Received: 2 May 2023 | Revised: 10 November 2023 | Accepted: 22 November 2023

DOI: 10.1002/prot.26650

PROTEINS WILEY

RESEARCH ARTICLE

Comparison of structural networks across homologous proteins

Vasam Manjveekar Prabantu | **Vasundhara Gadiyaram** | **Saraswathi Vishveshwara** | **Narayanaswamy Srinivasan**

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

Correspondence
Saraswathi Vishveshwara, Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.
Email: saraswathi@iisc.ac.in

Funding information
Department of Science and Technology; University grants commission; Department of Biotechnology; CSIR-RA fellowship

Abstract

Protein sequence determines its structure and function. The indirect relationship between protein function and structure lies deep-rooted in the structural topology that has evolved into performing optimal function. The evolution of structure and its interconnectivity has been conventionally studied by comparing the root means square deviation between protein structures at the backbone level. Two factors that are necessary for the quantitative comparison of non-covalent interactions are (a) explicit inclusion of the coordinates of side-chain atoms and (b) consideration of multiple structures from the conformational landscape to account for structural variability. We have recently addressed these fundamental issues by investigating the alteration of inter-residue interactions across an ensemble of protein structure networks through a graph spectral approach. In this study, we have developed a rigorous method to compare the structure networks of homologous proteins, with a wide range of sequence identity percentages. A range of dissimilarity measures that show the extent of change in the network across homologous structures are generated, which also includes the comparison of the protein structure variability. We discuss in detail, scenarios where the variation of structure is not accompanied by loss or gain of the overall network and its vice versa. The sequence-based phylogeny among the homologs is also compared with the lineage obtained from information from such a robust structure comparison. In summary, we can obtain a quantitative comparison score for the structure networks of homologous proteins, which also enables us to study the evolution of protein function based on the variation of their topologies.

KEYWORDS
homologous protein structure comparison, phylogeny, protein structural networks, structural variability, total network dissimilarity

1 | INTRODUCTION

The recognition of protein sequence-structure relationship by Anfinsen[1] is a landmark in structural biology. Similarly, the characterization

Abbreviations: PSN, Protein Structural Network; MC, Multiple Conformers of the same protein; HC, Homologous Conformers across homologous proteins; GDT, Global Distance Test total score; NDS, Network Dissimilarity Score; TND, Total Network Dissimilarity; RMSD, Root Mean Square Deviation; PCC, Pearson Correlation Coefficient.

of non-covalent interactions responsible for the secondary, super-secondary structures by GN Ramachandran,[2] is yet another milestone in exploring the 3-D structures of proteins. An exploration of structural and functional space by Cyrus Chothia, through the classification of protein families[3] has not only provided evolutionary insights but also has shown a remarkable limit of conformational space for folded proteins. Furthermore, the relationship between the evolution of sequences and function is being actively investigated to commensurate with the big leap in data and the development of newer methods of investigation.[4]

Currently, one of the approaches for protein structure data analysis is to focus on an ensemble of structural states that the protein structure populates,[5,6] rather than a single static 3-dimensional state. Thus, a paradigm shift has been made from single conformations to a conformational space.[7,8] Characterizing the protein ensemble has been an important step in determining the dynamics and functional aspects of proteins.[6] It is crucial in studying the activity and regulation of a protein which in turn impacts other fields such as biotechnology and drug design.[9,10]

The Protein Data Bank (PDB)[11] is a consortium that maintains a database of all submitted structures, obtained using protein structure determination experiments. X-ray crystallography, NMR, and cryo-EM are the most popular methods used to determine the structure of a protein. NMR can easily reveal the conformational fluctuations of a protein in a solution. However, using the x-ray crystallography method an abundant resource of atomic resolution protein structures that can describe their conformational space of proteins are made available on PDB. The method has been employed in capturing multiple substrates based on their inherent ability to form crystals or due to modifications induced in the protein to facilitate the crystal to form. Hence there is a certain bias in the composition of protein structures in the PDB[12,13] especially those of membrane proteins and structurally disordered proteins which are largely underrepresented.

From the structural coordinates of the multiple substates that a protein is able to populate, it is understood that protein structures are far better preserved than their sequence.[14,15] There are several reasons for this to occur but most importantly it is for the preservation of protein function. Understanding this structural plasticity in structures is the key to learning about the function and mechanisms of the protein. The alteration of protein conformation can be caused due to perturbations such as binding to different partners or even due to thermodynamic fluctuations. Crystallographic artifacts such as crystal lattice, pH, molecular packing, temperature, pressure, and crystallization conditions also contribute to variations in the structures. All these factors are not considered when representative states of a protein are used during the comparison of homologous proteins. Multiple structures of the same protein can have more variability than structural variations observed across a homolog with different sequences.[16]

The comparison of homologous protein structures is typically performed using structure deviation tools such as the root mean square deviation (RMSD). The backbone structure of one protein is superposed with the backbone of the homolog to find the deviation in Cα positions. The availability of high-resolution protein structures in abundance now allows us to look at the topology of the protein at the atomistic level. It is necessary to update to newer methods of protein structural analysis that are more pertinent in capturing atomistic details, including the side chain atoms. Protein Structural Networks (PSNs) are a node-edge representation of the protein structures where the residues are considered as nodes and an edge is drawn when there is a relationship that exists between them, the strength of which is judged based on factors like interaction energy and proximity.[17,18] This method has been advantageous over mere structural coordinates for its ease of mathematical comparison and analysis.[19]

The use of PSNs has been increasingly helpful in the field of protein structural analysis.[20,21] It is possible to use advanced graph-spectral-based methods for the comparison of proteins over conventional methods such as a direct comparison of structural topologies.[22–25] From recent work, it has been shown that the methods are robust and advantageous[26] due to reasons such as not having to be dependent on the structural superposition methods and also not based on proximity between superposed equivalent members.[27] The use of such a superimposition-independent, contact-based method is suitable in the comparison of homologous proteins that are diverged in sequence but have equivalent residues and conserved folds.

The crystal structure ensemble of a protein can be used to determine the structure variability of a protein by representing the connectivity of the residues in the static conformers as interactions that form a structural network. This has been defined in our earlier work where we observe the several native states of a protein that are captured by x-ray crystallography.[28] A structural network representation of the intra-connectivity of a structure was used to understand the variability observed from multiple crystal conformers of monomeric single-domain proteins of identical sequences. In reality, protein sequence variation influences the overall connectivity that exists in proteins. This has also been studied, specifically in human proteins that undergo disease-causing mutations.[29]

Likewise, there is a need to study the variation of structural networks of proteins with diverse sequences. This will aid the structure comparison of homologous proteins, which are predominantly performed for their evolutionary analysis. Furthermore, it is useful in obtaining insights into the fundamental concept of the limited number of protein folds accommodating a large number of sequences, put forward by Chothia-Lesk.[14] In this work, graph spectral methods are employed to compare the proteins of varying sequences that are classified as homologs, based on SCOPe definition.[3,30] Specifically, we have developed an extension of the graph spectral network comparison method, by accounting for the non-equivalent residues, to evaluate the total network dissimilarity across homologous protein structures. The change in PSN across homologs is brought into context, by including the diversity within its multiple conformers. We have evaluated the efficiency of the comparison method across several folds, with a wide range of sequence identities. Furthermore, the dendrograms obtained from comparison metrics are analyzed alongside their sequence-based phylogeny. The results obtained from these metrics are discussed in the context of new insights on the conservation of function.

