and Christopher James LangmeadLearning Generative Models for Protein Fold Families. *Proteins* 79 (4), 1061–1078. doi:10.1002/prot.22934

Baldassarre, F., Hurtado, D. M., Elofsson, A., and Azizpour, H. (2020). GraphQA: Protein Model Quality Assessment Using Graph Convolutional Networks. *Bioinformatics*, 37 (3), 360-366. doi:10.1093/bioinformatics/btaa714

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Bhattacharya, D., and Cheng, J. (2013a). 3Drefine: Consistent Protein Structure Refinement by Optimizing Hydrogen Bonding Network and Atomic-Level Energy Minimization. *Proteins* 81 (1), 119–131. doi:10.1002/prot.24167

Bhattacharya, D., and Cheng, J. (2013b). I3Drefine Software for Protein 3D Structure Refinement and its Assessment in CASP10. *PLOS ONE* 8 (7), e69648. doi:10.1371/journal.pone.0069648

Bhattacharya, D., and Cheng, J. (2013c). "Protein Structure Refinement by Iterative Fragment Exchange," in Proceedings Of the International Conference On Bioinformatics, Computational Biology And Biomedical Informatics. New York, NY, USA: BCB'13 Association for Computing Machinery.

Bhattacharya, D., Nowotny, J., Cao, R., and Cheng, J. (2016). 3Drefine: An Interactive Web Server for Efficient Protein Structure Refinement. *Nucleic Acids Res.* 44 (W1), W406–W409. doi:10.1093/nar/gkw336

Bhattacharya, D. (2019). RefineD: Improved Protein Structure Refinement Using Machine Learning Based Restrained Relaxation. *Bioinformatics* 35 (18), 3320–3328. doi:10.1093/bioinformatics/btz101

Bhattacharya, S., and Bhattacharya, D. (2019). Does Inclusion of Residue-residue Contact Information Boost Protein Threading? *Proteins* 87 (7), 596–606. doi:10.1002/prot.25684

Bhattacharya, S., and Bhattacharya, D. (2020). Evaluating the Significance of Contact Maps in Low-Homology Protein Modeling Using Contact-Assisted Threading. *Scientific Rep.* 10 (1), 2908. 10.1038/s41598-020-59834-2

Bhattacharya, S., Roche, R., and Bhattacharya, D. (2020). DisCovER: Distance- and Orientation-Based Covariational Threading for Weakly Homologous Proteins. *BioRxiv*. doi:10.1101/2020.01.31.923409

Bienkowska, J., and Lathrop, R. (2005). "Threading Algorithms," in *Encyclopedia Of Genetics, Genomics, Proteomics and Bioinformatics*. Editors L. B. Jorde, P. F. R. Little, M. J. Dunn, and S. Subramaniam (American Cancer Society). doi:10.1002/047001153X.g409202

Bowie, J., Luthy, R., and Eisenberg, D. (1991). A Method to Identify Protein Sequences that Fold into a Known Three-Dimensional Structure. *Science* 253 (5016), 164–170. doi:10.1126/science.1853201

Brylinski, M., and Skolnick, J. (2010). Comparison of Structure-Based and Threading-Based Approaches to Protein Functional Annotation. *Proteins* 78 (1), 18–134. doi:10.1002/prot.22566

Buchan, D. W. A., and Jones, D. T. (2017). EigenTHREADER: Analogous Protein Fold Recognition by Efficient Contact Map Threading. *Bioinformatics* 33 (17), 2684–2690. doi:10.1093/bioinformatics/btx217

Chen, C.-C., Jeong, H., Qian, X., and Yoon, B.-J. (2019). TOPAS: Network-Based Structural Alignment of RNA Sequences. *Bioinformatics* 35 (17), 2941–2948. doi:10.1093/bioinformatics/btz001

Cheng, J. (2008). A Multi-Template Combination Algorithm for Protein Comparative Modeling. *BMC Struct. Biol.* 8 (1), 18. doi:10.1186/1472-6807-8-18

Cheng, J., and Baldi, P. (2006). A Machine Learning Information Retrieval Approach to Protein Fold Recognition. *Bioinformatics* 22 (12), 1456–1463. doi:10.1093/bioinformatics/btl102

Di Lena, P., Fariselli, P., Margara, L., Vassura, M., and Casadio, R. (2010). Fast Overlapping of Protein Contact Maps by Alignment of Eigenvectors. *Bioinformatics* 26 (18), 2250–2258. doi:10.1093/bioinformatics/btq402

Dill, K. A., and MacCallum, J. L. (2012). The Protein-Folding Problem, 50 Years on. *Science* 338 (6110), 1042–1046. doi:10.1126/science.1219021

Ding, W., and Gong, H. (2020). Predicting the Real-Valued Inter-Residue Distances for Proteins. *Adv. Sci.* 7 (19), 2001314. doi:10.1002/advs.202001314

Du, Z., Pan, S., Wu, Q., Peng, Z., and Yang, J. (2020). CATHER: A Novel Threading Algorithm with Predicted Contacts. *Bioinformatics* 36 (7), 2119–2125. doi:10.1093/bioinformatics/btz876

Eismann, S., Suriana, P., Jing, B., Raphael, J., Townshend, L., and Dror, Ron. O. (2020). Protein Model Quality Assessment Using Rotation-Equivariant, Hierarchical Neural Networks [arXiv. 2011.13557]. http://arxiv.org/abs/2011.13557.

Ginalski, K., Pas, Jakub., Wyrwicz, L. S., von Grotthuss, M., Bujnicki, J. M., and Rychlewski, L. (2003). ORFeus: Detection of Distant Homology Using Sequence Profiles and Predicted Secondary Structure. *Nucleic Acids Res.* 31 (13), 3804–3807. doi:10.1093/nar/gkg504

Gniewek, P., Kolinski, A., Kloczkowski, A., and Gront, D. (2014). BioShell-Threading: Versatile Monte Carlo Package for Protein 3D Threading. *BMC Bioinformatics* 15 (1), 22. doi:10.1186/1471-2105-15-22

Greener, J. G., Kandathil, S. M., and Jones, David. T. (2019). Deep Learning Extends De Novo Protein Modelling Coverage of Genomes Using Iteratively Predicted Structural Constraints. *Nat. Commun.* 10 (1), 1–13. doi:10.1038/s41467-019-11994-0

Haas, J., Gumienny, R., Barbato, A., Ackermann, F., Tauriello, G., Bertoni, M., et al. (2019). Introducing "best Single Template" Models as Reference Baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins* 87 (12), 1378–1387. doi:10.1002/prot.25815

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2018). Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* 34 (23), 4039–4045. doi:10.1093/bioinformatics/bty481

He, B., Mortuza, S. M., Wang, Y., Shen, H.-B., and Zhang, Y. (2017). NeBcon: Protein Contact Map Prediction Using Neural Network Training Coupled with Naïve Bayes Classifiers. *Bioinformatics* 33 (15), 2296–2306. doi:10.1093/bioinformatics/btx164

Heo, L., and Feig, M. (2020). High-accuracy Protein Structures by Combining Machine-learning with Physics-based Refinement. *Proteins* 88 (5), 637–642. doi:10.1002/prot.25847

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., and Godzik, A. (2005). FFAS03: a Server for Profile-Profile Sequence Alignments. *Nucleic Acids Res.* 33 (Suppl. l_2), W284–W288. doi:10.1093/nar/gki418

Jones, D. T. (1999). GenTHREADER: an Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences. *J. Mol. Biol.* 287 (4), 797–815. doi:10.1006/jmbi.1999.2583

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: Combining Coevolution Methods for Accurate Prediction of Contacts and Long Range Hydrogen Bonding in Proteins. *Bioinformatics* 31 (7), 999–1006. doi:10.1093/bioinformatics/btu791

Jones, D. T., Taylort, W. R., and Thornton, J. M. (1992). A New Approach to Protein Fold Recognition. *Nature* 358 (6381), 86–89. doi:10.1038/358086a0

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-Based Protein Structure Modeling Using the RaptorX Web Server. *Nat. Protoc.* 7 (8), 1511–1522. doi:10.1038/nprot.2012.085

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci.* 110 (39), 15674–15679. doi:10.1073/pnas.1314045110

Kandathil, S. M., Greener, J. G., and Jones, D. T. (2019). Prediction of Interresidue Contacts with DeepMetaPSICOV in CASP13. *Proteins* 87 (12), 1092–1099. doi:10.1002/prot.25779

Karasikov, M., Pagès, G., and Grudinin, S. (2019). Smooth Orientation-dependent Scoring Function for Coarse-Grained Protein Quality Assessment. *Bioinformatics* 35 (16), 2801–2808. doi:10.1093/bioinformatics/bty1037

Lee, S. Y., and Skolnick, J. (2010). TASSER_WT: A Protein Structure Prediction Algorithm with Accurate Predicted Contact Restraints for Difficult Protein Targets. *Biophysical J.* 99 (9), 3066–3075. doi:10.1016/j.bpj.2010.09.007

Li, J., and Xu, J. (2020). "Study of Real-Valued Distance Prediction for Protein Structure Prediction with Deep Learning" *BioRxiv* doi:10.1101/2020.11.26.400523

Li, Y., Hu, J., Zhang, C., Yu, D.-J., and Zhang, Y. (2019). ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks. *Bioinformatics* 35 (22), 4647–4655. doi:10.1093/bioinformatics/btz291

Lobley, A., Sadowski, M. I., and Jones, D. T. (2009). PGenTHREADER and PDomTHREADER: New Methods for Improved Protein Fold Recognition and Superfamily Discrimination. *Bioinformatics* 25 (14), 1761–1767. doi:10.1093/bioinformatics/btp302

Ma, Jianzhu., Wang, Sheng., Wang, Zhiyong., and Xu, Jinbo. (2014). MRFalign: Protein Homology Detection through Alignment of Markov Random Fields. *PLOS Comput. Biol.* 10 (3), e1003500. doi:10.1371/journal.pcbi.1003500

Ma, J., Peng, J., Wang, S., and Xu, J. (2012). A Conditional Neural Fields Model for Protein Threading. *Bioinformatics* 28 (12), i59–i66. doi:10.1093/bioinformatics/bts213

Ma, J., Wang, S., Zhao, F., and Xu, J. (2013). Protein Threading Using Context-specific Alignment Potential. *Bioinformatics* 29 (13), i257–i265. doi:10.1093/bioinformatics/btt210

Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., et al. (2014). IMG/M 4 Version of the Integrated Metagenome Comparative Analysis System. *Nucl. Acids Res.* 42 (D1), D568–D573. doi:10.1093/nar/gkt919

Marks, D. S., Robert, S., Hopf, Ts. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* 6 (12), e28766. doi:10.1371/journal.pone.0028766

Marti, R. M. A., Madhusudhan, M. S., and Sali, A. (2004). Alignment of Protein Sequences by Their Profiles. *Protein Sci.* 13 (4), 1071–1087. doi:10.1110/ps.03379804

Meier, Armin., and Söding, Johannes. (2015). Automatic Prediction of Protein 3D Structures by Probabilistic Multi-Template Homology Modeling. *PLOS Comput. Biol.* 11 (10), e1004343. doi:10.1371/journal.pcbi.1004343

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. (2017). Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments. *Nucleic Acids Res.* 45 (D1), D170–D176. doi:10.1093/nar/gkw1081

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., et al. (2020). MGnify: The Microbiome Analysis Resource in 2020. *Nucleic Acids Res.* 48 (D1), D570–D578. doi:10.1093/nar/gkz1035

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., et al. (2018). EBI Metagenomics in 2017: Enriching the Analysis of Microbial Communities, from Sequence Reads to Assemblies. *Nucleic Acids Res.* 46 (D1), D726–D735. doi:10.1093/nar/gkx967

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., et al. (2011). Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt.Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proc. Natl. Acad. Sci.* 108 (49), E1293–E1301. doi:10.1073/pnas.1111471108

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., et al. (2017). Protein Structure Determination Using Metagenome Sequence Data. *Science* 355 (6322), 294–298. doi:10.1126/science.aah4043

Pearce, R., and Zhang, Y. (2021). Deep Learning Techniques Have Significantly Impacted Protein Structure Prediction and Protein Design. *Curr. Opin. Struct. Biol.* 68 (June), 194–207. doi:10.1016/j.sbi.2021.01.007

Peng, J., and Xu, J. (2009). "Boosting Protein Threading Accuracy," in *In* Research In Computational Molecular Biology, Editor S Batzoglou, 31–45. Lecture Notes in Computer Science (Berlin, Heidelberg: Springer Berlin Heidelberg). doi:10.1007/978-3-642-02008-7_3

Peng, J., and Xu, J. (2010). Low-Homology Protein Threading. *Bioinformatics* 26 (12), i294–300. 10.1093/bioinformatics/btq192

Peng, J., and Xu, J. (2011). A Multiple-Template Approach to Protein Threading. *Proteins: Struct. Funct. Bioinformatics* 79 (6), 1930–1939. doi:10.1002/prot.23016

Petrey, D., and Honig, B. (2005). Protein Structure Prediction: Inroads to Biology. *Mol. Cel* 20 (6), 811–819. doi:10.1016/j.molcel.2005.12.005

Ray, A., Lindahl, E., and Wallner, B. (2012). Improved Model Quality Assessment Using ProQ2. *BMC Bioinformatics* 13 (1), 224. doi:10.1186/1471-2105-13-224

Roche, Rahmatullah., Bhattacharya, Sutanu., and Bhattacharya, Debswapna. (2021). Hybridized Distance- and Contact-Based Hierarchical Structure Modeling for Folding Soluble and Membrane Proteins. *PLOS Comput. Biol.* 17 (2), e1008753. doi:10.1371/journal.pcbi.1008753

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000). Comparison of Sequence Profiles. Strategies for Structural Predictions Using Sequence Information. *Protein Sci.* 9 (2), 232–241. doi:10.1110/ps.9.2.232

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2019). Protein Structure Prediction Using Multiple Deep Neural Networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* 87 (12), 1141–1148. doi:10.1002/prot.25834

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* 577 (7792), 706–710. doi:10.1038/s41586-019-1923-7

