Computational and Structural Biotechnology Journal 19 (2021) 3599–3608

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Protein conformational switch discerned via network centrality properties

David Foutch [a], Bill Pham [b], Tongye Shen [b,c,*]

[a] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[b] Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA
[c] UT-ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

ARTICLE INFO

Article history:
Received 2 March 2021
Received in revised form 1 June 2021
Accepted 2 June 2021
Available online 5 June 2021

Keywords:
Protein structure network
Network analysis
Network centrality
Conformational switch

ABSTRACT

Network analysis has emerged as a powerful tool for examining structural biology systems. The spatial organization of the components of a biomolecular structure has been rendered as a graph representation and analyses have been performed to deduce the biophysical and mechanistic properties of these components. For proteins, the analysis of protein structure networks (PSNs), especially via network centrality measurements and cluster coefficients, has led to identifying amino acid residues that play key functional roles and classifying amino acid residues in general. Whether these network properties examined in various studies are sensitive to subtle (yet biologically significant) conformational changes remained to be addressed. Here, we focused on four types of network centrality properties (betweenness, closeness, degree, and eigenvector centralities) for conformational changes upon ligand binding of a sensor protein (constitutive androstane receptor) and an allosteric enzyme (ribonucleotide reductase). We found that eigenvector centrality is sensitive and can distinguish salient structural features between protein conformational states while other centrality measures, especially closeness centrality, are less sensitive and rather generic with respect to the structural specificity. We also demonstrated that an ensemble-informed, modified PSN with static edges removed (which we term PSN*) has enhanced sensitivity at discerning structural changes.

1. Introduction

The functions of biomolecular systems are frequently tied to the spatial organization of the components of the system and how these components dynamically change upon perturbation [1–3]. When the structural information is rendered as binary contact information, the pairwise physical interactions between components (e.g., amino acid residues in the case of proteins) in spatial proximity are captured. These contacts can be further arranged systematically in an adjacency matrix form and displayed as a contact map. Specifically, network representations of amino acid residue interactions for proteins are referred to as protein structure networks (PSNs) or protein contact networks (PCNs) [4,5]. When the mapping from a 3D protein structure to an abstract graph is made, mathematical properties of the abstract graph [6–8] can be further utilized for inspecting the properties of amino acid residues of the protein. PSNs have been shown to have the potential for bridging graph theory concepts and mechanisms of protein systems. Such connection not only enriches the statistical analysis of networks by providing a class of practical networks with unique properties, but also assists a better (quantitative) understanding of structure–function relationships in proteins [9–22].

One category of important concepts in network analysis is network centrality. Measurements of centrality are a set of definitions that assigns a value to each node of the network (in the case of PSNs, each amino acid), which roughly quantifies the “connectivity” of each node with respect to a specific definition [6–8]. These nodes with high centrality values are considered to be noteworthy since they either directly have more neighboring nodes and/or on the path of major communication channels. In subsequent applications of PSNs, the amino acid residues with highly connected residues (measured using centrality values) are considered to be biologically significant in various studies [23–25].

There are a number of centrality measures [6–8] that have been explored to describe network structure and identify key nodes. The current work focuses on four frequently studied measures: betweenness centrality (BC), closeness centrality (CC), degree centrality (DC), and eigenvector centrality (EC). These measurements may be roughly classified into two groups: path-oriented (BC and

* Corresponding author at: Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA.

E-mail addresses: d.foutch@vumc.org (D. Foutch), cpham3@vols.utk.edu (B. Pham), tshen@utk.edu (T. Shen).

https://doi.org/10.1016/j.csbj.2021.06.004
2001-0370/

CC) and site-oriented (DC and EC) definitions. Both BC and CC are evaluated based on path lengths that end at (incidental to) or pass through each node, whereas site-based measures such as DC and EC are evaluated by adjacency or node degree. Given that there is an established structure–function relationship in structural biology and there is also a definitive topology-centrality relationship in network theory, it makes sense to use these concepts to explore PSNs for biologically relevant features.

All four PSN centrality measurements have been reported in a number of studies [23–28,5,29,30]. At times, a specific measurement was used in an application to specific PSNs and no rationale on choosing a measurement or comparison between measurements was given. Nevertheless, residues with high BC values have been identified in important biological roles such as the inter-subunit interfaces of oligomeric proteins [31], the stability of the enzymatic core of c-Abl and c-Src kinases [32], and the conserved (predominantly hydrophobic) residues of cyclophilin A [33]. Furthermore, two studies (using 178 and 283 structures respectively) found that catalytic sites have significantly higher CC values [34,35]. In one study, a strong positive correlation between surface accessibility, conservation and high CC values was found in a set of 128 proteins [36]. In another study, PSN properties (mean node degree in particular) are used to distinguish between correctly and incorrectly predicted protein structures, with correct structures reported to have high DC values [37]. Other studies reported that residues with significantly higher DC also contribute to protein thermal stability [38,39]. Additionally, residues with higher DC were found to be significant to the structural communication in the calcium sensors [40,41] and GPCRs [42]. It is also worth noting that in a study of 795 proteins, a strong negative correlation was found between evolutionary rate and residues with the highest BC, CC, and DC values [14]. EC has also been used to study PSNs in various protein systems, such as identifying important effector-binding residues of enzyme imidazole glycerol phosphate synthase [43], substrate-binding residues of enzyme cyclophilin A [33], and protein-glycan interaction of osteopontin–heparin complex [44]. One can argue whether there is a universal answer when it comes to choosing a centrality measurement or other network properties of nodes that can identify a strong contact interaction between residues (high connectivity of nodes).

In this work, we compare four frequently mentioned measurements and examine which metric(s) best discern biologically relevant structural changes. We focus on advancing PSN analysis from three new aspects. First, we examine whether network analysis results can discriminate between subtle (yet significantly different function-wise) protein conformations [45–47]. This perspective has not been extensively explored previously. It is well-established that a group of protein functions may be related by only subtle structural changes. Examples of such changes are conformational switches [48–50] triggered by ligand binding, oligomerization, and post-translational modification [51–55] and these events result in a change of conformations (inactive to active conformation, or vice versa). Frequently, such protein conformational switches are not dramatically different from a structural viewpoint. Exploring how drastic the corresponding differences are in network properties is one focus of this study. We report in the below sections that these network properties, when selected carefully and especially when applied to the modified networks (which we will later introduce), are sufficiently sensitive to discern subtle structural changes in proteins.

Second, we compare how different centrality measurements vary under conformational changes. Despite extensive studies, there has been little direct comparison between different types of centrality used in PSN analysis. As reported below, we found that different centralities are not equally useful in describing protein structures and structural changes. For example, closeness centrality is highly correlated to the distance between the residue and the center of the protein and it is a relatively poor indicator for conformational switches. On the other hand, some other centralities (especially EC) can be used to discern subtle changes of the protein structure and identify functionally important amino acid residues.

