JCTC Journal of Chemical Theory and Computation

Article

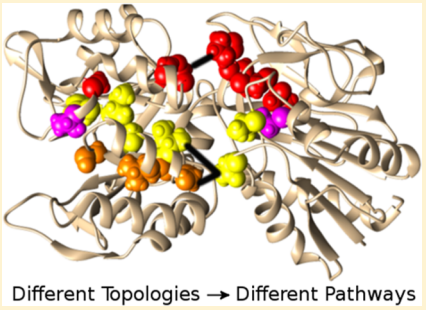
pubs.acs.org/JCTC

Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology

Andre A. S. T. Ribeiro and Vanessa Ortiz*

Department of Chemical Engineering, Columbia University, New York, New York 10027, United States

ABSTRACT: Network theory methods are being increasingly applied to proteins to investigate complex biological phenomena. Residues that are important for signaling processes can be identified by their condition as critical nodes in a protein structure network. This analysis involves modeling the protein as a graph in which each residue is represented as a node and edges are drawn between nodes that are deemed connected. In this paper, we show that the results obtained from this type of network analysis (i.e., signaling pathways, key residues for signal transmission, etc.) are profoundly affected by the topology of the network, with normally used determination of network edges by geometrical cutoff schemes giving rise to substantial statistical errors. We propose a method of determining protein structure networks by calculating inter-residue interaction energies and show that it gives an accurate and reliable description of the signal-propagation properties of a known allosteric enzyme. We also show that including covalent interactions in the network topology is essential for accurate results to be obtained.



1. INTRODUCTION

Allostery and other processes that involve signal propagation in proteins are extremely important in the regulation of cellular metabolism, and understanding these complex phenomena is a major goal of current biological research.[1−5] Computationally, the behavior of biological systems can be simulated by molecular dynamics (MD) simulations or related methods,[6−12] even though their size generally prevents the realization of MD simulations over the time scales that are characteristic of the biological processes in which they are involved.

Network theory has been successfully applied to gain insights into a wide variety of complex systems,[13] of which one example are proteins and other biological molecules.[14−21] A network's topology is defined by a set of elements (nodes) and connections (edges). The determined topology thus serves as an efficient way to map residue−residue contacts and can be used to compare different protein structures or conformations of the same molecule.[22] Networks can also be either weighted or unweighted. An unweighted protein structure network provides information on which residues are in contact, while, if weights are assigned based on the number of contacts, the corresponding weighted network provides information on how many contacts are formed between any pair of residues in the molecule.

Protein structure networks can also be used to study signal propagation in proteins. Signal propagation through a specific pathway means that the aminoacid residues along this path provide an optimal way for the structural changes to be propagated.[23] A related problem in network theory, for which several algorithms have been developed, is the determination of the shortest path between two nodes.[24−26] The solution to this problem provides the optimal way for traveling between points A and B. If an unweighted network is being analyzed, this optimal path will contain the least number of intermediate nodes. In the case of a weighted network, the optimal path will be the one with the shortest total distance.

In defining the network topology for a protein, aminoacid residues can be represented as nodes while edges can be drawn between neighboring residues, often subject to a given geometrically based distance cutoff value. If a single structure is being analyzed, residues that have atoms within the cutoff value of each other are considered connected. If an MD trajectory is being analyzed a second frequency-based cutoff needs to be defined, such that residues are considered connected if they have atoms within the distance cutoff value for most of the structures. In addition, it is reasonable to assume that signal propagation in proteins does not depend entirely on the topology and that usage of unweighted networks or euclidean distances as weights would provide poor results. Several studies have been recently published in which correlation coefficients calculated over MD trajectories are taken as weights.[27−34] Another possibility is to determine the weights based on the number of atom−atom contacts between residues.[35−39] This option comes with a computational advantage, since networks defined in this manner do not require the realization of demanding MD calculations. To determine whether these methods can provide a reliable account of signal propagation pathways in proteins, it is necessary to investigate the dependence of the analysis output on the network topology and employed weights.

Given the arbitrary nature of any cutoff value, it is important to question how the choice of cutoff values used to define the network topology affect the obtained results. A simple example

Received: November 15, 2013
Published: February 25, 2014

   dx.doi.org/10.1021/ct400977r | J. Chem. Theory Comput. 2014, 10, 1762−1769

of the implications of topological differences in networks is outlined in Figure 1. Figure 1A shows a graph that has a

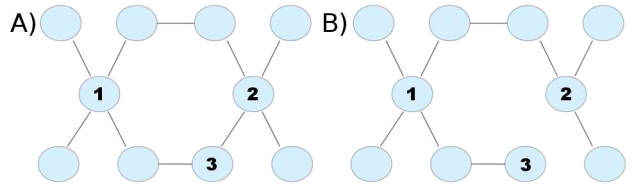


Figure 1. Two different graphs. Nodes are represented as circles and lines are drawn between connected nodes. Nodes 1 and 2 act as hubs of the network shown in panel A. This situation can be dramatically changed by the removal of one single connection, as can be seen in panel B.

