Computational and Structural Biotechnology Journal 20 (2022) 640–649

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

PSNtools for standalone and web-based structure network analyses of conformational ensembles

Angelo Felline *, Michele Seeber, Francesca Fanelli *

Department of Life Sciences, via Campi 103, 41125 Modena, Italy

ARTICLE INFO

Article history:
Received 17 October 2021
Received in revised form 22 December 2021
Accepted 30 December 2021
Available online 7 January 2022

Keywords:
Molecular simulations
Protein structure networks
Structural communication

ABSTRACT

Structure graphs, in which interacting amino acids/nucleotides correspond to linked nodes, represent cutting-edge tools to investigate macromolecular function.

The graph-based approach defined as Protein Structure Network (PSN) was initially implemented in the Wordom software and subsequently in the webPSN server. PSNs are computed either on a molecular dynamics (MD) trajectory (PSN-MD) or on a single structure. In the latter case, information on atomic fluctuations is inferred from the Elastic Network Model-Normal Mode Analysis (ENM-NMA) (PSN-ENM). While Wordom performs both PSN-ENM and PSN-MD analyses but without output post-processing, the webPSN server performs only single-structure PSN-EMN but assisting the user in input setup and output analysis.

Here we release for the first time the standalone software PSNtools, which allows calculation and post-processing of PSN analyses carried out either on single structures or on conformational ensembles. Relevant unique and novel features of PSNtools are either comparisons of two networks or computations of consensus networks on sets of homologous/analogous macromolecular structures or conformational ensembles. Network comparisons and consensus serve to infer differences in functionally different states of the same system or network-based signatures in groups of bio-macromolecules sharing either the same functionality or the same fold.

In addition to the new software, here we release also an updated version of the webPSN server, which allows performing an interactive graphical analysis of PSN-MD, following the upload of the PSNtools output.

PSNtools, the auxiliary binary version of Wordom software, and the WebPSN server are freely available at http://webpsn.hpc.unimo.it/wpsn3.php.

1. Introduction

Structure graphs, in which interacting amino acids/nucleotides correspond to linked nodes, represent cutting-edge approaches to investigate macromolecular function, including stability, recognition, folding, allostery [1–25]. Graphs are collections of vertices (or nodes) connected by edges (or links). In macromolecular structure graphs (or structure networks) residues (e.g. amino acids, nucleotides, small molecules, ions, etc) correspond to nodes. Links form on the basis of a non-covalent pairwise interaction strength, which is usually based on geometric criteria and can be used as a cutoff to build the structure network [3]. Links can be also weighted using force field-based interaction energies, thus producing protein energy networks (PENs) [26,27].

Structure networks can be computed either on a single structure or on conformational ensembles from molecular dynamics (MD) simulations, which account for link formation and breakage with atomic fluctuations. The majority of the tools for structure network analysis on conformational ensembles are standalone software packages such as Wordom [28], PSN-Ensemble [29], the PyMOL plugin xPyder [30], MD-TASK [31], PyInteraph [32] and gRINN [33]. Network analyses on MD trajectories provided by the user can be carried out by the MDN web portal [34] and the last version of the NAPS webserver [35].

The graph-based approach defined as protein structure network (PSN) analysis [3] is the one that we initially implemented in the Wordom software [28] and subsequently in the webPSN server [36,37]. The PSN is computed either on an MD trajectory (hereafter

* Corresponding author.
E-mail addresses: angelo.felline@unimore.it (A. Felline), fanelli@unimo.it (F. Fanelli).

https://doi.org/10.1016/j.csbj.2021.12.044

defined as PSN-MD) or on a single structure either downloaded from the protein databank (PDB) or uploaded from the local disc. The latter approach, hereafter defined as PSN-ENM, relies on Elastic Network Model-Normal Mode Analysis (ENM-NMA) to infer the cross-correlation of atomic fluctuations used to filter the shortest communication pathways (see below) [38]. While Wordom performs both PSN-ENM and PSN-MD analyses but without output post-processing, the webPSN server performs only single-structure PSN-ENM but assisting the user in input setup and output analysis, which can be interactively performed and graphically visualized on the webserver or on the local disk following data download [36,37].

Here we release the standalone software PSNtools, which allows both calculation and post-processing of either PSN-ENM or PSN-MD. Post-processing includes displays of the analysis output and comparisons of two or more networks. We present also a relevant extension of the webPSN server, which can now analyze and visualize also the PSNtools output of PSN-MD.

2. Implemented methodology

2.1. Building the structure network

The PSN analysis implemented in PSNtools is a product of graph theory applied to protein structures, based on the approach described by Vishveshwara and co-workers [3,39]. A graph or network is defined by a set of nodes connected by links. In a PSN, each linked residue (e.g. amino acid, nucleotide, small molecules, ion, etc) is a node [11]. Links form if the non-covalent interaction strength between pairs of nodes equals or overcomes a cutoff (I_{min}). Such interaction strength, expressed as a percentage, is computed by the Eq. (1) below:

$$I_{ij} = \frac{n_{ij}}{\sqrt{N_i N_j}} \times 100 \quad (1)$$

where I_{ij} is the percentage interaction between residues i and j; n_{ij} is the number of heavy atom–atom pairs between the side chains of residues i and j within a distance cutoff (4.5 Å); N_i and N_j are normalization factors for residue types i and j, which account for their propensities to make contacts with surrounding residues [3,39]. As for the normalization factors, both the PSNtools software and the webPSN server employ an internal database holding the normalization factors for the 20 standard amino acids and the 8 standard nucleotides (i.e. dA, dG, dC, dT, A, G, C, and U), as well as for ~34,000 molecules (e.g. small molecules, lipids, sugars, etc) and ions extracted from all the structures deposited to date in the Protein Data Bank. Normalization factors are computed as described in the relevant paper by Kannan and Vishveshwra [40]. In detail, the normalization factors for the 20 standard amino acids (N_r) were computed on a non-redundant data set of proteins with resolution higher than 2 Å, according to the following formula:

$$N_r = \frac{\sum_{k=1}^{p} \max(r_k)}{p} \quad (2)$$

where r is the residue type, k is the considered protein. The number of interaction pairs (i.e. the number of atom–atom pairs within 4.5 Å, considering both main-chain and side-chain) made by residue type r with all its surrounding residues in a protein k was evaluated. $\max(r_k)$ for residue type r, which represents the maximum number of interactions made by residue type r in protein k, was computed for each protein k in the data set. The final normalization factor for each amino acid residue type is the average of the maximum interaction value of residue type r over the whole data set of proteins p, in which residue type r occurs [40]. Accordingly, the normalization factor for a non standard amino acid residue (hereinafter referred to as non-aa for brevity sake) is defined as the number of interaction pairs made by the non-aa with all surrounding atoms, averaged over the total number of PDB structures, in which that residue is present. If a given non-aa is present more than once in the same PDB file, the maximum number of contacts is considered for calculating the average. When a PDB file is submitted, the software automatically retrieves all the normalization factors from the internal database and, if an un-parameterized non-aa is present, it transparently calculates the normalization factor of the new residue, by applying the method described above to the submitted coordinates.