Third, in previous studies of PSNs, each analysis was performed on a network derived from a single static structure or in the case of an ensemble of structures, an aggregated PSN using a single threshold. For any single protein conformation, one can directly obtain a set of contacts that are formed and convert them into a PSN composed of nodes (residues) and edges (residue-residue contacts). For an ensemble of structures, often sampled from modeling and simulations [56,57], a standard conversion scheme assigns edges based on a contact frequency with a single cutoff [58]. Here, we propose a modified scheme with a tandem cutoff, i.e., we not only remove the contacts that are rarely formed but also the contacts that are nearly always formed. We will demonstrate how this new scheme takes advantage of the ensemble average properties in a unique angle, and we term the newly converted network PSN*, an ensemble informed PSN.

The rationale behind removing the contacts that are nearly always formed is to remove the "static" contacts. Although these static contacts are a part of the protein structure network, they may not encode dynamic or function-related motion for a specific protein [59]. When one uses centralities to characterize the "communication" between different parts of a protein, the always-formed contacts may be less relevant as they do not represent the protein dynamics. Here, we report that the properties of the traditionally defined PSNs are less sensitive to conformational changes, whereas PSN*s are able to discern subtle conformational switches. Another motivation for studying network properties of PSN* is that the truncated network can be useful for studying protein contact dynamics [60,61] (a set of orthogonal dynamic modes indicating the concerted protein motions using contact forming and breaking) where the static edges are ignored [60,61].

2. Method and systems

In this section, we first recapitulate the conventional definition of PSN and our modified definition for PSN*. We then review four different centrality calculations performed on PSN and PSN*.

2.1. Contact matrices and network construction

A PSN is often obtained directly from a contact matrix u of a single protein conformation. For a given conformation, the contact between residue i and j is considered formed when any atoms of i and atoms of j are within a distance cutoff d_c, i.e., $u_{i,j} = 1$. When the contact is not considered formed, $u_{i,j} = 0$. Additionally, contacts between polymer sequence neighbors and self contact ($u_{i,i\pm1}$ and $u_{i,i}$) are ignored and set to zero, because they represent the generic nature of polymers and contribute little to conformational dynamics. The specific contact matrices used were obtained from previous studies [62,63], where a cutoff distance $d_c = 4.2$Å was used. By using the ligand-binding domain (LBD) of the constitutive androstane receptor (CAR) system as an example, we displayed the rendering from the 3D representation of two distinct CAR conformations to their contact matrices in Fig. 1. Note that the distance cutoff d_c used here was based on the previous studies on contact interaction energy [64]. The contact formation clearly depends on this cutoff value, and a more stringent cutoff leads to a network with fewer edges. In Supplementary Information (SI) Figure S1, we show the level of changes when we use an alternative distance cutoff.

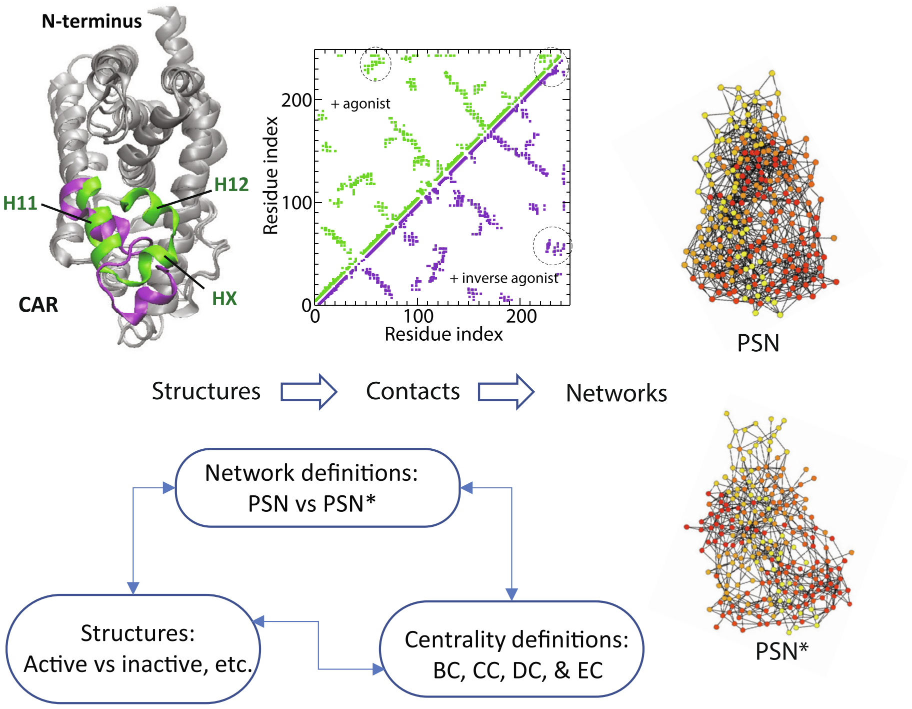

Fig. 1. The connection between protein structures, contact matrices, and networks is illustrated using the ligand-sensing domain of a nuclear hormone receptor CAR (PDBs: 1XLS and 1XNX). In the 3D representation, the regions with major structural changes are colored (green, agonist-bound and purple, inverse agonist-bound). In the network representation of the agonist-bound structure using PSN and PSN*, each node represents an amino acid residue and each edge indicates a contact formed between residues.

The connection to network theory is straightforward in the case of a single conformation (as demonstrated using single-frame PSN in Fig. 1), where the corresponding adjacency matrix of the graph follows the assignment $A_{ij} = u_{ij}$ to generate the PSN. When an ensemble of structures is available, as in the case of computer simulations, one can obtain multiple frames and further define a mean contact matrix $U = \langle u \rangle$ where each element represents the ensemble average (i.e., an average over snapshots of conformations) of contacts formed during the simulation. Thus, we define a mean contact $U_{ij} = \sum_{\alpha=1}^{N} u_{ij,\alpha}/N$, where N is the total number of frames. The mean contact values are ensemble averaged (average of the frames) and they range from 0 (never in contact) to 1 (always in contact). In this case, we define a lower bound cutoff U^l and include edges only when the contacts form more frequently than the cutoff. Therefore, the elements of the corresponding adjacency matrix $A_{ij} = 1$ when $U_{ij} \geqslant U^l$, and 0 otherwise. A practical value of $U^l = 0.02$ is used for the lower bound throughout this work. Note that the ideal lower bound value used in PSN depends on both the number of snapshots in the structural ensemble and the innate structural fluctuation of the system. A value too low would not filter discrete noise due to finite sampling, and vice versa, a value too high would cut off the subtle protein dynamics. Furthermore, in the case of PSN*, we filter out the static edges of the PSN. Together, we define $A_{ij} = 1$ when $U^l \leqslant U_{ij} \leqslant U^u$ and 0 otherwise. A practical value of the upper bound $U^u = 0.98$ is used throughout this work. As illustrated in Fig. 1, PSN* of the agonist-bound CAR is shown to be visibly thinner than PSN.

