SUPPLEMENTARY METHODS

WebPSN relies on the Wordom software (http://wordom.sourceforge.net/) (Seeber, et al., 2011) to compute the ENM-NMA, to build the PSG, and to search for the communication pathways. Additional, yet-unreleased programs and scripts are used for data mining and post-processing, the bulk of which concern communication path processing and analysis. Two modes of input setting have been implemented. On the one hand, the "basic input” mode requires the user to make only a limited number of choices, namely, the selection of the atom subset which has to be processed, the apical residues used to define the paths and, optionally, a list of important residues (conserved, mutated or else). On the other hand, the "advanced input” mode gives the user full control on the computation through a much more complex input panel.

Building of the PSG

Building of the PSG is carried out by means of the PSN module implemented in the Wordom software (Seeber, et al., 2011). PSN analysis is a product of graph theory applied to protein structures (Vishveshwara, et al., 2002). 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 (Vishveshwara, et al., 2009). The strength of interaction between residues i and j ($I_{ij}$) is evaluated as a percentage given by equation 1:

\[ I_{ij} = \frac{n_{ij}}{\sqrt{N_i N_j}} \times 100 \]

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. The normalization factors for the 20 amino acids were derived from the work by Kannan and Vishveshwara (Kannan and Vishveshwara, 1999). WebPSN automatically identifies un-parametrized non amino acid residues in the submitted PDB structure and offers to either ignore them, automatically compute the factor from the uploaded pdb, use the factor from another parametrized residue, or accept a value from the user. The last two options are actually discouraged, while submitting multiple structures with the unknown residue can improve the reliability of the automatic factor computing. The default setting, nevertheless, is to ignore the unparameterized residue (Supplementary Figure 1, Figure S1).

Thus, $I_{ij}$ are calculated for all node pairs. At a given interaction strength cutoff $I_{\text{min}}$, any residue pair $ij$ for which $I_{ij}$ $\geq$ $I_{\text{min}}$ 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 $I_{\text{min}}$. The WebPSN server automatically computes the $I_{\text{min}}$ cutoff based on a tested procedure (Raimondi, et al., 2013).

Residues making zero edges are termed as orphans and those that make four or more edges are referred to as hubs at that particular $I_{\text{min}}$. The definition of $I_{ij}$ for evaluating the hub character of a residue is slightly different from that given in equation 1:

\[ I_{ij} = \frac{n_{ij}}{N_i} \times 100 \]

where the denominator holds only the normalization value of the residue i whose hub behavior is being evaluated.
The advanced PSN options in the WebPSN server include the setting of: a) the distance cutoff for computing the interaction strength (default: 4.5 Å); b) the minimal number of links a node must be involved in to be considered as a hub (i.e. at least 4); c) the number of adjacent (in sequence) residues that will be skipped while computing the interaction strength (default: 1, which is equivalent to no skip); d) whether N-terminal and C-terminal amino- and carboxy-groups will be explicitly considered in the computation of the interaction strength or not.

Figure S1. Snapshot of the input setup web page.
Calculation of cross-correlations of motions through ENM-NMA

The ENM method is activated only if the “PSN-Path” option is set and serves to compute the cross-correlations of atomic motions and a number of indices.

In the updated version of the Wordom software recently released (Seeber, et al., 2011) we included the ENM approach that describes the system as Cα-atom coordinates (i.e. ENM-Cα) interacting by a Hookean harmonic potential (Tirion, 1996). In particular, the total energy of the system is described by the following Hamiltonian:

\[ E = \sum_{i \neq j} k_{ij} (d_{ij} - d_{ij}^0)^2 \]  

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. 4:

\[ k_{ij} = C \left( \frac{d_{ij}^0}{d_{ij}} \right)^6 \]  

where \( C \) is constant (with a default value of 40 Kcal/mol·Å²) (Kovacs, et al., 2004). We added in Wordom a variant of the ENM approach, the Rotation Translation Block (ENM-RTB) (Durand, et al., 1994; Raimondi, et al., 2013; Tama, et al., 2000). The difference between the basic ENM and RTB essentially resides in the strategies employed to reduce the dimensionality of the Hessian matrix for efficient diagonalization.

Both ENM methods are implemented in the WebPSN server (the ENM-Cα approximation being the default one).

ENM needs a reference atom for each residue. The Cα-atom is taken by default for the amino acids; for Cα-less residues, the atom in the middle of the residue-atom list is taken by default; the user can, however, select a specific atom. The advanced ENM setting allows to choose one out of the two available options concerning the force constant in the harmonic potential describing pairwise interatomic interaction: a) the “Kovacs” option, in which the force constant depends on the distance between the interacting particles; and b) the “linear” option, in which the force constant is equal to 1 for pairwise interactions between the Cα-atoms lying within a cutoff distance chosen by the user, and equal to 10 for adjacent Cα-atoms. Within the ENM advanced options, the user can also set: a) a “distance cutoff” that is the distance between atom pairs beyond which the interaction is neglected, which serves only for the “linear” option; and b) the number of Normal Modes used to compute the cross-correlation of the atomic motions instrumental in path filtering.