Thus, the interaction strengths (I_{ij}) are computed for all node pairs. At a given interaction strength cutoff, I_{min}, any residue pair ij for which $I_{ij} \geq I_{min}$ (see equation (1)) is considered to be interacting and hence is connected. Those residues making zero edges are termed as orphans and those that make at least four edges are referred to as hubs at the considered I_{min}. The four-link cutoff for hub definition relates to the intrinsic limit in the possible number of non-covalent connections made by an amino acid in protein structures, due to steric constraints, and it is close to its upper limit. Most amino acid hubs indeed make from 4 to 6 links. The I_{min} cutoff is set automatically according to the size of the largest node cluster. In detail, a cluster is an ensemble of nodes connected by at least one link. As for cluster identification, nodes are clustered together using an agglomerative clustering method based on a single-linkage-like criterion. In the beginning, each node is in its own cluster, and clusters are then iteratively merged into larger clusters. At each step, two clusters are merged together if there are at least two nodes, one per cluster, with an interaction strength $\geq I_{min}$. The process is repeated until no more merging can be performed. According to the study by Brinda and Vishveshwara on a set of 200 size-divergent protein structures, irrespective of protein size or fold, the normalized size of the largest cluster (in terms of number of nodes) in each protein undergoes a transition at a particular I_{min} value named $I_{critical}$ [3]. The $I_{critical}$ is therefore the I_{min} value, at which the size of the largest node cluster at I_{min} = 0% halves [3]. In our PSN analyses, the I_{min} cutoff is automatically set equal to the $I_{critical}$ approximated to the second decimal place. To avoid excessive network fragmentation, which would impair the search for shortest communication paths (see below), all clusters are iteratively connected by the link with the highest sub-$I_{critical}$ interaction strength.

The I_{min} employed for the PSN-MD analysis is the average over all the I_{min} computed on each trajectory frame.

Whereas clusters are ensembles of nodes involved in at least one link, node communities are densely linked portions of the network. Communities consist in fully interconnected sets of nodes so that intra-community nodes are densely linked between each other but poorly linked with nodes outside the community. Community building consists in merging sets of three fully interconnected nodes (i.e. k = 3-cliques) sharing at least one link.

2.2. Computing the shortest communication pathways

A meaningful way to exploit PSN analysis is prediction of allosteric communication between distal sites by computing communication pathways through the structure network. A pathway describes how signals are transferred between sites and consists of a set of residues in dynamic contact [4,41]. The allosteric communication depends on structure and dynamics, i.e. it involves correlated motions. Therefore, to infer the allosteric communication in a system, the PSNtools software searches for the shortest pathways between residue pairs (path extremities) while accounting for correlated motions, i.e. collective structural fluctuations. In this respect, the shortest path is the path, in which the two considered

extreme nodes are non-covalently connected by the smallest number of intermediate nodes.

The procedure for computing the shortest communication pathways, which has been previously described and validated [38], is based on Dijkstra's algorithm [42]. As stated above, in addition to being the shortest, a path should be also dynamically correlated [7].

The first step in path searching consists in computing the protein structure network. If the input is a single structure (i.e. in PSN-ENM), all links with $I_{ij} \geq I_{min}$ participate in the PSN; if the input is a conformational ensemble (i.e. in PSN-MD), only those links with $I_{ij} \geq I_{min}$ and with a frequency $\geq$ a cutoff participate in the network (Fig. 1). Thus, a relevant difference between PSN-ENM and PSN-MD analyses is that in the latter, link frequency (i.e. fraction of conformation ensemble, in which a given link occurs) is an additional criterion for link inclusion in the network.

Briefly, the procedure consists in searching for the shortest pathways between all node pairs (path extremities). Output pathways are then filtered so as to retain only the ones, in which at least one internal node holds correlated motions (i.e. bearing a correlation coefficient $\geq$ a given cutoff) with one of the two path extremities. PSN-ENM employs cross-correlation of atomic motion from ENM-NMA [38] (Fig. 1), whereas PSN-MD employs the linear mutual information (LMI) correlations [43] from MD trajectories.

Filtered paths can be used to compute consensus paths or metapaths made of the most recurrent (i.e. with a recurrence $\geq$ a given cutoff) nodes and links in the path pool. A metapath provides a coarse/global picture of the whole structural communication in the considered system (Fig. 1). By default, both PSNtools and webPSN compute the shortest paths between all node pairs (first option). However, the user can set two or a group of relevant residues as path extremities (second option). In the webPSN server this is possible through the path-filtering option in the result page. Whereas the first option is suitable to those cases, in which the allosteric sites are unknown, the second option is worth using when some knowledge either well defined or approximate on allosteric sites is available.

With the PSNtools software, the importance of each PSN link in a given metapath can be estimated by iteratively removing each link from the network and then recalculating the resulting metapath. The consequent perturbation can be expressed as a fraction of native metapath links missing in the new metapath.

3. Features of PSNtools

3.1. General information

PSNtools is a software for PSN analysis written in C++, running either via command-line or graphical interface. The software performs PSN analysis either on a single structure (PSN-ENM) or an MD trajectory (PSN-MD). It handles any kind of molecule. The software requires the auxiliary Wordom software to read the atomic coordinates, perform ENM-NMA for PSN-ENM, and compute the correlations of atomic fluctuations.

PSNtools computes: (a) single-molecule/ensemble PSN; (b) comparisons of PSNs (e.g. nodes, hubs, links, etc) or metapaths computed on two structures/ensembles (i.e. difference networks); and (c) consensus networks from a number of single-structures/ensembles. Network comparisons and consensus serve to infer differences in functionally different states of the same system or network-based signatures in groups of bio-macromolecules (e.g. protein mutants or protein homologues/analogues) sharing either the same functionality or the same fold.

As for network comparisons, the implemented approach, which requires labeling of structurally-equivalent nodes, allows to compare any link or node in two networks independent of the degree of network similarity. The current versions of PSNtools and of webPSN hold also the implementation of four additional approaches to graph comparisons, ultimately providing a global similarity index for each approach. Three of those approaches compute the average % of shared neighbors in two networks [44–46], whereas the fourth approach computes the graphlet degree-distribution agreement between two networks, by comparing the distribution of small connected induced non-isomorphic undirected subgraphs able to summarize network topology [47].