Shuvo, M. H., Bhattacharya, S., Bhattacharya, D., and Bhattacharya, Debswapna. (2020). QDeep: Distance-Based Protein Model Quality Estimation by Residue-Level Ensemble Error Classifications Using Stacked Deep Residual Neural Networks. *Bioinformatics* 36 (Suppl. ment_1), i285–i291. doi:10.1093/bioinformatics/btaa455

Skolnick, J., and Kihara, D. (2001). Defrosting the Frozen Approximation: PROSPECTOR? A New Approach to Threading. *Proteins* 42 (3), 319–331. doi:10.1002/1097-0134(20010215)42:3<319::aid-prot30>3.0.co;2-a

Söding, J. (2017). Big-Data Approaches to Protein Structure Prediction. *Science* 355 (6322), 248–249. 10.1126/science.aal4512

Söding, J. (2005). Protein Homology Detection by HMM-HMM Comparison. *Bioinformatics* 21 (7), 951–960. doi:10.1093/bioinformatics/bti125

Steinegger, M., and Söding, J. (2018). Clustering Huge Protein Sequence Sets in Linear Time. *Nat. Commun.* 9 (1), 2542. doi:10.1038/s41467-018-04964-5

Steinegger, M., Mirdita, M., and Söding, J. (2019). Protein-Level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold. *Nat. Methods* 16 (7), 603–606. doi:10.1038/s41592-019-0437-

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H., and the UniProt Consortium (2015). UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* 31 (6), 926–932. doi:10.1093/bioinformatics/btu739

Taylor, William. R. (1999). Protein Structure Comparison Using Iterated Double Dynamic Programming. *Protein Sci.* 8 (3), 654–665. doi:10.1110/ps.8.3.654

The UniProt Consortium (2019). UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47 (D1), D506–D515. doi:10.1093/nar/gky1049

Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., and Elofsson, A. (2017). ProQ3D: Improved Model Quality Assessments Using Deep Learning. *Bioinformatics* 33 (10), 1578–1580. doi:10.1093/bioinformatics/btw819

Uziela, Karolis., Shu, Nanjiang., Wallner, Björn., and Elofsson, Arne. (2016). ProQ3: Improved Model Quality Assessments Using Rosetta Energy Terms. *Scientific Rep.* 6 (1), 33509. doi:10.1038/srep33509

Wang, D., Geng, L., Zhao, Y.-J., Yang, Y., Huang, Y., Zhang, Y., et al. (2020). Artificial Intelligence-Based Multi-Objective Optimization Protocol for Protein Structure Refinement. *Bioinformatics* 36 (2), 437–448. doi:10.1093/bioinformatics/btz544

Wang, Sheng., Sun, Siqi., Li, Zhen., Zhang, Renyu., and Xu, Jinbo. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-deep Learning Model. *PLOS Comput. Biol.* 13 (1), e1005324. doi:10.1371/journal.pcbi.1005324

Wang, S., Li, W., Zhang, R., Liu, S., and Xu, J. (2016). CoinFold: A Web Server for Protein Contact Prediction and Contact-Assisted Protein Folding. *Nucleic Acids Res.* 44 (W1), W361–W366. doi:10.1093/nar/gkw307

Wang, Y., Shi, Q., Yang, P., Zhang, C., Mortuza, S. M., Xue, Z., et al. (2019). Fueling Ab Initio Folding with Marine Metagenomics Enables Structure and Function Predictions of New Protein Families. *Genome Biol.* 20 (1), 229. doi:10.1186/s13059-019-1823-z

Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., and Yang, J. (2020). Protein Contact Prediction Using Metagenome Sequence Data and Residual Neural Networks. *Bioinformatics* 36 (1), 41–48. doi:10.1093/bioinformatics/btz477

Wu, S., and Zhang, Y. (2007). LOMETS: A Local Meta-Threading-Server for Protein Structure Prediction. *Nucleic Acids Res.* 35 (10), 3375–3382. doi:10.1093/nar/gkm251

Wu, S., and Zhang, Y. (2008). "MUSTER: Improving Protein Sequence Profile–Profile Alignments by Using Multiple Sources of Structure Information. *Proteins: Struct. Funct. Bioinformatics* 72 (2), 547–556. doi:10.1002/prot.21945

Wu, S., and Zhang, Y. (2010). Recognizing Protein Substructure Similarity Using Segmental Threading. *Structure* 18 (7), 858–867. doi:10.1016/j.str.2010.04.007

Wu, Tianqi., Guo, Zhiye., Hou, Jie., and Cheng, Jianlin. (2021). DeepDist: Real-Value Inter-residue Distance Prediction with Deep Residual Convolutional Network. *BMC Bioinformatics* 22 (1), 30. doi:10.1186/s12859-021-03960-9

Xu, J. (2019). Distance-Based Protein Folding Powered by Deep Learning. *Proc. Natl. Acad. Sci. USA* 116 (34), 16856–16865. doi:10.1073/pnas.1821309116

Xu, J., Li, M., Kim, D., and Xu, Y. (2003). Raptor: Optimal Protein Threading by Linear Programming. *J. Bioinform. Comput. Biol.* 01 (01), 95–117. doi:10.1142/s0219720003000186

Xu, J., and Wang, S. (2019). Analysis of Distance-based Protein Structure Prediction by Deep Learning in CASP13. *Proteins* 87 (12), 1069–1081. doi:10.1002/prot.25810

Xu, Y., and Xu, D. (2000). Protein Threading Using PROSPECT: Design and Evaluation. *Proteins* 40 (3), 343–354. doi:10.1002/1097-0134(20000815)40:3<343::aid-prot10>3.0.co;2-s

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved Protein Structure Prediction Using Predicted Interresidue Orientations. *Proc. Natl. Acad. Sci. USA* 117 (3), 1496–1503. doi:10.1073/pnas.1914677117

Yang, Y., Faraggi, E., Zhao, H., and Zhou, Y. (2011). Improving Protein Fold Recognition and Template-Based Modeling by Employing Probabilistic-Based Matching between Predicted One-Dimensional Structural Properties of Query and Corresponding Native Properties of Templates. *Bioinformatics* 27 (15), 2076–2082. doi:10.1093/bioinformatics/btr350

Zhang, C., Zheng, W., Mortuza, S. M., Li, Y., and Zhang, Y. (2020). DeepMSA: Constructing Deep Multiple Sequence Alignment to Improve Contact Prediction and Fold-Recognition for Distant-Homology Proteins. *Bioinformatics* 36 (7), 2105–2112. doi:10.1093/bioinformatics/btz863

Zhang, Hg., and Shen, Y. (2020). Template-Based Prediction of Protein Structure with Deep Learning. *BMC Genomics* 21 (11), 878. doi:10.1186/s12864-020-07249-8

Zheng, W., Wuyun, Q., Yang, Li., Mortuza, S. M., Zhang, C., Pearce, R., et al. (2019a). Detecting Distant-Homology Protein Structures by Aligning Deep Neural-Network Based Contact Maps. *PLOS Comput. Biol.* 15 (10), e1007411. doi:10.1371/journal.pcbi.1007411

Zheng, W., Zhang, C., Wuyun, Q., Pearce, R., Li, Y., and Zhang, Y. (2019b). LOMETS2: Improved Meta-Threading Server for Fold-Recognition and Structure-Based Function Annotation for Distant-Homology Proteins. *Nucleic Acids Res.* 47 (W1), W429–W436. doi:10.1093/nar/gkz384

Zhou, H., and Zhou, Y. (2005). Fold Recognition by Combining Sequence Profiles Derived from Evolution and from Depth-dependent Structural Alignment of Fragments. *Proteins* 58 (2), 321–328. doi:10.1002/prot.20308

Zhu, J., Wang, S., Bu, D., and Xu, J. (2018). Protein Threading Using Residue Co-variation and Deep Learning. *Bioinformatics* 34 (13), i263–i273. doi:10.1093/bioinformatics/bty278

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 14 February 2022
doi: 10.3389/fmolb.2021.773234

Selection and Modelling of a New Single-Domain Intrabody Against TDP-43

Martina Gilodi[1,2], Simonetta Lisi[3], Erika F. Dudás[2], Marco Fantini[3], Rita Puglisi[2], Alexandra Louka[2], Paolo Marcatili[4], Antonino Cattaneo[3] and Annalisa Pastore[2]**

[1]Department of Molecular Medicine, University of Pavia, Pavia, Italy, [2]Dementia Research Institute at King's College London, The Wohl Institute, London, United Kingdom, [3]Bio@SNS Laboratory, Scuola Normale Superiore, Piazza dei Cavalieri, Pisa, Italy, [4]Department of Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disorder associated to deteriorating motor and cognitive functions, and short survival. The disease is caused by neuronal death which results in progressive muscle wasting and weakness, ultimately leading to lethal respiratory failure. The misbehaviour of a specific protein, TDP-43, which aggregates and becomes toxic in ALS patient's neurons, is supposed to be one of the causes. TDP-43 is a DNA/RNA-binding protein involved in several functions related to nucleic acid metabolism. Sequestration of TDP-43 aggregates is a possible therapeutic strategy that could alleviate or block pathology. Here, we describe the selection and characterization of a new intracellular antibody (intrabody) against TDP-43 from a llama nanobody library. The structure of the selected intrabody was predicted *in silico* and the model was used to suggest mutations that enabled to improve its expression yield, facilitating its experimental validation. We showed how coupling experimental methodologies with *in silico* design may allow us to obtain an antibody able to recognize the RNA binding regions of TDP-43. Our findings illustrate a strategy for the mitigation of TDP-43 proteinopathy in ALS and provide a potential new tool for diagnostics.

Keywords: antibody selection, hypervariable loops, intrabodies, modelling, misfolding proteins, ALS

OPEN ACCESS

Edited by:
Menico Rizzi,
University of Eastern Piedmont, Italy

Reviewed by:
Joost Schymkowitz,
VIB-KU Leuven Center for Brain and Disease Research, Belgium
Carlo Camilloni,
University of Milan, Italy

****Correspondence:***
Annalisa Pastore
annalisa.pastore@crick.ac.uk
Antonino Cattaneo
Antonino.cattaneo@sns.it

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *09 September 2021*
Accepted: *29 November 2021*
Published: *14 February 2022*

Citation:
Gilodi M, Lisi S, F. Dudás E, Fantini M, Puglisi R, Louka A, Marcatili P, Cattaneo A and Pastore A (2022) Selection and Modelling of a New Single-Domain Intrabody Against TDP-43. Front. Mol. Biosci. 8:773234. doi: 10.3389/fmolb.2021.773234

INTRODUCTION

Amyotrophic lateral sclerosis (ALS) and Frontotemporal dementia (FTD) are distinct but genetically correlated fatal neurodegenerative diseases. ALS is characterized by the selective degeneration of motor neurons that typically appears in middle-aged patients (average age 55 years) and progresses to muscle atrophy followed by complete paralysis. Death is caused by respiratory failure and typically intervenes within 3–5 years from diagnosis. The disease is predominantly (90%) sporadic, but familial cases (fALS) are found in ca. 10% of the cases (Prasad et al., 2019). FTD is also a midlife-onset disease that is clinically heterogeneous and characterized by changes in behaviour, personality, and/or speech (Mackenzie and Neumann, 2016). Because of a remarkable overlap in manifestations, the two diseases are now considered a disease continuum, with 50% of ALS patients presenting cognitive impairment (15–20% recognized as FTD), and 15% of FTD patients having motor impairments (Devenney et al., 2015; Burrell et al., 2016).

Several proteins have been implicated in these diseases. Among them is the TAR DNA-binding protein 43 (TDP-43), a DNA/RNA-binding protein ubiquitously expressed, and predominantly localized in the nucleus (Ayala et al., 2008; Prasad et al., 2019). TDP-43 is a modular protein that is

involved in different aspects of RNA metabolism including transcription, splicing, transport, and scaffolding (Buratti and Baralle, 2008; Cohen et al., 2011; Liu et al., 2017). The architecture of TDP-43 comprises a partially folded N-terminal domain, two RNA-binding RRM tandem domains (RRM1 and 2), and an unstructured C-terminus that contains a so-called prion-like motif (Buratti and Baralle, 2001; Winton et al., 2008; Lukavsky et al., 2013; Mompeán et al., 2016). An hallmark of the TDP-43 related pathologies is the mislocalization, accumulation and consequent aberrant aggregation of TDP-43 in the cytoplasm where the protein is heavily post-translationally modified (Suk and Rousseaux, 2020). TDP-43 aggregates are also associated to other diseases, such as Alzheimer's disease (AD), Parkinson's disease (PD), and Huntington's disease (HD) (Buratti and Baralle, 2009; Gao et al., 2018).

Clinical mutations of TDP-43 are rare and seem to occur mainly, but not exclusively, in the C-terminus of the protein (Pesiridis et al., 2009; Barmada et al., 2010). This observation had originally suggested that this region is the main cause of protein aggregation and misfolding. More recently TDP-43 fragments containing only the RRM domains or the whole region from the N-terminus to the end of RRM2 have been demonstrated to aggregate and misfold also in the absence of the C-terminus (Budini et al., 2015; Chen et al., 2019; Zacco et al., 2019) indicating that TDP-43 contains multiple aggregation-prone hotspots. Accordingly, clinically relevant mutations occurring in the two RRM domains have been described (Chen et al., 2019).

Despite the advancements made in understanding TDP-43 aggregation, too many details of the mechanism remain unclear. Lack of information partially arises from a lack of adequate research tools able to accurately probe aggregation. In this regard, antibodies constitute a ductile means widely used in research and in clinics, thanks to their high binding affinity and specificity. Antibody applications extend from quantitative *in vitro* measurements to *in vivo* studies. When expressed as intrabodies inside cells (Biocca et al., 1990; Cattaneo and Chirichella, 2019), they can for instance be used to sequester protein aggregates reducing cell toxicity (Meli et al., 2014). They are also great assets in diagnostics and basic science as they may be used in super-resolution microscopy, allowing visualization of protein aggregates at the nanoscale as in the recently developed DNA-PAINT methodology (Schermelleh et al., 2019; Sograte-Idrissi et al., 2019; Oi et al., 2020).