2 | MATERIALS AND METHODS

2.1 | A dataset of multiple structures of proteins and their homologs

All available multiple structures of proteins obtained by x-ray diffraction having resolution better than 3 Å along with R_{free} and R_{work} better than 0.3 and 0.25, respectively, are fetched from the RCSB Protein Data Bank (PDB). Only wild-type full-length protein conformers that do not have any mutations, modifications, missing residues in non-terminal regions, or any other crystallographic artifacts, are chosen. To ensure an optimal description of the structural variability of the protein only those proteins having more than five crystal conformers remained in the working dataset. Each protein, represented by its accession code (Uniprot ID), is clustered at the family level using SCOPe information of the PDB entry. The selection criteria that also involves eliminating structures based on GDT TS score (discussed in later Section 2.3) resulted in 14 different families each having at least two protein members. A list of all the 70 proteins clustered into SCOPe families makes up the working dataset and can be found in Table S1 (Data S1).

The most abundant is that of goblin family with 12 proteins. The human hemoglobin beta subunit has more than 150 crystal structures and is the protein with the highest number of structures in the dataset. Glutathione S-transferase (GST) and Tubulin are multidomain proteins having N-terminal and C-terminal domains with different SCOPe IDs (c.7.1.5 and a.45.1.1; c.32.1.1 and d.79.2.1, respectively). It should be noted that the Ferredoxin family, although having a distinct N-terminal and C-terminal domain, is listed by SCOPe as a single domain. This family has been retained in the dataset to check the effect of two interacting domains on the evaluation of the network metrics. The analysis of a few large proteins such as cyclin-dependent kinase and proteosomes with numerous structures, which would require significant computational time, will be very interesting to examine as a separate study.

2.2 | Protein structural network construction and comparison

Protein structural networks (PSN) are a model of the protein structure that captures its topology along with the interconnectivity between residues. The residues in the protein structure are modeled as nodes. The interactions between residues are manifested as weighted or binary edges between nodes that are interacting. Our all-atom model of the PSN is a weighted non-directed graph depicting interaction based on spatial proximity between atoms of sequentially non-adjacent residues. Any two atoms from different residues of the same protein chain that lie at a proximity lesser than 4.5 Å are termed as atom contacts. We then define the possibility of an edge between any pair of residues in the structure based on whether they make atom contacts. The maximum number of atom contacts between all pairs of amino acids across all structures in the dataset is computed, this data is also provided in Figure S1.

Weighted edges are drawn between nodes with an edge weight equivalent to the fraction of the number of atom contacts made between the residues and the maximum number of atom contacts found between the pair of corresponding amino acids in this dataset. No edge is drawn when no atom contacts are found, hence weight zero, and the maximum strength of the edge between any nodes in the network is one.

The graph spectra-based approach to the comparison of PSNs[24] is adopted in this work. The network dissimilarity score (NDS) is a robust method of delineating the dissimilarity between any pair of networks that have topologically equivalent nodes. The components of this score namely, edge difference score (EDS), eigenvalue weighted cosine score (EWCS), and correspondence score (CRS) describe different means of variation in the network. EDS is a difference in the Frobenius norm of the weighted matrices of the compared networks. EWCS and CRS point to the change in local and global clustering of nodes in the network, respectively.

2.3 | Pairwise comparisons of the protein conformers and their PSNs

First, all pairs of multiple conformers of the same protein are compared to bring out the variability in structural conformations by the above formalism. These comparisons are termed as pairs of multiple conformers (MC). In the comparison across homologs, each conformer of a protein is compared with all conformers of its homolog. These comparisons are termed as pairs of homologous conformers (HC). Say, in a given family there are three homologs, A, B, and C. A has x, B has y, and C has z number of structures. The following number of comparison pairs will be made:

- $\frac{x(x-1)}{2}$ MC of A;
- $\frac{y(y-1)}{2}$ MC of B;
- $\frac{z(z-1)}{2}$ MC of C;
- *xy* HC between A and B;
- *yz* HC between B and C;
- *xz* HC between A and C.

The Global distance test total score (GDT TS) between all pairs of MC is ensured to be greater than 90%. If any structure does not have >90% GDT TS in the majority of its comparisons the structure is dropped from further analysis. The backbone structure comparison is performed by computing the root mean square deviation (RMSD) between conformers at the Cα level. The all-atom RMSD is not considered as it may be biased to the method of superposition used, especially in the comparison of structures that are different in their sequence. TM-align[31] is used to compute the RMSD which also provides information of residue equivalences. The information on residue equivalences is necessary for the comparison of networks across homologs.

To perform network comparison, the protein structural networks (PSNs) are constructed as discussed in Section 2.2. All pairs of networks are subject to graph spectral comparison using the network dissimilarity method (NDS) discussed in Section 2.2. A NDS of zero would imply that the compared networks are exactly the same and any higher score implies proportional dissimilarity in the network. The same method of comparing the MC of a given protein was employed on a simpler dataset to understand the structure variability of proteins discussed in Prabantu et al.[32] While NDS works well for scoring different conformers of the same protein, a modification is also employed to score the conformers of the homologous proteins. This is discussed in Section 2.4.

2.4 | Total Network Dissimilarity (TND)

There are two types of comparisons being made in this work. First is the comparison of MC to determine structure variability in the ensemble of a given protein, as mentioned above. Second is the comparison of HC where the structures may have a poor number of topologically equivalent sites whose information is essential for structural comparison. In order to compare protein structures across homologs (HC), a metric that compares the entire topology, including the variations in non-aligned regions was required.

Prior to computing NDS between homologous proteins, structure-based sequence alignment is performed to obtain equivalent residues. The alignment results in aligned residues between the pair of protein sequences known as topologically equivalent residues and gaps against corresponding positions of non-aligned residues. The NDS method involves comparing the connectivity of only the equivalent residues and the remaining are ignored. Here, a new method of normalizing the comparison metric is devised for the comparison of homologous proteins so as to include the dissimilarity from the non-aligned residues. The obtained NDS is refined by incorporating fractions called gap parameters based on the length/number of non-aligned residues. The usage of this method is described below.

Suppose a protein with a sequence of length L_a is being compared with its homolog of length L_b. Using structure-based sequence alignment we find L_c number of residues to be aligned between the homologous sequences. Hence, the length of corresponding aligned residues is L_c, and all the remaining residues are non-aligned having gaps in the corresponding sequence. The adjacency matrices of the PSNs are generated as TN_1 and TN_2, respectively. The remaining network composed of only edges that connect to the non-aligned residues in the sequence is generated as GN_1 and GN_2, respectively. The following three gap parameters are computed:

$$NL_1 = \frac{\|GN_1\|_F}{\|TN_1\|_F}, NL_2 = \frac{\|GN_2\|_F}{\|TN_2\|_F}, NL_{com} = \frac{L_c}{\sqrt{L_a \times L_b}}.$$

The total network dissimilarity (TND) is obtained using the computed gap parameters as shown in Equation (1):

$$TND = \frac{NDS \times (1 + NL_1) \times (1 + NL_2)}{NL_{com}}. \quad (1)$$

A schematic for obtaining the TND from a pair of PSNs has been outlined in Figure 1. The dissimilarity score increases as the gap parameters (NL_1, NL_2, and NL_{com}) of the two sequences increases. Therefore, NDS is normalized by the fraction of unaligned portions in each of the sequences, which gives total network dissimilarity. The score is further explained by considering two sample networks NA and NB as shown in Figure S2. The observed change in NDS captured by TND as a result of having non-aligned residues is discussed in detail. It is to be noted that when the two sequences align completely (Case 1), NDS is equal to TND. As the number of non-aligned residues increases, the TND obtained by including gap parameters to the NDS is increased (Case 2) and when the change in aligned regions of the network is null TND will also be null as observed in Case 3.