As already stated in Section 2, PSNtools employs an internal database of normalization factors for the 20 standard amino acids and the 8 standard nucleotides (i.e. dA, dG, dC, dT, A, G, C, and U),

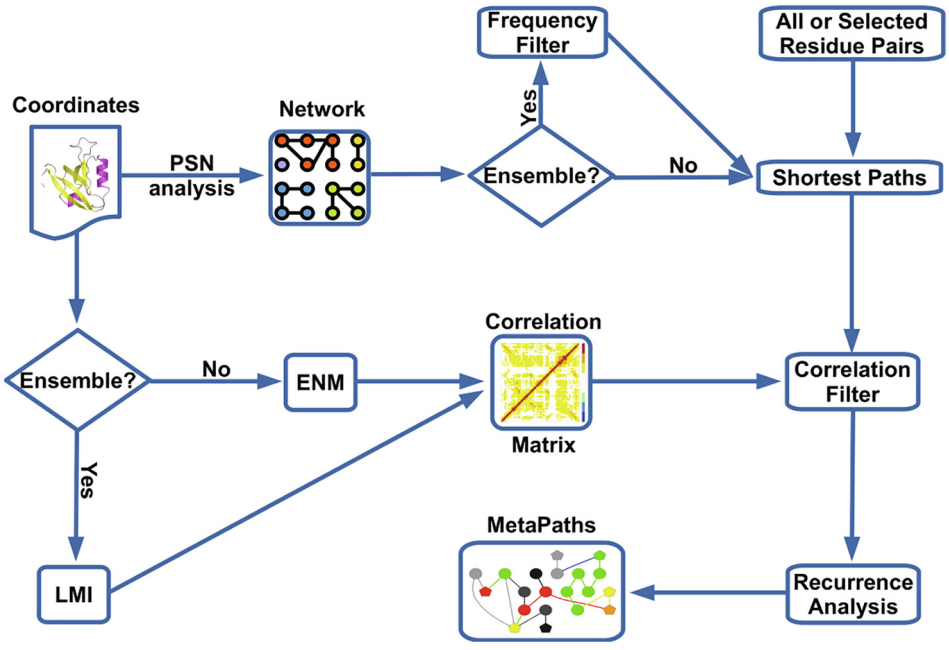

Fig. 1. Flowchart concerning the procedure for shortest path calculation.

as well as ~34,000 periodically updated small molecules and ions from the PDB. It also automatically computes the normalization factors of any residue not yet present in the internal database. Collectively, these relevant features of the software grant PSN calculation on any molecule.

The output of PSNtools consists in: (a) csv data files, (b) plots and 2D graphs, as well as (c) scripts for 3D molecular visualization by the Pymol (https://pymol.org/2/) and VMD (https://www.ks.uiuc.edu/Research/vmd/) software.

The PSNtool command-line and graphical-interface user guides can be downloaded from the webPSN site or read on the website.

3.2. Network element-based indices as markers of functionally different states

PNStools computes a number of PSN-based indices based on network elements (e.g. links, nodes, hubs, etc). An example of indices is listed in Table 1. Data in Table 1 derive from PSN analyses on previously run MD simulations of the PDZ2 domain from tyrosine phosphatase 1E (hereafter referred to as PDZ2) [38] and of the Ras GTPase (or G protein) RhoA [25].

As for PDZ2, it has been simulated in its peptide-bound (PDZ2-Bnd) and apo states (PDZ2-APO). The presence of the peptide increases network connectivity compared to the apo state as shown by the higher number of links, hubs, size of communities, and number of shortest paths (Table 1). Improvement in structural communication of PDZ2-Bnd compared to PDZ2-APO is also reflected by the lower average length of the shortest pathways (Table 1).

As for RhoA, data shown here concern previous PSN analyses on MD trajectories of the GDP-bound states either isolated (GDP) or in complex with the Rho-specific guanine nucleotide exchange factor (RhoGEF) Lbc (GDP') [25]. One of the most meaningful effects of RhoA binding to the RhoGEF Lbc is the pulling of an important loop in the nucleotide-binding site (i.e. the switch 1), which consequently looses contacts with the nucleotide [25]. This reflects on reduction in the number of hubs and their links as well as in the size of the largest node community, which involves the nucleotide itself (Table 1). RhoGEF binding also weakens the structural communication on RhoA as the number of shortest paths decreases while the average path length increases in the presence of the RhoGEF (Table 1).

Another relevant feature of PSNtools is that pairs of network-based indices from single, consensus, or difference PSNs computed by PSN-MD can be used as coordinates in distribution-surface plots. Such surfaces can be useful in discriminating different states based on dynamic network features (Fig. 2). PSNtools allows for the search of all possible network-based indices, which may serve as coordinates of protein function. In the example shown in Fig. 2, the network-based indices employed as coordinates are the number of links and the number of hubs in the interaction shell of the nucleotide GDP bound to RhoA. In more detail, the illustrative plot derives from previous PSN analysis of the MD trajectories of RhoA in its GDP-bound states either in the absence (GDP, orange) or in the presence of the RhoGEF (GDP', violet) [25]. As clearly shown by the distribution surface, the RhoGEF reduces the connections in the nucleotide interaction shell [25]. The latter includes nodes directly linked to GDP (first interaction shell) and nodes linked to the first interaction shell. As a consequence, the number of frequent links and the number of frequent hubs in the nucleotide interaction shell are effective as coordinates to distinguish the GDP and GDP' states of RhoA. Indeed, both indices diminish when RhoA is bound to the Lbc RhoGEF compared to the RhoA-free state (i.e. GDP' and GDP, respectively, Fig. 2).

3.3. Benchmarks and setup of default parameters

Default computational setting for PSN-ENM is based on benchmarks, evaluating the ability of the approach to predict amino acid residues likely involved in allosteric communication in five proteins in different functional states [37]. Selection of the five systems was based on the availability of *in vitro* information on residues involved in allosteric communication from ASD, a comprehensive database of allosteric proteins and modulators [48]. The systems included: (a) the peptide-bound state of the PDZ domain from the synaptic protein PSD-95 (PDZ3) [49,50]; (b) the agonist-bound state of vitamin D receptor (VDR) [51–53]; (c) the Pyridoxal-5-Phosphate-(PLP) bound state of the human Cystathionine β-Synthase (CBS) [54]; (d) the OFF- and ON-states of Bruton's tyrosine kinase (Btk) [55]; and (e) dimeric caspase-1 (Csp1) in the apo state and with a ligand either bound to the orthosteric site or to an allosteric site. Including the different functional states, the considered systems are eight. As previously reported, validation of the PSN-ENM method relied on comparison of those residues participating in the predicted metapath with those residues implicated in allosteric communication on the basis of *in vitro* experiments [37]. In synthesis, as for the building of the structure network, the following conditions were probed: (a) cluster merging by the link(s) with the highest sub-I_{critic} or no merging; and (b) a variable number or all possible ENM eigenvectors for computing motion correlations. As for the search of shortest communication pathways, the following conditions were probed: (a) link weighting by cross-correlation of motions or by interaction strength, or both, or no weighting; (b) different motion correlation cutoffs for path filtering; (c) several recurrence cutoffs (i.e. minimum % of paths a link must be present in to be part of the resulting metapath); and (d) two different ways to compute path-link recurrence. The Youden's index (J-index), combining in a single number sensitivity and specificity [56–58], was used to evaluate the predictive ability of the method. The J-index averaged over the J-indices of five systems (i.e. by automatically selecting the best performer state if more than one state per protein was present) was used to set the default conditions. In detail, average sensitivity, specificity, and J-index for the selected conditions were, respectively, 0.78, 0.93, and 0.72 [37].