Among the natural antibody scaffolds, variable domains of the heavy chain antibody (VHHs) (also named nanobodies) offer specific advantages over normal antibodies but also respect to single chain Fv (scFv) fragments (Bird et al., 1988) or domain antibodies (dAbs) (Ward et al., 1989) or other antibody mimetics. Natural VHHs were first identified in camelids (Saerens et al., 2005) which are typically single variable heavy chain domains of ca. 110 amino acids that are derived from heavy-chain-only antibodies (V_H), devoid of the light chain partners. A major advantage of camelid VHHs, with respect to immunoglobulin-derived dAbs (24), is their ability to specifically recognize antigens with affinities similar to those obtained by whole antibodies despite their smaller size, and the absence of the hydrophobic VH-VL interface. VHHs are also usually more stable, with melting temperatures as high as 90°C, and higher resilience to detergents and denaturants. Given their small size, good tissue penetration, and low immunogenicity, VHHs have been developed for different neurodegenerative disorders such as AD, Lewy body disease, PD, and HD, and in the attempt to block or prevent aggregation (Harmsen and De Haard, 2007; Khodabakhsh et al., 2018; Hoey et al., 2019; Messer and Butler, 2020).

Here, we describe a new naïve library of llama VHHs, and exploit it to select directly from TDP-43 cDNA a new anti-TDP-43 VHH, which we named VHH5. Usually, VHH libraries are obtained from immunized animals, and are used in different display platforms (phage, yeast, and ribosomal, etc.), that require the immunizing protein for antibody detection from the library. We constructed instead a *llama glaba* naïve VHH library in the SPLINT (Single Pot Library of Intracellular Antibodies) format in yeast, followed by antibody selection with the two-hybrid-based Intracellular Antibody Capture Technology (IACT) (Visintin et al., 1999; Visintin et al., 2002; Visintin et al., 2004). This approach allows direct selection of antibodies from antigen cDNA, with no need to express and purify the protein antigen (Meli et al., 2009). Based on the amino acid sequence deducted from the DNA sequence of the selected VHH5 intrabody, we performed an *in silico* prediction of the antibody structure. The resulting model was used to suggest mutations that optimized the expression of VHH5 in bacterial cells, enabling the experimental biochemical validation of the intrabody. We demonstrate that structure prediction is a powerful tool to guide carefully planned mutagenesis that can facilitate soluble intrabody production. To the best of our knowledge, this is the first detailed description of an anti-TDP-43 intrabody. This new VHH opens new avenues for diagnostic, to interfere with protein aggregation and for imaging applications by super-resolution microscopy (Messer and Joshi, 2013; Schermelleh et al., 2019).

MATERIALS AND METHODS

Llama Glaba VHH Library Construction

Naïve blood samples (40 ml) from two non immunized female llamas were kindly provided by the Biopark Zoom (Cumiana, Turin, and Italy) which is an approved public husbandry Zoo, which operates under the following law: legislative decree 21 March 2005, n. 73 (Gazzetta Ufficiale n. 100, 2 May 2005). The blood samples were taken from the two llama animals as part of the normal periodic blood testing of these animals. Periferal blood lymphocytes were separated by Ficoll-Histopaque-1077 (Sigma-Aldrich) discontinuous gradient centrifugation followed by washing with the phosphate buffered saline (PBS) solution and stored at −70°C. Total RNA was isolated from 10^7 leucocytes by acid guanidinium thiocyanate-phenol chloroform extraction (using TRIzol RNA Isolation Reagents, Thermo Fisher Scientific). RNA integrity was assayed by agarose gel electrophoresis. The total RNA (5 μg) was retrotranscribed in cDNA using the Reverse Transcriptase Core Kit (Eurogentec RT-RTCK-03), with the following thermocycles: 25°C for 10 min, 48°C for 30 min, 95°C for 5 min. The VHH sequences were

amplified from cDNA using previously described primers (van der Linden et al., 2000). We used a degenerate forward primer (VH1-Back BssHII) annealing to the hinge region of each heavy chain-only IgG isotype corresponding to the amino acid sequence (E/Q/K/*)V (Q/K)LQ (E/Q)SG), with the BssHII restriction site (underlined) VH1-Back BssHII: GC GCG CAT GCC VAG GTS MAR YTR GTN SAG TCW GG and two reverse primers Lam-07 NheI and Lam-08 NheI that respectively anneal the llama long-hinge heavy chain antibody (cIgG2), and the short-hinge antibody (cIgG3) (Hamers-Casterman et al., 1993) with the NheI restriction site (underlined) Lam-07 NheI: GCTAGC GGA GCT GGG GTC TTC GCT GTG GTG CG; Lam-08 NheI GCTAGC TGG TTG TGG TTT TGG TGT CTT GGG TT.

The PCR protocol consisted of an initial denaturation step at 98°C for 1 min followed by 10 cycles of 98°C for 10 s, 55°C for 30 s, and 72°C for 30 s, followed by 10 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 30 s, followed by 10 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s, and a final extension step at 72°C for 3 min. The resulting unique ~450 bp PCR fragment was purified from 1.5% highly pure agarose gel using the Wizard® SV Gel and the PCR Clean-Up System (Promega), digested with BssHII and NheI (New England Biolabs), re-purified and ligated (T4 DNA Ligase, NewEngland Biolabs) into BssHII, and NheI digested pLinker220 IACT plasmid (Visintin et al., 2004). This plasmid carries the LEU2 gene, involved in the synthesis of Leucine (L), the 2 μm origin of replication for transformation in yeast, and the selection marker (Ampicillin) and the origin of replication (ColE1 ori) for selection in bacteria. Ligation of the library (~1 μg) was transformed by electroporation into Max Efficiency *E. coli* DH5α cells (Invitrogen). Transformation efficiency was estimated by plating serial dilution aliquots on Luria Broth (LB)/ampicillin (100 μg/ml) agar plates, incubated overnight at 37°C, and assessed by colony count. ~1 million cells were inoculated the next day into 1 l of LB, Sea Prep Agar and ampicillin for library amplification (Elsaesser and Paysan, 2004). An aliquot of the inoculated mixture was plated on LB/ampicillin (100 μg/ml) agar plates to determine the effective colony count. The inoculated Sea Prep Agar was then poured in a pre-chilled sterile stainless-steel container (~200 × 300 × 50 mm^3; Neolab, Heidelberg, and Germany) on wet ice in a cold room and left on ice at 4°C for 1.5 h, and transferred to an incubator at 37°C for 40 h. The visible spherical bacterial colonies embedded in the semi-liquid gel were collected by centrifugation at 8,000 *g* for 20 min at room temperature. The pellet was washed with 100 ml of LB medium and centrifuged again at 8,000 g for 20 min at room temperature. Plasmid DNA from the pellet was extracted using a Qiagen GIGAprep kit, following the manufacturer's instructions.

NGS Llama Library Sequencing

The obtained llama library was sequenced as previously described (Fantini et al., 2017). To attach sequencing adapters to the VHH sequences, a ligation-based approach was designed. DNA adapters were synthesized harbouring overhangs complementary to the cleavage product of the restriction enzymes BssHII and NheI, used for excising the scFv fragment from the plasmid. The forward and reverse strands of the adapters were synthesized independently and annealed *in vitro* (1:1 ratio, 95°C 5 min, and 95→25°C in 5°C steps 1 min/step). Before annealing the reverse strand was phosphorylated (0.2 nmol of oligos, 10U PNK (NEB) at 37°C for 1 h, and at 65°C for 20 min) to allow ligation. The VHHs were excised from the library plasmid (~2 μg of the library were digested for 3 h at 37°C with 4U of NheI (NEB), and for 3 h at 50°C with 4U of BssHII (NEB)) and ligated to the adapters (forward adapter:VHH:reverse adapter in 10:1:10 ratio, ~200 ng library 400 U T4 ligase (NEB), and overnight at 16°C). Ligation was run on an agarose gel and the band corresponding to the single insert with the 5′ and 3′ adapters was resolved and purified with the MinElute Gel Extraction Kit (Qiagen).

The library was quantified by Qubit dsDNA HS Assay Kit (ThermoFisher Scientific), diluted to 4 nM, and denatured with 0.1 N NaOH (5 min at room temperature), neutralized and diluted again in buffer HT-1 (Illumina) to a final concentration of 12.5 pM. Equimolar denatured Phi-X Control V3 DNA (Illumina) was spiked-in 20% volume as an internal quality control and to increase the sample diversity according to Illumina guidelines. Sequencing was performed on the MiSeq system with the Reagent Kit v3 (Illumina), using 350 and 250 cycles for the forward,1 and reverse reads respectively.

Raw data were demultiplexed from. bcl files into separate. fastq files with bcl2fastq-1.8.4 (Illumina), using the following barcodes as indexes: i1 = TCAGCG, i2 = GATCAC, i3 = CTGAGA, and i4 = AGCTTT. To take into account the different lengths of shifter sequences introduced with the sequencing adapters, a specific number of nucleotides was discarded from the start of the reads (R1 index i1 = 0, i2 = 1, i3 = 7, and i4 = 8; R2 index i1 = 13, i2 = 12, i3 = 11, and i4 = 10). Reads were purged from adapter dimers, quality-filtered (Phred Score 32), and trimmed in sequences of the same length (R1: 320bp; R2: 220bp) with trimmomatic-0.32 (Bolger et al., 2014). All the sequences whose forward and reverse reads both survived from the previous step were selected, taking advantage of the Perl script fastq-remove-orphans.pl, and which is part of the fastq-factory suite (https://github.com/phe-bioinformatics/fastq-factory). The VHH nanobody library reads were merged using PEAR (van der Linden et al., 2000), a pair-end read merger available at http://sco.h-its.org/exelixis/web/software/pear/.

Intrabody Selection

The TDP-43 gene (residues 1–414) was cloned in pMicBD1 plasmid (pMicBD1-TDP-43 bait plasmid) and transformed in L40 yeast. The strain was grown in 1% Yeast Extract, 2% Bacto Peptone, 2% Glucose, and at pH 5.8 to an OD_{600} of 0.6. Cells were washed in 1xTE (10 mM Tris, 1 mM EDTA, and pH 7.5), and resuspended in 0.5 ml of 1xTE/1xLiAC (10 mM Tris, 1 mM EDTA, and 0.1 M Lithium acetate dehydrate pH 7.5). Cells (100 μl) were added to 100 μg of salmon tested DNA (STD) and 200 ng of pMicBD1-TDP-43 plasmid with 600 μl of 50% PEG/1xTE/1xLiAC (40% (w/v) PEG 4000, 10 mM Tris-HCl, 1 mM EDTA, and 0.1 M lithium acetate dehydrate pH 7.5) and spun at 150 rpm for 30 min at 30°C. DMSO (70 μl) was added and the cells were heat shocked at 42°C for 15 min, put in ice for 2 min, centrifuged, resuspended in 100 μl of 1 × TE and plated on Synthetic Designed liquid minimal medium lacking tryptophan (SD-W) plates.

For IACT screening, the strain expressing the LexA-TDP-43 bait was grown overnight at 30°C in SD-W media. The overnight culture was diluted in 1 l of pre-warmed rich medium YPAD (1% Yeast Extract, 2% Bacto Peptone, 0.01% Adenine, 2% Glucose, and pH 5.8) and cultured from OD_{600} 0.3–0.6. Cells were centrifuged, washed in 150 ml of 1 × TE, and resuspended in 15 ml of 1 × TE/1 × LiAC. Salmon tested DNA (STD) (10 mg), and the VHH llama DNA library (250 μg) cloned in the pLinker220 prey plasmid were added. The mixture was transferred in a flask with 140 ml of 50% PEG/1xTE/1xLiAC and incubated at 150 rpm for 30 min at 30°C. DMSO (17.6 ml) was added and the cells were heat shocked at 42°C for 15 min under gentle mixing. The flask was then put in ice for 5 min and the cells were washed three times with YPA (1% Yeast Extract, 2% Bacto Peptone, 0.01% Adenine, and at pH 5.8), and recovered in 1 l of YPAD for 1 h at 30°C. A quarter of the cells were washed three times with SD-WHL (SD without, Tryptophan, Histidine, and Leucine), resuspended in 5 ml of SD-WHL, and plated on SD-WHL Petri dishes. The remaining cells were washed in SD-WL (same of SD-WHL but with 0.05% Histidine), resuspended in 200 ml SD-WL and grown overnight at 30°C. The next morning the cells were washed and resuspended in SD-WHL, plated on SD-WHL Petri dishes, and incubated at 30°C for 4–5 days. Ninety nine clones were picked and re-streaked onto a SD-WHL and SD-WL plates. A liquid β-galactosidase (β-gal) assay, adapted from (Möckli and Auerbach, 2004), was performed using a 96-well plate. A small amount of the biomass from single colonies was resuspended in 50 μl of lysis buffer (20 mM Tris HCL pH 7.5, 333 U/ml lyticase) and incubated for 2 h at 37°C. 50 μl of a solution made of 60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, pH 7.0, X-gal at 20 mg/ml (170 μl), and β-mercaptoethanol (30 μl), was added to each well and incubated for 2 h at 37°C. Strong prey–bait interactions were identified by the development of blue color.

Colony PCR and Fingerprint Analysis

Colony PCR and fingerprint analysis were performed only on double positive colonies (His+/LacZ+). The clones were lysed using 10 μl of buffer (20 mM Tris HCl pH 7.5, 300 U/ml lyticase). The VHH of each clone was amplified by PCR using primers located at the 5′ and 3′ of the VHH in the pLinker220 plasmid. The primers were pL220 Fw (5′-AAG CTT ATT TAG GTG ACA CTA TAG-3′) and pL220 Rev (5′- CTT CTT CTT GGG TGC CAT G-3′). The PCR reaction was performed as follows: 3 min at 95°C, followed by 30 cycles at 95°C for 30 s, 50°C for 30 s and 72 °C for 40 s, 5 min at 72°C, and then 4°C to store. The PCR mixture (8 μl/20 μl) was digested with the restriction enzymes NlaIV and AluI, for 2 h at 37°C, to identify a specific pattern for each isolated VHH. Digested fragments were resolved using 8% polyacrylamide gel electrophoresis, followed by ethidium bromide staining. Once the different patterns were highlighted, six individual clones were selected to extract the prey DNA from yeast. Each plasmid was transformed by electroporation, using DH5α Emax cells into bacteria to obtain a pure and monoclonal preparation.