Cross-correlations of motions for path filtering are obtained from the covariance matrix \( C \) (Van Wynsberghge and Cui, 2006):

\[ C_{ij} = \frac{\sum_{l=1}^{M} V_{il} V_{jl}}{\left( \sum_{m=1}^{M} \frac{V_{im} V_{jm}}{\lambda_m} \right)^{1/2} \left( \sum_{n=1}^{M} \frac{V_{jn} V_{jn}}{\lambda_n} \right)^{1/2}} \]
where $C_{ij}$ denotes the correlation between particles $i$ and $j$, $M$ is the number of modes considered for computation (by default, the first 50 non-zero frequency modes), and $\nu_{xy}$ and $\lambda_y$ are, respectively, the $x$th element and the associated eigenvalue of the $y$th mode.

Additional ENM-based indices useful for path characterization are derivations of the inter-node distance fluctuations according to the following equation (adapted from the works by Hinsen (Hinsen, 1998) and by Wang and co-workers (Wang, et al., 2004)):

$$
(\Delta R_{ij})^2 = \frac{\sum_{m=1}^{M} \left( \frac{\bar{R}_{ij}^0 + \Delta \bar{R}_{mj} - \Delta \bar{R}_{mi}}{\lambda_m} \right)^2}{\lambda_m}
$$

where $\bar{R}_{ij}^0$ and $\Delta \bar{R}_{mj(i)}$ are, respectively, the distance vector between the $i$th and $j$th particles in the reference structure and the displacement of atom $i$ (or $j$) along the $m$th mode.

For the ENM-RTB approach, inter-block distance fluctuations can be obtained by the following equation:

$$
(\Delta R_{b1b2})^2 = \frac{1}{n_{b1} \cdot n_{b2}} \sum_{i=1}^{n_{b1}} \sum_{j=1}^{n_{b2}} (\Delta R_{ij})^2
$$

where $n_{b1}$ and $n_{b2}$ are, respectively, the number of atoms composing blocks 1 ($b1$) and 2 ($b2$). By summing the MSDF between all node pairs in a path $(p)$ the $\text{MSDF}^p$ index is obtained according to the following equation:

$$
\text{MSDF}^p = \frac{1}{L} \sum_{n=1}^{N} (\Delta R_{n,n+1})^2
$$

where $N$ and $L$ are, respectively, the number of nodes and links forming $p$, $\Delta R$ are the inter-residue distance fluctuations as determined by equations 6 and 7, $n$ and $n+1$ are consecutive nodes along $p$.

This index, whose formulation is reminiscent of the one shown in the work of Chennubhotla and Bahar (Chennubhotla and Bahar, 2007), accounts, at least in part, for the propensity in signal transfer through a given communication path. In this respect, the lower the inter-node distance fluctuations, i.e. lower $\text{MSDF}^p$, the higher the stiffness of the considered path and hence its propensity to sustain signal transfer. The WebPSN server allows the user to interactively rank paths according to the $\text{MSDF}^p$ and visualize them on the 3D structure.

**Combining PSN with ENM to search for the shortest communication paths**

The search for the shortest path(s) between pairs of nodes as implemented in Wordom relies on Dijkstra’s algorithm (Dijkstra, 1959). The search for the shortest communication pathways requires user’s specified residue pairs, which constitute the extremities of the path.

The correlation matrix obtained from ENM-NMA enters in the filtering stage of the search for the shortest paths between pairs of nodes belonging to the same network cluster (i.e. a collection of nodes connected by at least one link). Node pairs can be set by the user by selecting either pairs or ranges of residues or the whole protein. Thus, the ENM-based filtering stage consists in retaining all those paths, in which at least one intermediate node holds correlated motions with either one of the two extremities (i.e. the first and last amino acids in the path).

The relative number of residues holding correlated motions with either one of the two extremities is quantified by the correlation score.
Those paths that pass the filtering stage constitute the pool of paths of a system at given $I_{\text{min}}$, correlation coefficient, and minimum length cutoffs. The statistical analysis of such a pool of paths can lead to the building of global meta paths constituted by the most recurrent nodes and links in the pool.

The advanced PSN-Path options allow for the setting of the following parameters: a) minimum path length, excluding the two terminal nodes (default: 5); b) minimum correlation between a residue in the path and one of the apical residues for the path to be accepted (default: 0.6); c) sequence distance cutoff for two residues to be considered as path extremities, which serves in those cases in which residue ranges or the whole protein are selected for path calculation (default: 5); d) similarity cutoff for path clustering (default: 0.8 for clustering method 1, and 0.4 for clustering method 2); e) node and link recurrence cutoff used in building meta paths (default: 5).

