JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

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

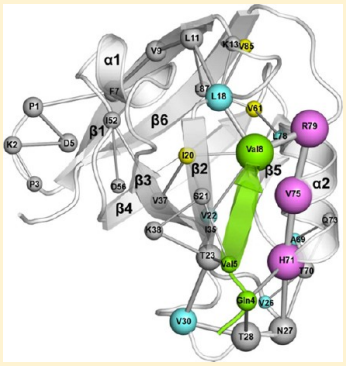
Francesco Raimondi,[†,‡] Angelo Felline,[†,‡] Michele Seeber,[†,‡] Simona Mariani,[†,‡] and Francesca Fanelli*[,†,‡]

[†]Department of Life Sciences, via Campi 183, 41125, Modena, Italy

[‡]Dulbecco Telethon Institute (DTI), via Campi 183, 41125, Modena, Italy

Supporting Information

ABSTRACT: Graph theory is being increasingly used to study the structural communication in biomolecular systems. This requires incorporating information on the system's dynamics, which is time-consuming and not suitable for high-throughput investigations. We propose a mixed Protein Structure Network (PSN) and Elastic Network Model (ENM)-based strategy, i.e., PSN-ENM, for fast investigation of allosterism in biological systems. PSN analysis and ENM-Normal Mode Analysis (ENM-NMA) are implemented in the structural analysis software Wordom, freely available at http://wordom.sourceforge.net/. The method performs a systematic search of the shortest communication pathways that traverse a protein structure. A number of strategies to compare the structure networks of a protein in different functional states and to get a global picture of communication pathways are presented as well. The approach was validated on the PDZ2 domain from tyrosine phosphatase 1E (PTP1E) in its free (APO) and peptide-bound states. PDZ domains are, indeed, the systems whose structural communication and allosteric features are best characterized both in vitro and in silico. The agreement between predictions by the PSN-ENM method and in vitro evidence is remarkable and comparable to or higher than that reached by more time-consuming computational approaches tested on the same biological system. Finally, the PSN-ENM method was able to reproduce the salient communication features of unbound and bound PTP1E inferred from molecular dynamics simulations. High speed makes this method suitable for high throughput investigation of the communication pathways in large sets of biomolecular systems in different functional states.



1. INTRODUCTION

Graph theory is a branch of discrete mathematics aimed at the study of graphs, mathematical structures used to model pairwise relations between objects. Concepts and methods borrowed from graph theory are being increasingly used to study several aspects of structural biology. The representation of biomolecular structures as networks of interacting amino acids/nucleotides has in fact been employed to investigate and elucidate complex phenomena such as protein folding and unfolding, protein stability, the role of structurally and functionally important residues, protein−protein and protein−DNA interactions, and intraprotein and interprotein communication and allosterism.[1−15]

These works rely on methods that differ in the set of graph construction rules. The graph-based approach proposed by Vishveshwara and co-workers[7] and defined as Protein Structure Network (PSN) is the one that we have recently implemented in the Wordom software.[16] It computes network features (e.g., nodes, hubs (i.e., hyper-connected nodes), links, etc.) and shortest communication pathways from Molecular Dynamics (MD) trajectories (herein indicated as PSN-MD). With this approach, information on the system's dynamics participates both in the building of the Protein Structure Graph (PSG) and in the search for the shortest communication paths. As for the PSG, the dynamics of the systems contributes in terms of recurrence of network components (e.g., hubs, links, etc.) in the frames constituting the MD trajectories. With respect to the search for the shortest communication pathways, the fluctuation dynamics of the system contributes in terms of both correlated motions and path occurrence in the trajectory frames (i.e., path frequencies). The computational costs in both achieving and analyzing MD trajectories make the PSN-MD prohibitive for a comparative usage on large data sets of proteins.

The evidence that functional dynamics of proteins relies on highly cooperative, low frequency, global/essential modes caused the diffusion of methods like the Normal Mode Analysis (NMA) to be able to infer such collective modes.[17,18] The observed robustness of global modes with respect to details in atomic coordinates or specific interatomic interaction and their insensitivity to the specific energy functions and parameters that define the force field provided support to the development of simplified, i.e., coarse-grained (CG), descrip-

Received: December 2, 2012
Published: April 4, 2013

tions of protein structures such as the Elastic Network Models (ENM; reviewed in ref 17). The latter rely on the fact that the property that apparently dominates the shape of global modes is the network of inter-residue contacts, which is a purely geometric quantity defined by the overall shape or native contact topology of the protein. In recent years, ENM-based NMA (ENM-NMA) contributed significantly to improving our understanding of the collective dynamics of a number of allosteric proteins (reviewed in ref 17). The ENM-NMA approach is implemented as well in the Wordom software.[16]

We propose a strategy for fast large-scale computations of the structural communication features of complex systems. The PSGs are computed on a single high resolution structure rather than an MD trajectory, and information on the system's dynamics (cross-correlation of atomic motions) is supplied by ENM-NMA (the method is hereafter indicated as PSN-ENM). A number of strategies to compare the structure networks in different functional states and to get a global picture of communication pathways will be presented as well.

The approach was validated on the PDZ2 domain from tyrosine phosphatase 1E (PTP1E) in its free (APO) state ($PDZ2^{APO}$; PDB code 3LNX) and in complex with the C-terminal peptide from RA-GEF2 (RA-GEF2-Ct) ($PDZ2^{PEP}$; PDB code 3LNY).[19] According to the CATH classification (http://www.cathdb.info/), this domain belongs to the class of mainly β proteins and holds a roll architecture made of six antiparallel β-strands (Figure 1). The PDZ topology includes also two α-helices. The binding pocket of RA-GEF2-Ct involves $\beta2$, $\alpha2$, and their preceding and following loops (Figure 1).

PDZs are protein−protein interaction domains typically involved in the targeting and assembly of multiprotein signaling complexes. Proteins generally recognize the PDZ domains through their C-terminal segments (four to seven amino acids in length).[20,21] In addition to passive scaffolding, a subset of these domains has also been demonstrated to be capable of allosterically regulating distal sites involved in effector binding.[22−24] Collectively, PDZ domains are the systems whose structural communication and allosteric features have been best characterized both in vitro[19,23,25−30] and in silico.[31−39]

Predictions by the PSN-ENM method were in remarkable agreement with experimental evidence on ligand-induced perturbations in the communication features of PTP1E. The agreement is comparable if not higher than that reached by other more time-consuming computational approaches tested on the same biological system. Finally, the method was able to reproduce the salient communication features of unbound and bound PTP1E inferred from MD simulations.

Being fast, it is suitable for high throughput comparative characterization of structural communication properties of large sets of biomolecular systems in different functional states.

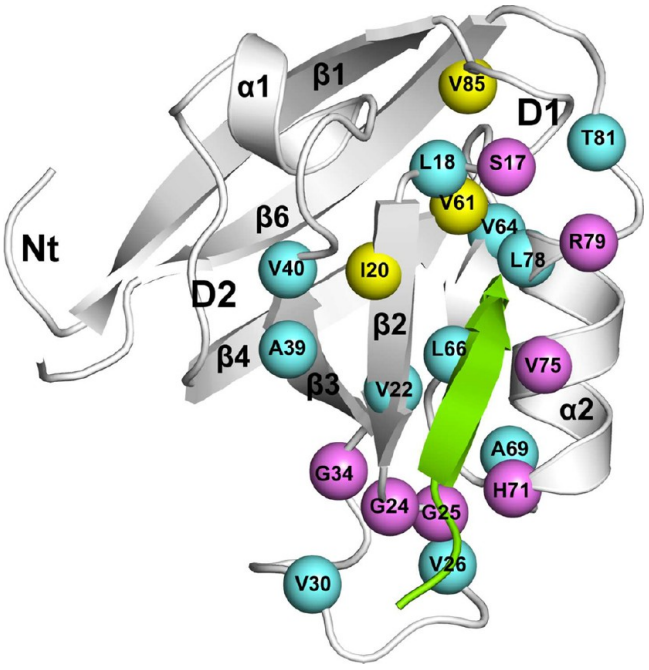


Figure 1. Amino acids involved in signal transfer. A cartoon representation of the crystallographic structure of the $PDZ2^{PEP}$ is shown; the peptide is green. The amino acids highlighted by TMCA,[25] NMR[26] experiments, or both are, respectively, represented as pink, light blue, and yellow spheres centered on the Cα atoms.

2. THE PROPOSED METHOD

Figure 2 shows a workflow of the PSN-ENM approach. The first step consists in performing the PSN analysis on a single high resolution structure, which is instrumental in computing the network components (e.g., nodes, hubs, links, etc.), and to build the PSG. The latter represents also the basis to search for the shortest paths between pairs of nodes, i.e., linked nodes connecting two extremities. In this framework, once the two extremities of interest have been specified, the algorithm first defines all possible shortest communication paths between such extremities and then filters the results according to the cross-correlation of atomic motions derived from ENM-NMA. The outcome of this stage is the total pool of paths for the system under study. Meta paths made of the most recurrent nodes and links in the path pool (i.e., global meta paths) are worth computing to infer a coarse/global picture of the structural communication in the considered system. More resolved information on the most likely communication pathways can be additionally inferred from cluster analysis of the path pool. Path clusters can be analyzed through cluster meta paths, cluster centers, as well as computational descriptors of path features such as the mean square distance fluctuations between all node pairs in a path (p) ($MSDF^p$), which is a measure of path stiffness (see section 2.2 for explanation). All parameters defined for the PSN and PSN-PATH analyses carried out in this case study are summarized in Table 1.

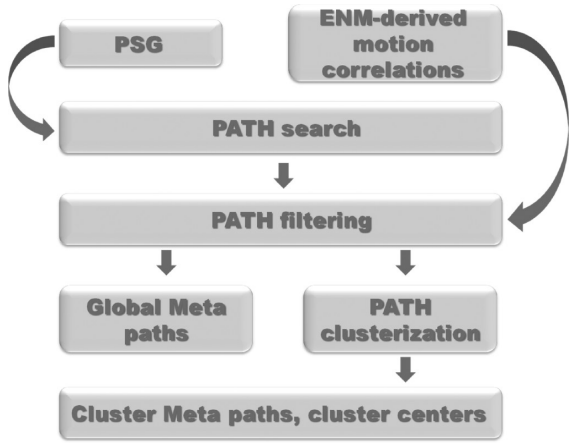


Figure 2. Flowchart of the approach.