In vivo Epitope Mapping of the anti-TDP-43 VHH5

To characterize the epitope recognized by the anti-TDP-43 VHH5 the original LexA-TDP-43 bait was truncated in two fragments named LexA-N-term + RRM1-2 (residues 1–258) and LexA-C-term (residues 259–414) and transformed in L40 yeast as described above. These strains were then transformed with the pLinker220 plasmid carrying the VHH5 with the same protocol and plating the cells on SD-WL or SD-WHL. To further narrow down the region carrying the epitope a second cycle was done, splitting the region found positive (1–258) into four smaller baits, the N-terminus (1–105), RRM1 (106–176), RRM2 (192–258), and a fragment of RRMs (160–208) which contains the linker between RRM1, and 2 (not to be confused with RRM1-2 which is represents a construct comprising the tandem domains). The anti-TDP-43 VHH5 was transformed in L40 yeast strains individually carrying one of the smaller baits.

Initial Model Generation

The most suitable template was identified by submitting the sequence of the target protein to the BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against the PDB database. Models were built both by the SWISS-MODEL (Waterhouse et al., 2018) and the ABodyBuilder (Leem et al., 2016) servers. The semi-automated procedure was used in SWISS-MODEL where alignment between the template and the target was fed manually.

Loop Generation

Modelling of the complementarity-determining region (CDR) H3 loop was carried out using the Sphinx algorithm (Marks et al., 2017). The input to Sphinx is a protein structure or a model (in PDB format) and the location and sequence of the loop to be modelled. We used the best SWISS-MODEL structure to model the loop region comprising residues 94–114. Once a complete set of decoys was generated, a statistical potential was used to reduce the set to only 500 structures, which were then scored using SOAP-Loop (Dong et al., 2013) to produce a ranking. SOAP-Loop was assessed by the average global root mean square deviation (RMSD) of the top ranked model for each loop. From the ranking that was generated based on the frequency of how often similar conformations were selected and the energy of single conformations, we selected ten models for the loop which we used as a mould to perform the docking between the nanobody and TDP-43.

Model Refinement

Molecular dynamics (MD) simulations were performed using the NAMD 2.13 package (Phillips et al., 2020) with the CHARMM36m force field. Input files were generated by CHARMM-GUI (Jo et al., 2008; Lee et al., 2016). The structures were solvated with the TIP3P water model in a rectangular box such that the minimum distance to the edge of the box was 10 Å under periodic boundary conditions. An appropriate number of Cl^- counterions were added to neutralize the protein charge. The time step was set to 2 fs throughout the

simulations. A cutoff distance of 12 Å for Coulomb and van der Waals interactions was used. Long-range electrostatics was evaluated through the Particle Mesh Ewald method. The two energetically best models—one provided by the SWISS-MODEL server homology modelling pipeline and one by the ABodyBuilder antibody modelling pipeline—were refined by energy minimization. 20,000 steps of conjugated gradient energy minimization were carried out 1) without constraints, 2) with positional constraints on the backbone heavy atoms of residues 1–70 and 77–135, and 3) with positional constraints on all heavy atoms of residues 1–70 and 77–135. Throughout these minimizations—providing replicas 1, 2, and 3 for each model—the applied force constant was 1.0 kcal $mol^{-1}Å^{-2}$. The energy minimization resulted in six models that after additional 10,000 steps of energy minimization were subjected to 1 ns of equilibration at 303.15 K and 1 atm. The production runs (100 ns) were performed under the same conditions except that all positional constraints were removed. A similar procedure was adopted on the energetically best model obtained after the H3 loop generation as ranked according to SOAP-Loop ranking. The model was subjected to 10,000 steps of energy minimization and 1 ns of equilibrations at 303.15 K and 1 atm. This was followed by an 80 ns production run.

Trajectories were visualized and analysed with the VMD program (Humphrey et al., 1996). Every tenth frame of each trajectory was loaded, for a total of 500 structures. Structural alignment was achieved on the whole molecule for the ABodyBuilder structures and on the region 1–121 for the SWISS-MODEL structures. Coordinates were extracted with a stride value of 10, resulting in 50 structures, and visualized in PyMOL.

ClusPro

Antigen-antibody binding was carried out based on the NMR structure of human TDP-43 tandem RRM1-2 in a complex with a UG-rich RNA (PDB code 4bs2) from which the RNA molecule was removed. Molecular docking was performed by using the ClusPro software (Kozakov et al., 2017). The standard inputs of ClusPro are two PDB files, one denoted as the ligand, and the other one as the receptor. To influence docking, an attractive force was set on the residues of H3 using default parameters. The calculations were repeated on each of the ten best structures obtained by Sphinx. Cluster selection was made to exclude solutions that did not show any contact between the CDR loops and the TDP-43 ligand. An additional filtering step was included to remove all the solutions in which less than ten CDR residues were involved in molecular interactions with the antigen. A residue was defined as interacting if any of its atoms was at less than 4 Å distance from any antigen atom. Similarly, each solution was annotated based on the number of contacts with the first, and second domain in the TDP-43 structure. All the representative structures from then ten ClusPro runs were then pooled together and analysed to identify conserved interaction patterns with the antigen. The interface RMSD (iRMSD) between each pair of solutions was then computed, by superimposing the antigen structure, and measuring the RMSD of the Cα atoms in the CDR regions of the respective interacting antibody. Clustering of the solution was then performed on the complete distance matrix, by using the DBScan algorithm from the Python package SciKit-Learn (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html), using the parameters eps = 9, and min_clust = 3 (Pedregosa et al., 2011). The clustering results were then visualised by transforming the distance matrix to a two-dimensional space using the t-SNE algorithm in SciKit-Learn (Van Der Maaten and Hinton, 2008) (https://scikit-learn.org/stable/auto_examples/index.html). The models were visualised by the Pymol software.

Sequence Analysis

AGGRESCAN (Conchillo-Solé et al., 2007) was used to predict the aggregation properties of VHH5. The standard input for AGGRESCAN is the polypeptide sequence(s) consistent with FASTA format. In the output, the regions of the sequence with the highest predicted aggregation propensity are highlighted in red in the peptide sequence column and appear as peaks in the profile graphs. The position of the CDR loops was obtained by the http://cao.labshare.cn/AbRSA/abrsa.php server (Li et al., 2019).

VHH5 Production

Preliminary attempts to produce the protein in *E. coli* were done using a pET-17b which encoded a fusion protein with an N-terminal PelB leader sequence and a C-terminal $(His)_7$-tag. Since this strategy proved unsuccessful, VHH5 was recloned by PCR into a pET-SUMO plasmid, and expressed in BL21 (DE3) pLysS cells as a fusion protein with an N-terminal SUMO solubilization domain and a $(His)_6$-tag. Cells transformed with the plasmid were grown overnight at 37°C in LB medium containing 50 μg/ml kanamycin. Cell cultures were diluted 1:50 in fresh LB with 50 μg/ml kanamycin and grown to an OD_{600} of 0.6, before adding 0.5 mM IPTG to induce protein expression for 4 h at 37°C. The cells were collected by centrifugation at 4,000 rpm for 20 min at 4°C, resuspended in lysis buffer (10 mM potassium phosphate buffer at pH 7.2, 150 mM KCl, 5 mM imidazole, 5% v/v glycerol, 1 mg/ml lysozyme, a cOmplete™ EDTA-free Protease Inhibitor tablet (Roche), and 1 μg/ml DNase I), and lysed by sonication. The soluble protein was recovered in the supernatant by centrifugation at 20,000 rpm for 50 min at 4°C, and purified by nickel affinity chromatography (Super Ni-NTA agarose resin, Generon) at 4°C, eluting the $(His)_6$-SUMO tag with 10 mM potassium phosphate buffer at pH 7.2, 150 mM KCl with 250 mM imidazole. The tag was cleaved by incubating the construct with tobacco etch virus protease (1:5 protein construct/tobacco etch virus molar ratio) overnight at 4°C, while dialyzing the mixture with 10 mM potassium phosphate at pH 7.2, 1 M KCl. A second nickel column at 4°C was applied. The flow-through was collected and dialyzed at 4°C against 10 mM potassium phosphate buffer at pH 7.2 and 15 mM KCl. Pure VHH5 was obtained after a further step of size-exclusion chromatography on an Äkta pure system (HiLoad 16/60 Superdex 75 prep grade column, GE Healthcare). The protein was eluted in 10 mM potassium phosphate buffer at pH 7.2 and 15 mM KCl, aliquoted, and

flash-frozen and stored at −20°C. The protein purity was assessed by SDS-PAGE and size-exclusion chromatography.

Circular Dichroism and NMR Measurements

Far-UV CD spectra of VHH5 (50 μM) was acquired at 25°C in 10 mM potassium phosphate buffer at pH 7.2 and 15 mM KCl. CD spectra were recorded on a JASCO-1100 spectropolarimeter equipped with a temperature control system, and averaged over 10 scans. Measurements were carried out in 1 mm path-length quartz cuvettes (type S3/Q/1; Starna Scientific), applying a constant N_2 flush at 4.0 l/min. NMR experiments were carried out at 800 MHz on an Avance Bruker spectrometer equipped with a cryogenic probe. The sample (160 μM) was in 10 mM potassium phosphate at pH 7.2 with 15 mM KCl and 10% D_2O. 1D spectra were acquired at 25°C.

ELISA Assays

For the Sandwich ELISA, purified VHH5 were coated in triplicates onto a 96-well plate at concentrations of 1 μM, 3 μM, 5 μM, and 10 μM (corresponding to 15–150 μg/ml), left overnight at 4°C, and in carbonate buffer at pH 9.6. After coating, 2 h blocking at room temperature was performed in PBS/BSA at 1% and pH 7.4. Purified RRM1-2, RRM1, and RRM2 (10 μg/ml) were used to capture the VHH5 prey. The solution was incubated for 2 h at room temperature, followed by a further 2 h incubation in the presence of rabbit anti-TDP-43 polyclonal antibodies (Proteintech) at a 1:2000 dilution. Detection of the retained antigen was performed with goat anti-rIgG [HRP] antibody (Cell Signaling) at a 1:2000 dilution. After a 2 h incubation at room temperature in PBS/BSA 1%, with 3,3′,5,5′Tetramethylbenzidine (TMB) (ThermoFisher, cat. No. 34021) the absorbance was read at 450 nm. Antibody dilutions were in PBS/BSA 1%, pH 7.4. The wells were washed three times between steps with PBST at 0.05% and pH 7.4. Wells that did not contain VHH5 but all the other components were used as negative controls.

For the indirect Elisa, purified RRM1-2, RRM1, and RRM2 were coated in triplicates in a 96-well plate at a concentration of 1 μM (corresponding to 10 μg/μl), left overnight at 4°C in carbonate buffer at pH 9.6. After coating, the reaction was blocked for 2 h at room temperature by PBS/BSA at 1%, and pH 7.4. Purified VHH5 (1 μM, 3 μM, 5 μM, and 10 μM, corresponding to 15–150 μg/ml) was used to capture the antigen by a 2 h incubation at room temperature. Detection of VHH5 was performed with rabbit anti-camelid VHH [HRP] antibody (GenScript) at a 1:5,000 dilution. After 2 h incubation at room temperature in PBS/BSA 1%, with 3,3′,5,5′Tetramethylbenzidine (TMB) (ThermoFisher, cat. No.34021) the absorbance at 450 nm was detected. The antibody dilutions were in PBS/BSA 1%, pH 7.4. The wells were washed three times between steps with PBST at 0.05% and pH 7.4. Wells that did not contain the antigen (TDP-43 fragments) but all the other components were used as negative controls.

RESULTS

Naïve Llama VHH SPLINT Library Construction

A VHH library was created from cDNA derived from peripheral blood lymphocyte RNA isolated from two not immunized (naïve) *llama glaba* animals and cloned in SPLINT format (Visintin et al., 2004) for further use. In this format, the VHH antibody domains are fused in frame to the activation domain of the transcription factor VP16. The VHH DNA library was amplified in bacteria obtaining a complexity of 1.7×10^7, defined as the number of total transformants, determined through colony forming unit (CFU) count. The library was sequenced by Next Generation Sequencing. From a total number of 6,322,129 sequences the sequence diversity resulted to be 1.15×10^6. The library complexity was estimated by the truncated Negative Binomial distribution (Fantini et al., 2017) to fit the number of sequences as a function of sequence cardinality (**Figure 1A**). Most of the sequences (93%) were full-length and did not contain premature stop codons or frameshifts. The VHH lengths fit a normal Gaussian distribution centered on 120.7725 amino acids with a standard deviation of 4.8723 (**Figure 1B**). The diversity of the SPLINT library is in line with our previous mouse or human libraries, which were shown to contain antibody domains able to effectively bind their corresponding protein antigen intracellularly.