symmetrical topology. Calculation of the shortest path between nodes 2 and 3 will obviously show no intermediate nodes if the edges are unweighted, while the shortest path between nodes 1 and 2 is doubly degenerate. Figure 1B shows that the removal of a single edge completely changes the network topology. Regardless of the weight assigned to the internode distance between nodes 2 and 3, their shortest path will need to go through node 1. It should also be noted that, depending on the topology, graphs may exhibit *hubs*, which are nodes that are highly connected to others and are thus critical for information transfer through the network. In Figure 1B, node 2 ceases to function as a prominent hub of the network, while the opposite holds for node 1. If a MD trajectory of a protein is being analyzed, how will atomic fluctuations affect the topology? It can certainly be expected that contact distances that fluctuate closely to the chosen cutoff value will result in topological changes depending on the analyzed frames of the trajectory. We also expect that the implications for the signaling properties of the network will be severe if the affected edges are important for signal propagation. Previous studies have assumed that a solution to this problem is to employ relatively large cutoff values and rely on the employed weights to discriminate between the different signaling pathways. In this paper we show that correlation coefficients do not accomplish this task. We also propose a different method of assigning topologies and internode distances and show that this method is able to determine residues that are important for signal propagation in a known allosteric enzyme.

2. METHODS

2.1. System Preparation and MD Simulations. Structure networks were determined for the allosteric enzyme imidazole glycerol phosphate synthase (IGPS). The crystal structure of the apo state of IGPS from *T. Maritima* (PDB: 1GPW)[40] was used as input for the calculations. This structure contains monomers HisF and HisH. For the rest of this manuscript, specific aminoacid residues will be identified by their identity, residue number and monomer they belong to. For example, Asp98 of monomer HisF may be referred to as Asp98f or D98f.

The protein structure was solvated with TIP3P water molecules[41] in a truncated octahedral box with minimum distance between the box edges and any protein atoms set to 10 Å. The AMBER-03 forcefield was used to describe the protein system.[42] Long-range electrostatic interactions were treated with the particle mesh Ewald (PME) method, with a cutoff of 12 Å, while van der Waals interactions were switched off between 10 and 12 Å. The system was minimized with the *steepest descents* algorithm, followed by a constant temperature MD simulation at 300 K for 500 ps and an equilibration run at (T = 300 K, p = 1 bar) for a further nanosecond. Finally, a production run of 50 ns was performed. The GROMACS simulation package[43] in its version 4.5.5 was used to perform the simulations. The thermostat of Bussi and co-workers[44] was used to maintain the temperature of the system, while pressure control was achieved with the Parrinello−Rahman algorithm.[45] An integration time step of 2 fs was used in the MD simulations, with bond lengths of protein and water molecules constrained with the LINCS[46] and SETTLE[47] algorithms, respectively.

2.2. Correlation Coefficients. Correlated motions were extracted from the 50-ns trajectory by calculating the correlation coefficients using the following equation:

$$r_{\mathrm{P}}^{ij} = \frac{\langle x_i x_j \rangle}{(\langle x_i^2 \rangle \langle x_j^2 \rangle)^{1/2}} \tag{1}$$

where $\mathbf{x}_i$ is the fluctuation vector of residue i, defined as the deviation from the average coordinates ($\mathbf{x}_i \equiv \mathbf{r}_i - \langle \mathbf{r}_i \rangle$, where $\mathbf{r}_i$ is the position vector of residue i). The trajectory frames are fitted to a reference structure to remove translational and rotational motions. Equation 1 defines residue−residue correlation coefficients. In order to apply this equation to the analysis of the atomic-level trajectory, the center of mass coordinate of each residue was used to define its position.

The correlation coefficients defined by eq 1 are commonly referred to as Pearson correlation coefficients. They are well-suited to quantify correlations if $\mathbf{x}_i$ and $\mathbf{x}_j$ are colinear vectors, however, they completely fail to identify correlated motions if the two vectors are perpendicular to each other. Furthermore, nonlinear correlations are also not taken into account. To overcome these difficulties, a formulation based on the mutual information concept[48−51] has been used to define generalized correlation coefficients.[49] The calculation of the mutual information is based on the comparison of the joint and marginal probability distributions of a set of random variables. For the particular case of two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the joint probability $p(\mathbf{x}_i,\mathbf{x}_j)$ will only be equal to the product of the marginal probabilies $p(\mathbf{x}_i)$ and $p(\mathbf{x}_j)$ if the two vectors are independent (uncorrelated). The mutual information can then be defined as

$$I_{ij} \equiv \int p(\mathrm{x}_i, \mathrm{x}_j) \ln \frac{p(\mathrm{x}_i, \mathrm{x}_j)}{p(\mathrm{x}_i) p(\mathrm{x}_j)} \, \mathrm{dx}_i \, \mathrm{dx}_j \tag{2}$$

We note that the mutual information is zero for two independent vectors. Lange and Grubmüller have noted that Pearson correlation coefficients capture all correlations for the special case of colinear Gaussian distributions of unit variance.[49] These authors have defined generalized correlation coefficients r_{MI} as the Pearson coefficient of a distribution of this kind, whose mutual information is equal to the actual distribution that is being analyzed:

$$r_{\mathrm{MI}}^{ij} = (1 - \mathrm{e}^{-2I_{ij}/d})^{1/2} \tag{3}$$

where d is the dimensionality of the vector space.