Cluster analysis may provide finer information on the predicted pathways. The WebPSN server implements two path clustering methods, both based on a different path Similarity Score ($S$). The first method, hereafter referred to as clustering method 1, relies on a similarity score between paths $a$ and $b$, computed according to the following equation:

$$ S_{ab} = \frac{C_L}{\min(L_a, L_b)} $$

(9)

where $C_L$ is the number of common links in both paths, and $L_a$ and $L_b$ are the number of links in path $a$ and $b$, respectively. $S$ ranges from 0, for two totally different paths, to 1, when the smaller path is completely included in the longer one, i.e. the smaller path is a subset of the longer path. Once the $S$ cutoff is selected, the path clustering procedure is such that a path is assigned to a cluster if there is at least one path in such a cluster with a similarity score $\geq S$. A path not assignable to existing clusters initiates a new cluster and the procedure continues until all paths are assigned. Finally, all the clusters made up of a single path are discarded and their paths are considered as un-clustered paths.

The second, recently reported (Raimondi, et al., 2013), method, referred to as clustering method 2, relies on a different similarity score and employs clustering method 2 employed the Quality Threshold (QT) algorithm, according to a similarity score (Heyer, et al., 1999). This method relies on a similarity score ($S$) between paths $a$ and $b$, computed according to the following equation:

$$ S_{ab} = \left( \frac{2C_N}{N_a + N_b} \right) \cdot 0.15 + \left( \frac{2\max(C_p)}{N_a + N_b} \right) \cdot 0.4 + \left( \frac{2C_L}{L_a + L_b} \right) \cdot 0.45 $$

(10)

Where $C_N$ is the number of common nodes in both paths; $N_a$ and $N_b$ are the number of nodes in paths $a$ and $b$, respectively; $\max(C_p)$ is the greatest number of nodes at the same position in the path as obtained by sliding the nodes of path $a$ over the nodes of path $b$ by one position at a time and then inverting the two paths (i.e. sliding path $b$ over path $a$); $C_L$ is the number of common links in both paths (i.e. those links connecting pairs of identical nodes); and $L_a$ and $L_b$ are the number of links in path $a$ and $b$, respectively. The similarity score ranges from 0, for two totally different paths, and 1, for two identical paths (Raimondi, et al., 2013). Irrespective of the clustering method, for each cluster, following a pairwise comparison of all cluster members, the center is computed as well, which is the path with the highest average $S$ among all the paths in the cluster. With method 1, cluster centers may be longer than the majority of paths in the clusters, whereas this does not happen with method 2.

Clustering method 1 relies on a simpler $S$ score and on a faster algorithm than method 2; therefore, it has been set as a default method in the WebPSN server.

Meta paths and cluster centers can be all visualized on the 3D structure through the graphical interface.
**Graphical representation of the Output**

Results can be both interactively visualized on the web page and downloaded. Web 3D visualizations (i.e. by the Jmol applet) include PSG, hub distribution, meta paths, and centers of the three most populated clusters, which can be selected through a predisposed box.

As for the 3D representations, in general, nodes are shown as spheres centered on the C\textalpha-atom for amino acid residues, and on the user-selected atom for non amino acid residues. In deep detail: a) in the PSG representation, nodes and links are colored according to the three most populated node clusters they belong to (where a node cluster is an ensemble of nodes connected by at least one link (Raimondi, et al., 2013)); red, green, and blue indicate clusters 1, 2, and 3, respectively. The diameter of the sphere is proportional to the number of links made by the considered node (Figure S2). b) The “hubs” representation shows the 3D distribution of all hubs. c) The “global meta path” and “meta path” representations display, respectively, the meta path from all paths and the meta paths from the paths in each of the first three most populated path clusters. The diameter of the spheres and the width of the links are proportional to the correspondent recurrences in the considered path pool. Finally, d) the “cluster center” representation shows the centers of the three most populated clusters. In general, the representation of pathways is such that each node is represented as a sphere, whereas links between node pairs are represented by sticks.

The result page shows also the distribution of: a) path length; b) MSDF\textsuperscript{P} index; c) hub content; d) correlation score; f) node frequency in paths; and g) frequency of four-amino acid fragments in paths. A table is shown as well listing the number of all nodes, hubs, and links in the PSG or in each node cluster as well as the number of paths and of nodes and links in those paths (Figure S2). Finally, a box in the output page allows the user to provide a list of residues (which can also be taken from the "important residues" field of the input module) to be visualized in the context of the protein structure network and the communication pathways.
Figure S2. Selected graphical representations of the output are shown.
REFERENCES