Intrabody Selection

The yeast two hybrid based IACT system was used to select intracellular specific intrabodies against TDP-43 from the VHH SPLINT library (Visintin et al., 2002; Visintin et al., 1999). IACT screening works by exploiting yeast L40 strains co-transformed with antigen-bait/antibody-prey pairs, in which the antigen-bait is fused to a DNA binding domain (LexA-DBD) that is challenged with a library of natural recombinant antibody domains fused to the VP16 activation domain (the prey). The TDP-43 gene (amino acids 1–414) was cloned in fusion with LexA and used to challenge the llama antibody library (Visintin et al., 2004). A positive interaction between a prey and the bait activates transcription of the *HIS3* gene, allowing survival on selective media (SD-WHL), and of the *LacZ* gene as a second marker of interaction. After a primary selection, a second round of selection pointed to a lead candidate (VHH5) as a positive TDP-43 interactor. The specificity of VHH5 was analysed for survival on selective media (SD-WHL) using either the screening bait (LexA-TDP-43) or an unrelated bait (LexA-Synuclein) to exclude interactions between VHH5 and the LexA domain of the fusion protein bait (**Figure 2A**). Activation of the second reporter marker *LacZ* was assessed in a liquid β-gal assay. VHH5 interaction with LexA-TDP-43 gave positive β-gal assay as compared to the positive control of the assay (interaction of LexA-TDP-43 with the Y1 anti Lex A nanobody) and the negative control (interaction of LexA-TDP-43 with a scFv anti p-Tau) (**Figure 2B**). Analysis of the intrabody sequence

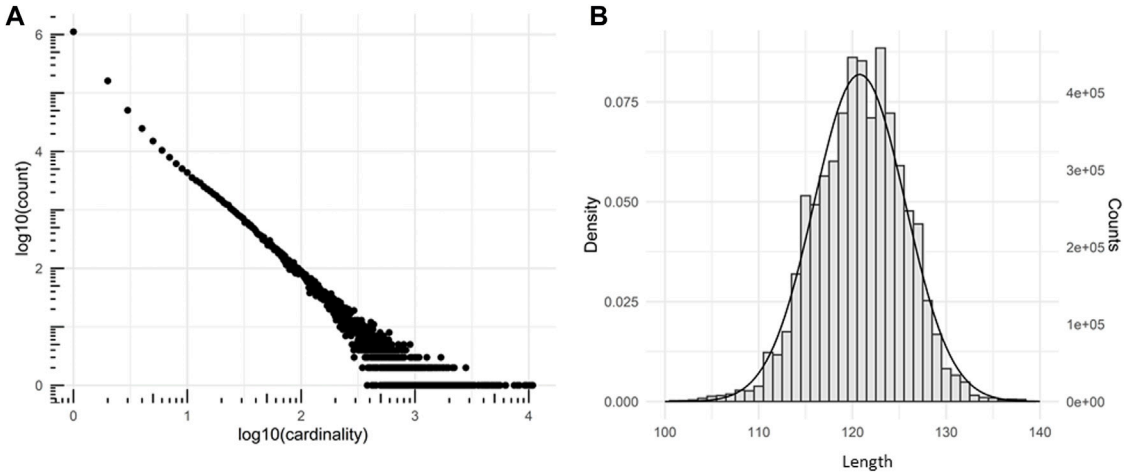


FIGURE 1 | Characterization of the SPLINT library. **(A)** Cardinality plot of the sequenced library. Log-log plot showing the number of time a group of identical n sequences (*n* = cardinality) was found in the sequencing. **(B)** VHH proteins length distribution. Distribution of the number of residues observed in the peptide chains of the of translated llama VHHs (amino acid sequence length) and gaussian fit.

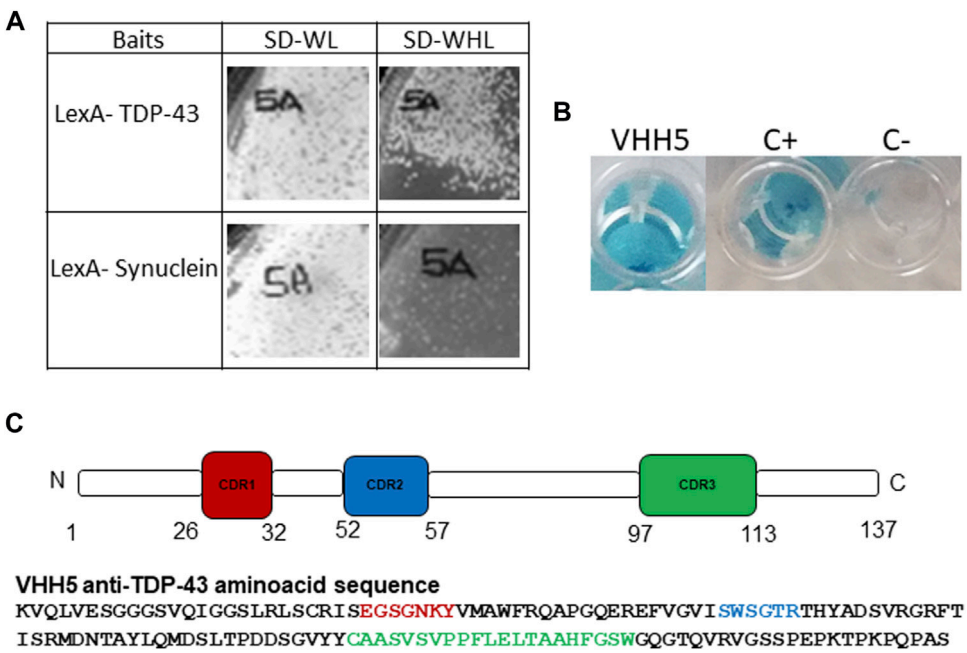


FIGURE 2 | Selection of VHH5. **(A)** Growth on selective plate SD-WHL of the VHH5 co-transformed with the LexA-TDP-43 bait and the unrelated LexA-Synuclein bait. The images of growth on plates were acquired using Chemidoc XRS (Biorad). **(B)** Liquid β-gal assay of yeast co-expressing the LexA-TDP-43 bait and the VHH5 intrabody, C+: LexA-TDP-43+ Y1, an anti-LexA intrabody, and C–: LexA-TDP-43+ scFv anti-pTau. The images were acquired using HUAWEI Mate 10 lite. **(C)** Amino acid sequence of VHH5 and schematic representation of VHH5 with the position of the CDRs, as defined using the Chothia and Lesk numbering scheme (Chothia et al., 1989) in the http://cao.labshare.cn/AbRSA/abrsa.php server (Li et al., 2019).

revealed a short charged H1 loop, a shorter H2 loop containing a Trp, and a rather long H3 loop, comprising 17 residues according to Chothia, and Lesk numbering system (Chothia et al., 1989). This loop is circa ten residues longer than the average of the H3 in antibodies, but within average for intrabodies (**Figure 2C**). It does however contain many degrees of freedom, making prediction of its structure not straightforward.

Attempts to Characterise Recombinant VHH5 by *E. Coli* Overexpression

In the attempt to characterize the anti-TDP-43 VHH5, we tried to express and purify the protein in *E. coli*. VHH5 was first inserted into a pET-17b expression vector fused with the PelB leader sequence that directs proteins to the periplasmic space allowing disulfide bridge formation. The construct was transformed in *E. Coli* BL21 (DE3) cells but resulted poorly overexpressed (**Supplementary Figure S1A**). We then re-cloned the protein in a pET-17b plasmid as fusion protein with an N-terminal SUMO solubilization domain and a $(His)_6$-tag to enhance protein solubility. We also changed the *E. coli* strain and expressed it in BL21 (DE3) pLysS cells. The expression yield appreciably increased but the highly expressed protein accumulated in the cytoplasm as inclusion bodies (data not shown). All attempts to avoid precipitation failed, including changes of the induction temperature. Inclusion body formation has been proven to result from the conflict between aggregation and protein fold and it is a well-known impediment particularly in antibody production (Ventura and Villaverde, 2006).

To predict which residues/regions of the protein could contribute to aggregation, we analysed the sequence by AGGRESCAN (Conchillo-Solé et al., 2007). This is a web-based software that allows prediction of the aggregation properties of a protein on the basis of its sequence. We found several regions predicted to be aggregation prone, some of which in the CDR loops (**Supplementary Figure S1B**). As an alternative strategy, we resorted to model the structure of the intrabody by comparative modelling to have an independent insight based on a 3D model of the structure of VHH5 and a more solid idea of the expected structural features.

Modelling the Antibody Scaffold

The structure of the antibody main scaffold, that is the β-sandwich that holds the antigen recognizing CDR loops, can be easily predicted as this region is highly conserved amongst antibodies, and their derivatives (Narciso et al., 2011). A BLAST search over the PDB database identified 5wcc as the closest sequence-wise template for comparative modelling. This is the crystal structure of the broadly neutralizing Influenza A antibody VRC 315 02-1F07 Fab. We used in parallel both the SWISS-MODEL (Waterhouse et al., 2018) and the ABodyBuilder (Leem et al., 2016) servers for the prediction. SWISS-MODEL relies on ProMod3, an in-house comparative modelling engine based on OpenStructure (Biasini et al., 2013). The ABodyBuilder algorithm also follows template selection, orientation prediction, and CDR loop modelling and side chain prediction. ABodyBuilder then annotates the "confidence" of the model as the probability that a component of the antibody (e.g., a loop or a strand) is modelled within a RMSD threshold. We obtained models that were closely evaluated. The two energetically best structures from each of the two programs could be superposed with a RMSD of 0.45 Å (**Supplementary Figure S2**). The template and target structures were of similar lengths with two one-residue insertions in the H2 and H3 CDR loops and a deletion in another loop.

The two energetically best structures from each of the two programs were then refined by energy minimization using the CHARMM36m force field that has extensively been shown to be robust in simulations of globular proteins. Twenty thousand steps of conjugated gradient energy minimization were applied using no constraints or with positional constraints on the backbone heavy atoms and on the heavy atoms of the solute in the regions 1–70 and 77–135 for both the SWISS-MODEL and ABodyBuilder VHH5 structures 1, 2, and 3. The resulting models were then used as the input to model the CDR loops of VHH5.

H3 Modelling and Structure Refinement

The challenge in antibody structure prediction is the design of the CDR loops. Of the three loops, H1 and H2 can easily be classified according to the canonical structures first described in 1987 by Chothia and Lesk, and their structures can confidently be predicted (Al-Lazikani et al., 1997). The problematic loop is H3 because of the high variability of its sequence, length, and conformation that makes difficult to build a high-quality structure with ordinary modelling techniques. Modelling of the H3 loop (residues 94–114) was carried out using the Sphinx algorithm, a combination of the FREAD knowledge-based method (Deane and Blundell, 2001; Choi and Deane, 2010) and an *ab initio* algorithm. Given the overall similarity between the two structural bundle and to reduce the number of structures to analyse, we restricted the prediction only to the best structure from SWISS-MODEL. We obtained a bundle of 500 structures from which we selected 10 energetically best structures. In most of the solutions the loop turned out not to contain any regular structural element with the loop mostly protruding out from the rest of the molecule (**Supplementary Figure S3**). Only in one model, the loop contains a short 1-turn helical element in the middle of the loop. In seven out of ten structures, and the first two residues of the loop pair with a close-by strand.

We then refined the energetically most favourable structure from the H3 loop modelling (Model 1) by MD simulations, also to obtain information on the conformational space covered by the long H3 loop. Throughout the 80 ns production run, this loop adopted two significantly different conformations: protruding out from the rest of the molecule (open form, 1.4–39.8 ns) or bending closer to the beta strands encompassing residues 33–38 and 46–52 (closed form, 43.2–80.0 ns) (**Figure 3**). This potential variability was also reflected in the time evolution of the total RMSD calculated for the N, CA, and C backbone atoms (**Supplementary Figure S4**). When the RMSD of the individual residues was separately calculated along the trajectory for each of the open and closed forms (**Figure 4**), variability was noticed at the three CDR loops, and especially at H3. The C-terminus (residues 122–135) is completely disordered.

The predicted models were validated by PROCHECK (PDBSum) (Laskowski et al., 1996; Laskowski, 2001). According to this analyser, the Ramachandran plot contained 90% of the residues in the most favoured regions, and 10% in additional/generously allowed regions (**Supplementary Figure S6**). Gly and Pro residues were also located in allowed regions.

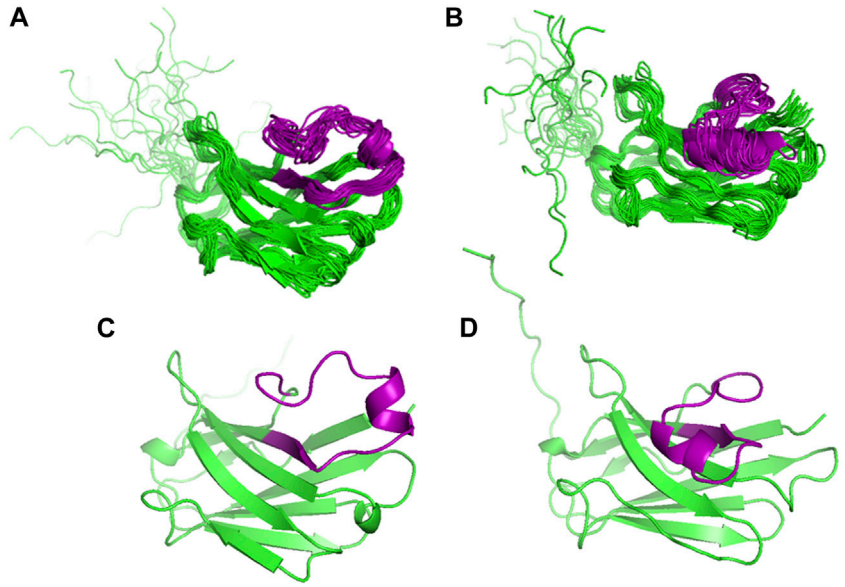


FIGURE 3 | Comparison of the MD derived ensembles of VHH5 Model 1 from the H3 loop generation. **(A)** Twenty structures from 1.4 to 39.8 ns; **(B)** Nineteen structures from 43.2 to 80.0 ns of the simulation time. The H3 loop conformations obtained from the MD simulations; **(C)** open conformation; **(D)** closed conformation. The H3 loop is colored in purple.

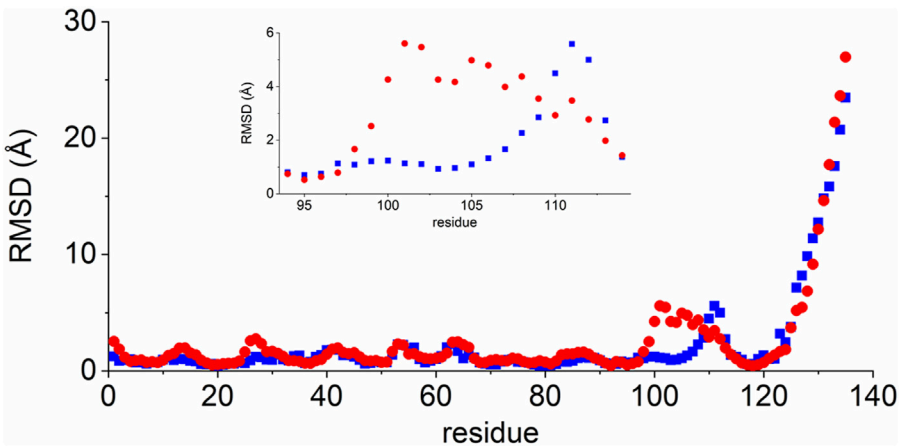


FIGURE 4 | RMSD values of VHH5 Model 1 after the H3 loop generation for the two loop conformations. The RMSD was calculated for the CA, C′, and N backbone atoms of each residue. *Blue rectangles*: open conformation, 1.4–39.8 ns; *Red dots*: closed conformation, 43.2–80.0 ns. RMSD values for the residues in the H3 loop region are shown in the insert.