2.5 | Dendrograms for phylogenetic analysis

The obtained TND between all pairs of conformers describes the variability in network connectivity across homologs. For the comparison of a pair of homologous proteins, we take the average TND value to depict the variation of connectivity in the several conformers of the protein. A very low value can arise from a scenario where the connectivity is well preserved across conformers of the given pair of proteins. Likewise, this value will be relatively higher when comparing homologs with altered connectivity.

The evolutionary relationships obtained from sequence-based phylogeny are compared with those obtained from topological dissimilarity and structure deviation.[33] To do this, a dendrogram is constructed using the values of sequence mismatch (i.e., 100—sequence identity), total network dissimilarity (TND), and structure deviation (RMSD). The differences in lineages are analyzed to draw conclusions on the evolution of proteins and function. The Kitsch program that carries out Fitch–Margoliash and least squares method is used to build the dendrogram using the PHYLIP package.[34]

3 | RESULTS

The 3D structure captured using x-ray diffraction is a snapshot of the protein in a fixed conformation referred to as a single conformer. An ensemble of multiple conformers for each protein is gathered to prepare the dataset consisting of proteins that are classified into SCOPe families each having two or more protein members and each of these proteins having more than five crystal conformers available. There are a total of 70 proteins in the dataset consisting of 1244 crystal conformers, all of which are listed in Table S2 (Data S1). Every protein is paired with all other homologs in the family and in each pair of homologous proteins, every single conformer of the given protein is compared with all the conformers of its homolog. A total of 112 596 pairs of homologous conformers (abbreviated HC) are compared.

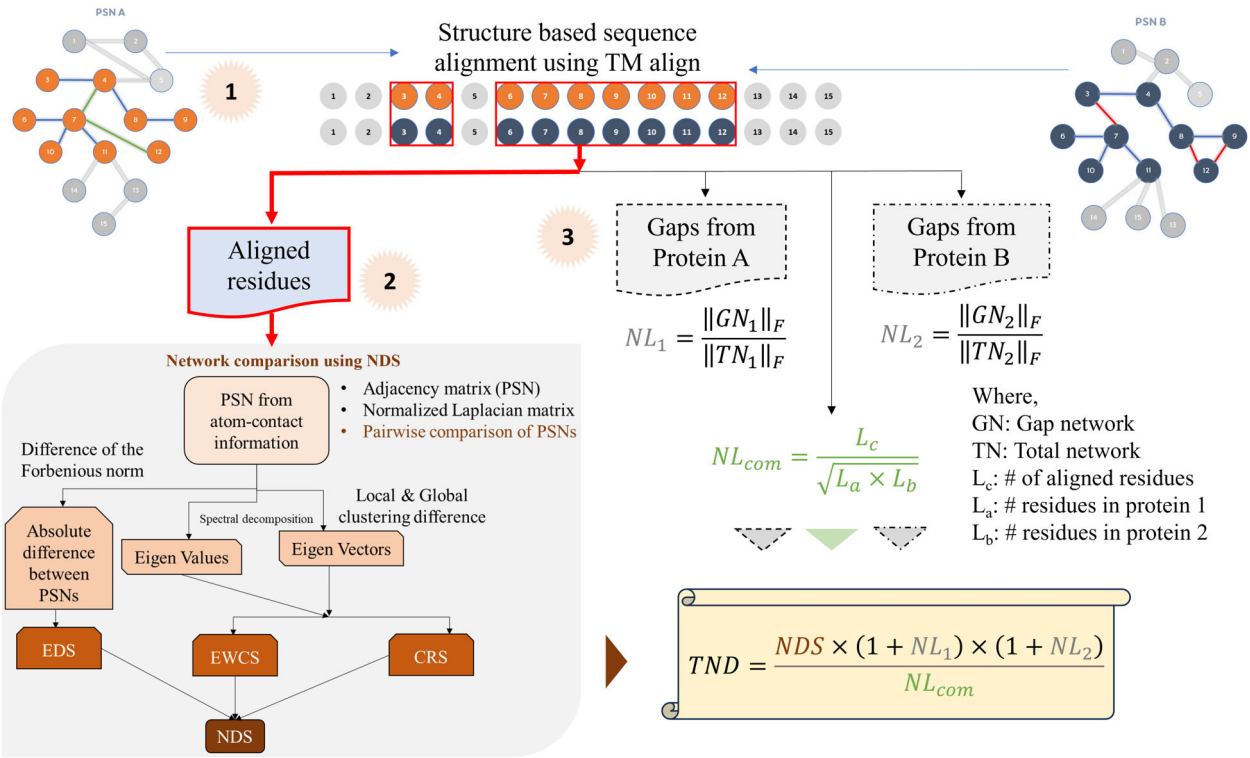


FIGURE 1 Schematic for obtaining the total network dissimilarity (TND) between a pair of homologous conformers. The initial step (1) in the network comparison of homologous conformers is the identification of aligned residues (topologically equivalent) from structure-based sequence alignment. The following step (2) is to compute the NDS from the aligned regions. Those residues that are not aligned make up the gap networks. In the sequence comparison graphic, aligned residues are shown using red outlines and gaps are shown using dashed outlines. The final step (3) involves equating the gap parameters to normalize the network dissimilarity from aligned regions, thereby obtaining the TND metric.

Furthermore, the multiple conformers of a given protein are compared among themselves as well. A total of 32 723 pairwise comparisons of multiple conformers (abbreviated MC) are made. The global distance test total score (GDT TS) of all pairs of MC is computed and it is ensured that each conformer has greater than 90% GDT TS with a majority of its multiple conformers. Information on the extent of structural change is obtained by computing Cα Root Mean Square Deviation (RMSD) in all pairs, followed by the Network Dissimilarity Score (NDS) that quantifies the extent of dissimilarity in networks between the compared conformers.