The selected default setting comprises: (a) the application of cluster merging by sub-optimal I_{min} while computing the structure graph; (b) no link weighting; (c) employment of 10 ENM-eigenvectors, which are sufficient to describe almost the entirety of total variance while accounting for higher correlated motions; (d) a motion-correlation-coefficient cutoff equal to 0.7; and (e) a link-recurrence cutoff of 10%. We recommend such setting for PSN on single structure (PSN-ENM), which has been also implemented in the webPSN server [37]. Variation of the J-index with motion-correlation-coefficient and link-recurrence cutoffs, by fixing the other conditions listed above, shows that the link-recurrence cutoff is the limiting parameter (Fig. 3). Indeed, whereas the motion-correlation-coefficient cutoff may vary from 0 to 0.7, link-recurrence cutoff should not overcome 10% for the J-index to be significantly high (Fig. 3).

Whereas the five proteins above served to benchmarking, other systems served as case studies by PSN-ENM. The latter was, indeed, used to infer commonalties and differences concerning the structural communication in homologous proteins such as RhoGEFs of the Dbl family [59] and the β3 head piece of integrins [22].

Parameter setting for PSN-MD relied on benchmarks carried out on the MD trajectory of peptide-bound PDZ2 as well as on a number of case studies, aimed at unraveling and predicting functionality of different biosystems. PDZs are protein–protein interaction domains typically involved in the assembly of multiprotein signaling complexes. Proteins generally recognize the PDZ domains

Table 1
Examples of structure network-based indices provided by PSNtools.

Indices	PDZ2-Bnd	PDZ2-APO	GDP	GDP'
I_{min}[a]	4.63	4.57	3.39	3.34
Number of Linked Nodes[b]	100	94	178	177
Number of Links[c]	149	130	205	192
Number of Hubs[d]	31	26	21	16
Number of Links mediated by Hubs[e]	103	85	114	103
Number of Communities[f]	4	7	7	6
Number of Nodes involved in Communities[g]	50	36	39	26
Number of Links involved in Communities[h]	76	47	53	30
Number of Nodes in the largest Community[g]	27	12	14	8
Number of Links in the largest Community[h]	42	19	24	10
Number of Nodes in the ligand Community[g]	27	–	14	6
Number of Links in the ligand Community[h]	42	–	24	8
Number of Nodes in the MetaPath[i]	72	89	14	18
Number of Links in the MetaPath[j]	71	88	12	17
Number of Shortest Paths[k]	7152	6419	1915	1247
Length of the Shortest Path[l]	3	3	3	3
Average Path Length[m]	8.37	8.88	12.66	14.92
Length of the Longest Path[n]	18	17	20	23
Minimum Path Force[o]	1.41	1.73	2.70	3.70
Average Path Force[p]	5.15	5.31	5.60	5.39
Maximum Path Force[q]	10.43	10.68	11.36	10.70
Minimum Path Correlation[r]	0.81	0.80	0.70	0.70
Average Path Correlation[s]	0.88	0.89	0.85	0.88
Maximum Path Correlation[t]	0.93	0.94	0.94	0.94
Minimum % Of Corr. Nodes[u]	6.25	7.14	5.55	4.76
Average % Of Corr. Nodes[v]	28.03	27.08	14.22	11.05
Maximum % Of Corr. Nodes[w]	100	100	100	100
Minimum Path Hubs %[x]	0	0	0	25
Average Path Hubs %[y]	49.87	40.66	40.90	51.25
Maximum Path Hubs %[z]	100	100	87.50	77.78

[a] The minimum interaction strength needed to connect two nodes.
[b] Total number of nodes with at least one link.
[c] Total number of links with an interaction strength $\geq I_{min}$. Links with a lower value may have been added to avoid excessive network fragmentation.
[d] Total number of nodes with at least 4 links.
[e] Total number of links mediated by hubs.
[f] Total number of communities.
[g] Number of nodes in all communities, in the largest community and in the community involving the small ligand (if any).
[h] Number of links in all communities, in the largest community and in the community involving the small ligand (if any).
[i] Total number of nodes in the global metapath.
[j] Total number of links in the global metapath.
[k] Total number of paths in the global path pool.
[l] Number of nodes in the shortest path.
[m] Average number of nodes in the global path pool.
[n] Number of nodes in the longest path.
[o] Lowest average interaction strength of links in the global path pool.
[p] Average of the average interaction strengths of links in the global path pool.
[q] Highest average interaction strength of links in the global path pool.
[r] Lowest average motion correlation between each node and the two extreme nodes in a path from the global path pool.
[s] Average of the average motion correlations between each node and the two extreme nodes in a path from the global path pool.
[t] Highest average motion correlation between each node and the two extreme nodes in a path from the global path pool.
[u] Lowest percentage of internal nodes with a motion correlation $\geq$ the cutoff with one or both the two extremities in a path from the global path pool.
[v] Average percentage of internal nodes with a motion correlation $\geq$ the cutoff with one or both the two extremities in a path from the global path pool.
[w] Highest percentage of internal nodes with a motion correlation $\geq$ the cutoff with one or both the two extremities in a path from the global path pool.
[x] Lowest percentage of hubs in the global path pool.
[y] Average percentage of hubs in the global path pool.
[z] Highest percentage of hubs in the global path pool.

through their C-terminal segments (four to seven amino acids in length) [60,61]. In addition to passive scaffolding, a subset of these domains is implicated in allosteric regulation of distal sites involved in effector binding [62–64]. PDZ domains are proteins of the mainly-β class and hold a roll architecture made of six antiparallel β-strands (Fig. 4). The structure includes also two α-helices. The binding pocket of the C-terminal portion of the interacting protein involves the β-strand #2, the α-helix #2, and their preceding and following loops (Fig. 4).

Benchmarks were based on the fit between the metapath nodes computed on the MD trajectory and the corresponding amino acid residues on PDZ3 likely involved in allosteric communication according to combined computational and *in vitro* experiments [50] (Fig. 4). In this respect, sensitivity, specificity, and J-index are 0.9, 0.86, and 0.76, respectively. The predicted metapath accounts for the existence of an allosteric communication between peptide binding site and distal amino acid residues in the N-term of the β-strand #6, mediated by the β-strands #3 and #4 (Fig. 4).

Those parameters that may be worth varying include: (a) the link frequency cutoff for building the structure graph (default = 50%); (b) the motion correlation coefficient cutoff (default = 0.8); and (c) the link recurrence cutoff for metapath building (default = 20%). Variation of the J-index with those three parameters is shown in Fig. 5.