Our practical values of U^l and U^u are fairly arbitrary but they can be justified by the sensible results from studying various protein dynamics systems including the examples used in this study. To provide an idea of how our results depend on the boundary cutoffs, we compare the centrality results with an alternative scheme of different cutoff parameters, and the results can be found in SI Figure S2. We conclude that both the distance and the boundary cutoff parameters mainly affect the overall density of the network (measured by the number of edges of the network). However, the networks are robust enough and still able to reveal consistent essential feature differences between the centrality measurements.

2.2. Definitions of centrality measurements

Once PSN and PSN* are obtained, four types of network centrality measurements (Fig. 2) were studied here [6]. Centrality values are coefficients that are assigned to nodes based on their position in the network. Therefore, each centrality measure reflects certain characteristic features about the connectivity of the network under consideration.

As illustrated in Fig. 2b, betweenness centrality (BC) of residue i (values range from 0 to 1) is defined as the proportion of the shortest paths (formed between any two residues) passing through residue i among all the shortest paths in the network. The rationale that betweenness might be an important indicator of residue function stems from the assumption that removing of such residues will eliminate important nodes that sit on multiple shortest paths and impair the "communication" between residues in PSN. Similarly, a second centrality definition, closeness centrality (CC) of residue i (values range from 0 to 1), measures the inverse of the mean distance of residue i to all other residues [6–8]. Thus, the amino acid residues with the highest CC values are the set of nodes positioned at the geodesic "center" of the graph. Both BC and CC are considered path-oriented definitions. The principle difference between them is how the distances are measured. BC evaluates the shortest path from residue j to k through i and then assigns a value for residue i. CC emphasizes paths that begin at residue i and evaluates the distance to every other j and then assigns a value for residue i.

Another centrality, degree centrality (DC) of residue i, simply counts the number of edges (degree) connected to residue i in the PSN. Since the values are normalized by dividing by the maxi-

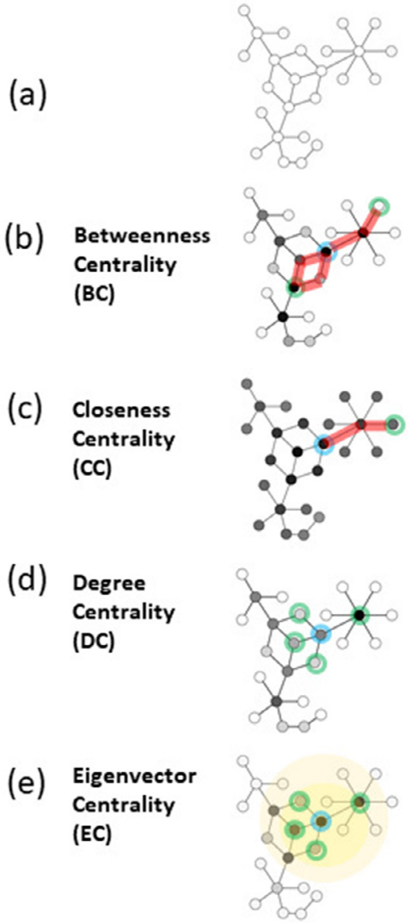
(a)
(b)
Betweenness Centrality (BC)
(c)
Closeness Centrality (CC)
(d)
Degree Centrality (DC)
(e)
Eigenvector Centrality (EC)

Fig. 2. (a) A pedagogical graph composed of 25 nodes and 27 edges is shown. (b-d) The same graph is used to illustrate different centrality measurements. The darker nodes correspond to the relatively higher values. BC and CC are path-oriented measurements. BC counts the paths (such as red line connecting two ending green circled nodes) that go through a target node (blue circle) while CC focuses on the red paths terminating at the target node. In contrast, DC and EC focus on the number of neighboring nodes of the target node. DC is a direct count of the number of the nearest neighbors (green circles), whereas EC is a weighted version with further neighbors being considered.

mum possible contact formation in principle $(N-1)$, they also range from 0 to 1. Because DC only counts the contacts to the immediate neighbors, it is considered a short-ranged, local property. A modified version of DC that recursively weighs other neighbors so that longer-ranged connectivity can be included is called eigenvector centrality (EC). If the nodes that are the most well-connected (highest ranked by degree centrality) exert the most influence on the network, then the nodes immediately adjacent to these are likely vertices that are also "influential". Operation-wise, the EC of residue i satisfied $E_i = \frac{1}{\lambda}\sum_{j=1}^{N} A_{ij}E_j$. Here, λ is the largest eigenvalue of the adjacency matrix, $Av = \lambda v$ [6–8]. Thus, the centrality of ith residue is connected to all other residues.

Practically, all four centrality calculations were performed using the software NetworkX [65]. Although our work focuses on comparing these four commonly studied centrality measures, it is important to point out that there are other generalized centrality concepts such as Katz centrality [6,66]. In addition, more generalized concepts and definitions can be used to describe the connectivity of nodes in a network such as cluster coefficients and many variant definitions. Even though they are not necessarily classified as centralities, these generalized node properties may link to the four types that we studied here.

3. Results and discussion

3.1. Different centrality measures emphasize a range of local to global structural properties

We illustrate the differences between the ligand-free (apo) and ligand-bound conformations by evaluating four centrality measures of the PSN and PSN* representations of each system. Here, we use the constitutive androstane receptor (CAR) as the first test protein system. The mean contact matrices of ligand-free and ligand-bound CARs were obtained from a previous study [62]. CAR plays a crucial role as a xenobiotic-sensing nuclear receptor involved in regulating hepatic drug metabolism and cancer development [67,68]. The structure of the LBD of CAR contains 12 helices, with helices H11, HX, and H12 comprising the C-terminal region. The conformation of the C-terminus changes drastically depending on whether CAR binds to an agonist or an inverse agonist, as shown from the canonical viewing angle in Fig. 1a. Further structural biology information of this system can be found in [62]. The agonist-bound form makes the protein bind more effectively to a coactivator and become active, whereas the inverse agonist-bound form suppresses activation by binding to a corepressor. Without ligand binding, CAR is still partially activated, and the unliganded CAR shares a similar conformation to the agonist-bound CAR [62]. The two forms are still quite different and the ligand-bound form has stronger intraprotein contacts induced by ligand. The differences between bound and unbound structures, though subtle, can have an impact on CAR activity. One of the objectives of this work is to determine whether the corresponding network properties can distinguish between these subtle but salient differences. Specifically, two ensembles of structures generated from MD simulation were used in this work: the ligand-bound structure ensemble is the ligand-binding domain of murine CAR bound to the agonist ligand TCPOBOP, and the starting point of the corresponding apo system simulation comes from the same TCPOBOP-bound murine CAR system but with the ligand removed (PDB 1XLS) [62].