The information obtained from all pairwise comparisons is plotted on a scatter plot with RMSD on the x-axis and NDS on the y-axis. Figure 2 shows the scatter plot of these MC and HC pairs. The computed RMSD ranges from 0.01 to 1.88 Å for MC and between 0.12 and 8.04 Å in the case of HC pairs. The mean RMSD recorded in the case of MC pairs is 0.54 Å and that of HC pairs is 1.7 Å. NDS ranges from 0.001 to 0.315 in the case of pairs across MC and 0.061–0.52 in the case of HC pairs. Likewise, the mean NDS of MC pairs is 0.123 and that of HC pairs is 0.249 in the dataset. Overall, the comparison of conformers across homologs is generally observed with higher comparison scores than those of MC pairs, as expected. Also, the correlation between the obtained RMSD and NDS scores across the dataset in the HC pairs is 0.816 which is better than the correlation of 0.582 in the MC pairs. From the information on correlation, we can suggest that there are scenarios in the comparison of MCs where the network is highly variable, but the backbone structure has not deviated proportionally and vice versa. This has been reported earlier in our analysis of structural variability among MC of the same protein.[28]

3.1 | Variability of protein across multiple conformers as compared to homologous conformers

A table of all comparisons made across HC and their comparison metrics is tabulated in Table S3. Table 1 is a list of families that have been analyzed in the dataset along with information on the mean score of all comparisons within the family. It is evident in all families that the mean score is much higher while comparing across HC than the comparison of MC pairs. An interesting scenario would be when the variability that is being observed across MC is higher than what is being observed across HC. This would point to cases where the structural variability of the protein is so extremely high that there can be a scenario where the comparison of conformers from two different proteins can have a lower structural change as opposed to within multiple conformers of the same protein.

There are five proteins from the family of Ubiquitin-conjugating enzyme (UBC) related proteins (SCOPe ID: d.20.1.1) in our dataset.

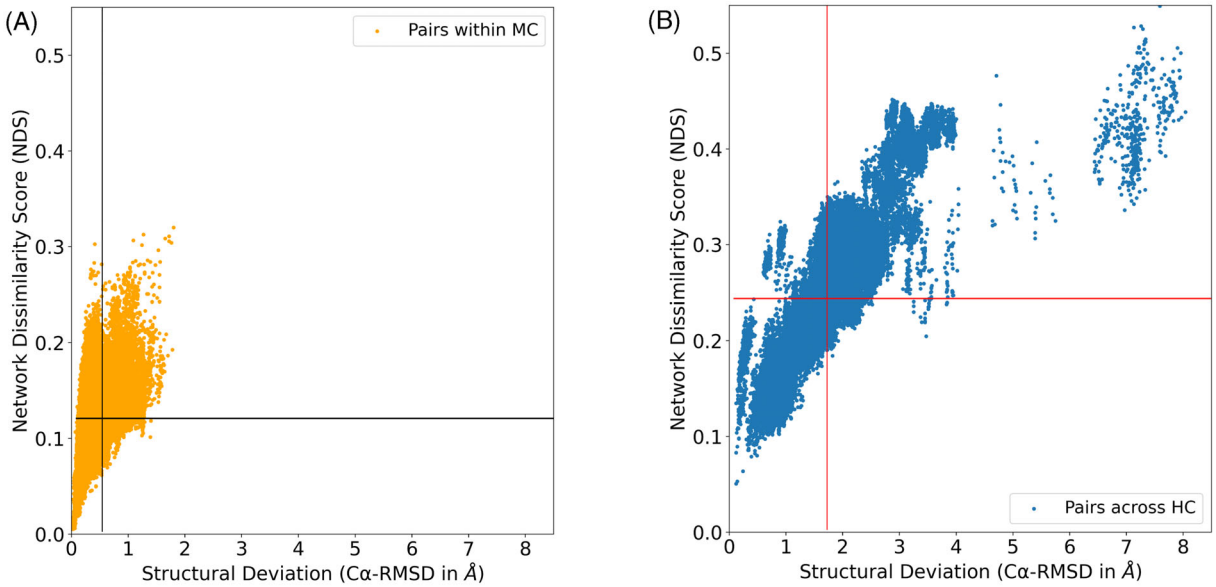


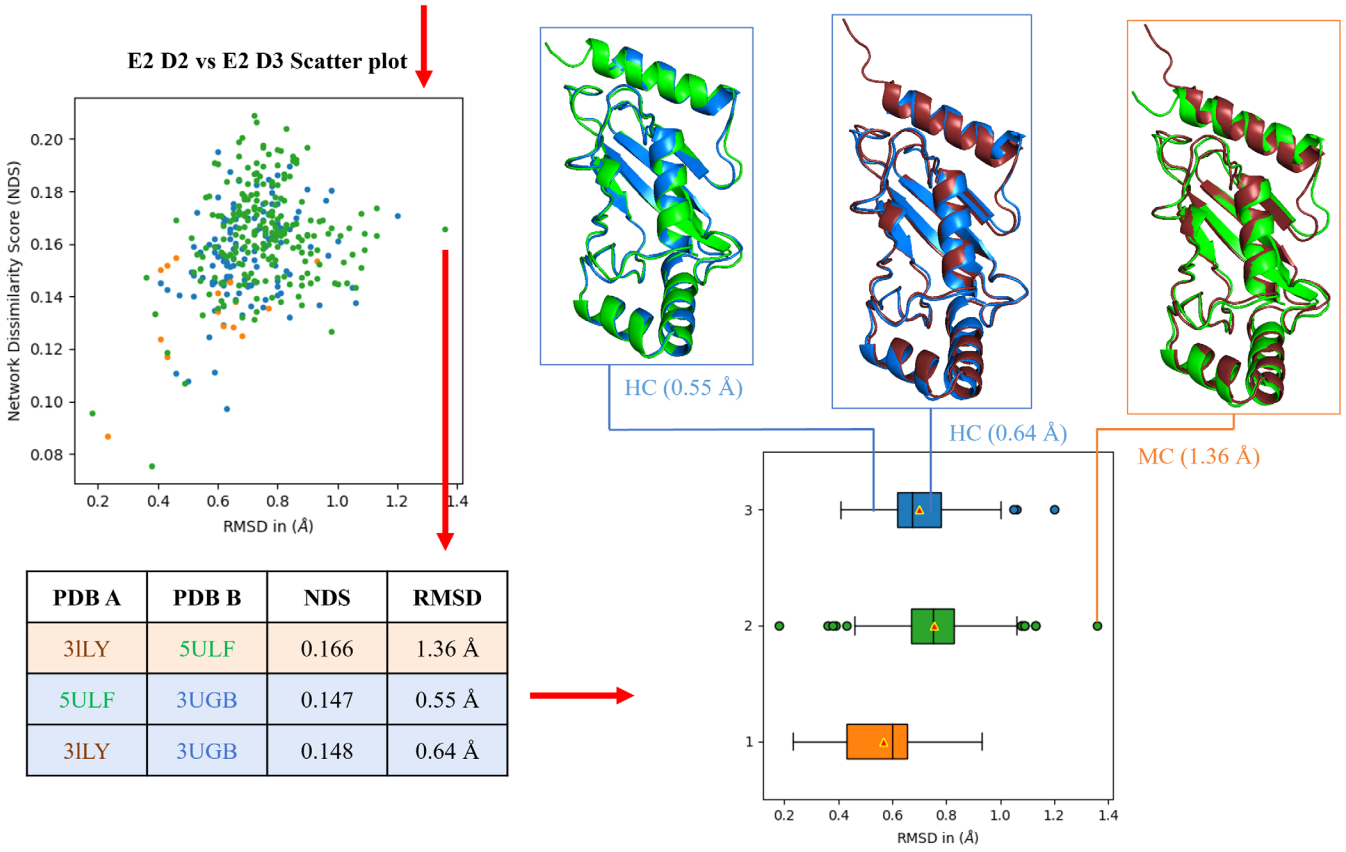
FIGURE 2 Scatter plot of structure deviation and network dissimilarity. The information on structure deviation using RMSD and alteration of networks using NDS is illustrated on the scatter plot. RMSD is plotted on the x-axis and NDS is plotted on the y-axis. (A) datapoints are obtained by the comparison of multiple conformers (MC) of the same protein. (B) The scatter shown in blue corresponds to the comparison of homologous conformers (HC). It can be observed that the mean of the comparison scores obtained across HC is significantly much higher than that obtained between MC.