Table 1. Network Components and Parameters

	$PDZ2^{APO}$	$PDZ2^{PEP}$
I_{min}[a]	1.79	2.61
nodes[b]	65	74
hubs[c]	8	14
links[d]	70	89
node pairs[e]	4000	4500
CCcutoff[f]	0.6	0.6
paths[g]	770	845
length[h]	12.44 ± 3.77	10.45 ± 2.50
score[i]	0.21 ± 0.12	0.22 ± 0.12
MSDF[j]	$4 \times 10^{-4} \pm 0.80 \times 10^{-4}$	$3 \times 10^{-4} \pm 0.90 \times 10^{-4}$
$Sim1_{cutoff}$[k]	0.8	0.8
Cluster1_M1[l]	769	828
Sim2cutoff[m]	0.4	0.4
Cluster1_M2[n]	492	322
Cluster2_M2[n]	64	189
Cluster3_M2[n]	58	112

[a]Interaction strength cutoff (%). [b]Total number of nodes. [c]Total number of hubs. [d]Total number of links. [e]Number of node pairs employed for path search. [f]Correlation coefficient cutoff concerning atomic motions employed as a path filter. [g]Total number of paths. [h]Path length averaged over the total number of paths. [i]Correlation score averaged over the total number of paths. The score is the ratio between the number of correlated amino acids and path length; the latter excludes the two extremities. [j]Path Mean Square Distance Fluctuation averaged over the total number of paths. [k]Similarity cutoff employed for path clusterization by method 1. [l]Population of the first cluster achieved by clusterization method 1. [m]Similarity cutoff employed for path clusterization by method 2. [n]Population of the first three clusters achieved by clusterization method 2.

2.1. Building of the PSG. Building of the PSG is carried out by means of the PSN module implemented in the Wordom software.[16] PSN analysis is a product of graph theory applied to protein structures.[40] 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 noncovalent interactions between residues.[7] The strength of interaction between residues i and j (I_{ij}) is evaluated as a percentage given by eq 1:

$$I_{ij} = \frac{n_{ij}}{\sqrt{N_i N_j}} \times 100 \tag{1}$$

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), and 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.[41] The possibility to compute the normalization factor of non-amino acid molecules has been implemented as well in Wordom.

Thus, I_{ij} is calculated for all node pairs. 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. The 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_{min}. The definition of I_{ij} for evaluating the hub character of a residue is slightly different from that given in eq 1:

$$I_{ij} = \frac{n_{ij}}{N_i} \times 100 \tag{2}$$

where the denominator holds only the normalization value of the residue i whose hub behavior is being evaluated. The rationale behind the employment of different equations for computing interaction strength and hub behavior is as follows. While eq 1 is used to calculate how strong the interaction between two residues is, eq 2 assesses the propensity of a residue to establish noncovalent interactions with the surrounding side-chains. Therefore, the goal of eq 2 is not to link nodes in the network but rather to identify residues with high interaction propensity. However, as already demonstrated by Ghosh and co-workers for other systems,[42] for both $PDZ2^{APO}$ and $PDZ2^{PEP}$, all those hubs defined by eq 2 are found involved in ≥4 links in the network built by eq 1.

Node interconnectivity is finally used to highlight cluster-forming nodes, 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 (reviewed in ref 7). Such transition generally occurs in the 2−5% I_{min} range. Because this transition is system-dependent, when comparing two or more states of the same protein (like the unbound and bound states of PDZ2 in the present study) a homogeneous treatment of such states requires selection, for each state, of an I_{min} value close to the transition of the largest cluster, rather than setting the same I_{min} for all. In this respect, the optimal I_{min} values, i.e., I_{crit}, were approximated to the second decimal place. This treatment allows most nodes and connectivities between them to fall into a single cluster, which is the largest node cluster at the selected I_{min}. Thus, at I_{min} 1.79% and 2.61%, the largest cluster turned out to hold all the hubs as well as 80% and 69% of the total nodes characterizing the networks of $PDZ2^{APO}$ and $PDZ2^{PEP}$, respectively (Table 1).

Perturbations in a network due to ligand binding can be inferred by plotting nodes and links peculiar to each state, as well as by highlighting in three dimensions commonalties and differences between the PSGs of the unbound and bound states.

2.2. Calculation of Cross-Correlations of Motions through ENM-NMA. In the updated version of the Wordom software recently released,[16] we included the ENM approach that describes the system as Cα-atom coordinates (i.e., ENM-Cα), interacting by a Hookean harmonic potential.[43] In particular, the total energy of the system is described by the following Hamiltonian:

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

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 4:

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

where C is constant (with a default value of 40 kcal/mol·Å^2).[44]

We added in Wordom two less coarse ENM approaches, the Vibrational Subsystem Analysis (ENM-VSA)[45] and the Rotation Translation Block (ENM-RTB).[46,47] The difference between the basic ENM and the other two methods essentially resides in the strategies employed to reduce the dimensionality of the Hessian matrix for efficient diagonalization.

According to the ENM-VSA, the total system is divided into two components: (a) the subsystem, which is defined as the region of interest (i.e., part of the system that controls functionality) and is hence subjected to the vibrational analysis, and (b) the environment, which consists of the less important remaining portions of the molecule, whose effects on the subsystem are implicitly taken into account. The Hessian matrix is thus decomposed as follows:

$$H = \begin{matrix} H_{ss} & H_{se} \\ H_{es} & H_{ee} \end{matrix} \quad (5)$$

where H_{ss}, H_{se}, and H_{ee} are the subsystem−subsystem, subsystem−environment, and environment−environment Hessians, respectively.

At a minimum of the potential energy, the environmental degrees of freedom can be integrated out by eq 6:

$$H_{ss}^{eff} = H_{ss} - H_{se}H_{ee}^{-1}H_{es} \quad (6)$$

where H_{ss}^{eff} is the effective Hessian matrix of the subsystem which takes into account the environmental effect.

With the ENM-RTB model, the less coarse of the three models, the system is made of n_b rigid blocks, and the normal modes are expressed as rigid body rotations of the constituent blocks, each of them having six degrees of freedom (three translational and three rotational). In this study, the heavy atoms of each amino acid residue constitute a different block. A $3N \times 6n_b$ matrix, P, is used to project the original Hessian matrix (H) from a $3N$-dimensional space to a $6n_b$-dimensional one through the following transformation:

$$H_b = P^T H P \quad (7)$$

H_b is then diagonalized:

$$V_b^T H_b V_b = \Lambda_b \quad (8)$$

with V_b^T and Λ_b being, respectively, the eigenvectors and eigenvalues in the reduced subspace. The resulting eigenvectors can be reverted back into the full $3N$-dimensional space by applying the inverse projection:

$$V = P^T V_b \quad (9)$$

The three different analyses ultimately provide a number of normal modes holding 3N dimensions where N is the number of Cα atoms for ENM and ENM-VSA, and the number of all heavy atoms for ENM-RTB.

In this case study, the cross-correlations between experimental and calculated B-factors considering the first 50 modes yielded better results for the ENM-RTB (0.72 and 0.6 for PDZ2^{APO} and PDZ2^{PEP}, respectively) than both the ENM-Cα (0.65 and 0.53 for PDZ2^{APO} and PDZ2^{PEP}, respectively) and ENM-VSA (0.69 and 0.56 for PDZ2^{APO} and PDZ2^{PEP}, respectively) approaches. Therefore, although we probed all three ENM-NMA approximations, the case study shown herein is based on the employment of the ENM-RTB to derive the correlation of motions for path calculation as well as a number of descriptors for path characterization. In this respect, cross-correlations of motions were obtained from the covariance matrix C:[48]

$$C_{ij} = \frac{\frac{\sum_{l=1}^{M} \nu_{il}\nu_{jl}}{\lambda_l}}{\left(\frac{\sum_{m=1}^{M} \nu_{im}\nu_{im}}{\lambda_m}\right)^{1/2}\left(\frac{\sum_{n=1}^{M} \nu_{jn}\nu_{jn}}{\lambda_n}\right)^{1/2}} \quad (10)$$

where C_{ij} denotes the correlation between particles i and j, M is the number of modes considered for computation (the first 50 nonzero frequency modes), ν_{xy} and λ_y are, respectively, the xth element and the associated eigenvalue of the yth mode.

Additional ENM-based indices useful for path characterization are derivations of the internode distance fluctuations according to the following equation (adapted from the works by Hinsen[49] and by Wang and co-workers[50]):

$$(\Delta R_{ij})^2 = \frac{\sum_{m=1}^{M} (|\vec{R}_{ij}^{\,0} + \Delta\vec{R}_{mj} - \Delta\vec{R}_{mi}| - |\vec{R}_{ij}^{\,0}|)^2}{\lambda_m} \quad (11)$$

where $\vec{R}_{ij}^{\,0}$ and $\Delta\vec{R}_{m_{i(j)}}$ are, respectively, the distance vector between the ith and jth particles in the reference structure and the displacement of atom i (or j) along the mth mode.

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

$$(\Delta R_{b1b2})^2 = \frac{1}{n_{b1}n_{b2}} \sum_{i=1}^{n_{b1}} \sum_{j=1}^{n_{b2}} (\Delta R_{ij})^2 \quad (12)$$

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 MSDF^p index is obtained according to the following equation:

$$\mathrm{MSDF}^p = \frac{1}{L} \sum_{n=1}^{N-1} (\Delta R_{nn+1})^2 \quad (13)$$

where N and L are, respectively, the number of nodes and links forming p, ΔR is the inter-residue distance fluctuations as determined by eqs 11 and 12, 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 Bahr,[13] relates, at least in part, to the propensity in signal transfer through a given communication path. In this respect, we postulate that the lower the internode distance fluctuations, i.e., the lower the MSDF^p, the higher the stiffness of the considered path, and the easier the structural communication through it.