Data suggest that two of the three parameters, the motion correlation coefficient and the link recurrence cutoffs, must be kept at their default values whereas the link frequency cutoff may vary from 30% to 70%. The latter index should be kept around 30% for

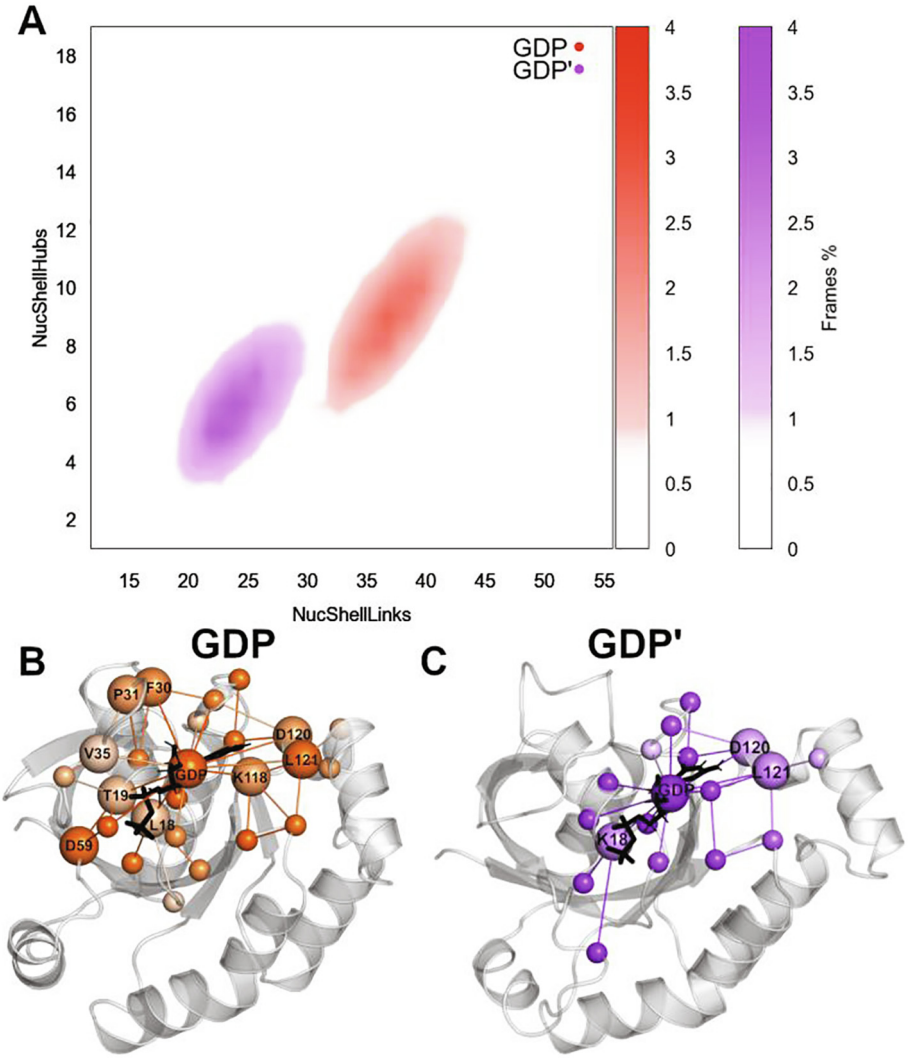

Fig. 2. Links and hubs in the nucleotide-binding site as markers of RhoA functional states. PSN analyses were done on the MD trajectories of the Ras GTPase RhoA simulated in the GDP-bound states either isolated (GDP, orange) or in complex with the RhoGEF Lbc (GDP', violet) [25]. A. The GDP interaction shell includes nodes directly linked to GDP (first interaction shell) and nodes linked to the first interaction shell. The number of links (NucShellLinks) and hubs (NucShellHubs) in such shell computed on each frame of the MD trajectories and plotted as distribution surfaces discriminate well the two different states of the G protein. B. Nodes and links in the nucleotide interaction shell of the GDP state are shown here. Nodes behaving as hubs are labeled and are represented as big spheres centered on the Cα-atoms. Hub and link colors range from dark to light orange with decrease in frequency of those elements. C. Nodes and links in the nucleotide interaction shell of the GDP' state are shown here. Hub and link colors range from dark to light violet with decrease in frequency of those elements. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

large systems in terms of atom number and conformational ensemble. This and other information contributing to PSNtools setup for conformational ensembles have been based on a number of case studies aimed at addressing different aspects of function. In deep detail, PSN analysis on the G protein coupled receptor (GPCR) luteinizing hormone receptor (LHR) in its wild type and two constitutively active LHR mutant forms (D564G and D578H), in combination with *in vitro* mutational analysis, allowed to identify the regulatory amino acid network responsible for the structural communication between the extracellular and intracellular poles of the receptor. Such network relied on highly conserved amino acids behaving as hubs and recurring in the majority of communication pathways [65]. An analogous role of highly conserved amino acids was found in the structural communication between the GPCR V2 vasopressin receptor and the intracellular protein arrestin 1 [66]. Comparative PSN analyses on representative members of the Ras GTPase superfamily inferred the central role of the nucleotide in dictating the allosteric communication in the G protein [67]. PSN analysis also served to infer those links, which maintain the structure network of the G protein transducin in its resting state and are weakened under the effects of activating mutations. Those links involve nodes in the ultraconserved nucleotide-binding regions, which loose connections under the effects of activating mutations [21] or of a GEF [25]. PSN analysis allowed to gain insights into the structural determinants of the Nougaret Congenital Night Blindness linked to a missense mutation in the G protein transducin [68]. Last but not least, PSNtools served to study a conformational disease, the autosomal dominant Retinitis Pigmentosa (adRP) linked to mutations in the GPCR rod opsin [13,24,69]. Thermal or mechanical unfolding simulations coupled to the PSN analysis were, indeed, combined with *in vitro* subcellular localization analyses to infer the effects of 33 adRP rod opsin mutations on stability and transport of the protein in the absence and presence of the natural ligand 11-*cis*-retinal [24]. The definition of an index of structure network perturbation relying on hubs and links was instrumental in clustering the adRP rod opsin mutants and in

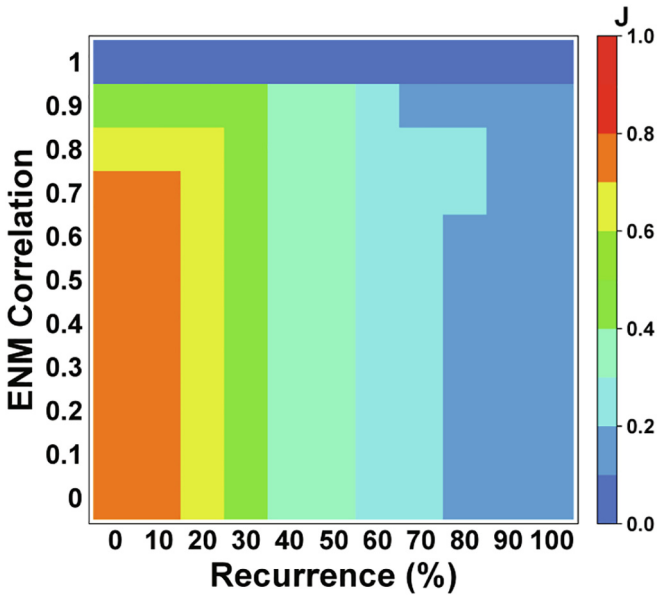

Fig. 3. PSN-ENM benchmark. Changes in J-index (J) with the cutoffs of motion-correlation coefficient (i.e. cross-correlations of atomic motions by ENM-NMA) and link-recurrence are shown.

building a computational model for algorithmic prediction of the structural/functional effects of novel adRP mutations and for aiding the design of small chaperones with therapeutic potential [24,69]. The model allowed also to infer a structure network-based landscape of rod opsin misfolding by mutation [69].