Before we provide a detailed comparison of conformational changes of CAR reflected by network centrality properties, we first demonstrate that certain network centrality measurements, such as closeness centrality, may reflect the overall geometry of the protein and thus are insensitive to subtle protein conformational switches. To quantify the position of a given residue in the structure, we used the distance measurement d_{CM}, which is defined as the distance from the given residue to the mass center of the protein calculated using the average structure of the protein. Note that the calculated distance is the real Euclidean distance measured in Angstrom, not a topological distance measurement in a network. As shown in Fig. 3a, there is a strong correlation between the d_{CM} and the CC values of the residues in the apo murine CAR system. For PSN, the absolute value of the correlation coefficient is 0.89. For PSN*, they are slightly less correlated with a corresponding value of 0.73. We further examined the correlation of d_{CM} vs BC, d_{CM} vs EC, and d_{CM} vs DC (scatter plots shown in SI Figure S3), whose corresponding correlation coefficients are 0.60, 0.16, 0.17 (0.34, 0.19, and 0.12 for PSN*) respectively. The PSN* results are mainly following the PSN trend but less correlated overall, which suggests that removing the static contacts breaks apart the global geometry aspect of a network. We only used the ligand-free CAR system for this task and all the other protein systems (introduced in the next subsection) also show similar results on CC and BC being global properties of PSN and that they may not be sensitive to conformational switches.

There are several contributing factors to the strong correlation between closeness centrality and d_{CM}. The main reason is that

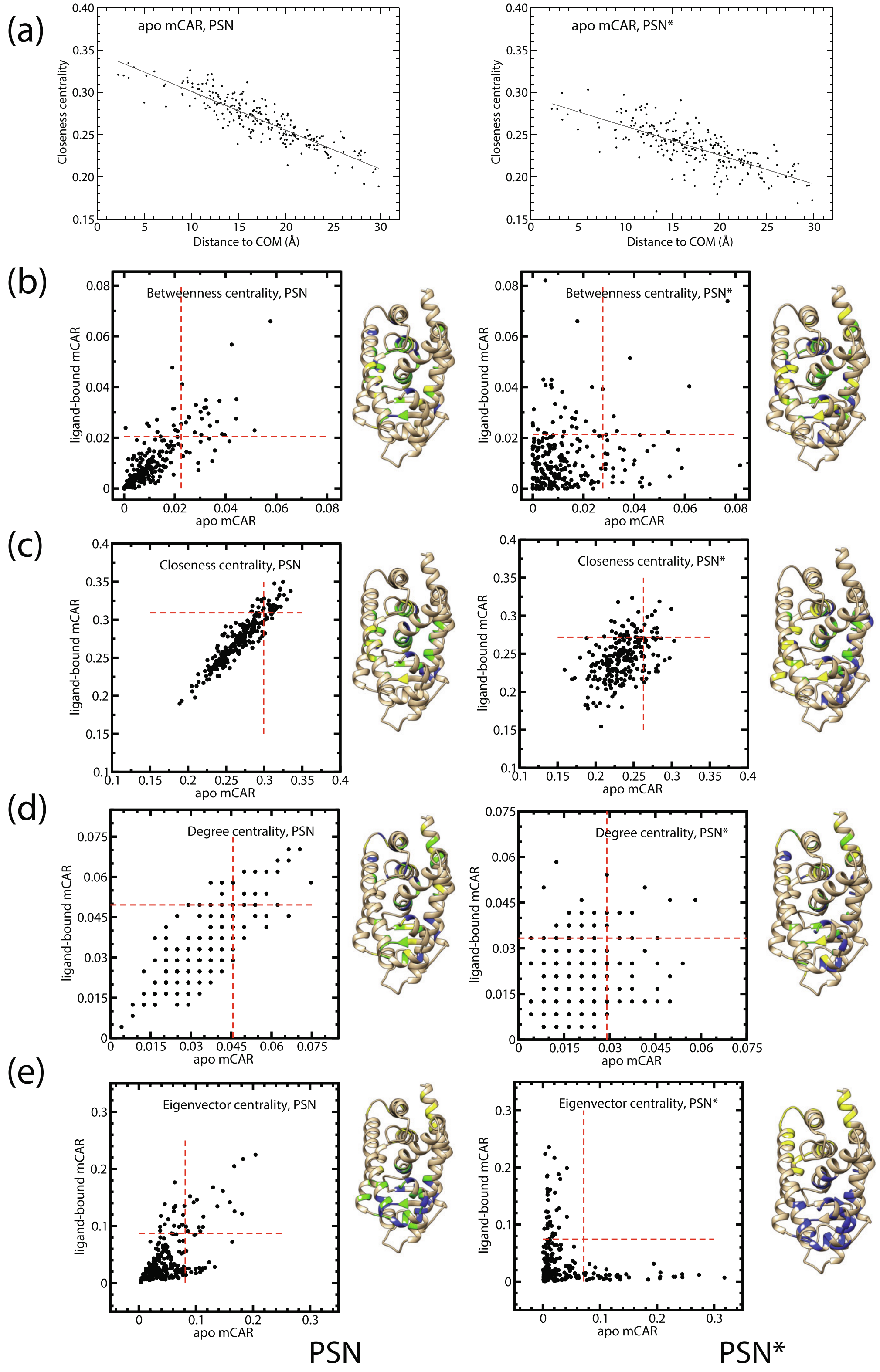

Fig. 3. Network centrality properties of nuclear receptor CAR. (a) The scatter plot of individual residue's distance to the protein center versus the CC value of the corresponding residue in PSN (left panel) and PSN* (right panel) for the ligand-free protein. The scatter plots for the apo versus agonist-bound CAR systems are shown for BC, CC, DC, and EC in the respective (b-e) panels. In the 3D representation, the top 15% ranked residues are color-coded: yellow indicates residues that are top ranked for the apo system only, blue for the ligand-bound only, and green residues are top ranked for both networks. Additionally, these top ranked residues are indicated by red dashed lines, which separate the scatter plot into four quadrants. The residues in the lower-right, upper-left, and upper-right quadrants are interpreted as yellow, blue, and green residues respectively.

PSN is generally a homogeneous network where residues are nearly uniformly distributed and the network path length (highly influential on CC) is quite comparable to the physical distance in such a case. BC is also seen to be correlated with d_{CM} (SI Figure S3), but this correlation is weaker than the correlation between CC and d_{CM}. Given that BC depends on the count of the pairwise shortest paths, the residues with the highest BC values would also be geometrically central to the PSN. Both DC and EC show much less (if any) correlations with the d_{CM}. This follows the fact that these properties weigh more on the local features of PSN, such as edge counts (Figs. 2(d,e)).

Figure panels 3b-e show scatter plots comparing four centrality values (BC, CC, DC, and EC respectively) between the apo and agonist-bound mCAR systems for PSN (left panels) and PSN* (right panels). In addition to using the scatter plots, we also labelled amino acids with the highest centrality values (top 15%) directly onto the corresponding 3D structures in Fig. 3b-e, which helps enhance visual display and add spatial information on the distribution of residues with high connectivity. Together, one can observe a degree of correlation between the two structure ensembles for each centrality property, especially for PSN (left panels). The stronger correlation on the left scatter plots compared to the right ones in Fig. 3b-e indicates that the inclusion of static elements increases the degree of structural similarity between the apo and ligand-bound CAR systems. One can observe that in both the PSN and the PSN*, the CC values are more aligned between the two systems than the correlation for other centralities. This is consistent with the geometric nature of CC. Out of the four centralities, EC values (Fig. 3e) exhibit the least similarity between the ligand-free and ligand-bound conformations. In the PSN, the EC values for the two conformations are poorly correlated, whereas the corresponding scatter plot for PSN* shows that they are uncorrelated. The loss of correlation demonstrates that the static contacts strongly influence the topology of the resulting EC values, since they emphasize the consensus features of both conformations and potentially obscure the information associated with conformational switches and contact rearrangements.