The ENM module of Wordom allows one to compute also theoretical B-factors:[51]

$$B_n^T = \frac{8\pi^2 kT}{3} \sum_{m=1}^{M} \frac{\nu_{mn}^2}{\lambda_m} \quad (14)$$

where k is the Boltzmann constant and T is the temperature in K (300 K).

2.3. 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.[52]

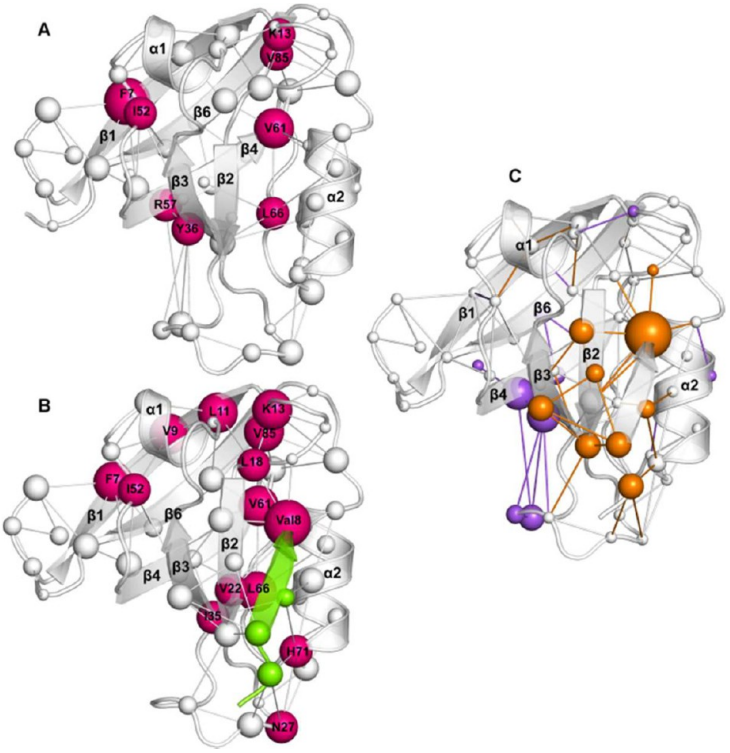


Figure 3. 3D PSG representations of the PDZ2^{APO} (A) and PDZ2^{PEP} (B). Nodes are spheres centered on the Cα atoms, whose diameter is proportional to the number of links made by the node. Magenta spheres are the hubs. (C) The difference 3D PSG between PDZ2^{APO} and PDZ2^{PEP} is shown. Nodes and links shared by the two functional states are white, whereas those peculiar to the unbound and bound states are violet and orange, respectively. The radii of the white spheres are proportional to the difference of node interconnectivity between unbound and bound states, whereas the radii of violet and orange spheres are proportional to the number of links made by the considered specific node.

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. In this case study, all residue pairs except those at sequence distance ±4 were set, leading, respectively, to 4000 and 4500 pairs for PDZ2^{APO} and PDZ2^{PEP} (Table 1). Thus, the ENM-based filtering stage consists of retaining all those paths in which at least one node holds correlated motions with either one of the two extremities (i.e., the first and last amino acids in the path). The results shown herein were produced at a 0.6 correlation coefficient cutoff (Table 1). The relative number of residues holding correlated motions with either one of the two extremities is quantified by the correlation score, i.e., the ratio between number of correlated residues and path length; the latter excludes the two extremities.

The paths that pass the filtering stage constitute the pool of paths of a system at given I_{min} and correlation coefficient 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. In this study, we built meta paths to describe and compare the global communication features of PDZ2^{APO} and PDZ2^{PEP}. Those meta paths are made of nodes present in ≥5% of the considered path pool (i.e., "frequent nodes") and of links satisfying both conditions of being present in ≥5% of the paths and of connecting "frequent nodes."

Cluster analysis may provide finer information on the predicted pathways. In this study, we employed two path clusterization methods: (a) a new method, clusterization method 1, and (b) the one recently reported,[15] clusterization method 2.

Clusterization method 1 relies on a similarity score (S) between paths a and b, computed according to the following equation:

$$S_{a,b} = \frac{C_L}{\min(L_a, L_b)} \tag{15}$$

where C_L is the number of common links in both paths and L_a and L_b are the number of links in paths 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 clusterization 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 ≥S. A path not assignable to existing clusters initiates a new cluster, and the procedure continues until all paths are assigned. Finally, all of the clusters made up of a single path are discarded, and their paths are considered as unclustered paths. The selected S cutoff for the present analysis is 0.8 as it represents the maximum score at which path clusters are not redundant.

Clusterization method 2 employed the Quality Threshold (QT) algorithm, according to a similarity score.[53] This method relies on a similarity score (S) between paths a and b, computed according to the following equation:

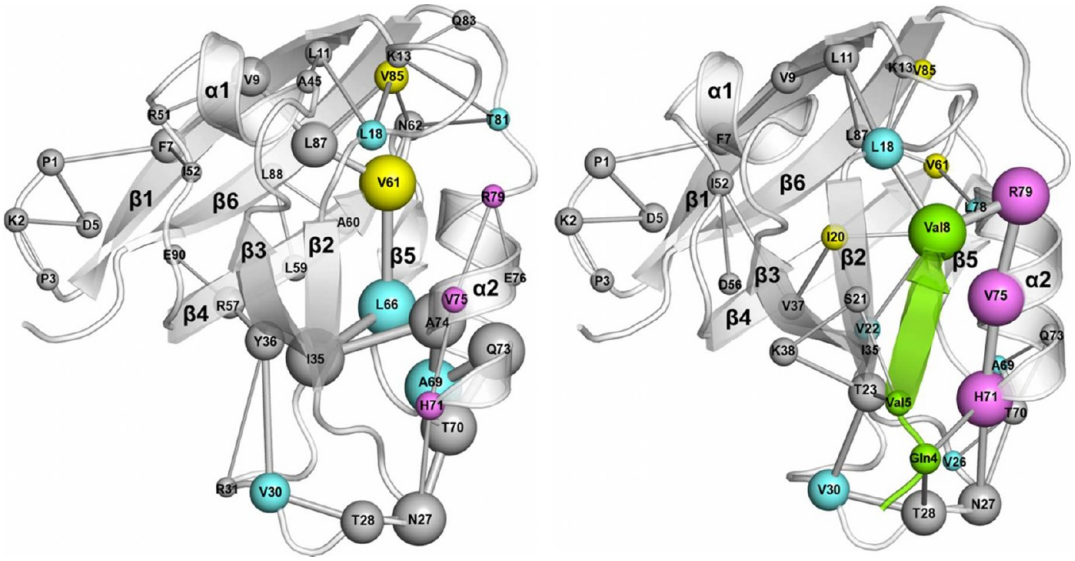


Figure 4. Global meta paths of PDZ2^{APO} (left) and PDZ2^{PEP} (right) achieved by PSN-ENM. The description of meta paths has been made in the text. In the 3D representation of global meta paths, sphere diameters and link thickness are proportional to node and link frequencies in the pool of paths. Pink, light blue, and yellow colors mark those amino acids highlighted by in vitro experiments (see the legend to Figure 1).

$$S_{a,b} = \left(\frac{2C_N}{N_a + N_b}\right)\cdot 0.15 + \left(\frac{2\mathrm{Max}(C_P)}{N_a + N_b}\right)\cdot 0.4 + \left(\frac{2C_L}{L_a + L_b}\right)\cdot 0.45 \quad (16)$$

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, $\mathrm{Max}(C_P)$ 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, to 1, for two identical paths. The three coefficients 0.15, 0.40, and 0.45 are the result of an empirical tuning aimed at giving more weight to similarities in node connectivity than to the mere node composition of the two compared paths. We employed an S cutoff equal to 0.4.

Irrespective of the clusterization method, for each cluster, following a pairwise comparison of all the 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.

Clusterization method 1 relies on a simpler S score and is faster than method 2.

3. RESULTS OF THE PSN ANALYSIS: PTP1E AS A MODEL SYSTEM

3.1. Peptide Binding Perturbs the PSG of PTP1E. In the PDZ2 domain, proximal and distal sites are those protein portions participating in the peptide binding site or far from it, respectively. We define as proximal sites $\beta 2$, the $\beta 2/\beta 3$ loop, $\beta 5$, the $\beta 5/\alpha 2$ loop, and $\alpha 2$, whereas distal sites essentially include the N-term, $\beta 1$, the initial portion of the $\beta 1/\beta 2$ loop, the C-term of the $\alpha 2/\beta 6$ loop, and $\beta 6$.

The PSG of PDZ2^{PEP} is richer in nodes, links, and hubs compared to PDZ2^{APO} (Table 1). As expected, upon ligand binding, the PDZ2 domain acquires hubs in the peptide binding site (residues L18, V22, N27, and H71). However, hubs peculiar to the peptide-bound state also concern distal sites (V9, L11, and I35), indicating signal propagation (Figure 3).

The structural perturbation caused by peptide binding can be inferred by comparing the PSGs of the unbound and bound forms, which means computing changes in node, hub, and link distribution upon ligand binding. A map of such perturbation can be built on the 3D structure of one of the two compared systems, by distinguishing nodes and links peculiar to each system from those in common (Figure 3). In the case of the APO and bound states of PTP1E, ligand-induced perturbations essentially consist of a gain of intermolecular links in the peptide binding site and, to a lesser extent, a loss of nodes and links between the $\beta 2/\beta 3$ loops and both $\beta 3$ and $\beta 4$ as well as between $\beta 4$ and $\beta 6$. Remarkably, the gain and loss of links also concerns distal sites on $\beta 1$, $\beta 4$, $\beta 6$, $\alpha 1$, and the $\alpha 1/\beta 4$ loop, indicative of signal propagation upon ligand binding (Figure 3). About 34% of the nodes involved in such perturbations (i.e., S17, L18, I20, T23, V30, I52, L66, A69, H71, V75, L78, R79 and V85) were found to be implicated in the allosteric effects of ligand binding by in vitro experiments.[19,23,25−30]

3.2. Peptide Binding Perturbs the Communication Pathways in PTP1E. The search for the shortest communication paths led to a total of 770 and 845 paths for PDZ2^{APO} and PDZ2^{PEP}, respectively, thus indicating an increase in the possible pathways upon ligand binding (Table 1). The average length of such paths and the average MSDF^{p} index decrease in the bound state, suggestive of more effective information transfer compared to the unbound form (Table 1). Such a trend is clearly seen from the path distributions based upon the two scores (Supporting Figure 1 (Figure S1)) and indicates more dispersion in terms of path lengths in the APO form, which in turn corresponds to higher organization in the bound state. Remarkably, only three paths are found in common between the two functional states.