TABLE 1 Mean comparison scores of all families in the dataset. The list of 14 families in the dataset is presented along with the mean of comparison scores, RMSD, and NDS. MC stands for comparison of multiple conformers of the same proteins. The mean of all vs all pairs of conformers of each protein in the family is computed. HC stands for comparison of each conformer of a protein with all the conformers of its homologous protein. This is computed between all pairs of homologs in a given family. It is interesting to note that the mean of all comparison scores computed across HC is always greater than those that are computed within MC.

#	Family name	<NDS_MC>	<RMSD_MC>	<NDS_HC>	<RMSD_HC>
1	Tubulin, C-terminal domain	0.189	0.29	0.188	0.28
2	UBC-related	0.152	0.68	0.255	1.67
3	Globins	0.114	0.56	0.227	1.59
4	Fatty acid binding protein-like	0.123	0.41	0.241	1.52
5	Dihydrofolate reductases	0.154	0.6	0.297	1.74
6	Purine and uridine phosphorylases	0.182	0.78	0.327	2.27
7	V set domains (antibody variable domain-like)	0.094	0.32	0.247	1.63
8	Ferritin	0.109	0.31	0.275	2.13
9	G proteins	0.143	0.7	0.318	2.4
10	Tyrosine-dependent oxidoreductases	0.149	0.68	0.345	2.6
11	Glutathione S-transferase (GST), C-terminal domain	0.11	0.28	0.311	2.07
12	Ferredoxin domains from multidomain proteins	0.147	0.25	0.375	5.69
13	Cytochrome P450	0.174	0.38	0.403	2.99
14	Ribonucleotide reductase-like	0.138	0.26	0.398	2.97

Ubiquitin-conjugating enzymes are reported to catalyze the SUMOylating or the ubiquitin conjugation (polyubiquitination) of specific substrates in the mitochondria. These proteins commonly termed as E2, accept ubiquitin from the E1 complex and along with the E3 subunit are able to perform ligation of ubiquitin molecules onto a substrate which are later targeted for protein degradation. While comparing closely related proteins in this family, namely the UBC-E2-D2 (Uniprot ID: P62837) and UBC-E2-D3 (Uniprot ID: P61077) it is observed that the mean score across HC is lower than that of MC both in terms of network and structure (from Table S4). Figure 3 illustrates the scatter plot and corresponding data in a boxplot to show the range and statistics of RMSD obtained between these pairs of

Family	Uniprot	Protein Name	#	PDB structures
d.20.1.1 (5)	P62837	Ubiquitin-conjugating enzyme **E2 D2**	21	**3ILY, 5ULF,** 2esk, 3tgd, 4wz3, 4ldt, 4a4b, 4a49, 4auq, 3a33, 5ulh, 5ulk, 5mnj, 5d0k, 5d0m, 5d1k, 5d1m, 5d1l, 5vzw, 4v3k, 4v3l
	P61077	Ubiquitin-conjugating enzyme **E2 D3**	6	**3UGB,** 3rpg, 4bvu, 4s3o, 5egg, 5ifr



PDB A	PDB B	NDS	RMSD
3ILY	5ULF	0.166	1.36 Å
5ULF	3UGB	0.147	0.55 Å
3ILY	3UGB	0.148	0.64 Å

FIGURE 3 Network dissimilarity and structure deviation in MC are higher than those from HC between a pair of UBC-related proteins. The scatter plot depicting RMSD and NDS obtained from a pair of homologs, UBC-E2-D2 protein compared with the UBC-E2-D3 protein is shown. The range and statistics of RMSD are shown using boxplots. It can be observed that the mean score within MC of UCB-E2-D2 is higher than across its HC. From the boxplot, the sample scenario where the structure deviation between a pair of multiple structures of UBC-E2-D2 is higher than when conformers across its homologs are compared is shown.

proteins. It is observed that the mean NDS of comparing MC of only UBC-E2-D2 protein is higher than the mean NDS of comparing HC. One may also suggest that the network differences observed across the E2-D2 protein and the E2-D3 are smaller than what is observed within conformers of the E2-D2 protein alone. This is an interesting scenario where the network connectivity between conformers of the same protein is more dissimilar than across their homologs.

Using the information of mean HC NDS and mean MC NDS (tabulated in Table S4), similar variability is observed in the tubulin and globin families. Four other cases are found in the families of ribonucleotide reductase, fatty acid binding protein, tubulin, and globins families where mean MC RMSD is higher than mean HC RMSD. These results bring about the need for including the structural variability during the comparison of proteins across homologs so as to measure the quantity of topological difference.

3.2 | Analyzing the total network dissimilarity between homologs

The Total Network Dissimilarity (TND) method is devised to normalize the quantification of network dissimilarity by also considering topologically non-equivalent positions in the comparison of homologous structures. The computation of this quantity is elucidated in Section 2.3. TND is computed for all HC pairs in the entire dataset and averaged at the protein level such that a single quantity of the average TND is able to describe the variability within as well as across the protein conformers.

In this work, the information of average TND between homologous proteins is used to study the evolutionary relationships between proteins of a family. The table of average TND value between all homologous proteins compared are presented in Table S5. Sequence identity information and the average RMSD are also presented in this

TABLE 2 Comparing the dendrograms obtained from the different metrics. The alteration of protein structural networks across proteins of the same family is analyzed alongside their sequence and structure deviation. The information of sequence mismatch, network dissimilarity and structure deviation, obtained from the information of their sequence identity, total network dissimilarity and root mean square deviation are used to plot dendrograms that help to study relationships between the proteins. Dendrograms for each family are provided in Data S1 and Tables S3–S5. The following is a table of agreement between the studied dendrograms.

Family Name	Sequence versus Network	Network versus Structure	Structure versus Sequence
Globins	Equal	Equal	Equal
Glutathione S transferase	Equal	Equal	Equal
Purine/Uridine phosphorylases	Equal	Equal	Equal
UBC related proteins	Equal	Equal	Equal
G proteins	Equal	Equal	Equal
Tubulin family	Equal	Equal	Equal
Fatty acid binding protein like	Equal	Partially equal	Partially equal
Ribonucleotide reductase like	Partially equal	Partially equal	Partially equal
Tyrosine dependent oxidoreductases	Partially equal	Equal	Partially equal
Ferritin	Partially equal	Equal	Partially equal
V set domains	Not equal	Not equal	Equal
Cytochrome P450	Not equal	Not equal	Partially equal
Dihydrofolate reductase	Not equal	Partially equal	Not equal
Ferredoxin domains	Not equal	Partially equal	Not equal

table. One important advantage of studying the phylogeny from such a structural network-based metric is the aspect of including the alteration of connectivity across the homologs. Preserved inter-residue connectivity that stems from conserved sequence and structure may also pertain to the conservation of protein function.