The G-factors on dihedral angles, that provide a measure of how unusual, or out-of-the-ordinary, a property is, were all above the −0.5 threshold or positive, and indicating good quality. The overall average value was −0.14.

Structure-Guided Optimization of VHH5 Expression

We used the predicted structures to analyse the protein surface and identify exposed hydrophobic residues not contributing to the hydrophobic cores or to the CDR loops that could be mutated to reduce the risk of the proteins to be in inclusion bodies. We both visually inspected the models and analysed the coordinates with the DSSP software which provides per residue accessible surface areas. As the result of this analysis, we found that the regions that could mostly promote aggregation could be H3 which is indeed rather hydrophobic with four bulky hydrophobic residues and two uncharged aromatics. This region cannot however be mutated as it may be essential for epitope recognition. Additionally, we found a few exposed

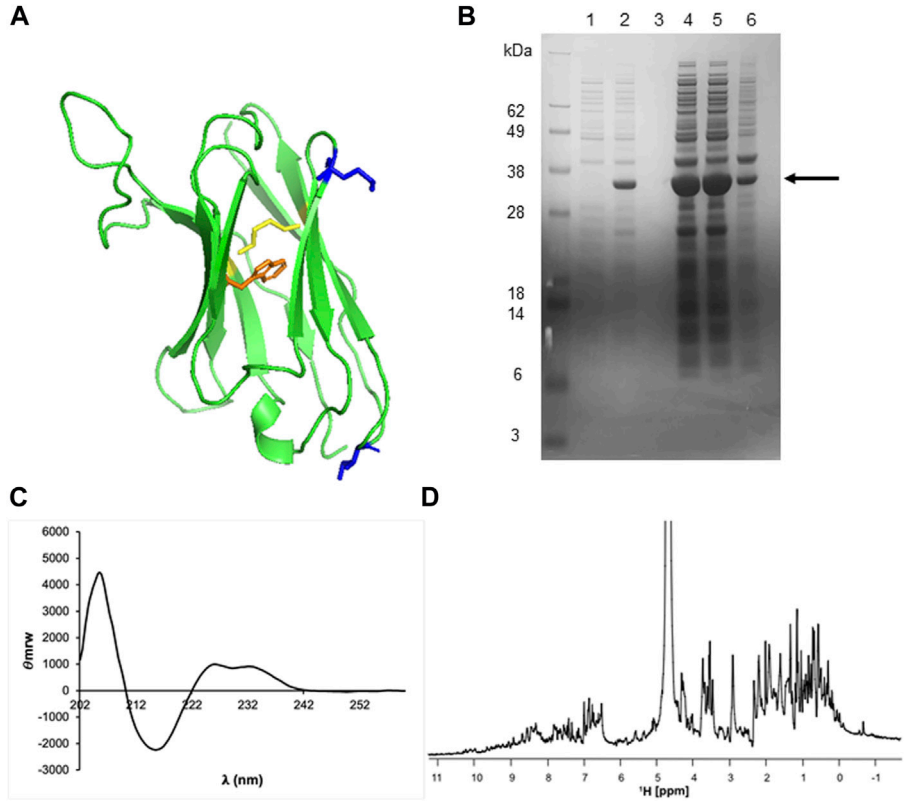


FIGURE 5 | Structural analysis, production, and characterization of VHH5. **(A)** VHH5 structure. The disulfide bond is highlighted in yellow and the side chain of the tryptophan is in orange. The two hydrophobic residues, I15 and M74, that were hypothesized to help inclusion body formation are highlighted in blue. **(B)** Overexpression of VHH5 in *E. Coli* BL21 (DE3)pLysS cells as a soluble protein. SDS-PAGE analysis of SUMO + VHH5 (29 kDa) shows the soluble protein and a high overexpression. The columns correspond to: lane 1, pre-induction; lane 2, after induction with IPTG; lane 3, pre-lysis supernatant; lane 4, pre-lysis pellet; lane 5, post-lysis supernatant; lane 6, post-lysis pellet. **(C)** CD and **(D)** ^{1}H NMR spectra of VHH5 recorded at room temperature.

hydrophobic residues such as I15 and M74 that could potentially interfere with protein folding leading to inclusion bodies (**Figure 5A**). We thus decided to mutate I15 to alanine and M75 to lysine creating the double mutant VHH5-I15A_M75K and attempted to express this mutant in *E. coli*.

We found that protein production switched from being all in the inclusion bodies to being mostly soluble **(Figure 5B)**. This strategy allowed us to obtain suitable quantities of VHH5-I15A_M75K. After purification, we managed to typically obtain ca. 13 ml (1.96 mg/ml or 132 μM) of >98% pure protein after cleaving it from the tag. The protein identity was confirmed by mass spectrometry which also confirmed disulfide formation (data not shown). We also confirmed the state of fold by far-UV circular dichroism (CD), a technique able to detect the secondary structure of proteins. The CD spectrum of VHH5- I15A_M75K recorded at room temperature has a maximum at 205 nm and a single minimum around 215 nm which are features typical of the β-sheet conformation expected for an antibody (**Figure 5C**). The positive contribution at 225–235 nm is usually diagnostic of the presence of stacking interactions between aromatic residues (Budyak et al., 2013). The mono-dimensional NMR spectrum of the unlabelled protein presented well dispersed resonances as expected for a folded monomeric protein of the size of VHH5 (**Figure 5D**). We thus concluded that the protein obtained was folded and well-behaved.

Epitope Mapping

To characterize the epitope of TDP-43 recognized by VHH5, we first performed *In Vivo* Epitope Mapping (IVEM) in yeast (Visintin et al., 2002) by truncating the original LexA-TDP-43 bait into two fragments, LexA-N-term + RRM1-2 (residues 1–258) and LexA-C-term (residues 259–414). The epitope recognized by the VHH5 resulted to be located in the N-terminal half of the protein. To further narrow the region carrying the epitope, a second IVEM was carried out by splitting this region into four smaller baits containing the N-terminus (1–105), RRM1 (106–176), RRM2 (192–258), and a fragment of RRMs (160–208). The epitope seemed to be mainly located in RRM2, since growth on SD-WHL plates was detected both with the LexA-RRM2 and the RRMs baits (**Figure 6A**). To substantiate these results with further evidence, we used the purified recombinant VHH5- I15A_M75K for ELISA experiments. We performed both sandwich and indirect ELISA assay using a rabbit anti-TDP-43 polyclonal antibody (Proteintech) and a

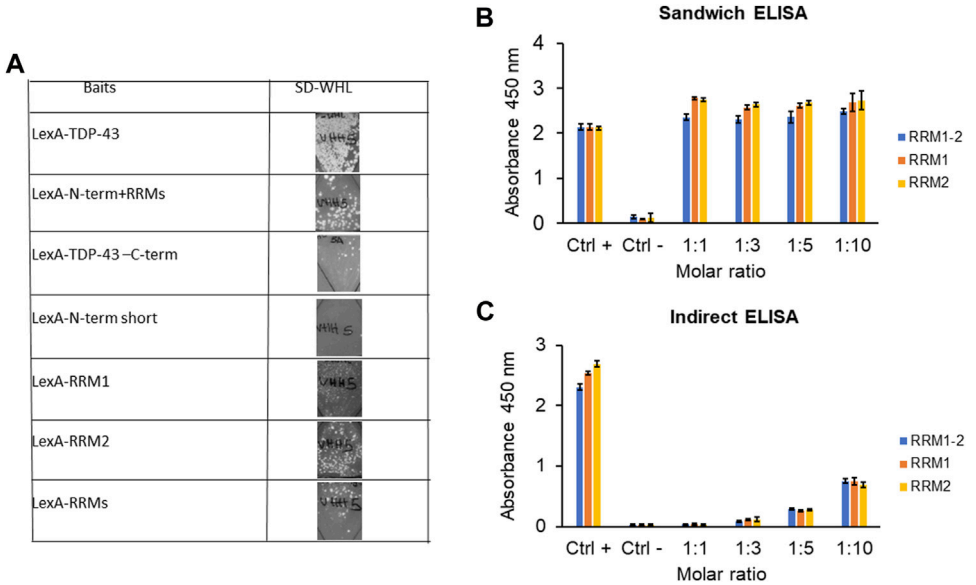


FIGURE 6 | Epitope mapping of VHH5 on TDP-43. **(A)** *In vivo* Epitope Mapping. The VHH5 was transfected in L40 yeast strain expressing the baits LexA-TDP-43 full length (residues 1–414), LexA- N-Term + RRM1-2 (1–258), N-term short (1–105), RRM1 (106–176), RRM2 (192–258), and RRM1-2 (160–208). Interaction is detected by growth on–WHL plates. **(B)** Sandwich ELISA assay. Coating antibody: VHH5 (final molar ratio coating antibody: binding antigen 1:1, 1:3, 1:5, and 1:10); Binding antigen: RRMs, RRM1, and RRM2 (1 μM); Detection: anti-TARDBP and then anti-hIgG-HRP. The assay shows an interaction of VHH5 with all the TDP-43 fragments. **(C)** Indirect ELISA assay. Coating antigen: RRM1-2, RRM1, and RRM2 (1 μM); Binding antibody: VHH5 (molar ratio 1:1, 1:3, 1:5, and 1:10); Detection: anti-VHH-HRP. The assay shows an interaction of VHH5 with all the TDP-43 fragments. The interaction increases as the molar ratio increases.

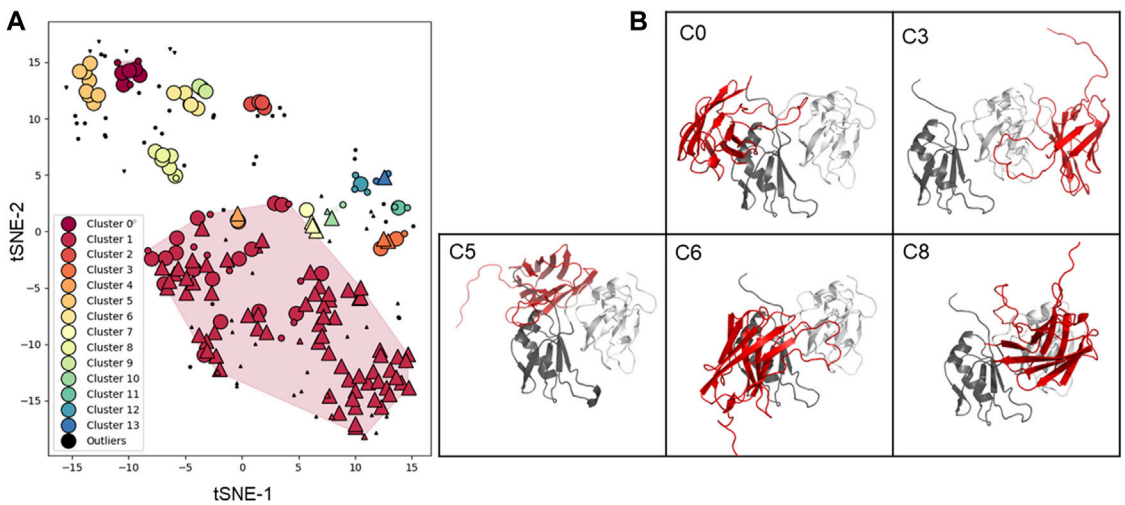


FIGURE 7 | Clustering and structure of the docking solutions. **(A)** Clustering is represented as a 2D map that preserves local similarity. Each dot corresponds to a docking solution, and is coloured according to the cluster it belongs to. Dots depicted as upward triangles, downward triangles, and circles represent solutions where the antibody interacts with the first (RRM1), second (RRM2), and or both antigen domains (RRM1-2), respectively. Solutions depicted in black are considered outliers by the clustering algorithm, small dots, and large dots are core and reachable elements, respectively. **(B)** Representative solutions from clusters with more than five elements, excluding cluster 1. The antibody is represented in red, and the first and second antigen domains in dark grey and white, respectively.

rabbit anti-camelid VHH [HRP] antibody (GenScript) respectively. In both cases, we observed response to RRM1, RRM2, and RRM1-2, indicating that the epitope involves both domains (**Figures 6B,C**). This result could mean that VHH5 recognises each of the repeats which share some homology. However, while the homology is fairly high, and the sequence identity is only 26%. It is thus fairly unlikely that there are two independent epitopes one in each repeat. It is more likely that the epitope is conformational and involves both domains. We also noticed that only the indirect ELISA

showed a dependence on the antibody to protein ratios. This could be explained by considering that the difference between the two assays is that in the indirect ELISA, the target protein is fixed and the intrabody is added at increasing concentrations. No concentration dependence in the latter assay could easily be explained by the assumption that when the intrabody is fixed it could adopt a conformation that makes it more competent for binding. Viceversa, when the target protein is fixed, the epitope may be partially masked. This means that the detected affinity can be different in the two cases. Thus, the signal can appear saturated in **Figure 6B** but not in the indirect ELISA done with the intrabody in solution.

Using this information, we then performed molecular docking. Although docking carried out on low resolution structures and without experimental restraints has only very limited reliability, we reasoned that it could provide a visual impression of epitope binding and inform future studies. Models of the antigen-antibody complexes were generated by the ClusPro software using each of the ten energetically best Sphinx structures and the NMR structure of the putative antigen (PDB 4bs2). This calculation resulted in 228 models which were further analysed. After the filtering procedure described in the Materials and Methods section, a total of 14 clusters were identified (**Figure 7A**). The complex structures with the lowest score and binding free energy were selected and analysed (**Figure 7B**). Cluster 1 contains the vast majority of the solutions, in which the antibody only interacts with a single domain of the antigen. However, upon closer inspections, we realised that these solutions were likely the result of an artefact of the docking procedure: the H3 loop of the antibody would encircle the C-terminus of the antigen, in a configuration that would result in a knot or a lasso in the complete antigen. Excluding these solutions, cluster 0, 2, 5, 6, 8, 9, 11, 12, and 13 mainly contained solutions in which the interaction involved both domains. In total, 51 out of the 61 solutions that were not outliers nor part of cluster 1, and contained interactions to both domains (**Figure 7B**). These models, that are only indicative and low resolution, will need experimental validation through fine epitope-mapping at the level of the individual residues.