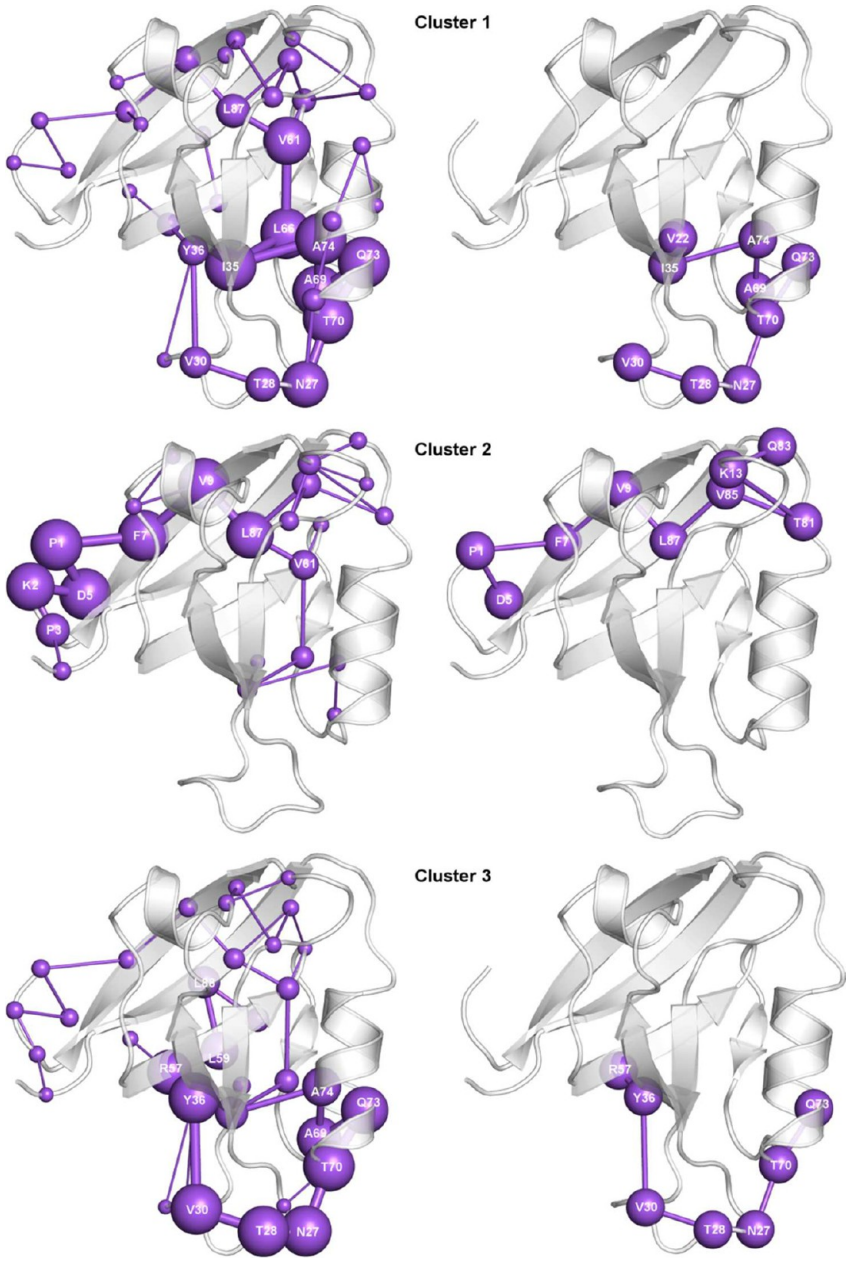


Figure 5. Pathways concerning the first three most populated clusters obtained for PDZ2^{APO} by clusterization method 2. Left panels show the meta paths of clusters 1, 2, and 3. The labels indicate the centers of each cluster. In the right panels, the paths with the lowest MSDF^{p} in each cluster are shown.

The analysis of the most frequent nodes and links in the pools of paths concerning PDZ2^{APO} and PDZ2^{PEP} (i.e., those nodes and links with recurrence ≥5%) shows a clear change in the structural communication upon peptide binding (Figures 4 and S2). The most marked ligand-induced perturbations essentially consist of a loss or frequency-reduction of nodes contributed by β4, the β5/α2 loop, and β6 and in a gain of nodes by the β1/β2 loop and β2 (Figures 4 and S2). Moreover, portions like β1, β3, and α2 undergo redistributions in terms of node contribution following ligand binding. As expected, peptide binding shifts the core of the communication on the ligand binding site, the peptide itself contributing with a number of nodes (i.e., Gln4, Val5, and Val8; it is worth noting that, to distinguish between protein and ligand amino acids, the three-letter code is exclusively used for the latter). Remarkably, some recurrent nodes are peculiar to a protein state. In fact, whereas Y36, R57, and L59 are specific to the unbound state, I20, S21, T23, and K38 are specific to the bound one.

The central role of the RA-GEF2 ligand in the communication pathways of the bound form is indicated by the presence of Val8, at the C-term of the peptide, in 78% of paths (Figure S2). Such an amino acid makes links with nodes on the α1/β2 loop and β2 (L18, I20, and V22, respectively) and/or with nodes on the C-term of α2 (i.e., L78 and R79).

The histograms shown in Figure S2 correspond to the global meta paths mapped on the structures of PDZ2^{APO} and PDZ2^{PEP}, in which sphere diameters and link thickness are, respectively, proportional to the node and link frequencies (Figures 4 and S2; see The Proposed Method for meta path definition). In fact, the meta path of the unbound form is

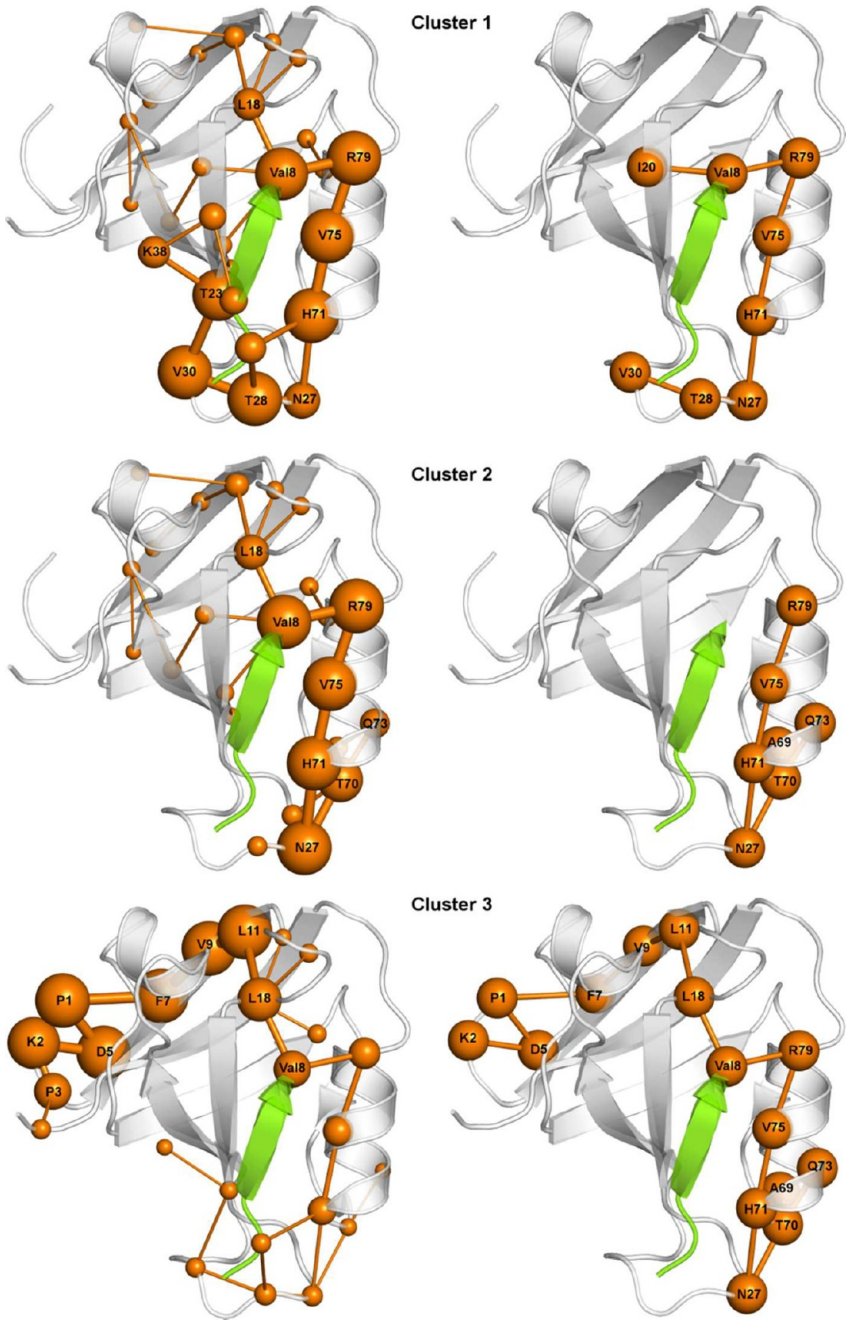


Figure 6. Pathways concerning the first three most populated clusters obtained for PDZ2^{PEP} by clusterization method 2. Left panels show the meta paths of clusters 1, 2, and 3. The labels indicate the centers of each cluster. In the right panels, the paths with the lowest MSDF^{P} in each cluster are shown.