3.3 | Phylogenetic analysis of proteins using TND

The network variation between homologous proteins using the average TND of all pairwise comparisons is deployed, similar to the use of sequence identity information to obtain a phylogeny. The measure of sequence identity is an extent of match between the compared entities whereas the average TND is a measure of dissimilarity between the compared entities. The extent of mismatch between the sequences (computed as 100—Sequence identity in %) is compatible with the other information captured such as network dissimilarity and structure deviation. The evolutionary information obtained from a dendrogram using the sequence method is analyzed alongside that obtained from the network and structure comparison methods.

A detailed list of all the dendrograms obtained is listed in Figure S3. Table 2 depicts the observation of preserved relationships between the different dendrograms obtained for each family based on the different metrics. It is observed in seven of the fourteen families that the dendrogram obtained from the sequence input exactly matches the dendrogram obtained from the network input trees. The scales of how closely the proteins are related may vary; however, the relationships shared between the homologs in the overall dendrogram remain conserved. Apart from Tubulin family and G-proteins family which consist of only two members each, the relationships between all members of the families of Globins, Purine and Uridine phosphorylases, UBC-related, and Glutathione S-transferases are observed from their dendrograms to be exactly sharing the same relationships. There are no preserved relationships observed between the sequence vs network dendrograms of the four other families of the V set domain, cytochrome P450, dihydrofolate reductases, and ferredoxin domains. Hence, in some proteins, even when the proteins are closely related in terms of their sequence, their structural network would vary significantly.

A general observation is that the lineages obtained from either of the methods have preserved relationships, which is expected. What is interesting is where they are partially preserved or entirely different. In the ribonucleotide reductase-like family, it is observed that the lineages are partially preserved as shown in Figure 4. The relationships between members within the dotted boxes retain their relationship among each other but their relationship with other members of the family has changed. A similar partial conservation of the dendrogram is observed in the case of tyrosine-dependent oxidoreductases family and the Ferritin family. Fatty acid binding protein-like family has preserved network relationships similar to their sequence however their structure-based phylogeny is different. This is an interesting scenario to highlight the advantage of the structural network-based method over structure which will be discussed in Section 4.

4 | DISCUSSION

The progress in structural biology in the past two decades has been exponential, particularly in the case of protein structures, providing huge datasets of high-resolution structures. This has made it possible

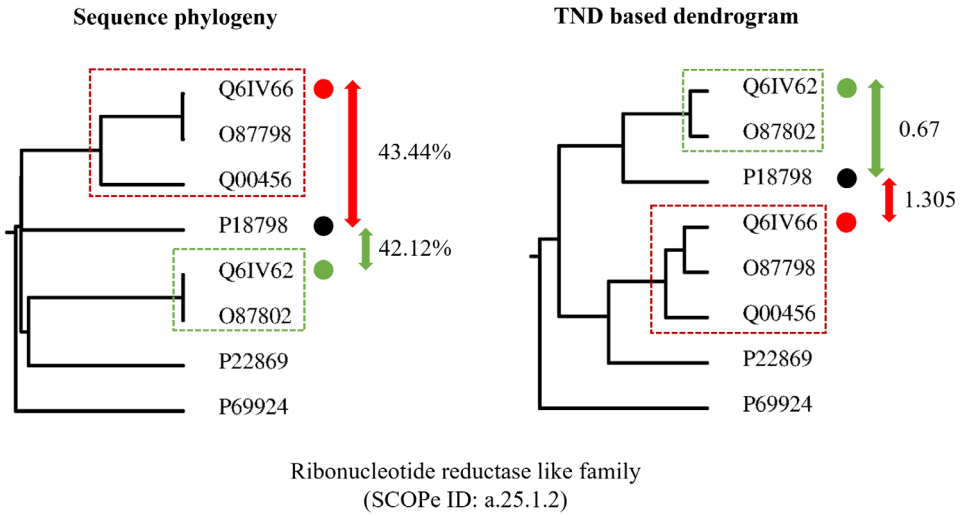


FIGURE 4 Dendrograms obtained from sequence and structural network-based information. The dendrograms obtained for the ribonucleotide reductase-like family are shown. On the left panel is the tree obtained from sequence-based input and on the right panel is that obtained from the network dissimilarity input. Each protein is represented using its uniport ID. Conserved relationships between proteins are grouped using dotted red and green boxes. The sequence identity (left panel) and average TND (right panel) between chosen proteins Q6IV66, P18798, and Q6IV62 are shown alongside their dendrograms. It is found that P18798 is much closer to Q6IV62 (0.67) than Q6IV66 (1.305) in terms of their structural network, contrary to that of sequence information.

for us to view the structures as an ensemble of different states rather than a static structure. Instances of these are captured in the form of crystal conformers solved by x-ray diffraction, after modulating several experimental parameters. Most of these instances that are of high resolution are resolved at the atomic level and hence there is significant improvement in the accuracy of their coordinates. In this study, the high-resolution crystal conformers obtained from the same sequence are recognized as multiple conformers of the same protein (MC). The structural variability of this protein is defined by the extent of diversity that is being observed in the several structural coordinates of the multiple conformers. There is evidence from the analysis of certain families of proteins that it is necessary to include the structure variability when one is performing a topological comparison of homologous proteins. Hence, we will need to compare all conformers of a given protein with all conformers of its homolog, which is the comparison of homologous conformers (HC).

It is a common practice to compare the backbone structure (using RMSD) of homologs in order to distinguish and classify proteins into groups for their functional and evolutionary analysis. Other dimensionality reduction methods such as t-SNE and principal component analysis are also used in grouping the protein ensembles.[35–37] A brief note on the improvements brought about by earlier methods of protein structure comparison is discussed in the appendix (Data S1).[19,31,38–41] There is a need for a more robust method of comparing proteins which can capture accurate differences between atomistic models of the protein structures. Hence, the structural network approach of network dissimilarity score (NDS) is employed.[24] The structure and network comparison of ensembles from a dataset of chosen proteins is performed. The alteration of their topologies across MC and HC are analyzed.

The structural variability computed from a comparison of MC of the entire dataset is found to be low (mean RMSD and NDS) as compared to earlier studies. This is partly due to the GDT TS cut-off of 90% that is applied while choosing the multiple conformers. The GDT TS cut-off ensures that protein structures that are being compared are in the same major conformation. When the protein undergoes a large conformational change then the structure can be understood as existing in different states. It would be trivial to find that the variation across HC is much higher than what has been observed across MC. However, when we compare the pairs across homologs, we might come across those structures of the homologous proteins in a similar conformational state. Several cases have been studied in detail when their global comparison scores are found to be atypical. In a few proteins, the network variability is high whereas the backbone structure remains preserved and vice versa. These are interesting instances where the topology across the homolog is well retained while the structural variability of the protein is high.