DISCUSSION

The use of antibodies in misfolding diseases is in principle a flexible and ductile strategy to control protein aggregation, because, by binding to a monomeric protein, they prevent self-assembly by steric hindrance. There are now several different strategies that allow screening (Hanes and Pluckthun, 1997; Smith and Petrenko, 1997; Ho and Pastan, 2009; Uchanski et al., 2019), ab initio design (Hardin et al., 2002; Zhu and Day, 2013) or evolutionary selection of antibodies, and smaller derivatives (Visintin et al., 2002). A problem remains however the production of the antibody by bacterial expression once a potentially effective sequence has been identified. Unfortunately, the large molecular weight (typically ~150,000) and hetero-tetrameric composition of antibodies with two different polypeptides (a heavy and a light chain) and a total of up to 15 disulfide bridges make difficult when not prohibitive their production in bacteria or in the cytoplasm of eukaryotic cells. This is why scFv fragments, that contain only one copy of the variable domains of immunoglobulin motif, offer undiscussable advantages. However, also in this case, it is difficult to predict *a priori* whether an intrabody obtained by library screening can easily be produced in *E. coli*, and problems in successfully refolding the intrabody from inclusion bodies have been described (Vaks and Benhar, 2014; Bao et al., 2016).

In the present study, we used a composite approach in which we screened an intrabody for TDP-43 recognition, and produced it in bacteria and characterised it for epitope recognition. We first described a new naïve library of llama VHHs, and exploited it to select a new anti-TDP-43 VHH directly from the TDP-43 cDNA. A significant advantage of SPLINT-derived antibodies, as the anti-TDP-43 VHH5 described here, is that the genes coding for the antibody domains are by definition well validated as intrabodies, since the IACT selection is performed under conditions of intracellular expression in yeast cells. SPLINT-derived antibody domains are well suited to be used as intrabodies (Biocca et al., 1990), possibly coupled to effector domains for targeted degradation (Melchionna and Cattaneo, 2007; Schapira et al., 2019) or for imaging purposes.

We then modelled the structure of the intrabody to get a visual impression of its structure. The model suggested exposed hydrophobic residues that could be mutated to reduce the risk of inclusion body formation. We found that it was sufficient to mutate two exposed hydrophobic residues to have a soluble protein that could be purified in suitable amounts for proper direct characterization. We demonstrated by CD and NMR studies that the protein is folded and monomeric and that has all the features expected for the expected β-rich structure. We then demonstrated by ELISA experiments that the double mutant is still able to recognise the TDP-43 epitope. This conclusion was far from being obvious, since it is known that regions outside the CDR loops can contribute to epitope recognition of intrabodies (Sela-Culang et al., 2013). We mapped the epitope binding regions first coarsely by *in vivo* epitope mapping and then, more specifically, and by ELISA experiments with individual or tandem domains of TDP-43. We found that the anti-TDP-43 VHH5 intrabody binds both RRM1 and RRM2. This is in agreement with structural studies that have revealed that VHHs often tend to recognize concave surfaces of their antigens with high shape-complementarity. Based on these experimental findings, we modelled the interaction by *in silico* docking. Despite their overall diversity, in most of the solution we found the long H3 of VHH5 protruding out from the body of the antibody and docks into the cleft formed by the interface between the two domains. This arrangement would permit recognition of the antigen with high shape complementarity. A similar type of recognition has been described in a structural study that compared the binding mode of VHH with that of Fvs using hen egg lysozyme (HEL) as a model antigen (Akiba et al., 2019). Several more studies have also revealed that VHHs usually target concave surfaces on the antigen molecule (Kromann-Hansen et al., 2016; Rossey et al., 2017; Gulati et al., 2018). It is believed

that in this way, VHHs compensate for the limitations of their small size, while maintaining the high affinity and specificity that constitute the hallmarks of antibodies.

It is interesting to compare our intrabody with previously developed anti-TDP-43 antibodies. A systematic survey in 2015 revealed the existence of 29 antibodies, many of which were generated in house (Goossens et al., 2015). Amongst the ten highest-ranking primary antibodies, one has two distinct epitopes, that recognize TDP-43 N-terminus and RRM2. Two other antibodies are directed at RRM2, and three have epitopes in the C-terminus of TDP-43. The remaining four antibodies also map in the C-terminus but are specific for phosphorylated serine residues. The majority of these antibodies are polyclonal and therefore their genes cannot be available for further downstream engineering. A single chain antibody against RRM1 was generated in 2019 (Pozzi et al., 2019). Two more monoclonal antibodies were recently described that were raised against an epitope within the RRM2 domain of TDP-43 (residues 198–216) (Trejo-Lopez et al., 2020).

The novel intrabody will aid in diagnostic and research efforts within the context of TDP-43 proteinopathies. Availability of this intrabody opens new avenues to the diagnosis and treatment of ALS.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Written informed consent was obtained from the veterinarian (Dr Sara Piga) of the Biopark Zoom (Cumiana, Turin, and Italy) for the participation of the animals in this study.

AUTHOR CONTRIBUTIONS

MG did most of the molecular biology and wrote the first draft, EFD was responsible for the MD simulations, RP and AL helped with the biophysics and protein production respectively. SL and MF made and characterized the SPLINT selection, SL and MG performed the selection of anti TDP-43 antibodies, AC conceived the library, AP and AC supervised the research and acquired resources. AP conceived the project and wrote the final version of the manuscript. All authors contributed with comments and criticisms.

FUNDING

The research was funded by the United Kingdom DRI funding scheme (grant REI 3556) and Alzheimer United Kingdom (grant ARUK-PG2019B-020) to AP and by the Human Brain Project EU Flasghip (grant No. 604102) to AC. The authors acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk/), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London and Maudsley and Guy's and St. Thomas' NHS Foundation Trusts, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's and St. Thomas' Charity (TR130505). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care.

ACKNOWLEDGMENTS

The authors are grateful to Tamás Földes for his helpful advice for the MD simulations. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. AC and SL gratefully acknowledge the contribution of Martina Goracci and Ottavia Vitaloni to the construction and characterization of the llama nanobody library.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.773234/full#supplementary-material

REFERENCES

Akiba, H., Tamura, H., Kiyoshi, M., Yanaka, S., Sugase, K., Caaveiro, J. M. M., et al. (2019). Structural and Thermodynamic Basis for the Recognition of the Substrate-Binding Cleft on Hen Egg Lysozyme by a Single-Domain Antibody. *Sci. Rep.* 9 (1), 15481. doi:10.1038/s41598-019-50722-y

Al-Lazikani, B., Lesk, A. M., and Chothia, C. (1997). Standard Conformations for the Canonical Structures of Immunoglobulins 1 1Edited by I. A. Wilson. *J. Mol. Biol.* 273 (4), 927–948. doi:10.1006/jmbi.1997.1354

Ayala, Y. M., Zago, P., D'Ambrogio, A., Xu, Y. F., Petrucelli, L., Buratti, E., et al. (2008). Structural Determinants of the Cellular Localization and Shuttling of TDP-43. *J. Cell Sci* 121 (Pt 22), 3778–3785. doi:10.1242/jcs.038950

Bao, X., Xu, L., Lu, X., and Jia, L. (2016). Optimization of Dilution Refolding Conditions for a Camelid Single Domain Antibody against Human Beta-2-Microglobulin. *Protein Expr. Purif.* 117, 59–66. doi:10.1016/j.pep.2015.09.019

Barmada, S. J., Skibinski, G., Korb, E., Rao, E. J., Wu, J. Y., and Finkbeiner, S. (2010). Cytoplasmic Mislocalization of TDP-43 Is Toxic to Neurons and Enhanced by a Mutation Associated with Familial Amyotrophic Lateral Sclerosis. *J. Neurosci.* 30 (2), 639–649. doi:10.1523/jneurosci.4988-09.2010

Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., et al. (2013). OpenStructure: an Integrated Software Framework for Computational Structural Biology. *Acta Crystallogr. D Biol. Crystallogr.* 69 (Pt 5), 701–709. doi:10.1107/S0907444913007051

Biocca, S., Neuberger, M. S., and Cattaneo, A. (1990). Expression and Targeting of Intracellular Antibodies in Mammalian Cells. *EMBO J.* 9 (1), 101–108. doi:10.1002/j.1460-2075.1990.tb08085.x

Bird, R. E., Hardman, K. D., Jacobson, J. W., Johnson, S., Kaufman, B. M., Lee, S.-M., et al. (1988). Single-chain Antigen-Binding Proteins. *Science* 242 (4877), 423–426. doi:10.1126/science.3140379

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Budini, M., Romano, V., Quadri, Z., Buratti, E., and Baralle, F. E. (2015). TDP-43 Loss of Cellular Function through Aggregation Requires Additional Structural Determinants beyond its C-Terminal Q/N Prion-like Domain. *Hum. Mol. Genet.* 24 (1), 9–20. doi:10.1093/hmg/ddu415

Budyak, I. L., Zhuravleva, A., and Gierasch, L. M. (2013). The Role of Aromatic-Aromatic Interactions in Strand-Strand Stabilization of β-Sheets. *J. Mol. Biol.* 425 (18), 3522–3535. doi:10.1016/j.jmb.2013.06.030

Buratti, E., and Baralle, F. E. (2009). Chapter 1 the Molecular Links between TDP-43 Dysfunction and Neurodegeneration. *Adv. Genet.* 66, 1–34. doi:10.1016/s0065-2660(09)66001-6

Buratti, E., and Baralle, F. E. (2001). Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator ofCFTR Exon 9. *J. Biol. Chem.* 276 (39), 36337–36343. doi:10.1074/jbc.m104236200

Buratti, E., and Baralle, F. E. (2008). Multiple Roles of TDP-43 in Gene Expression, Splicing Regulation, and Human Disease. *Front. Biosci.* 13, 867–878. doi:10.2741/2727

Burrell, J. R., Halliday, G. M., Kril, J. J., Ittner, L. M., Götz, J., Kiernan, M. C., et al. (2016). The Frontotemporal Dementia-Motor Neuron Disease Continuum. *The Lancet* 388 (10047), 919–931. doi:10.1016/s0140-6736(16)00737-6

Cattaneo, A., and Chirichella, M. (2019). Targeting the Post-translational Proteome with Intrabodies. *Trends Biotechnol.* 37 (6), 578–591. doi:10.1016/j.tibtech.2018.11.009

Chen, H.-J., Topp, S. D., Hui, H. S., Zacco, E., Katarya, M., McLoughlin, C., et al. (2019). RRM Adjacent TARDBP Mutations Disrupt RNA Binding and Enhance TDP-43 Proteinopathy. *Brain* 142 (12), 3753–3770. doi:10.1093/brain/awz313

Choi, Y., and Deane, C. M. (2010). FREAD Revisited: Accurate Loop Structure Prediction Using a Database Search Algorithm. *Proteins* 78 (6), 1431–1440. doi:10.1002/prot.22658

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., et al. (1989). Conformations of Immunoglobulin Hypervariable Regions. *Nature* 342 (6252), 877–883. doi:10.1038/342877a0

Cohen, T. J., Lee, V. M. Y., and Trojanowski, J. Q. (2011). TDP-43 Functions and Pathogenic Mechanisms Implicated in TDP-43 Proteinopathies. *Trends Mol. Med.* 17 (11), 659–667. doi:10.1016/j.molmed.2011.06.004

Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN: a Server for the Prediction and Evaluation of "hot Spots" of Aggregation in Polypeptides. *BMC Bioinformatics* 8, 65. doi:10.1186/1471-2105-8-65

Deane, C. M., and Blundell, T. L. (2001). CODA: a Combined Algorithm for Predicting the Structurally Variable Regions of Protein Models. *Protein Sci.* 10 (3), 599–612. doi:10.1110/ps.37601

Devenney, E., Vucic, S., Hodges, J. R., and Kiernan, M. C. (2015). Motor Neuron Disease-Frontotemporal Dementia: a Clinical Continuum. *Expert Rev. Neurotherapeutics* 15 (5), 509–522. doi:10.1586/14737175.2015.1034108

Dong, G. Q., Fan, H., Schneidman-Duhovny, D., Webb, B., and Sali, A. (2013). Optimized Atomic Statistical Potentials: Assessment of Protein Interfaces and Loops. *Bioinformatics* 29 (24), 3158–3166. doi:10.1093/bioinformatics/btt560

Elsaesser, R., and Paysan, J. (2004). Liquid Gel Amplification of Complex Plasmid Libraries. *Biotechniques* 37 (2), 200202–202. doi:10.2144/04372bm04

Fantini, M., Pandolfini, L., Lisi, S., Chirichella, M., Arisi, I., Terrigno, M., et al. (2017). Assessment of Antibody Library Diversity through Next Generation Sequencing and Technical Error Compensation. *PLoS One* 12 (5), e0177574. doi:10.1371/journal.pone.0177574

Gao, J., Wang, L., Huntley, M. L., Perry, G., and Wang, X. (2018). Pathomechanisms of TDP-43 in Neurodegeneration. *J. Neurochem.* doi:10.1111/jnc.14327

Goossens, J., Vanmechelen, E., Trojanowski, J. Q., Lee, V. M., Van Broeckhoven, C., van der Zee, J., et al. (2015). TDP-43 as a Possible Biomarker for Frontotemporal Lobar Degeneration: a Systematic Review of Existing Antibodies. *Acta Neuropathol. Commun.* 3, 15. doi:10.1186/s40478-015-0195-1

Gulati, S., Jin, H., Masuho, I., Orban, T., Cai, Y., Pardon, E., et al. (2018). Targeting G Protein-Coupled Receptor Signaling at the G Protein Level with a Selective Nanobody Inhibitor. *Nat. Commun.* 9 (1), 1996. doi:10.1038/s41467-018-04432-0