Collectively, the PSNtools software has been probed on a number of proteins holding different architectures including: (a) up-down and orthogonal bundles of class α; (b) β-roll and β-sandwich of class β; and c) two-layer αβ- and three-layer αβα-sandwiches of class αβ. The wide variety of systems and case studies faced by PSNtools supports usage of default setup, which, with the exception of the three parameters tuned in Figs. 3 and 5, is identical for PSN-ENM and PSN-MD. For PSN-MD, the only parameter worth changing may be link frequency. For large systems the default, which works with equilibrium simulations, should be kept around 30% or 33% (i.e. 1/3 of the whole trajectory frames) [25,66]. For non-equilibrium simulations such as, for example, mechanical unfolding, lower frequency cutoffs (e.g. 20–25%) are worth using [13,69].

4. webPSN-based visualization of PSNtools output

The PSNtool output can be analyzed and visualized on webPSN as a relevant novel feature of the webserver. The updated version of webPSN also plots, as a surface, the distributions of the trajectory frames as a function of two coordinates, consisting in network elements interactively selected by the user (see Fig. 2 as an example). In this respect, the user can choose among 24 available indices, which are incremented by 10 additional indices for each ligand present (e.g. the example shown in Fig. 2). Coordinate pairs can be interactively tested by the user in their ability to discriminate two functionally different states of the same macromolecule or to act as common signatures of a given functional state in a set of homologous macromolecules. In the context of network comparisons or consensus such plots may be used as valuable signatures of given functional states.

Examples on PDZ2 are available on webPSN.

5. Concluding remarks

We release the standalone software PSNtools and the updated version of webPSN, which allow PSN analysis both on single structures or on conformational ensembles. Relevant features of the software are comparisons of two or more structure networks, which already proved fundamental in inferring: (a) the landscape of protein point mutations also linked to disease [13,21,24,69]; (b) the determinants of functional differences in the same protein

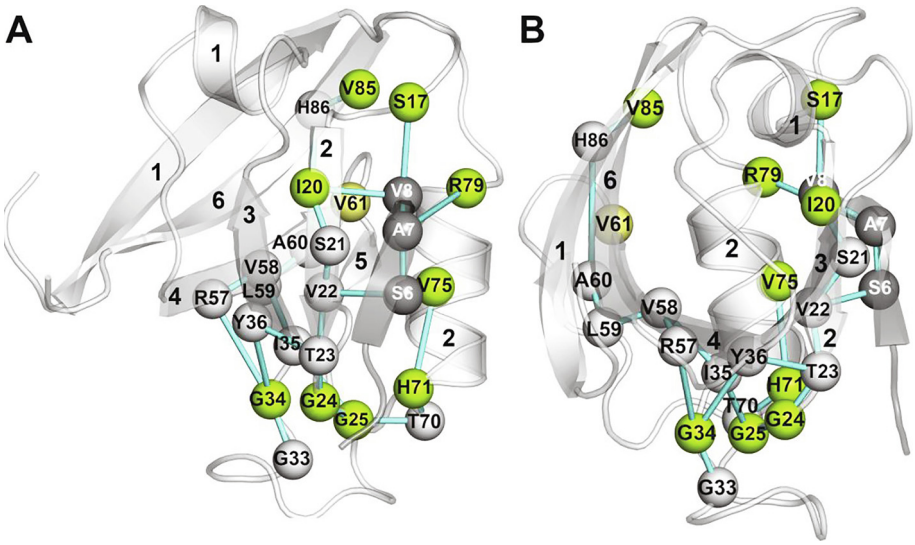

Fig. 4. Nodes participating in the metapath. In A and B, two side views of the predicted metapath are shown. The metapath was inferred from the MD trajectory employing the 3NLY crystal structure of PDZ2 as an input [38]. Paths were searched between any residue-pair in the following two sets of amino acid residues: S17, I20, V61, R79, V85 and G24, G25, G33, G34, H71. Green spheres indicate those amino acids corresponding to the ones predicted as involved in allosteric communication by computational and *in vitro* experiments on PDZ3 (i.e. S17, I20, G24, G25, G34, H71, V75, R79, V85) [50]. The yellow sphere indicates the only amino acid (V61) found *in vitro* but not in the predicted metapath. White spheres correspond to residues participating in the metapath but not found *in vitro* (S21, T23, Y36, H86, I35, A60, T70, V22, G33, R57, V58, L59). The cartoons of the bound peptide as well as peptide nodes participating in the metapath are grey; those nodes did not participate in the determination of the J-index. The color of metapath links is light blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