The differences between four types of network centrality calculations suggest that these calculations emphasize different aspects of a network structure. For example, some emphasize local structural properties and others are more global. DC is the most local property as the value of DC only depends on its contacting neighbors. At the other end of the spectrum, CC was found to be the most global and its values are highly correlated with the spatial geometry of the network as shown in Fig. 3a. Meanwhile, EC and BC lie in the middle. EC is a variant of DC and less local since it depends on the values of neighbours. BC, on the other hand, focuses on the available paths of the network and it is quite global to a similar degree as CC.

Another way of interpreting network properties is in terms of robustness versus sensitivity. It is desirable to define properties that best distinguish two sets of related structures differing in biological functions. However, one wants to avoid overly sensitive properties that may point out irrelevant changes between two structures. In the case of protein conformations, we think CC is a robust property that locates where the residue is relative to the geometric center of the protein. Its value is insensitive to subtle structure changes. As shown by the correlation from the scatter plots in Fig. 3b-e of the structural changes of CAR and other examples below, one can observe that CC and BC are the most consistent between two sets of structures. We believe that the property in the middle of the spectrum, EC, is more ideal in terms of describing structural changes in biomolecular structures.

One can examine the robust-vs-sensitive nature of centrality definitions by directly tracking how residues with high values (top 15% nodes) shift from one structure to the other in the 3D representation of Fig. 3. Fig. 3d shows that DC is quite sensitive (and it also only takes on discrete values) as it does not show as one "cluster" on the 3D structure. On the other hand, BC has a smaller movement and CC nearly stays at the center as expected. It is interesting to point out that residues with high EC values form large and well-defined spatial regions. These "patch" regions likely contain important biological functions. Upon ligand binding to CAR, the highest EC region transitions from a wide region loosely centered around the N-terminus to a more focused region at the C-terminus around helices H11, HX and H12. This C-terminal region is known to be important for the function of the CAR protein [62], since its ligand-induced conformation determines whether the protein becomes active or inactive.

3.2. Residues with high eigenvector centrality values are related to significant structural differences in proteins

We examined a second and more complex protein, the α subunit of ribonucleotide reductase (RNR), which can help us assess our conclusions drawn from CAR. The α subunit of RNR is a relatively larger protein (comprising of 742 residues) than the ligand-binding domain of CAR (242 residues). RNR also has a more complex ligand-binding status and ligand-induced conformational changes which can further test network centrality measurements. Function-wise, RNR catalyzes the synthesis of building blocks of DNA, which is essential for all known life-forms [63]. As a promiscuous enzyme, not surprisingly RNR has an extremely complicated sensing and regulation scheme in order to balance the population of various nucleotide products dXDP, where X = A, U, G, C. The human RNR has two ligand-binding sites (allosteric sites), the specificity site (s-site) which can sense different types of small nucleotides and direct which substrate to be catalyzed and the activity site (a-site) for an additional overall control. For example, when dTTP is bound at the s-site and ATP is bound at the a-site at the same time, this protein is considered to be active. In the current work, we compare four human RNR systems, where RNR1 has no ligand association whereas RNR2, RNR3, and RNR4 are different ligand-bound states. Fig. 4a shows the main structural differences between the unliganded (RNR1) and dTTP-bound forms (RNR2), which are present in three different regions known as the a-site (top), the s-site (bottom), and the s*-site (middle). The s*-site contains an important region called loop2, which in this case, refers to the loop region which the s-site interacts with in the dimer counterpart of RNR. As shown in Fig. 4a, all three liganded systems have dTTP bound to the s-site, and additionally, RNR3 and RNR4 have a second ligand effector (ATP and dATP, respectively) bound to the a-site. The details of the structural modeling (based on a series of PDBs: 3HNC, 3HNE, and 3HNF) and the molecular dynamics simulation that leads to the ensemble-averaged contact matrices can be found in Table I of Ref. [63].

We display in Fig. 4b-e the four centrality properties (BC, CC, DC and EC) for the two RNR systems, RNR1 versus RNR2. The top 15% ranked residues for each centrality are also visualized as 3D representations in SI Figure S4. Compared to CAR (Fig. 3b-e), the scatter plots in Fig. 4b-e show more pronounced dissimilarities in centrality values between the apo (RNR1) and the ligand-bound (RNR2) systems, yet one can see the shared patterns between these two groups of scatter plots. We can also observe the features of centrality described in the previous subsection, such as CC being a global geometry measurement and EC being sensitive to structural changes in the RNR systems. We found that EC is the most sensitive property to structural rearrangement and the residues with the highest EC values form spatial regions that have biological significance. Since we have two other different ligand-bound forms for