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, C., Songa, E. B., et al. (1993). Naturally Occurring Antibodies Devoid of Light Chains. *Nature* 363 (6428), 446–448. doi:10.1038/363446a0

Hanes, J., and Pluckthun, A. (1997). *In Vitro* selection and Evolution of Functional Proteins by Using Ribosome Display. *Proc. Natl. Acad. Sci.* 94 (10), 4937–4942. doi:10.1073/pnas.94.10.4937

Hardin, C., Pogorelov, T. V., and Luthey-Schulten, Z. (2002). Ab Initio protein Structure Prediction. *Curr. Opin. Struct. Biol.* 12 (2), 176–181. doi:10.1016/s0959-440x(02)00306-8

Harmsen, M. M., and De Haard, H. J. (2007). Properties, Production, and Applications of Camelid Single-Domain Antibody Fragments. *Appl. Microbiol. Biotechnol.* 77 (1), 13–22. doi:10.1007/s00253-007-1142-2

Ho, M., and Pastan, I. (2009). Mammalian Cell Display for Antibody Engineering. *Methods Mol. Biol.* 525, 337–352. doi:10.1007/978-1-59745-554-1_18

Hoey, R. J., Eom, H., and Horn, J. R. (2019). Structure and Development of Single Domain Antibodies as Modules for Therapeutics and Diagnostics. *Exp. Biol. Med. (Maywood)* 244 (17), 1568–1576. doi:10.1177/1535370219881129

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14 (1), 3327–3388. doi:10.1016/0263-7855(96)00018-5

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: a Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* 29 (11), 1859–1865. doi:10.1002/jcc.20945

Khodabakhsh, F., Behdani, M., Rami, A., and Kazemi-Lomedasht, F. (2018). Single-Domain Antibodies or Nanobodies: A Class of Next-Generation Antibodies. *Int. Rev. Immunol.* 37 (6), 316–322. doi:10.1080/08830185.2018.1526932

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., et al. (2017). The ClusPro Web Server for Protein-Protein Docking. *Nat. Protoc.* 12 (2), 255–278. doi:10.1038/nprot.2016.169

Kromann-Hansen, T., Oldenburg, E., Yung, K. W. Y., Ghassabeh, G. H., Muyldermans, S., Declerck, P. J., et al. (2016). A Camelid-Derived Antibody Fragment Targeting the Active Site of a Serine Protease Balances between Inhibitor and Substrate Behavior. *J. Biol. Chem.* 291 (29), 15156–15168. doi:10.1074/jbc.m116.732503

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for Checking the Quality of Protein Structures Solved by NMR. *J. Biomol. NMR* 8 (4), 477–486. doi:10.1007/BF00228148

Laskowski, R. A. (2001). PDBsum: Summaries and Analyses of PDB Structures. *Nucleic Acids Res.* 29 (1), 221–222. doi:10.1093/nar/29.1.221

Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., et al. (2016). CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theor. Comput.* 12 (1), 405–413. doi:10.1021/acs.jctc.5b00935

Leem, J., Dunbar, J., Georges, G., Shi, J., and Deane, C. M. (2016). ABodyBuilder: Automated Antibody Structure Prediction with Data-Driven Accuracy Estimation. *MAbs* 8 (7), 1259–1268. doi:10.1080/19420862.2016.1205773

Li, L., Chen, S., Miao, Z., Liu, Y., Liu, X., Xiao, Z. X., et al. (2019). AbRSA: A Robust Tool for Antibody Numbering. *Protein Sci.* 28 (8), 1524–1531. doi:10.1002/pro.3633

Liu, E. Y., Cali, C. P., and Lee, E. B. (2017). RNA Metabolism in Neurodegenerative Disease. *Dis. Model. Mech.* 10 (5), 509–518. doi:10.1242/dmm.028613

Lukavsky, P. J., Daujotyte, D., Tollervey, J. R., Ule, J., Stuani, C., Buratti, E., et al. (2013). Molecular Basis of UG-Rich RNA Recognition by the Human Splicing Factor TDP-43. *Nat. Struct. Mol. Biol.* 20 (12), 1443–1449. doi:10.1038/nsmb.2698

Mackenzie, I. R. A., and Neumann, M. (2016). Molecular Neuropathology of Frontotemporal Dementia: Insights into Disease Mechanisms from Postmortem Studies. *J. Neurochem.* 138 (Suppl. 1), 54–70. doi:10.1111/jnc.13588

Marks, C., Nowak, J., Klostermann, S., Georges, G., Dunbar, J., Shi, J., et al. (2017). Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve

Protein Loop Prediction. *Bioinformatics* 33 (9), 1346–1353. doi:10.1093/bioinformatics/btw823

Melchionna, T., and Cattaneo, A. (2007). A Protein Silencing Switch by Ligand-Induced Proteasome-Targeting Intrabodies. *J. Mol. Biol.* 374 (3), 641–654. doi:10.1016/j.jmb.2007.09.053

Meli, G., Lecci, A., Manca, A., Krako, N., Albertini, V., Benussi, L., et al. (2014). Conformational Targeting of Intracellular Aβ Oligomers Demonstrates Their Pathological Oligomerization inside the Endoplasmic Reticulum. *Nat. Commun.* 5, 3867. doi:10.1038/ncomms4867

Meli, G., Visintin, M., Cannistraci, I., and Cattaneo, A. (2009). Direct *In Vivo* Intracellular Selection of Conformation-Sensitive Antibody Domains Targeting Alzheimer's Amyloid-β Oligomers. *J. Mol. Biol.* 387 (3), 584–606. doi:10.1016/j.jmb.2009.01.061

Messer, A., and Butler, D. C. (2020). Optimizing Intracellular Antibodies (Intrabodies/nanobodies) to Treat Neurodegenerative Disorders. *Neurobiol. Dis.* 134, 104619. doi:10.1016/j.nbd.2019.104619

Messer, A., and Joshi, S. N. (2013). Intrabodies as Neuroprotective Therapeutics. *Neurotherapeutics* 10 (3), 447–458. doi:10.1007/s13311-013-0193-6

Möckli, N., and Auerbach, D. (2004). Quantitative β-galactosidase Assay Suitable for High-Throughput Applications in the Yeast Two-Hybrid System. *Biotechniques* 36 (5), 872–876. doi:10.2144/04365pt03

Mompeán, M., Romano, V., Pantoja-Uceda, D., Stuani, C., Baralle, F. E., Buratti, E., et al. (2016). The TDP-43 N-Terminal Domain Structure at High Resolution. *FEBS J.* 283 (7), 1242–1260. doi:10.1111/febs.13651

Narciso, J. E. T., Uy, I. D. C., Cabang, A. B., Chavez, J. F. C., Pablo, J. L. B., Padilla-Concepcion, G. P., et al. (2011). Analysis of the Antibody Structure Based on High-Resolution Crystallographic Studies. *New Biotechnol.* 28 (5), 435–447. doi:10.1016/j.nbt.2011.03.012

Oi, C., Mochrie, S. G. J., Horrocks, M. H., and Regan, L. (2020). PAINT Using Proteins: A New brush for Super-resolution Artists. *Protein Sci.* 29 (11), 2142–2149. doi:10.1002/pro.3953

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Pesiridis, G. S., Lee, V. M.-Y., and Trojanowski, J. Q. (2009). Mutations in TDP-43 Link Glycine-Rich Domain Functions to Amyotrophic Lateral Sclerosis. *Hum. Mol. Genet.* 18 (R2), R156–R162. doi:10.1093/hmg/ddp303

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* 153 (4), 044130. doi:10.1063/5.0014475

Pozzi, S., Thammisetty, S. S., Codron, P., Rahimian, R., Plourde, K. V., Soucy, G., et al. (2019). Virus-mediated Delivery of Antibody Targeting TAR DNA-Binding Protein-43 Mitigates Associated Neuropathology. *J. Clin. Invest.* 129 (4), 1581–1595. doi:10.1172/jci123931

Prasad, A., Bharathi, V., Sivalingam, V., Girdhar, A., and Patel, B. K. (2019). Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis. *Front. Mol. Neurosci.* 12, 25. doi:10.3389/fnmol.2019.00025

Rossey, I., Gilman, M. S. A., Kabeche, S. C., Sedeyn, K., Wrapp, D., Kanekiyo, M., et al. (2017). Potent Single-Domain Antibodies that Arrest Respiratory Syncytial Virus Fusion Protein in its Prefusion State. *Nat. Commun.* 8, 14158. doi:10.1038/ncomms14158

Saerens, D., Pellis, M., Loris, R., Pardon, E., Dumoulin, M., Matagne, A., et al. (2005). Identification of a Universal VHH Framework to Graft Non-canonical Antigen-Binding Loops of Camel Single-Domain Antibodies. *J. Mol. Biol.* 352 (3), 597–607. doi:10.1016/j.jmb.2005.07.038

Schapira, M., Calabrese, M. F., Bullock, A. N., and Crews, C. M. (2019). Targeted Protein Degradation: Expanding the Toolbox. *Nat. Rev. Drug Discov.* 18 (12), 949–963. doi:10.1038/s41573-019-0047-y

Schermelleh, L., Ferrand, A., Huser, T., Eggeling, C., Sauer, M., Biehlmaier, O., et al. (2019). Super-resolution Microscopy Demystified. *Nat. Cell Biol* 21 (1), 72–84. doi:10.1038/s41556-018-0251-8

Sela-Culang, I., Kunik, V., and Ofran, Y. (2013). The Structural Basis of Antibody-Antigen Recognition. *Front. Immunol.* 4, 302. doi:10.3389/fimmu.2013.00302

Smith, G. P., and Petrenko, V. A. (1997). Phage Display. *Chem. Rev.* 97 (2), 391–410. doi:10.1021/cr960065d

Sograte-Idrissi, S., Oleksiievets, N., Isbaner, S., Eggert-Martinez, M., Enderlein, J., Tsukanov, R., et al. (2019). Nanobody Detection of Standard Fluorescent Proteins Enables Multi-Target DNA-PAINT with High Resolution and Minimal Displacement Errors. *Cells* 8 (1). doi:10.3390/cells8010048

Suk, T. R., and Rousseaux, M. W. C. (2020). The Role of TDP-43 Mislocalization in Amyotrophic Lateral Sclerosis. *Mol. Neurodegeneration* 15 (1), 45. doi:10.1186/s13024-020-00397-1

Trejo-Lopez, J. A., Sorrentino, Z. A., Riffe, C. J., Lloyd, G. M., Labuzan, S. A., Dickson, D. W., et al. (2020). Novel Monoclonal Antibodies Targeting the RRM2 Domain of Human TDP-43 Protein. *Neurosci. Lett.* 738, 135353. doi:10.1016/j.neulet.2020.135353

Uchanski, T., Zögg, T., Yin, J., Yuan, D., Wohlkönig, A., et al. (2019). An Improved Yeast Surface Display Platform for the Screening of Nanobody Immune Libraries. *Sci. Rep.* 9 (1), 382. doi:10.1038/s41598-018-37212-3

Vaks, L., and Benhar, I. (2014). Production of Stabilized scFv Antibody Fragments in the *E. coli* Bacterial Cytoplasm. *Methods Mol. Biol.* 1060, 171–184. doi:10.1007/978-1-62703-586-6_10

van der Linden, R., de Geus, B., Stok, W., Bos, W., van Wassenaar, D., Verrips, T., et al. (2000). Induction of Immune Responses and Molecular Cloning of the Heavy Chain Antibody Repertoire of *Lama glama*. *J. Immunol. Methods* 240 (1-2), 185–195. doi:10.1016/s0022-1759(00)00188-5

Van Der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-SNE. *J. Machine Learn. Res.* 9, 2579–2625.

Ventura, S., and Villaverde, A. (2006). Protein Quality in Bacterial Inclusion Bodies. *Trends Biotechnol.* 24 (4), 179–185. doi:10.1016/j.tibtech.2006.02.007

Visintin, M., Meli, G. A., Cannistraci, I., and Cattaneo, A. (2004). Intracellular Antibodies for Proteomics. *J. Immunol. Methods* 290 (1-2), 135–153. doi:10.1016/j.jim.2004.04.014

Visintin, M., Settanni, G., Maritan, A., Graziosi, S., Marks, J. D., and Cattaneo, A. (2002). The Intracellular Antibody Capture Technology (IACT): towards a Consensus Sequence for Intracellular Antibodies. *J. Mol. Biol.* 317 (1), 73–83. doi:10.1006/jmbi.2002.5392

Visintin, M., Tse, E., Axelson, H., Rabbitts, T. H., and Cattaneo, A. (1999). Selection of Antibodies for Intracellular Function Using a Two-Hybrid *In Vivo* System. *Proc. Natl. Acad. Sci.* 96 (21), 11723–11728. doi:10.1073/pnas.96.21.11723

Ward, E. S., Güssow, D., Griffiths, A. D., Jones, P. T., and Winter, G. (1989). Binding Activities of a Repertoire of Single Immunoglobulin Variable Domains Secreted from *Escherichia coli*. *Nature* 341 (6242), 544–546. doi:10.1038/341544a0

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* 46 (W1), W296–W303. doi:10.1093/nar/gky427

Winton, M. J., Igaz, L. M., Wong, M. M., Kwong, L. K., Trojanowski, J. Q., and Lee, V. M.-Y. (2008). Disturbance of Nuclear and Cytoplasmic TAR DNA-Binding Protein (TDP-43) Induces Disease-like Redistribution, Sequestration, and Aggregate Formation. *J. Biol. Chem.* 283 (19), 13302–13309. doi:10.1074/jbc.m800342200

Zacco, E., Graña-Montes, R., Martin, S. R., de Groot, N. S., Alfano, C., Tartaglia, G. G., et al. (2019). RNA as a Key Factor in Driving or Preventing Self-Assembly of the TAR DNA-Binding Protein 43. *J. Mol. Biol.* 431 (8), 1671–1688. doi:10.1016/j.jmb.2019.01.028

Zhu, K., and Day, T. (2013). Ab Initiostructure Prediction of the Antibody Hypervariable H3 Loop. *Proteins* 81 (6), 1081–1089. doi:10.1002/prot.24240

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

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership