Supplementary Methods

Workflow of the PSN-ENM approach

The first step in PSN analysis consists in computing the PSG, i.e. an ensemble of nodes and links on a single high-resolution structural model. Building of the PSG is carried out by means of the PSN module implemented in Wordom (1). A number of network features including hubs or node communities are computed as well.

Building the PSG provides the basis to search for the shortest paths between pairs of nodes, i.e. linked nodes connecting two extremities. The server starts computing the shortest communication pathways between all node pairs selected for calculation on job submission and retains only those pathways, in which at least one central node holds correlated motions with either one of the two path extremities. The server provides a global metapath made up by the most recurrent links in the pool of filtered paths. Metapaths represent a coarse/global picture of the structural communication in the considered system. In the result page, the user can filter those paths that begin and end at a given residue pair or that pass through a residue. Such a path filtering provides a novel metapath.

Deep detail of the theory is provided here below and on the webserver.

Building the PSG

Building of the PSG is carried out by means of the PSN module implemented in the Wordom software (1). PSN analysis is a product of graph theory applied to protein structures (2). A graph is defined by a set of vertices (nodes) and connections (edges) between them. In a PSG, each amino acid residue is represented as a node and these nodes are connected by edges based on the strength of non-covalent interactions between residues (3,4). The strength of interaction between residues i and j (I_{ij}) is evaluated as a percentage given by the following equation:

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

where I_{ij} is the percentage interaction between residues i and j; n_{ij} is the number of 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 the differences in size of the amino acid side chains and their propensity to make the maximum number of contacts with other amino acids in protein structures. Glycines, are now included in the PSN analysis. The webPSN server has an internal database with 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 >30,000 biologically relevant molecules and ions (small molecules, lipids, sugars, etc) from the PDB. Additionally, the server automatically identifies un-parametrized molecules in the submitted PDB files and

automatically calculates their normalization factors transparently. In general, the normalization factors are computed as described in the relevant paper by Kannan and Vishveshwra (5). As an example, the normalization factors for the 20 standard amino acids (N_r) was 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 was the average of the maximum interaction value of residue type r over the whole data set of proteins p, in which residue type r had occurred (5).

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. For example, in the crystal structure of bovine rhodopsin holding a 1U19 PDB code, there are two molecules of 11-cis-retinal, which make 132 and 143 contacts, respectively. In that case the considered value is 143. When a PDB file is submitted, the server 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 uploaded coordinates. The internal database of the server is updated monthly and the normalization factors of un-parameterized non-aa residues submitted by the users are not integrated in the database.

I_{ij} are calculated for all node pairs. At a given interaction strength cutoff, I_{min}, any residue pair ij for which $I_{ij} \geq I_{min}$ is considered to be interacting and hence is connected. Node interconnectivity is used to highlight node clusters, where a cluster is a set of connected nodes in a graph. Cluster size, i.e., the number of nodes constituting a cluster, varies as a function of the I_{min}, and the size of the largest cluster is used to calculate the I_{critic} value. The latter is defined as the I_{min}, at which the size of the largest cluster is half the size of the largest cluster at I_{min} = 0.0%. Studies by Vishveshwara's group found that optimal I_{min} corresponds to the one at which the largest cluster undergoes a transition (4). An interaction strength cutoff I_{min} is then chosen and any residue pair ij for which $I_{ij} \geq I_{min}$ is considered to be interacting and hence is connected in the PSG. Therefore, it is possible to obtain different PSGs for the same protein structure depending on the selected I_{min}. Consequently, I_{min} can be varied to obtain graphs with strong or weak interactions

forming the edges between the residues. Finally, to avoid excessive network fragmentation, which would impair the search for shortest communication paths, all resulting clusters were iteratively re-connected by the link with the highest sub-I_{critic} interaction strength. Thus, cluster merging would compensate, at least in part, for the fact that side chain fluctuations are neglected with the PSN-ENM method.

The residues making zero edges are termed as orphans and those that make at least four edges are referred to as hubs at that particular I_{min}. Such 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 (6). The cutoff 4 is close to the upper limit. The majority of amino acid hubs indeed make from 4 to 6 links, with 4 being the most frequent value.

Finally, links are used to highlight network communities, which are sets of highly interconnected nodes such that nodes belonging to the same community are densely linked to each other and poorly connected to nodes outside the community. Communities can be considered as fairly independent compartments of a graph. They were built by identifying all the k=3-cliques, i.e. sets of 3 fully interconnected nodes, and then merging all those cliques sharing at least one node.

Different states of a molecular system, e.g. free or bound, wild type or mutated, inactive or active, monomeric or oligomeric, etc. can be compared in terms of PSG differences or consensus.