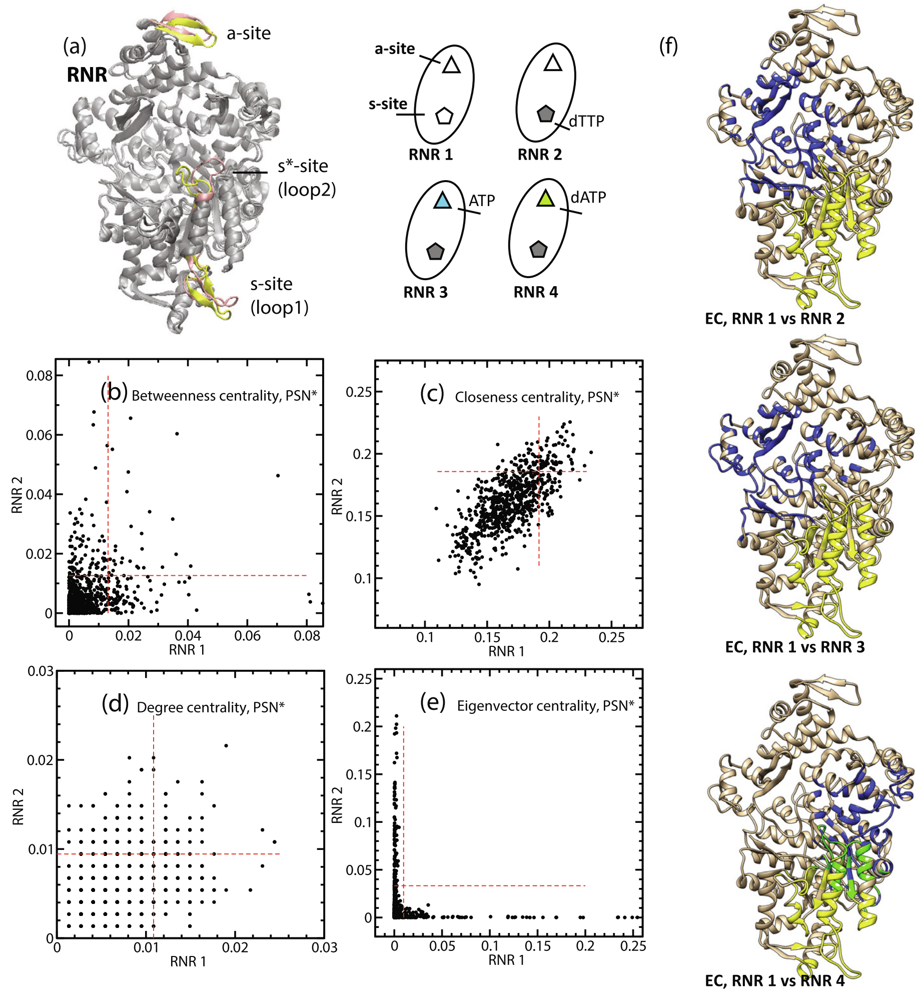

Fig. 4. Network centrality properties of RNR. (a) The 3D representation of RNR shows the structural differences between the ligand-free form (yellow) vs the ligand-bound form (pink, dTTP at the s-site). The ligand-binding status of the four systems, which we term RNR1-4, are indicated on the right panel. The scatter plots of PSN* comparing RNR1 vs RNR2 for BC, CC, DC, EC are in (b-e), respectively. The top 15% ranked residues are also displayed. (f) The comparison of the top 15% residues for EC values of RNR1 versus RNR2 (top), RNR3 (middle), or RNR4 (bottom). Here, the yellow-blue-green color scheme for the apo only, the ligand-bound only, and both networks is applied similarly to Fig. 3.

RNR, we would like to show how EC describes the changes besides RNR2. Thus, in the 3D presentation (Fig. 4f), we focus on the EC comparison for RNR1 versus RNR2, RNR3, and RNR4. Specifically, we found that the top regions for the unliganded RNR are mostly near the s-site, including the important loop1, and partially the s*-site. Upon dTTP binding at the s-site, the top region moves away from the s-site. In the case of the dATP-bound RNR (at the a-site), the region of highest EC values shifts towards the region surrounding the s*-site (loop2) and near the s-site. This movement indicates that dATP reduces the flexibility of the s-site and the s*-site binding pockets and in turn, locks the specificity allosteric effector in place as suggested from a previous study [63].

To describe the degree of similarity and dissimilarity between PSNs, we provide a general method using centrality measures and cross comparing different protein systems. Here, an overlap function is developed to distinguish between the two conformational ensembles, which essentially contains the information of a Venn diagram with a running cutoff. To compare the similarity between two networks, we first obtained two sets of ordered residues ranked by their PSN* centrality measurement. We then define a cutoff value, above which we consider the residues in the subset to be of high centrality value. We further evaluate the number of residues in both subsets, i.e., the number of residues that overlap between the two networks. The overlap function $y(x)$ is defined with the following criteria. At a given cutoff ratio (x-axis, bound between 0 and 1), two groups of residues are selected for their high centrality values (top $x \times 100\%$-ranked residues), i.e., $x = N_h/N_t$ where N_h is the number of residues that is considered to be in the subset of residues with high centrality values and N_t is the total number of residues of the protein system. The overlap ratio (y-axis, bound between 0 and 1) is defined as the number of overlapping residues (residues that both networks consider to be in the high value set), N_o, normalized by N_t, i.e., $y = N_o/N_t$. In terms of set theory, the overlap function describes the probability of observing an intersecting subset of elements given the selection of an equal number of elements from two identical sets. As shown in Fig. 5, each overlap function plot always passes through (0,0) and (1,1). On the one hand, when the overlap function passes through (0,0), this indicates that when no residues are selected, there is no overlap. On the other hand, when the function passes through (1,1), this indicates that all residues are selected and the two sets are identical. There are three ideal, or limiting, scenarios for the overlap function: (i) the upper bound of similarity, where the ranking between two conformations is identical. (ii) the independent case, where there is no correlation between the rankings, and (iii) the lower bound, where the residues are selected to minimize the

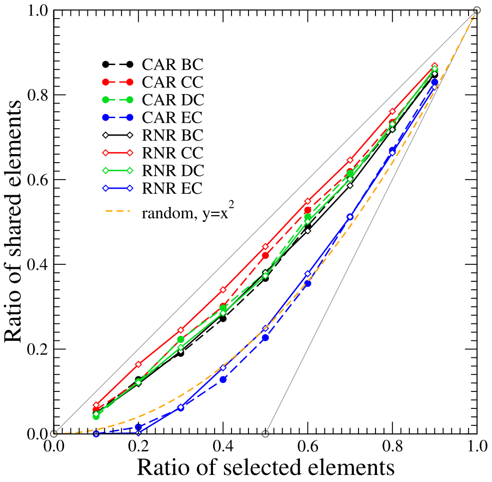

Fig. 5. The overlap function compares how (dis) similar two structural networks are. The normalized number of the common residues between the two networks (y-axis) increases when the selecting criterion (x-axis) is more relaxed, i.e, selecting residues that are the top ranked residues (using the specified centrality value).

overlap. For cases with a strong correlation between two PSN* results, the data trend would be near the diagonal line $y = x$, which is the upper bound scenario. For the scenario ii, when two sets of residues are totally uncorrelated, the overlap function approaches $y = x^2$. In the worst case scenario (the most negative correlation possible), when one tries to arrange the rankings in such a way to avoid overlapping, one finds that the lower bound function is a piece-wise function of $y = 0$ for $0 \leqslant x \leqslant 0.5$ and $y = 2(x - 0.5)$ for $0.5 < x \leqslant 1$. Practically, all curves are bound between the upper bound of $y = x$ (the scenario i) and the lower bound (the scenario iii). Curves that fall in the area bound between the scenarios i and ii indicate networks that are more similar than the random networks and centralities that are relatively insensitive to the conformational changes between the two systems. Conversely, curves in the area bound by the scenarios ii and iii indicate less similar networks and more sensitive centralities.

In Fig. 5, the CAR comparison is between two conformational ensembles, apo mCAR vs mCAR(TCPOBOP), while the RNR comparison is between RNR1 (apo form) vs RNR2 (dTTP-bound). The results demonstrated that for the same centrality method, a similar trend is observed for both CAR and RNR systems. It is important to discuss whether the similar features of the overlap functions observed between murine CAR and human RNR systems in Fig. 5 are applicable to other protein systems. We suspect that the features of a specific centrality measure are not sensitive to the choice of the system. We have evidence that it is certainly true for systems that are fairly close to the examples that we used here, such as human CAR and the LBDs of other nuclear receptors. Though we have not applied this analysis to other distinct protein systems, it is quite likely that the centrality features are generally consistent for a large number of proteins. The main reason is that the CAR and RNR systems are drastically different from each other. The centrality property features that we identified are not simply binary information and it is highly unlikely that the reported features come from purely random factors. We also provided in SI Figure S5 additional features of the network centrality measurements for these two systems to demonstrate the similar trend between different protein network centrality measurements.