characterized by a central core of almost equally frequent nodes, located on the $\beta2/\beta3$ loop (N27), $\beta3$ (I35), $\beta4$ (V61), the $\beta5/\alpha2$ loop (L66, A69, and T70), and $\alpha2$ (Q73 and A74; Figures 4 and S2). Such a core of the most recurrent nodes may involve the $\beta2/\beta3$ loop, on one extremity, and/or more distal regions such as $\beta1$ and the N-term, on the other extremity (Figures 4 and S2). Remarkably, the most likely communication pathways of the unbound state do not pass through $\beta2$ (Figures 4 and S2). The latter, instead, enters in the pathways characterizing the bound state as it is directly involved in contacts with the peptide (i.e., through the I20−Val8 and T23−Val5 links; Figures 4 and S2). In the bound state, the core of the meta path is represented by the peptide itself (i.e.Val8) and three nodes from $\alpha2$ (H71, V75, and R79). Thus, $\alpha2$ plays a central role in the communication pathways of both PDZ2^{APO} and PDZ2^{PEP}, though by different nodes. While in PDZ2^{APO}, nodes from $\alpha2$ lie at the N-term of the helix, in PDZ2^{PEP} they distribute along the main axis of the helix. In fact, at the N-term of the helix, H71 establishes links with nodes of the $\beta2/\beta3$ loop and with Gln4 of the peptide, while at the C-term of the helix, R79 establishes contacts with Val8 of the peptide. The latter mediates in turn the communication with nodes on $\beta1$ and the N-term via L18 on the $\beta1/\beta2$ loop. Differently from PDZ2^{APO}, $\beta4$ and $\beta5$ do not intervene in the communication pathways of PDZ2^{PEP} (Figures 4 and S2). The inferences above on meta paths are robust and are not artifacts of crystal packing. Indeed,

dx.doi.org/10.1021/ct400096f | J. Chem. Theory Comput. 2013, 9, 2504−2518

the same analyses carried out on the alternative chains or side chains found in the 3LNX and 3LNY structures show a substantial overlap concerning those nodes contributing the most to the global meta paths of the two functional states (Figures S3 and S4).

As stated in section 2.3, this study profited from the implementation of two different methods for path clusterization to make an in-depth investigation of the structural communication in PDZ2 (see section 2.3 for detailed methodological description).

Path clusterization with the more recent method 1 grouped almost the totality of the communication pathways in a single cluster (Table 1 and Figure S5). The center of such a cluster, i.e., the path with the highest average S among all the paths in the cluster, is the longest one, being made of 21 and 17 nodes for the unbound and bound states, respectively. As expected, it draws, in a more defined way, the communication profile shown by the global meta path (Figures 4 and S5). The most frequent fragment of nodes expressing a communication between proximal and distal sites is indeed comprised in the center of cluster 1 and is formed by the following nodes: T70−Q73−A69−A74−I35−L66−V61−L87−V9, shared by 26% of paths in cluster 1 for the unbound state, and H71−V75−R79−Val8−L18−L11−V9, shared by 22% of paths in cluster 1 for the bound state. For both functional states, pathways starting from the proximal sites and protruding in the N-term (i.e., including P1) represent 8% of the paths belonging to cluster 1. Whereas for the bound state pathways expressing a communication between proximal and distal sites (e.g., Q73−T70−N27−H71−V75−R79−Val8−L18−K13−Q83) can be found among the 10 pathways characterized by the lowest $MSDF^p$, this does not happen for the unbound form, which on average shows higher $MSDF^p$ than the bound one (Table 1). This is likely due to the L18−Val8−R79 links which expand the stiffness of the network in the proximal region.

According to clusterization method 2, 80% and 74% of the total paths found for the unbound and bound forms, respectively, distribute in the first three most populated clusters, which remark the differences between the two PDZ2 states. The path population of these clusters is properly accounted for by the respective meta paths; cluster centers are labeled inside cluster meta paths (Figure 5). For the unbound state, 64% of the paths group in the first cluster, which is similar to cluster 3 (i.e., holding 7.5% of paths); therefore, the meta path of this cluster summarizes the salient traits of the structural communication between proximal sites and distal nodes on $\beta6$ (Figure 5). A rather different communication is shown by the members of cluster 2 (representing 8% of the total pathways), in which the last amino acid of $\beta4$, V61, communicates with the N-term via nodes on $\beta6$ (L87) and $\beta1$ (V9 and F7; Figure 5). Differently from the unbound state, the pathways that characterize the bound state distribute essentially in three clusters that hold 38%, 22%, and 13% of the total pathways (Figure 6). All pathways in these three clusters involve the peptide (Figure 6). Pathways in clusters 1 and 2 pass through $\alpha2$, being parallel to the main axis of the helix, and express a peptide-mediated long-distance communication between the $\beta2/\beta3$ loop and $\beta1$. Pathways in cluster 3 tend to start at Val8 and end in the N-term (Figure 6). For the unbound state, pathways with the highest stiffness (i.e., lowest $MSDF^p$) in a cluster localize either on the proximal sites (e.g., $\beta2/\beta3$ loop, $\beta2$, $\beta3$, and the N-term of $\alpha2$) or on the distal sites (e.g., N-term, $\beta1$, $\beta6$, and the $\beta1/\beta2$ and $\alpha2/\beta6$ loops). Differently from the unbound state, the pathways with the highest stiffness in each of the three clusters all involve nodes distributed along the main axis of $\alpha2$; two of these pathways also involve Val8 of the peptide, always linked to R79 (Figure 6). Remarkably, the most stiff communication pathway in the third cluster of the bound state describes a communication between distal and proximal sites. This is likely a peculiarity of the bound state related to the establishment of the L18−Val8−R79 intermolecular links (Figure 6).

Collectively, the analysis of path clusterizations suggests that, for each of the two functional states of PDZ2, the salient structural communication tends to converge into a few defined pathways. In this framework, in $PDZ2^{PEP}$, Val8 of the ligand establishes a L18-mediated communication with distal nodes in the N-term and/or $\beta1$. In this respect, the Val8−L18 and Val8−R79 intermolecular links contribute to put in communication two regions, i.e. the $\beta2/\beta3$ loop and $\alpha2$ on one side and the N-term and/or $\beta1$ on the other side, which are characterized by intrinsic stiffness (i.e., relatively low internode distance fluctuations).

In conclusion, although the communication between proximal and distal sites seems to be an intrinsic feature of the PDZ topology, peptide binding is expected to facilitate such communication. Thus, signal transfer in the bound state is expected to involve the main axis of $\alpha2$, and this is likely a consequence of the intermolecular links established by the ligand rather than an intrinsic feature of the PDZ domain.

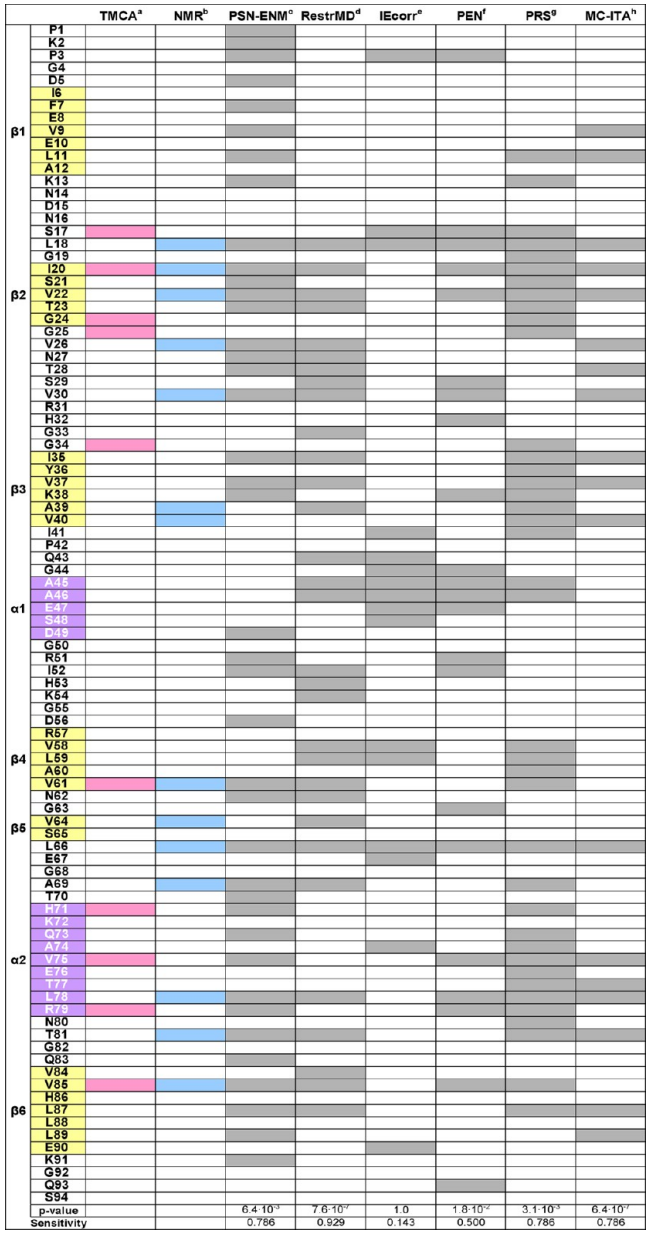
From comprehensive pathway analysis, we could draw the following methodological guidelines. The most meaningful picture of the communication features of a system as well as of the salient differences between two functionally different states is captured by the global meta path, which can be read in tandem with the center of cluster 1 obtained by clusterization method 1 (Figures 4 and S5). The meta paths of the most populated clusters arising from clusterization method 2, accompanied by the inspection of cluster centers as well as of the pathways with the lowest $MSDF^p$ in each cluster, may help complete the picture, possibly highlighting a number of more defined pathways.

4. VALIDATION OF THE METHOD

4.1. Agreement with in Vitro Evidence. The structural communication and allosteric features of PDZ domains have been largely characterized by in vitro experiments.[19,23,25−30] For this reason, PDZ domains have been widely used to validate computational approaches designed to predict the allosteric communication in biomolecular systems.