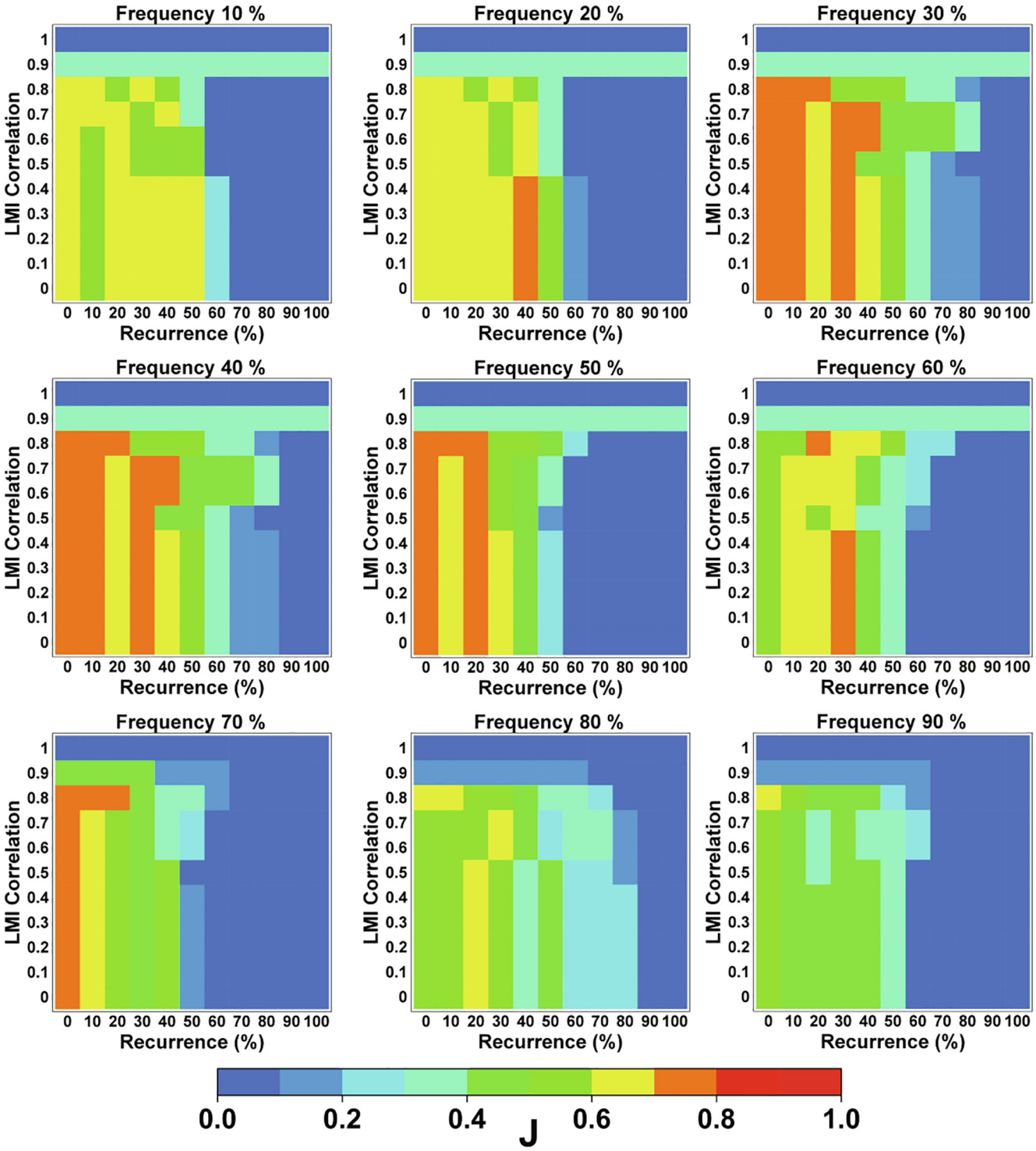

Fig. 5. PSN-MD. For each value of link frequency cutoff, changes in J-index (J) with the cutoffs of motion correlation coefficient (i.e. by LMI) and link recurrence are shown.

[22,25,38]; and (c) the structural communication signatures in a set of homologous or analogous proteins [22].

The computation setup has been extensively tested and benchmarked, therefore, the user is not required to change default setting, even if the possibility exists in the standalone software.

Tools for structure network analyses of MD trajectories essentially consist in standalone software packages such as Wordom [28], PSN-Ensemble [29], the PyMOL plugin xPyder [30], MD-TASK [31], PyInteraph [32] and gRINN [33]. The PSNtools software proposed here is singular in a relevant number of features, compared to the existing tools. Unique features and added values of PSNtools include: (a) structure-dependent and user-independent setting of calculation parameters and approach; (b) the possibility to include all kind of residues in the structure network; (c) a user-independent incorporation of information on system's dynamics in computation of communication pathways; (d) the possibility to identify allosteric sites in an unbiased manner, by automatically computing the shortest communication pathways between all

node pairs in the structure network; (e) computation of difference and consensus networks; (f) extension to nucleic acids of the same computational approach employed for proteins; and (g) high speed.

The ability to compare two or more networks inferred either from high-resolution structures of homologous/analogous proteins or function-related conformational ensembles is an invaluable e unique feature of PSNtools and the updated version of webPSN. The software released here is a very powerful and comprehensive PSN analysis tool.

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.

Acknowledgements

This study was supported by a PRIN2017 MIUR grant [2017R5ZE2C] and a University of Modena and Reggio Emilia grant [FAR2018] to FF.

Author statement

A.F. wrote and benchmarked the PSNtools code and implemented the webPSN server. MS contributed to writing the auxiliary Wordom software. F.F. conceived and supervised the study and contributed to benchmarks. A.F. and F.F. wrote the manuscript.