Among the four centrality measures, CC (red solid and dashed curves) shows the greatest degree of correlation between the two PSN* results, which indicates that the method was least able to distinguish between two protein conformations. This is also consistent with the characterization of CC as stated earlier (Fig. 3a and 3c). On the other hand, the PSN* results of the two conformations provided by EC (blue solid and dashed curves that are shown the lowest in the overlap function) are the most distinct, showing that EC is the most sensitive method in detecting protein conformational changes. Since there is no overlap between the most dominant EC residues, the EC curves near (0,0) coalesce with the lower bound, which corroborates with the EC results in Fig. 3e. Both the curves for BC and DC fall between the CC and EC curves.

One may wonder the applicable range and limitations of EC in discerning structural network differences. Clearly, when the structural differences become so subtle that they are not reflected in the PSN, EC is unable to discern any changes. However, our protein test systems indicate that the biologically relevant level of conformational changes can provide evidence for the changes that are noticeable enough. On the other hand, if the structural differences become too drastic (such as the folding and unfolding of a protein), one may not need the network centrality properties to discern the changes. It is intriguing to apply this type of analysis to study the conformations along a conformational transition path to detect the extent of subtle structural changes can be, and corroborate the results from other data analysis methods, such as principle component analysis and advanced machine learning algorithms [69]. So far, we have demonstrated the conformational differences by only using the ligand binding-induced changes, so it would be interesting to expand this analysis to other environment-induced changes. We suspect that the conclusion can be generalized to other types of physical and chemical perturbations that result in a biologically relevant level of conformational changes.

4. Concluding Remarks

In this work, we provide insights on the network analysis of protein conformational switches using two systems, a relatively small protein (the ligand-binding domain of a signaling protein CAR) and a larger protein (an allosteric enzyme RNR). Despite having different functions, sizes, and complexity of regulation, these proteins demonstrate similar patterns for each network centrality property (BC, CC, DC, and EC) when it comes to discerning protein conformations. Some centralities such as closeness centrality describe global geometrical properties and they are less sensitive to protein conformational changes. On the other hand, some centrality properties can reflect subtle yet biologically significant conformational changes of proteins, especially when one compares the corresponding PSN* (the subnetwork of PSN with static edges removed) and when one uses eigenvector centrality.

CRediT authorship contribution statement

David Foutch: Conceptualization, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Bill Pham:** Investigation, Data curation, Writing - review & editing. **Tongye Shen:** Conceptualization, Investigation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in parts by the National Institutes of Health (USA) under award number R15 GM123469.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2021.06.004.