A number of in vitro experiments could identify energetically or dynamically coupled amino acids likely involved in the allosteric communication in PDZ domains (Figure 1 and Table 2). In this respect, the thermodynamic mutant cycle analysis (TMCA) on the third PDZ domain from PSD-95 ($PDZ3^{PSD\text{-}95}$) measured coupling energies for a mutation at position 76 (H76Y) against mutations at a set of 14 PDZ domain positions and two peptide positions.[25] The mutations chosen were designed to test predictions from the Statistical Coupling Analysis (SCA); a set of sites that are not significantly statistically coupled were included as well (Figure 1 and "TMCA" headed column in Table 2).[25] The nine sites predicted to be statistically coupled to H71(H76) (i.e., S17(G26), I20(F29), G24(G33), G25(G34), G34(G39), V61(V66), V75(A80), R79(K84), V85(V90)) were also found to be thermodynamically coupled through mutagenesis

Table 2. Comparisons between in Vitro and in Silico Experiments



[a]Residues highlighted by the thermodynamic mutant cycle analysis (TMCA) on PDZ3^{PSD-95}.[25] [b]Dynamically coupled amino acid side chains (in the picoseconds-nanosecond range) upon peptide binding inferred from side-chain (^{2}H-methyl) and backbone (^{15}N) NMR spin relaxation determinations.[26] [c]Residues participating in the mata path of the bound form as inferred from the PSN-ENM method herein proposed. [d]Amino acid residues highlighted by the RestrMD approach.[34] [e]Amino acid residues highlighted by the IEcorr approach.[35] [f]Amino acid residues highlighted by the PEN pproach.[37] [g]Amino acid residues highlighted by the PRS approach.[38] [h]Amino acid residues highlighted by the MC-ITA approach.[39]

(the first label refers to the PTP1E sequence; in contrast, as for the labeling in parentheses, the amino acid letter refers to the PDZ3^{PSD-95} sequence, whereas the number refers to the position in a structure-based alignment of 274 PDZ domains by Lockless and Ranganathan).[25] The majority of these residues lie in the peptide binding site, whereas only two are more distally located (i.e., V61(V66) in β4 and V85(V90) in β6).

Side-chain (^{2}H-methyl) and backbone (^{15}N) NMR spin relaxation determinations highlighted 14 dynamically coupled amino acid side chains (in the picosecond−nanosecond range) upon peptide binding ("NMR" headed column in Table 2).[26] The study highlighted a structural communication between a set of residues circumscribing the peptide binding site (i.e., L18, I20, V22, V26, V30, and L78) and two sets in distal sites, V61, V64, L66, A69, T81 and V85 (distal surface 1 (DS1) on β4, β5, β5/α2 loop, α2/β6 loop and β6), and A39 and V40 (distal surface 2 (DS2) on β3) (Figures 1 and 4, "NMR" column in Table 2).[26] In a subsequent study, the impact on side-chain dynamics was further tested with a C-terminal peptide from APC, which showed results nearly identical to those with the RA-GEF2-Ct peptide.[19]

The 14 dynamically coupled amino acid side chains inferred from NMR determinations on unbound and RA-GEF2-Ct-bound PDZ2[26] have been used as a reference to evaluate the accuracy of predictions by a number of computational approaches including our PSN-ENM method (Table 2; this and the following paragraph). The p values from the Fisher's exact test concerning the significance of predictions as well as the sensitivity values are listed in Table 2. Statistics against the TMCA data has not been done because such determinations were carried out on PDZ3^{PSD-95} and not PDZ2. Moreover, they concern only those protein sites for which a correlation was found between statistical (i.e., predictions by the SCA method) and thermodynamic coupling.[25] TMCA data will be, therefore, considered only for a qualitative comparison with computational predictions.

Comparisons between PSN-ENM predictions and in NMR data were based on the nodes constituting the meta path of the bound form (Figure 4 and Table 2). A two-tailed Fisher's exact test gave a p value of 6.4×10^{-3}, indicating the statistical significance of predictions (Table 2, "PSN-ENM" column). Indeed, 83% of the peptide binding site and DS1 amino acids from NMR determinations coincide with recurrent nodes in the pathways of the bound state (Table 2, Figures 4 and S2). Although we could not find any of the two β3 residues defined as DS2, we found four out of the six nodes in the same strand (I35, I36, V37, and K38; Table 2, Figures 4 and S2). It is worth recalling that 18 of the 29 false positives found in our analysis (i.e., 62%) do not hold methyl groups and could not be tested by NMR. Remarkably, we found four out of the five amino acids in the N-term as implicated in the long distance communication characterizing the bound state. Since no one of the amino acids in the N-term holds a methyl group, this region was constitutively undetectable by the NMR experiments of Fuentes and co-workers.[27]

Node composition of the global meta path used to evaluate the significance of predictions by the PSN-ENM method essentially depends on the motion correlation coefficient employed as a cutoff for path filtering as well as on the recurrence cutoff for a node to participate in the meta path. The analysis of sensitivity variation as a function of these two cutoffs shows that for recurrence cutoffs ranging between 1% and 5%, any correlation coefficient cutoff between 0.1 and 0.8 gives sensitivity values comprised between 0.643 and 0.786, the latter value being reached by 80% of combinations, including the one employed in this study, i.e. 5% and 0.6, which represent the highest motion correlation coefficient and node recurrence cutoffs producing in tandem the highest and most frequent sensitivity value (Figure S6).

Incidentally, 60% of the amino acids highlighted by TMCA on PDZ3^{PSD-95} were found as recurrent in the predicted global meta path of the bound form by the PSN-ENM method (Figure 4 and Table 2).

Collectively, predictions by the PSN-ENM method remark the salient ligand-induced perturbations in the communication features of PTP1E highlighted by in vitro experiments. Both significance and sensitivity of predictions by the PSN-ENM method are high. Added value of our predictions is, however, the increased resolution in the knowledge of the internode links that participate in the communication pathways.

4.2. Comparison with Predictions by the Existing Computational Methods. A number of computational approaches aimed at unveiling the structural communication in biomolecular systems were challenged on PTP1E.

In this respect, Ho and Agard developed the Rotamerically Induced Perturbation (RIP) method, which identifies strong couplings between residues by analyzing the pathways of heat-flow resulting from thermal excitation of rotameric rotations at individual residues.[36] Application of the method to five PDZ domains, including PTP1E, found 17 pair positions in the general PDZ fold, in which buried tertiary couplings are found in at least one of the five PDZ domains. In this framework, only two pair positions resulted to be coupled in the APO form of PDZ2, i.e., I52−D56 and V61−V85.[36] The Anisotropic Thermal Diffusion (ATD) method applied to PDZ3^{PSD-95} predicted a pathway from H76/H71 to I45/I40 via I31/I22 and F29/F20.[32]

Dhulesia and co-workers employed MD simulations with ensemble averaged NMR restraints to infer the effects of RA-GEF2-Ct binding to PTP1E.[34] Such restraints incorporated also the NMR determinations by Fuentes and co-workers. The analysis of the free and bound states of the ensemble of conformations led to map changes in structure and dynamics revealing the presence of two interconnected networks of residues associated with the response to ligand binding. The first network, the structural network, is formed by a set of 20 strongly interacting residues that undergo changes in the distribution of their rotameric states. The second one, the dynamic network, is composed by a set of residues that experience a variation in their picosecond to nanosecond dynamics. The latter network, made by the same amino acids identified experimentally by Fuentes and co-workers[26] plus six additional residues (T23, A45, A46, I52, V58, and L87), shared eight amino acids in common with the "structural network" (Table 2, "RestrMD" column).[34] The 7.6×10^{-7} and 0.929 of p value and sensitivity, respectively, demonstrate the remarkably high predictive power of the approach. We cannot exclude, however, the possibility that this result is due in part to the fact that sampling was primed by the same NMR data we are using for validation.

Possible communication pathways in the PDZ2 domain were also inferred by Kong and Karplus.[35] According to the method, a residue correlation matrix was constructed from the interaction energy correlations between all residue pairs obtained from MD simulations. Two continuous interaction pathways, starting at the ligand binding pocket, were identified by a hierarchical clustering analysis of the residue correlation matrix. One proposed pathway was mainly localized at the N-terminal side of α1 and the adjacent C-terminus of the β1/β2 loop. The other pathway was perpendicular to the central β-sheet extending toward the side of the PDZ2 domain opposite to the ligand binding pocket.[35] The residues involved in such pathways were discussed as showing some overlap with the ones highlighted by in vitro analysis (Table 2, "IEcorr" column). The results of the approach were judged complementary to the sequence-based approach by Lockless and Ranaganathan.[25] In line with the analysis by Cilia and co-workers,[39] the quality of the IEcorr predictions is low (p value = 1.0, sensitivity = 0.143, Table 2).

The studies reported above identified amino acids in dynamical and/or structural communication but did not attempt predictions of connectivities between nodes either in a global network or in communication pathways. In contrast, the three approaches discussed and evaluated below did so.[37−39] In this respect, changes in the structural communication of PTP1E upon ligand binding were investigated by a variant of the PSN method called Protein Energy Network (PEN), which employs interaction energies instead of distances to link pairs of nodes.[37] The authors employed the closeness index of a given node (Ci), which accounts for the centrality of a node in the network, to mark the allosteric effects of ligand binding. They found that the residues whose Ci decreases upon ligand binding derive from α1, the β1/β2 loop, as well as from the β2/β3 loop and C- and N-terminal residues. The decrease of Ci in only a fraction of residues upon ligand binding was interpreted as a channeling of information. They also computed the shortest communication pathways between selected node pairs from the regions found to undergo conformational changes upon ligand binding[26,27] or predicted to be energetically coupled although distal.[35] In this framework, communication paths were searched between S29 in the β2/β3 loop and distal residues in α1 (i.e., G44, A45, A46, E47, and S48) as well as between A69 in the C-terminus and distal residues in the β1/β2 loop (i.e., K13, N14, D15, N16, N17, and L18). The analysis showed changes in the type and length of communication pathways upon ligand binding.[37] A number of amino acids highlighted by the PEN analysis overlap with those inferred from in vitro evidence (Table 2, "PEN" column). The significance of such overlap is, however, not that high (p value = 1.8×10^{-2}, sensitivity = 0.500, Table 2). The PEN analysis-based method is the most similar to the PSN-ENM presented in this study, being a derivation of the PSN analysis. However, differently from what we did in this investigation, Vijayabaskar and Vishveshwara did not systematically search for the shortest communication pathways between all possible residue pairs in the protein, but they limited the search to a few selected residue pairs. Therefore, the degree of overlap between their and our results in terms of path prediction cannot be evaluated. Collectively, the significance and sensitivity of predictions by PSN-ENM is higher compared to PEN (Table 2).