While comparing homologous proteins, there are regions that are well superposable and those that are completely from a different fold. So, a general trend is that as the amino acid sequence diverges, the regions with the same fold are reduced with an increased expectation of structural change.[14] Likewise, in this dataset it is observed that the Chothia-Lesk theory of having a limited number of folds for a large number of sequences holds true. Figure 5 shows the scatter of topological variance as compared to preservation of sequence. A strong inverse correlation is observed between the sequence identity of the homologs and the (A) structure deviation or (B) network dissimilarity. The Pearson correlation while comparing NDS is much stronger than that while comparing RMSD with sequence identity. However, there is a caveat to the results obtained from both these methods which is

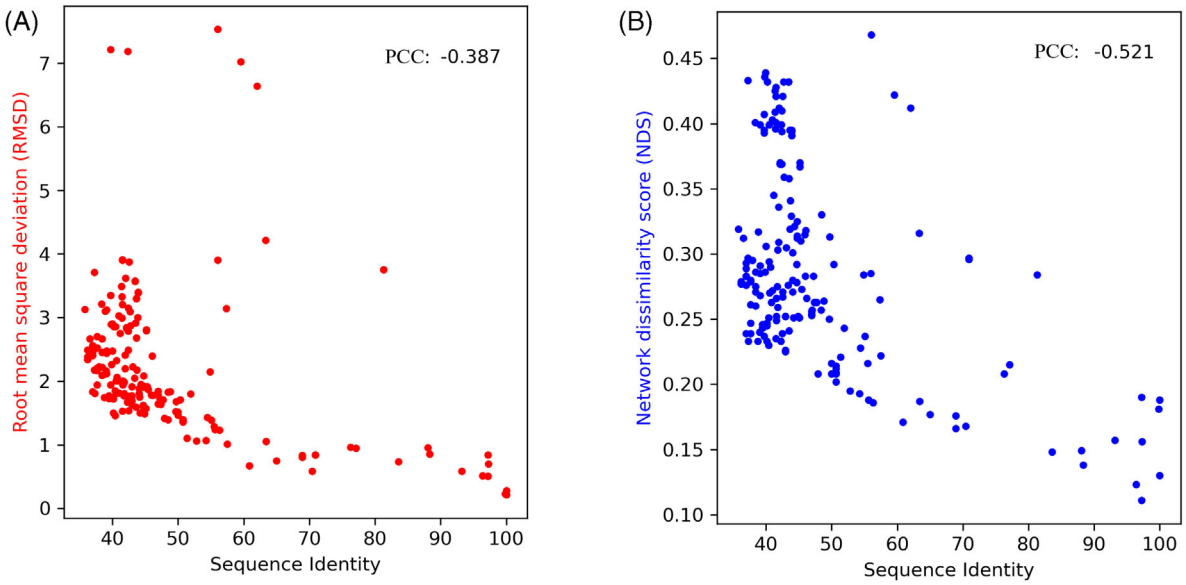


FIGURE 5 Correlation between sequence identity and topological difference. Scatter plots showing sequence identity on the x-axis, alongside the average of (A) structure deviation (RMSD) and (B) network dissimilarity (NDS) on the y-axis obtained from the dataset. A general inverse correlation is observed between sequence identity and the average structure or network deviation scores. This is similar to what was observed in the Chothia-Lesk analysis done on a smaller dataset of homologs. A relatively stronger inverse correlation is observed while comparing structural networks as compared to the comparison of just the backbone structure.

that only topologically equivalent positions are being considered for the comparison. Ferredoxin protein from the multidomain family is one such example, in which only a small fraction of the proteins are well superposable. Yet the measure of network dissimilarity obtained is too small and still comparable with the variability observed in other families. Hence a new method was necessary for comparing homologous proteins that would account for those loop, coils, and non-aligned regions.

The variability across homologous proteins is strongly dependent on their structural alignment. While comparing proteins of different sequences factors such as the non-alignment of certain loop regions or protein length difference due to extended terminals may bring about changes to the topology. This will affect the observed variability between the proteins. In order to correct these effects, here we have implemented a new scoring scheme on the existing network comparison method, so that a normalized metric is used to directly compare proteins. The structural ensemble of a protein is compared with the structure ensemble of its homolog, and the information of the total network dissimilarity (TND) in all pairs of conformers across the homologs are deduced. The averages of all these pairs of comparisons are used in building a phylogeny between the compared homologs.

Phylogeny of proteins is generally obtained by comparing the protein sequences. The sequence identity between all pairs of proteins is used to obtain its dendrogram which can provide information on how the members are related. In this study three separate trees are constructed for each family: (1) The sequence-based tree, which is the common phylogeny obtained from the sequence mismatch information (100—sequence identity in %); (2) a network-based tree is obtained using the information of TND computed between all pairs of conformers between all the homologs in a family; and (3) similarly, a structure-based tree which takes structure deviation information as input. A detailed case study of a ribonucleotide reductase-like family is performed since the dendrograms obtained for this family are neither identical nor completely different. It is observed that a group of proteins that are closely related by sequence may have preserved network information closer to a different member in the family that may be very distant in terms of sequence.

The conservation of sidechain network around the functional site plays an important role in preserved network topology. In several cases, especially where the functions of the proteins are preserved such as in the fatty acid binding protein family, it is found that the overall topology of the structural networks is similar. This is also evident in their dendrograms obtained from the network-based methods. The preservation of the structural network around the functional site can be probed further to analyze how the overall topology of the protein structure remains intact even when the sequence conservation is very poor. Similar analysis can be possible where it is extended to the comparison of complexes with conserved domains, where the inter-domain and inter-protein network may also be studied.

5 | CONCLUSION

An extended method of the graph spectral comparison of protein structural networks that is independent of structure superposition is adopted for the comparison of homologous protein structures. The structural network comparison involves comparing the overall topology at the atomic level including the sidechain connectivity. While

mere backbone structure deviation captures information about the change in overall topologies, the structural network comparison captures minute details such as the change in clustering and alteration of essential connectivity between residues.

Dendrograms obtained from information on topological change show that the preservation of network connectivity correlates well with proteins performing similar functions rather than proteins with preserved sequence. Preservation of protein function by explicitly retaining the required connectivity is observed from the analysis, where homologs participating in similar function may have largely different sequences, but their structural networks are quite similar. Hence, the method developed here has the potential for rigorous investigation of the evolution of protein function, through phylogenetic analysis paired with atomic resolution structural details.

AUTHOR CONTRIBUTIONS

Vasam Manjveekar Prabantu: Writing – original draft; methodology; data curation; writing – review and editing. **Vasundhara Gadiyaram:** Writing – original draft; methodology; data curation; writing – review and editing. **Saraswathi Vishveshwara:** Conceptualization; writing – original draft; methodology; writing – review and editing. **Narayanaswamy Srinivasan:** Conceptualization; methodology.

ACKNOWLEDGMENTS

Vasam Manjveekar Prabantu thanks Dr Himani Tandon, Dr Sankaran Sandhya and Prof Ramanathan Sowdhamini for providing their valuable suggestions and mentoring. This article is dedicated to one of the authors, late Prof. N. Srinivasan. Research from NS group was supported by the Department of Science and Technology (DST), University grants commission (UGC), Department of Biotechnology (DBT), Government of India. Vasundhara Gadiyaram was supported by CSIR-RA fellowship. Saraswathi Vishveshwara is an Honorary Scientist of NASI (National Academy of Sciences, Allahabad, India).