References

[1] Fersht A. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. New York: W.H. Freeman; 1999.
[2] Jackson MB. Molecular and cellular biophysics. 1st edn. Cambridge University Press; 2006.
[3] C.-I. Branden, J. Tooze, Introduction to protein structure, Garland Pub, New York, ISBN 9780815303442 9780815302704 9780815304869, 1991..
[4] Chakrabarty B, Parekh N. NAPS: network analysis of protein structures. Nucleic Acids Res 2016;44(W1):W375–382.
[5] Di Paola L, Giuliani A. Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 2015;31:43–8.
[6] Newman MEJ. Networks an introduction. Oxford University Press; 2018.
[7] Lewis TG. Network science: theory and applications. Wiley; 2013.
[8] van Steen M. Graph theory and complex networks: an introduction. Maarten van Steen 2010.
[9] Daily MD, Upadhyaya TJ, Gray JJ. Contact rearrangements form coupled networks from local motions in allosteric proteins. Proteins 2008;71(1):455–66.
[10] Gasper PM, Fuglestad B, Komives EA, Markwick PR, McCammon JA. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. Proc Natl Acad Sci USA 2012;109(52):21216–22.
[11] Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA: protein complexes. Proc Natl Acad Sci USA 2009;106(16):6620–5.
[12] Yao XQ, Momin M, Hamelberg D. Elucidating allosteric communications in proteins with difference contact network analysis. J Chem Inf Model 2018;58(7):1325–30.
[13] Brysbaert G, Mauri T, de Ruyck J, Lensink MF. Identification of key residues in proteins through centrality analysis and flexibility prediction with RINspector. Curr Protoc Bioinformatics 2019;65(1):e66.
[14] Fokas AS, Cole DJ, Ahnert SE, Chin AW. Residue geometry networks: a rigidity-based approach to the amino acid network and evolutionary rate analysis. Sci Rep 2016;6:33213.
[15] Ghalmane Z, Cherifi C, Cherifi H, Hassouni ME. Centrality in complex networks with overlapping community structure. Sci Rep 2019;9(1):10133.
[16] Jalili M, Salehzadeh-Yazdi A, Gupta S, Wolkenhauer O, Yaghmaie M, Resendis-Antonio O, Alimoghaddam K. Evolution of centrality measurements for the detection of essential proteins in biological networks. Front Physiol 2016;7:375.
[17] Karain WI, Qaraeen NI. The adaptive nature of protein residue networks. Proteins 2017;85(5):917–23.
[18] Pritykin Y, Singh M. Simple topological features reflect dynamics and modularity in protein interaction networks. PLoS Comput Biol 2013;9(10):e1003243.
[19] Lindsay RJ, Pham B, Shen T, McCord RP. Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. Nucleic Acids Res 2018;46(16):8143–52.
[20] Hicks M, Bartha I, di Iulio J, Venter JC, Telenti A. Functional characterization of 3D protein structures informed by human genetic diversity. Proc Natl Acad Sci USA 2019;116(18):8960–5.
[21] Runnels CM, Lanier KA, Williams JK, Bowman JC, Petrov AS, Hud NV, Williams LD. Folding, assembly, and persistence: the essential nature and origins of biopolymers. J Mol Evol 2018;86(9):598–610.
[22] Tiberti M, Invernizzi G, Lambrughi M, Inbar Y, Schreiber G, Papaleo E. PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. J Chem Inf Model 2014;54(5):1537–51.
[23] Di Paola L, De Ruvo M, Paci P, Santoni D, Giuliani A. Protein contact networks: an emerging paradigm in chemistry. Chem Rev 2013;113(3):1598–613.
[24] Yan W, Zhou J, Sun M, Chen J, Hu G, Shen B. The construction of an amino acid network for understanding protein structure and function. Amino Acids 2014;46(6):1419–39.
[25] Taylor NR. Small world network strategies for studying protein structures and binding. Comput Struct Biotechnol J 2013;5:e201302006.
[26] O'Rourke KF, Gorman SD, Boehr DD. Biophysical and computational methods to analyze amino acid interaction networks in proteins. Comput Struct Biotechnol J 2016;14:245–51.
[27] Atilgan C, Okan OB, Atilgan AR. Network-based models as tools hinting at nonevident protein functionality. Annu Rev Biophys 2012;41:205–25.
[28] Krishnan A, Zbilut JP, Tomita M, Giuliani A. Proteins as networks: usefulness of graph theory in protein science. Curr Protein Pept Sci 2008;9(1):28–38.
[29] McKnight W, Chapter twelve - graph databases: when relationships are the data, in: W. McKnight (Ed.), Information Management, Morgan Kaufmann, Boston, 120–131, ISBN 978-0-12-408056-0, 2014..
[30] Oldham S, Fulcher B, Parkes L, Iute AA, Suo C, Fornito A. Consistency and differences between centrality measures across distinct classes of networks. PLoS One 2019;14:e0220061.
[31] Wako H, Endo S. Dynamic properties of oligomers that characterize low-frequency normal modes. Biophys Physicobiol 2019;16:220–31.
[32] Tse A, Verkhivker GM. Molecular dynamics simulations and structural network analysis of c-Abl and c-Src kinase core proteins: capturing allosteric mechanisms and communication pathways from residue centrality. J Chem Inf Model 2015;55(8):1645–62.
[33] Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D. Dynamical network of residue-residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci USA 2016;113(17):4735–40.
[34] Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344(4):1135–46.
[35] Chea E, Livesay DR. How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinformatics 2007;8:153.
[36] Thibert B, Bredesen DE, del Rio G. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. BMC Bioinformatics 2005;6:213.
[37] Praznikar J, Tomic M, Turk D. Validation and quality assessment of macromolecular structures using complex network analysis. Sci Rep 2019;9(1):1678.
[38] Brinda KV, Vishveshwara S. A network representation of protein structures: implications for protein stability. Biophys J 2005;89(6):4159–70.
[39] Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, Csermely P. Network analysis of protein dynamics. FEBS Lett 2007;581(15):2776–82.
[40] Borsatto A, Marino V, Abrusci G, Lattanzi G, Dell'Orco D, Effects of Membrane and Biological Target on the Structural and Allosteric Properties of Recoverin: A Computational Approach, Int J Mol Sci 20 (20)..
[41] Marino V, Dell'Orco D. Evolutionary-Conserved Allosteric Properties of Three Neuronal Calcium Sensor Proteins. Front Mol Neurosci 2019;12:50.
[42] Fanelli F, Felline A, Raimondi F. Network analysis to uncover the structural communication in GPCRs. Methods Cell Biol 2013;117:43–61.
[43] Negre CFA, Morzan UN, Hendrickson HP, Pal R, Lisi GP, Loria JP, Rivalta I, Ho J, Batista VS. Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci USA 2018;115(52):E12201–8.
[44] Kurzbach D. Network representation of protein interactions: theory of graph description and analysis. Protein Sci 2016;25(9):1617–27.
[45] Toussi CA, Soheilifard R. A better prediction of conformational changes of proteins using minimally connected network models. Phys Biol 2017;13(6):066013.
[46] Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. Proc Natl Acad Sci USA 2002;99(13):8637–41.
[47] Heal JW, Bartlett GJ, Wood CW, Thomson AR, Woolfson DN. Applying graph theory to protein structures: an atlas of coiled coils. Bioinformatics 2018;34(19):3316–23.
[48] Anderson TA, Cordes MH, Sauer RT. Sequence determinants of a conformational switch in a protein structure. Proc Natl Acad Sci U S A 2005;102(51):18344–9.
[49] Ha JH, Loh SN. Protein conformational switches: from nature to design. Chemistry 2012;18(26):7984–99.
[50] Sannigrahi A, Nandi I, Chall S, Jawed JJ, Halder A, Majumdar S, Karmakar S, Chattopadhyay K. Conformational Switch Driven Membrane Pore Formation by Mycobacterium Secretory Protein MPT63 Induces Macrophage Cell Death. ACS Chem Biol 2019;14(7):1601–10.
[51] Tinoco I, Sauer K, Wang JC, Puglisi JD, Harbison G, Rovnyak D. Physical chemistry: principles and applications in biological sciences. Prentice Hall; 2013.
[52] Miesfeld RL, McEvoy MM. Biochemistry. New York: W.W. Norton; 2017.
[53] Walsh C. Posttranslational modification of proteins: expanding nature's inventory. Publishers, Greenwood Village, Colorado: Roberts and Co.; 2006.
[54] Wacker D, Stevens RC, Roth BL, How ligands illuminate GPCR molecular pharmacology, Cell 170 (3) (2017) 414–427, ISSN 0092–8674..
[55] Huang W, Manglik A, Venkatakrishnan AEA, Structural insights into mu-opioid receptor activation, Nature 524 (2015) 315–321..
[56] Benson NC, Daggett V. A comparison of multiscale methods for the analysis of molecular dynamics simulations. J Phys Chem B 2012;116(29):8722–31.
[57] Grazioli G, Martin RW, Butts CT. Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. Front Mol Biosci 2019;6:42.
[58] Vishveshwara S, Ghosh A, Hansia P. Intra and inter-molecular communications through protein structure network. Curr Protein Pept Sci 2009;10(2):146–60.
[59] Chang CA, Huang YM, Mueller LJ, You W. Investigation of structural dynamics of enzymes and protonation states of substrates using computational tools. Catalysts 2016;6. 82(1–21).
[60] Johnson QR, Lindsay RJ, Shen T. CAMERRA: An analysis tool for the computation of conformational dynamics by evaluating residue-residue associations. J Comput Chem 2018;39(20):1568–78.
[61] Lindsay RJ, Siess J, Lohry DP, McGee TS, Ritchie JS, Johnson QR, Shen T. Characterizing protein conformations by correlation analysis of coarse-grained contact matrices. J Chem Phys 2018;148(2):025101.
[62] Pham B, Arons AB, Vincent JG, Fernandez EJ, Shen T. Regulatory mechanics of constitutive androstane receptors: basal and ligand-directed actions. J Chem Inf Model 2019;59(12):5174–82.
[63] Pham B, Lindsay RJ, Shen T. Effector-binding-directed dimerization and dynamic communication between allosteric sites of ribonucleotide reductase. Biochemistry 2019;58(6):697–705.
[64] Kolinski A, Godzik A, Skolnick J. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. J Chem Phys 1993;98(9):7420–33.

[65] Hagberg AA, Schult DA, Swart PJ, Exploring network structure, dynamics, and function using NetworkX, in: G. Varoquaux, T. Vaught, J. Millman (Eds.), Proceedings of the 7th Python in Science Conference, Pasadena, CA USA, 11–15, 2008..

[66] Sharkey KJ. A control analysis perspective on Katz centrality. Sci Rep 2017;7(1):17247.

[67] McMahon M, Ding S, Jimenez LA, Terranova R, Gerard MA, Vitobello A, Moggs J, Henderson CJ, Wolf CR. Constitutive androstane receptor 1 is constitutively bound to chromatin and 'primed' for transactivation in hepatocytes. Mol Pharmacol 2019;95(1):97–105.

[68] Guo Z, Dai B, Jiang T, Xu K, Xie Y, Kim O, Nesheiwat I, Kong X, Melamed J, Handratta VD, Njar VC, Brodie AM, Yu LR, Veenstra TD, Chen H, Qiu Y. Regulation of androgen receptor activity by tyrosine phosphorylation. Cancer Cell 2006;10(4):309–19.

[69] Brunton SL, Kutz JN. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press 2019.