Gerek and Ozkan developed an advancement of perturbation response scanning (PRS), which couples ENM with linear response theory (LRT) to predict key residues in allosteric transitions in PTP1E. With PRS, they first identified the residues that give the highest mean square fluctuation response upon perturbing the binding sites, observing that the residues with the highest mean square fluctuation response agree with experimentally determined residues involved in allosteric transitions (Table 2, "PRS" column).[38] Second, they constructed the allosteric pathways by linking the residues giving the same directional response upon perturbation of the binding sites. For unbound PTP1E, they found that the most highly weighted pathway of PTP1E follows through connections S17 → V22 → G25 → R31 → I35 → V61 → V64 → T70 → A74

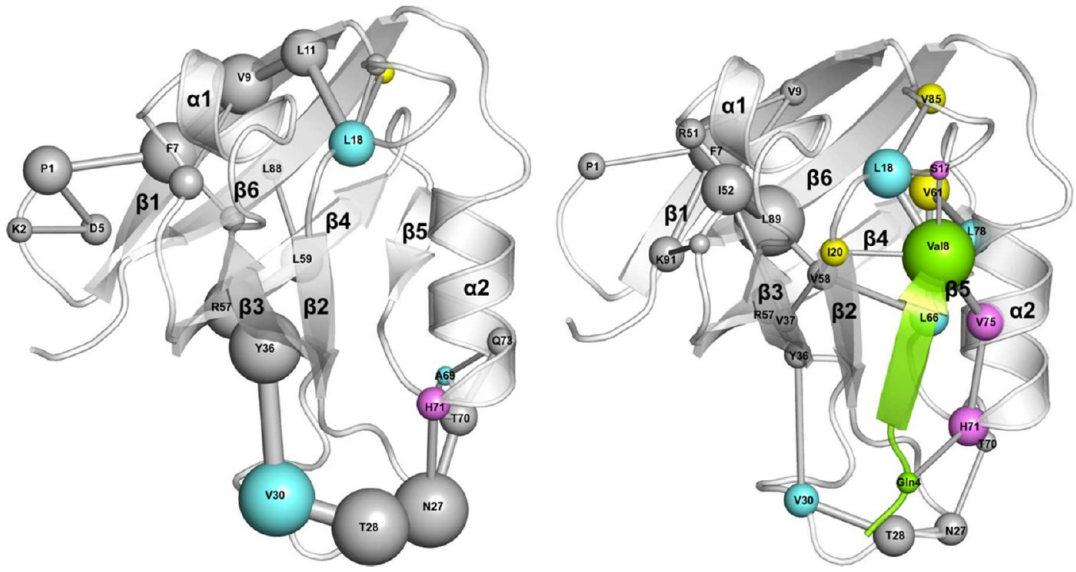


Figure 7. Global meta paths found in this study by the PSN-MD method. Sphere diameters and link thickness are proportional to node and link frequencies in the pool of paths. Pink, light blue, and yellow colors mark those amino acids highlighted by in vitro experiments (see the legend to Figure 1). As specified in the text, a merge of the shortest paths derived by three distinct MD replicas contribute to the meta paths shown in this figure. The I_{min} cutoffs set for the PSN analysis on the three MD replicas concerning $PDZ2^{APO}$ are 2.62%, 2.55%, and 2.53%, respectively, whereas those concerning $PDZ2^{PEP}$ are 2.85%, 2.65%, and 2.79%, respectively.

→ L78 → T81 → L88. The PSN analysis approach that we use would never have found such a path since a number of residue pairs, e.g. S17−V22, I35−V61, V64−T70, and T81−L88 are too far apart (the inter-Cα distances being 12.84, 11.38, 14.60, and 14.22 Å, respectively) to be linked together in the structure graph. The predictive ability of the approach is significant and comparable to that of PSN-ENM (p value = 3.1×10^{-3}, sensitivity = 0.786, Table 2).

Finally, Cilia and co-workers very recently proposed an approach that combines Monte Carlo (MC) sampling of the side chain conformational space and an information theoretical analysis (ITA; "MC-ITA," Table 2). The approach determined the PDZ2 residues affected by the binding of the C-terminus of APC by quantifying the change in conformational coupling between residue side-chains within the major states of the domain, i.e., the bound and unbound states.[39] All the structures in the NMR ensemble concerning the unbound (PDB code: 1GM1) and bound (PDB code: 1VJ6) states were used as inputs of calculations. In a first step, an MC sampling process was used to determine the conformational freedom of each residue, represented by a probability mass function over a fine-grain discrete set of side-chain conformations for each protein state. In a second step, these probability mass functions were used to calculate the mutual information (MI) between every residue pair, again for both protein states. A high MI value designates a high degree of coupling between the side-chain conformations of two residues, whereas a low value shows the opposite. In the final step, the absolute differences between the MI values of the bound and unbound forms were calculated (DMI), producing a matrix of mutual information changes. Residue pairs displaying the higher absolute DMI value were predicted as involved in the dynamical changes induced by the peptide on the domain or protein structure. According to our statistical analysis, predictions made on all residues were quite significant and better than those considering only methyl-group containing residues (p value = 6.4×10^{-7}; sensitivity = 0.786 in the first case and p value = 5.5×10^{-5} and sensitivity = 0.643, in the second one).

Collectively, the RestrMD and MC-ITA approaches showed the best accuracies in predicting the amino acids highlighted by NMR determinations (Table 2). Among the methods that computed inter-residue connectivities, MC-ITA, PRS, and PSN-ENM show exactly the same sensitivity (Table 2). According to the p value, all three methods made significant predictions even if MCA-ITA performed better than PRS and PSN-ENM, the latter two behaving the same in that respect (Table 2).

The advantage of our method with respect to the others is that, except for the PRS approach, they rely on MD or MC simulations and, therefore, are more time-consuming than PSN-ENM. Moreover, some methods like MCA-ITA start from an ensemble of conformations rather than a single crystallographic structure, the latter being the most frequent situation in the Protein Data Bank.

The only method that, like ours, is not based on fine-grained simulations is PRS. PSN-ENM and PRS show comparable accuracies. Assuming that they hold the same performance in terms of computing time, we think that the advantage of PSN-ENM compared to PRS resides in the higher physical meaning of the internode links in the communication pathways.

4.3. Overlap between Predicted Pathways by PSN-ENM and PSN-MD. The communication pathways inferred by the mixed PSN-ENM method introduced in this study were compared with those by PSN done on MD trajectories (i.e., PSN-MD). In this respect, for each of the $PDZ2^{APO}$ and $PDZ2^{PEP}$ forms, three replicas of 20 ns equilibrated MD trajectories were computed (see the Supporting Information for the calculation setup). Collectively, the bound form is more stable in terms of both Cα-Root Mean Square Deviations and Fluctuations (Cα-RMSDs and Cα-RMSFs, respectively) than the APO one (Figure S7). The most marked differences in Cα-RMSFs between $PDZ2^{APO}$ and $PDZ2^{PEP}$ essentially concern the $\beta1/\beta2$ loop, which is significantly more flexible in the absence

of the peptide. Other regions that undergo decreases in flexibility upon ligand binding include the $\beta3/\alpha1$, $\beta4/\beta5$, and $\beta5/\alpha2$ loops and $\alpha2$ (Figure S7). The Cα-RMSF profiles from MD simulations are quite comparable with that from ENM-NMA (Figure S8).

The setup of path search on MD trajectories and the subsequent building of meta paths are the same as the ones employed in a recent study.[15] In this respect, it is worth remarking that the PSN-ENM and PSN-MD approaches are significantly different. The main differences between them include the way in which the cross-correlations of atomic motions for path filtering are computed. Whereas with PSN-ENM such correlations are extracted from the covariance matrix of the deformation modes computed by ENM-NMA (see methods above), with PSN-MD they are computed on the trajectory frames by means of the Linear Mutual Information (LMI) method.[54] Another remarkable difference is that, PSN-ENM considers all those paths that pass the motion correlation filter, while PSN-MD refilters those paths that pass the motion correlation filter by finally keeping only those that exceed a recurrence cutoff (i.e., presence in a given number of trajectory frames). This second filter was introduced to take into account as much as possible the information contained in the ensemble of conformations constituting MD trajectories.[8,10,15] In this case study, a path frequency cutoff $\geq$10% was used.

A peculiarity of the PSN-MD is that, for each functional state of PTP1E, the global meta path was computed on those paths obtained by merging the three MD replica-specific path sets. This strategy was aimed at drawing a structural communication profile of PDZ2^{APO} and PDZ2^{PEP} relying on all three MD replicas. Thus, those paths from the three MD replicas that passed both motion-correlation and frequency filters were merged together to calculate the meta paths of the free and bound states (Figure 7). Those paths were 262 and 267 for the free and bound forms, respectively, about 3-fold less than those achieved by the PSN-ENM method (i.e., 770 and 845, respectively, Table 1). This is likely due, at least in part, to the additional frequency-based filtering singular to PSN-MD.

Despite the differences in the two methodologies, the overall agreement between them is significant. Similar to the results of PSN-ENM, with PSN-MD the average path length tends to decrease in the bound state, though the difference between the two states is lower compared to the one inferred from PSN-ENM. For both the PDZ2^{APO} and PDZ2^{PEP} forms, PSN-MD gives lower average path lengths compared to PSN-ENM (i.e., 7.77 ± 1.06 and 7.28 ± 0.56 vs 12.44 ± 3.77 and 10.45 ± 2.50, respectively).