ENM-NMA

ENM-NMA is ever increasingly used to study the collective dynamics of complex systems. ENM-NMA is a coarse grained normal mode analysis technique able to describe the vibrational dynamics of protein systems around an energy minimum.

The ENM approach actually implemented in the webserver describes the protein system as Cα-atom coordinates (i.e. ENM-Cα) interacting by a Hookean harmonic potential (7). In particular, the total energy of the system is described by the following Hamiltonian:

$$E = \sum_{i \neq j} k_{ij}(d_{ij} - d_{ij}^0)^2 \qquad (3)$$

where d_{ij} and d_{ij}^0 are respectively the instantaneous and equilibrium distances between Cα-atoms i and j, while k_{ij} is a distance dependent force constant defined by eq. 3:

$$k_{ij} = C\left(\frac{d_{ij}^0}{d_{ij}}\right)^6 \qquad (4)$$

where C is constant (with a default value of 40 Kcal/mol·Å^2) (8).

The approach has been also adapted to handle any molecular system. In this respect, the nucleic acid or small molecule structure is described by the atom nearest to the geometric center.

The cross-correlations of motions for path filtering are obtained from the covariance matrix C (9):

$$C_{ij} = \frac{\sum_{l=1}^{M} \frac{v_{il} v_{jl}}{\lambda_l}}{\left(\sum_{m=1}^{M} \frac{v_{im} v_{im}}{\lambda_m} \right)^{1/2} \left(\sum_{n=1}^{M} \frac{v_{jn} v_{jn}}{\lambda_n} \right)^{1/2}} \quad (5)$$

where C_{ij} denotes the correlation between particles *i* and *j, M* is the number of modes considered for computation (by default, the first 10 non-zero frequency modes), and v_{xy} and λ_y are, respectively, the x^{th} element and the associated eigenvalue of the y^{th} mode.

Search for the shortest communication pathways

The search for the shortest path(s) between pairs of nodes as implemented in the PSN-path module of Wordom relies on the Dijkstra's algorithm (10). Paths are searched by combining PSN data with cross-correlation of atomic motions calculated by the covariance matrix inferred from ENM-NMA.

Following calculation of the PSG and of correlated motions, the procedure to search for the shortest path(s) between each residue pair consists of (a) searching for the shortest path(s) between each selected amino acid pair based upon the PSN connectivities, and (b) selecting the shortest path(s) that contains at least one residue correlated (i.e. with a correlation coefficient $\geq$0.7) with either one of the two extremities. All residues selected for building the PSG are employed for path search.

Thus, the paths that pass the filtering stage(s) constitute the pool of paths of a system at given I_{min} and correlation coefficient cutoffs. The statistical analysis of such pool of paths can lead to the building of global metapaths constituted by the most recurrent nodes and links in the pool (a recurrence cutoff of 10% was set in the webserver). Metapaths represent a coarse/global picture of the structural communication in the considered system. In the result page, the user can filter those paths that begin and end at given residue pair(s) or that pass through a residue. Such a path filtering provides a novel metapath.

References

1. Seeber, M., Felline, A., Raimondi, F., Muff, S., Friedman, R., Rao, F., Caflisch, A. and Fanelli, F. (2011) Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J Comput Chem*, **32**, 1183-1194.
2. Vishveshwara, S., Brinda, K.V. and Kannan, N. (2002) Protein structure: insights from graph theory. *J. Theor. Comput. Chem.*, **1**, 187-211.
3. Vishveshwara, S., Ghosh, A. and Hansia, P. (2009) Intra and inter-molecular

communications through protein structure network. *Curr Protein Pept Sci*, **10**, 146-160.
4. Brinda, K.V. and Vishveshwara, S. (2005) A network representation of protein structures: implications for protein stability. *Biophys J*, **89**, 4159-4170.
5. Kannan, N. and Vishveshwara, S. (1999) Identification of side-chain clusters in protein structures by a graph spectral method. *J Mol Biol*, **292**, 441-464.
6. Bhattacharyya, M., Upadhyay, R. and Vishveshwara, S. (2012) Interaction signatures stabilizing the NAD(P)-binding Rossmann fold: a structure network approach. *PLoS One*, **7**, e51676.
7. Tirion, M.M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett*, **77**, 1905-1908.
8. Kovacs, J.A., Chacon, P. and Abagyan, R. (2004) Predictions of protein flexibility: First-order measures. *Proteins*, **56**, 661-668.
9. Van Wynsberghe, A.W. and Cui, Q. (2006) Interpreting correlated motions using normal mode analysis. *Structure*, **14**, 1647-1653.
10. Dijkstra, E.W. (1959) A Note on Two Problems in Connexion with Graphs. *Numer. Math.*, **1**, 269-271.