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/prot.26650.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Vasam Manjveekar Prabantu https://orcid.org/0000-0001-8024-8708

Saraswathi Vishveshwara https://orcid.org/0000-0002-5035-7433

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. *Science (1979)*. 1973;181(4096):223-230. doi:10.1126/science.181.4096.223
2. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol*. 1963;7(1):95-99. doi:10.1016/S0022-2836(63)80023-6
3. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247(4):536-540. doi:10.1016/S0022-2836(05)80134-2
4. Bordin N, Sillitoe I, Lees JG, Orengo C. Tracing evolution through protein structures: nature captured in a few thousand folds. *Front Mol Biosci*. 2021;8:408. doi:10.3389/fmolb.2021.668184
5. Hilser VJ, García-Moreno B, Oas TG, Kapp G, Whitten ST. A statistical thermodynamic model of the protein ensemble. *Chem Rev*. 2006;106(5):1545-1558. doi:10.1021/CR040423
6. Wei G, Xi W, Nussinov R, Ma B. Protein ensembles: how does nature harness thermodynamic fluctuations for life?: the diverse functional roles of conformational ensembles in the cell. *Chem Rev*. 2016;116(11):6516-6551. doi:10.1021/acs.chemrev.5b00562
7. Burra PV, Zhang Y, Godzik A, Stec B. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc Natl Acad Sci U S A*. 2009;106(26):10505-10510. doi:10.1073/pnas.0812152106
8. Miller MD, Phillips GN. Moving beyond static snapshots: protein dynamics and the protein data Bank. *J Biol Chem*. 2021;296:100749. doi:10.1016/j.jbc.2021.100749
9. Eisenmesser EZ, Millet O, Labeikovsky W, et al. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*. 2005;438(7064):117-121. doi:10.1038/nature04105
10. Knoverek CR, Amarasinghe GK, Bowman GR. Advanced methods for accessing protein shape-shifting present new therapeutic opportunities. *Trends Biochem Sci*. 2019;44(4):351-364. doi:10.1016/j.tibs.2018.11.007
11. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235
12. Xie L, Bourne PE. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput Biol*. 2005;1(3):e31. doi:10.1371/journal.pcbi.0010031
13. Fenwick RB, van den Bedem H, Fraser JS, Wright PE. Integrated description of protein dynamics from room-temperature x-ray crystallography and NMR. *Proc Natl Acad Sci USA*. 2014;111(4):E445-E454. doi:10.1073/pnas.1323440111
14. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;5(4):823-826. doi:10.1002/j.1460-2075.1986.tb04288.x
15. Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*. 1990;247(4948):1306-1310. doi:10.1126/science.2315699
16. Vetrivel I, de Brevern AG, Cadet F, Srinivasan N, Offmann B. Structural variations within proteins can be as large as variations observed across their homologues. *Biochimie*. 2019;167:162-170. doi:10.1016/j.biochi.2019.09.013
17. Vijayabaskar MS, Vishveshwara S. Interaction energy based protein structure networks. *Biophys J*. 2010;99(11):3704-3715. doi:10.1016/j.bpj.2010.08.079
18. Bhattacharyya M, Ghosh S, Vishveshwara S. Protein structure and function: looking through the network of side-chain interactions. *Curr Protein Pept Sci*. 2015;17(1):4-25. doi:10.2174/1389203716666150923105727
19. Schieber TA, Carpi L, Díaz-Guilera A, Pardalos PM, Masoller C, Ravetti MG. Quantification of network structural dissimilarities. *Nat Commun*. 2017;8(1):1-10. doi:10.1038/ncomms13928
20. Amitai G, Shemesh A, Sitbon E, et al. Network analysis of protein structures identifies functional residues. *J Mol Biol*. 2004;344(4):1135-1146. doi:10.1016/j.jmb.2004.10.055

21. Brinda KV, Vishveshwara S. A network representation of protein structures: implications for protein stability. *Biophys J*. 2005;89(6):4159-4170. doi:10.1529/biophysj.105.064485
22. Bhattacharyya M, Bhat CR, Vishveshwara S. An automated approach to network features of protein structure ensembles. *Protein Sci*. 2013;22(10):1399-1416. doi:10.1002/pro.2333
23. Chakrabarty B, Parekh N. PRIGSA: protein repeat identification by graph spectral analysis. *J Bioinform Comput Biol*. 2014;12(6):1442009. doi:10.1142/s0219720014420098
24. Gadiyaram V, Ghosh S, Vishveshwara S. A graph spectral-based scoring scheme for network comparison. *J Complex Netw*. 2017;5(2):219-244. doi:10.1093/comnet/cnw016
25. Gadiyaram V, Vishveshwara S, Vishveshwara S. From quantum chemistry to networks in biology: a graph spectral approach to protein structure analyses. *J Chem Info Model*. 2019;59(5):1715-1727. doi:10.1021/acs.jcim.9b00002
26. Ghosh S, Gadiyaram V, Vishveshwara S. Validation of protein structure models using network similarity score. *Proteins*. 2017;85(9):1759-1776. doi:10.1002/prot.25332
27. Kufareva I, Abagyan R. Methods of protein structure comparison. *Methods Mol Biol*. 2012;857:231-257. doi:10.1007/978-1-61779-588-6_10
28. Prabantu VM, Gadiyaram V, Vishveshwara S, Srinivasan N. Understanding structural variability in proteins using protein structural networks. *Curr Res Struct Biol*. 2022;4:134-145. doi:10.1016/j.crstbi.2022.04.002
29. Prabantu VM, Naveenkumar N, Srinivasan N. Influence of disease-causing mutations on protein structural networks. *Front Mol Biosci*. 2021;7:492. doi:10.3389/fmolb.2020.620554
30. Chandonia JM, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res*. 2019;47(D1):D475-D481. doi:10.1093/nar/gky1134
31. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33(7):2302-2309. doi:10.1093/nar/gki524
32. Deb D, Vishveshwara S, Vishveshwara S. Understanding protein structure from a percolation perspective. *Biophys J*. 2009;97(6):1787-1794. doi:10.1016/j.bpj.2009.07.016
33. Agarwal G, Rajavel M, Gopal B, Srinivasan N. Structure-based phylogeny as a diagnostic for functional characterization of proteins with a Cupin fold. *PloS One*. 2009;4(5):e5736. doi:10.1371/journal.pone.0005736
34. Felsenstein J. PHYLIP (Phylogeny Inference Package) Version 3.6. Distributed by Author. Department of Genome Sciences, University of Washington, Seattle. Scientific Research Publishing. 2005 Accessed April 18, 2023. https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=89604
35. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.
36. Rydzewski J, Nowak W. Machine learning based dimensionality reduction facilitates ligand diffusion paths assessment: a case of cytochrome P450cam. *J Chem Theory Comput*. 2016;12(4):2110-2120. doi:10.1021/acs.jctc.6b00212
37. Zhou H, Wang F, Tao P. T-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations. *J Chem Theory Comput*. 2018;14(11):5499-5510. doi:10.1021/acs.jctc.8b00652
38. Olechnovič K, Kulberkyte E, Venclovas Č. CAD-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins*. 2013;81(1):149-162. doi:10.1002/prot.24172
39. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003;31(13):3370-3374. doi:10.1093/nar/gkg571
40. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013;29(21):2722-2728. doi:10.1093/bioinformatics/btt473
41. Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol*. 2012;8:8. doi:10.1038/msb.2011.99

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Prabantu VM, Gadiyaram V, Vishveshwara S, Srinivasan N. Comparison of structural networks across homologous proteins. *Proteins*. 2023;1-12. doi:10.1002/prot.26650