References

[1] Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature 2001;409 (6820):641–5.
[2] Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, et al. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344(4):1135–46.
[3] Brinda KV, Vishveshwara S. A network representation of protein structures: implications for protein stability. Biophys J 2005;89(6):4159–70.
[4] del Sol A, Fujihashi H, Amoros D, Nussinov R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Syst Biol 2006;2006(2):0019.
[5] Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, Csermely P. Network analysis of protein dynamics. FEBS Lett 2007;581(15):2776–82.
[6] Eyal E, Chennubhotla C, Yang LW, Bahar I. Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. Bioinformatics 2007;23(13):i175–84.
[7] Ghosh A, Vishveshwara S. A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis. Proc Natl Acad Sci U S A 2007;104(40):15711–6.
[8] Chennubhotla C, Bahar I, Levitt M. Signal propagation in proteins and relation to equilibrium fluctuations. PLoS Comput Biol 2007;3(9):1716–26.
[9] Chennubhotla C, Yang Z, Bahar I. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. Mol Biosyst 2008;4(4):287–92.
[10] Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA: protein complexes. Proc Natl Acad Sci U S A 2009;106(16):6620–5.
[11] Vishveshwara S, Ghosh A, Hansia P. Intra and inter-molecular communications through protein structure network. Curr Protein Pept Sci 2009;10(2):146–60.
[12] Bhattacharyya M, Ghosh A, Hansia P, Vishveshwara S. Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. Proteins 2010;78(3):506–17.
[13] Fanelli F, Seeber M. Structural insights into retinitis pigmentosa from unfolding simulations of rhodopsin mutants. FASEB J 2010;24(9):3196–209.
[14] Doncheva NT, Klein K, Domingues FS, Albrecht M. Analyzing and visualizing residue networks of protein structures. Trends Biochem Sci 2011;36 (4):179–82.
[15] Pandini A, Fornili A, Fraternali F, Kleinjung J. Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. FASEB J 2012;26(2):868–81.
[16] Papaleo E, Lindorff-Larsen K, De Gioia L. Paths of long-range communication in the E2 enzymes of family 3: a molecular dynamics investigation. Phys Chem Chem Phys 2012;14(36):12515–25.
[17] Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. Nature 2013;494 (7436):185–94.
[18] Sethi A, Tian J, Derdeyn CA, Korber B, Gnanakaran S, Shakhnovich EI. A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein. PLoS Comput Biol 2013;9(5):e1003046.
[19] 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.
[20] 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.
[21] Felline A, Mariani S, Raimondi F, Bellucci L, Fanelli F. Structural determinants of constitutive activation of gα proteins: transducin as a paradigm. J Chem Theory Comput 2017;13(2):886–99.
[22] Felline A, Ghitti M, Musco G, Fanelli F. Dissecting intrinsic and ligand-induced structural communication in the beta3 headpiece of integrins. BBA 2017;1861:2367–81.
[23] Salamanca Viloria J, Allega MF, Lambrughi M, Papaleo E. An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. Sci Rep 2017;7(1):2838.
[24] Behnen P, Felline A, Comitato A, Di Salvo MT, Raimondi F, Gulati S, et al. A small chaperone improves folding and routing of rhodopsin mutants linked to inherited blindness. iScience 2018;4:1–19.
[25] Felline A, Belmonte L, Raimondi F, Bellucci L, Fanelli F. Interconnecting flexibility, structural communication, and function in RhoGEF oncoproteins. J Chem Inf Model 2019;59(10):4300–13.
[26] Vijayabaskar MS, Vishveshwara S. Interaction energy based protein structure networks. Biophys J 2010;99(11):3704–15.
[27] Sladek V, Tokiwa H, Shimano H, Shigeta Y. Protein residue networks from energetic and geometric data: are they identical? J Chem Theory Comput 2018;14(12):6623–31.
[28] Seeber M, Felline A, Raimondi F, Muff S, Friedman R, Rao F, et al. Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. J Comput Chem 2011;32(6):1183–94.
[29] Bhattacharyya M, Bhat CR, Vishveshwara S. An automated approach to network features of protein structure ensembles. Protein Sci 2013;22 (10):1399–416.
[30] Pasi M, Tiberti M, Arrigoni A, Papaleo E. xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. J Chem Inf Model 2012;52(7):1865–74.
[31] Brown DK, Penkler DL, Sheik Amamuddy O, Ross C, Atilgan AR, Atilgan C, et al. MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics. 2017; 33(17):2768–71.
[32] 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.
[33] Sercinoglu O, Ozbek P. gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations. Nucleic Acids Res 2018;46(W1):W554–62.
[34] Ribeiro AST, Ortiz V. MDN: a web portal for network analysis of molecular dynamics simulations. Biophys J 2015;109(6):1110–6.
[35] Chakrabarty B, Naganathan V, Garg K, Agarwal Y, Parekh N. NAPS update: network analysis of molecular dynamics data and protein-nucleic acid complexes. Nucleic Acids Res 2019;47(W1):W462–70.
[36] Seeber M, Felline A, Raimondi F, Mariani S, Fanelli F. WebPSN: a web server for high-throughput investigation of structural communication in biomacromolecules. Bioinformatics 2015;31(5):779–81.
[37] Felline A, Seeber M, Fanelli F. webPSN v2.0: a webserver to infer fingerprints of structural communication in biomacromolecules. Nucleic Acids Res 2020;48 (W1):W94–W103.
[38] Raimondi F, Felline A, Seeber M, Mariani S, Fanelli F. A mixed protein structure network and elastic network model approach to predict the structural communication in biomolecular systems: the PDZ2 domain from tyrosine phosphatase 1E as a case study. J Chem Theory Comput 2013;9(5):2504–18.
[39] Vishveshwara S, Brinda KV, Kannan N. Protein structure: insights from graph theory. J Theor Comput Chem 2002;01(01):187–211.
[40] Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 1999;292(2):441–64.
[41] del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci 2006;15(9):2120–8.
[42] Dijkstra EW. A note on two problems in connexion with graphs. Numer Math 1959;1(1):269–71.
[43] Lange OF, Grubmuller H. Generalized correlation for biomolecular dynamics. Proteins 2006;62(4):1053–61.
[44] Vijaymeena MK,, Kavitha K. A survey on similarity measures in text mining. Machine Learning Appl. 2016;3(1):19–28.
[45] Jaccard P. The distribution of the Flora in the Apline zone. New Phytol 1912;11:37–50.
[46] Romesburg HC. Cluster analysis for researchers. Am. Political Sci. Rev. 1984;78.
[47] Przulj N. Biological network comparison using graphlet degree distribution. Bioinformatics 2007;23(2):e177–83.

[48] Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, et al. ASD: a comprehensive database of allosteric proteins and modulators. Nucleic Acids Res 2011;39 (Database):D663–9.
[49] Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 1996;85(7):1067–76.
[50] Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286(5438):295–9.
[51] Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 2004;116(3):417–29.
[52] Tocchini-Valentini G, Rochel N, Wurtz JM, Mitschler A, Moras D. Crystal structures of the vitamin D receptor complexed to superagonist 20-epi ligands. PNAS 2001;98(10):5491–6.
[53] Yamamoto K, Abe D, Yoshimoto N, Choi M, Yamagishi K, Tokiwa H, et al. Vitamin D receptor: ligand recognition and allosteric network. J Med Chem 2006;49(4):1313–24.
[54] Yadav PK, Xie P, Banerjee R. Allosteric communication between the pyridoxal 5'-phosphate (PLP) and heme sites in the H2S generator human cystathionine beta-synthase. J Biol Chem 2012;287(45):37611–20.
[55] Joseph RE, Xie Q, Andreotti AH. Identification of an allosteric signaling network within Tec family kinases. J Mol Biol 2010;403(2):231–42.
[56] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.
[57] Le CT. A solution for the most basic optimization problem associated with an ROC curve. Stat Methods Med Res 2006;15(6):571–84.
[58] Böhning D, Böhning W, Holling H. Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. Stat Methods Med Res 2008;17(6):543–54.
[59] Raimondi F, Felline A, Fanelli F. Catching functional modes and structural communication in Dbl family Rho guanine nucleotide exchange factors. J Chem Inf Model 2015;55(9):1878–93.
[60] Nourry C, Grant SGN, Borg J-P. PDZ domain proteins: plug and play! Sci STKE 2003;2003:RE7.
[61] Sheng M, Sala C. PDZ domains and the organization of supramolecular complexes. Annu Rev Neurosci 2001;24(1):1–29.
[62] Bezprozvanny I, Maximov A. PDZ domains: more than just a glue. Proc Natl Acad Sci U S A 2001;98(3):787–9.
[63] Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL. Hidden dynamic allostery in a PDZ domain. Proc Natl Acad Sci U S A 2009;106(43):18249–54.
[64] Zhang M. Scaffold proteins as dynamic switches. Nat. Chem Biol 2007;3 (12):756–7.
[65] Angelova K, Felline A, Lee M, Patel M, Puett D, Fanelli F. Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor. Cell Mol Life Sci 2011;68(7):1227–39.
[66] Bellucci L, Felline A, Fanelli F. Dynamics and structural communication in the ternary complex of fully phosphorylated V2 vasopressin receptor, vasopressin, and beta-arrestin 1. Biochim Biophys Acta Biomembr 2020;1862(9):183355.
[67] Raimondi F, Felline A, Portella G, Orozco M, Fanelli F. Light on the structural communication in Ras GTPases. J Biomol Struct Dyn 2013;31(2):142–57.
[68] Mariani S, Dell'Orco D, Felline A, Raimondi F, Fanelli F, Dunbrack RL. Network and atomistic simulations unveil the structural determinants of mutations linked to retinal diseases. PLoS Comput Biol 2013;9(8):e1003207.
[69] Felline A, Schiroli D, Comitato A, Marigo V, Fanelli F. Structure network-based landscape of rhodopsin misfolding by mutations and algorithmic prediction of small chaperone action. Comput Struct Biotechnol J 2021;19:6020–38.