As for PDZ2^{APO}, similar to the PSN-ENM method, PSN-MD found two major distal pathways, one involving nodes in the N-term, $\beta1$, the $\beta1/\beta2$ loop, and $\beta6$, and the other involving nodes in the N-term of $\alpha2$ and the $\beta2/\beta3$ loop (Figures 4, 7, S2, and S9). With PSN-ENM, the two pathways are linked together via nodes A74, I35, L66, V61, and L87, whereas with PSN-MD they are not connected, likely due to the double filtering of paths peculiar to PSN-MD. As for PDZ2^{PEP}, the two different methods agree in finding ligand-mediated communication between distal sites like the $\beta2/\beta3$ loop and $\beta6$ through the Val8−L18 link (Figures 4, 7, S2, and S9). However, according to PSN-ENM, such communication may continue along $\beta1$ till the N-term.

Collectively, by merging the results concerning the two functional states, it turns out that 76% of nodes and 61% of links in the meta paths by PSN-MD are shared with meta paths by the PSN-ENM method (Figures S2 and S9). Vice versa, 81% of nodes and 62% of links in the meta paths by PSN-ENM are shared with the meta paths by PSN-MD. These numbers are indicative of a significant overlap between the results of the two methods.

Similar to the results of PSN-ENM, those of PSN-MD suggest that signal transfer upon peptide binding uses the main axis of $\alpha2$ to reach distal sites like $\beta6$.

5. CONCLUDING REMARKS

We combined a graph theory-based approach with ENM to infer the communication pathways of complex biomolecular systems.

Predictions by the PSN-ENM method are significant and valuable especially if we compare their performance in terms of CPU time with those of the exceedingly more time-consuming computational approaches tested on the same biological system. Indeed, on a protein complex made of 100 amino acids (PDZ2^{pep}), an Intel XEON CPU 2.4 GHz (1 core out of 6) takes less than 10 s to compute the PSG and all the shortest communication pathways between 4500 amino acid pairs. Therefore, the method allows for a fast systematic scanning of the shortest communication pathways between all residue pairs in a biomolecular system, which cannot be achieved by any MD/MC simulation-based method.

Finally, the PSN-ENM approach was able to reproduce the salient communication features of PTP1E inferred from MD simulations, but predicting a more extended communication, likely due to the less heavy filtering applied to the communication pathways.

High speed and accuracy make this approach suitable for high throughput investigation of the communication pathways in large sets of biomacromolecular systems in different functional states.

■ ASSOCIATED CONTENT

Supporting Information

Methods and seven figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +39- 059-2055114. Fax: +39- 059-373543. E-mail: fanelli@unimo.it.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This study was supported by an Airc-Italy grant [IG10740] and a Telethon-Italy grant [GGP11210/S00068TELC] to F.F.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The employment of the PyMOL 0.99rc6 software for the realization of all drawings is acknowledged.

■ REFERENCES

(1) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. Three key residues form a critical contact network in a protein folding transition state. *Nature* **2001**, *409*, 641−645.

(2) 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*, 1135−1146.
(3) Brinda, K. V.; Vishveshwara, S. A network representation of protein structures: implications for protein stability. *Biophys. J.* **2005**, *89*, 4159−4170.
(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**, *2*, 0019.
(5) Bode, C.; Kovacs, I. A.; Szalay, M. S.; Palotai, R.; Korcsmaros, T.; Csermely, P. Network analysis of protein dynamics. *FEBS Lett.* **2007**, *581*, 2776−2782.
(6) Tang, S.; Liao, J. C.; Dunn, A. R.; Altman, R. B.; Spudich, J. A.; Schmidt, J. P. Predicting allosteric communication in myosin via a pathway of conserved residues. *J. Mol. Biol.* **2007**, *373*, 1361−1373.
(7) Vishveshwara, S.; Ghosh, A.; Hansia, P. Intra and inter-molecular communications through protein structure network. *Curr. Protein Pept. Sci.* **2009**, *10*, 146−160.
(8) 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*, 1227−1239.
(9) Fanelli, F.; Seeber, M. Structural insights into retinitis pigmentosa from unfolding simulations of rhodopsin mutants. *FASEB J.* **2010**, *24*, 3196−3209.
(10) Fanelli, F.; Felline, A. Dimerization and ligand binding affect the structure network of A(2A) adenosine receptor. *Biochim. Biophys. Acta* **2011**, *1808*, 1256−1266.
(11) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* **2012**, *26*, 868−881.
(12) 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*, 12515−12525.
(13) Chennubhotla, C.; Bahar, I. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput. Biol.* **2007**, *3*, 1716−1726.
(14) 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*, 287−292.
(15) Raimondi, F.; Felline, A.; Portella, G.; Orozco, M.; Fanelli, F. Light on the structural communication in Ras GTPases. *J. Biomol. Struct. Dyn.* **2013**, *31*, 142−157.
(16) Seeber, M.; Felline, A.; Raimondi, F.; Muff, S.; Friedman, R.; Rao, F.; Caflisch, A.; Fanelli, F. Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J. Comput. Chem.* **2011**, *32*, 1183−1194.
(17) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.* **2010**, *110*, 1463−1497.
(18) Kitao, A.; Go, N. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164−169.
(19) Zhang, J.; Sapienza, P. J.; Ke, H.; Chang, A.; Hengel, S. R.; Wang, H.; Phillips, G. N.; Lee, A. L. Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of protein tyrosine phosphatase 1E. *Biochemistry* **2010**, *49*, 9280−9291.
(20) Nourry, C.; Grant, S. G. N.; Borg, J.-P. PDZ domain proteins: plug and play! *Sci. STKE* **2003**, *2003*, RE7.
(21) Sheng, M.; Sala, C. PDZ domains and the organization of supramolecular complexes. *Annu. Rev. Neurosci.* **2001**, *24*, 1−29.
(22) Bezprozvanny, I.; Maximov, A. PDZ domains: More than just a glue. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 787−789.
(23) Petit, C. M.; Zhang, J.; Sapienza, P. J.; Fuentes, E. J.; Lee, A. L. Hidden dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 18249−18254.
(24) Zhang, M. Scaffold proteins as dynamic switches. *Nat. Chem. Biol.* **2007**, *3*, 756−757.
(25) Lockless, S. W.; Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **1999**, *286*, 295−299.
(26) Fuentes, E. J.; Der, C. J.; Lee, A. L. Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J. Mol. Biol.* **2004**, *335*, 1105−1115.
(27) Fuentes, E. J.; Gilmore, S. A.; Mauldin, R. V.; Lee, A. L. Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. *J. Mol. Biol.* **2006**, *364*, 337−351.
(28) Gianni, S.; Walma, T.; Arcovito, A.; Calosci, N.; Bellelli, A.; Engstrom, A.; Travaglini-Allocatelli, C.; Brunori, M.; Jemth, P.; Vuister, G. W. Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure* **2006**, *14*, 1801−1809.
(29) Chi, C. N.; Elfstrom, L.; Shi, Y.; Snall, T.; Engstrom, A.; Jemth, P. Reassessing a sparse energetic network within a single protein domain. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 4679−4684.
(30) Gianni, S.; Haq, S. R.; Montemiglio, L. C.; Jurgens, M. C.; Engstrom, A.; Chi, C. N.; Brunori, M.; Jemth, P. Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity. *J. Biol. Chem.* **2011**, *286*, 27167−27175.
(31) De Los Rios, P.; Cecconi, F.; Pretre, A.; Dietler, G.; Michielin, O.; Piazza, F.; Juanico, B. Functional dynamics of PDZ binding domains: a normal-mode analysis. *Biophys. J.* **2005**, *89*, 14−21.
(32) Ota, N.; Agard, D. A. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.* **2005**, *351*, 345−354.
(33) Sharp, K.; Skinner, J. J. Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Proteins* **2006**, *65*, 347−361.
(34) Dhulesia, A.; Gsponer, J.; Vendruscolo, M. Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein. *J. Am. Chem. Soc.* **2008**, *130*, 8931−8939.
(35) Kong, Y.; Karplus, M. Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins* **2009**, *74*, 145−154.
(36) Ho, B. K.; Agard, D. A. Conserved tertiary couplings stabilize elements in the PDZ fold, leading to characteristic patterns of domain conformational flexibility. *Protein Sci.* **2010**, *19*, 398−411.
(37) Vijayabaskar, M. S.; Vishveshwara, S. Interaction energy based protein structure networks. *Biophys. J.* **2010**, *99*, 3704−3715.
(38) Gerek, Z. N.; Ozkan, S. B. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput. Biol.* **2011**, *7*, e1002154.
(39) Cilia, E.; Vuister, G. W.; Lenaerts, T. Accurate Prediction of the Dynamical Changes within the Second PDZ Domain of PTP1e. *PLoS Comput. Biol.* **2012**, *8*, e1002794.
(40) Vishveshwara, S.; Brinda, K. V.; Kannan, N. Protein structure: insights from graph theory. *J. Theor. Comput. Chem.* **2002**, *1*, 187−211.
(41) Kannan, N.; Vishveshwara, S. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **1999**, *292*, 441−464.
(42) Ghosh, A.; Brinda, K. V.; Vishveshwara, S. Dynamics of lysozyme structure network: probing the process of unfolding. *Biophys. J.* **2007**, *92*, 2523−2535.
(43) Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(44) Kovacs, J. A.; Chacon, P.; Abagyan, R. Predictions of protein flexibility: First-order measures. *Proteins* **2004**, *56*, 661−668.
(45) Zheng, W.; Brooks, B. R. Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: myosin versus kinesin. *Biophys. J.* **2005**, *89*, 167−178.
(46) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* **2000**, *41*, 1−7.

(47) Durand, P.; Trinquier, G.; Sanejouand, Y.-H. A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers* **1994**, *34*, 759−771.
(48) Van Wynsberghe, A. W.; Cui, Q. Interpreting correlated motions using normal mode analysis. *Structure* **2006**, *14*, 1647−1653.
(49) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins* **1998**, *33*, 417−429.
(50) Wang, Y.; Rader, A. J.; Bahar, I.; Jernigan, R. L. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* **2004**, *147*, 302−314.
(51) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des.* **1997**, *2*, 173−181.
(52) Dijkstra, E. W. A Note on Two Problems in Connexion with Graphs. *Numer. Math.* **1959**, *1*, 269−271.
(53) Heyer, L. J.; Kruglyak, S.; Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **1999**, *9*, 1106−1115.
(54) Lange, O. F.; Grubmuller, H. Generalized correlation for biomolecular dynamics. *Proteins* **2006**, *62*, 1053−1061.