The generalized correlation coefficients were calculated for the IGPS trajectory, using the center of mass coordinates to define residue fluctuations $\mathbf{x}_i$, as in the Pearson analysis. To efficiently estimate the quantities I_{ij}, a k-nearest neighbor

distances algorithm was implemented using a k value of 6 and the first algorithm as described by Kraskov and co-workers.[52]

2.3. Implementation and Further Details. Computational routines for calculating Pearson correlation coefficients, generalized correlation coefficients, minimum paths, and several analysis tools were implemented using Perl Data Language (PDL). These routines may be obtained from the authors upon request. The protein visualization figures presented in this paper were prepared with the software UCSF Chimera.[53]

3. RESULTS AND DISCUSSION

3.1. Shortest Path Analysis. Correlated motions are essential for biological function as complex biological phenomena frequently involve significant structural fluctuations.[54] Microscopically, the motional coupling of different degress of freedom results in an entropy loss that must be compensated by some energetic contribution. Perturbations of the biomolecular system, such as the binding of an effector molecule, may change this balance of atomic interactions and result in increased correlations between different parts of the molecule. The employment of correlation coefficients to characterize signaling pathways thus seems a reasonable approach. The question that needs to be addressed is whether, when used as weights in network analysis, they are specific enough to overcome relatively loose topological definitions dictated by the usage of arbitrary cutoff values. More specifically, it is desirable to not only identify regions of the protein that are important for signaling but also specific residues and inter-residue interactions.

To evaluate the signaling properties of networks based on correlation coefficients, we have simulated imidazole glycerol phosphate synthase (IGPS) for 50 ns and analyzed the resulting correlation coefficients in an equivalent manner to a previous study of the same system.[33] IGPS is a heterodimeric enzyme (Figure 2), with the HisH monomer catalyzing glutamine hydrolysis into ammonia, and the HisF monomer catalyzing a chemical reaction between the produced ammonia and an effector molecule. The active site of HisH is a catalytic triad composed by residues Glu180h (Figure 2, black), His178h (Figure 2, black), and Cys84h (Figure 2, magenta). This enzyme is allosterically regulated, since the binding of the effector molecule to HisF results in a dramatic increase in glutamine hydrolysis in HisH. Mutational analysis revealed a salt-bridge between two residues in the HisF-HisH interface, namely Asp98f (Figure 2, blue) and Lys181h (Figure 2, red), to be essential for allosteric function.[55,56] This is also the only conserved salt-bridge across the HisF−HisH interface.[55] Computational studies indicate that there is an increase in correlations between residues close to this salt-bridge upon effector binding,[33,34] further highlighting the importance of these residues.

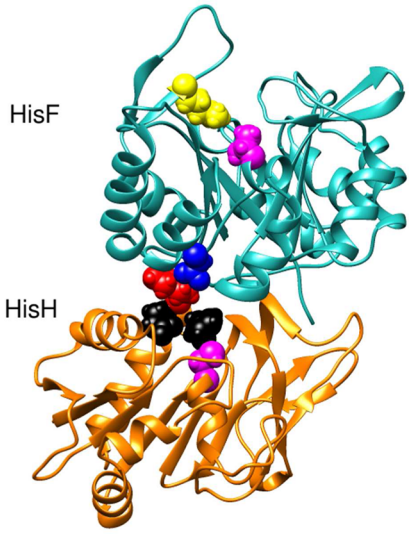


Figure 2. Structure of the enzyme imidazole glycerol phosphate synthase. The allosteric mechanism involves signal propagation from the effector binding site in the HisF domain to the active site in the HisH domain. Residues K19f (yellow) and T104f (magenta) are part of the effector binding site. Residues E180h (black), H178h (black), and C84h (magenta) constitute the HisH active site. Residues D98f (blue) and K181h (red) form a salt bridge that is essential for allosteric function.

The effects of the topology on the resulting signaling pathways were studied by changing both the cutoff value and the chemical nature of the atoms considered when analyzing the 50 ns MD trajectory. We have performed this analysis with four different methods. The first two, which we shall refer to as PA-3.0 and PA-4.5, use the all-atom representation of the trajectory to calculate the center of mass of each residue and to identify which residues are in contact. Edges are drawn between nodes if the respective residues have any atoms within the cutoff values of 3.0 and 4.5 Å for methods PA-3.0 and PA-4.5, respectively. In addition, a frequency cutoff is employed, such that edges between residues that are not in contact for at least 75% of the trajectory are disregarded. The remaining two methods (PH-4.0 and PH-4.5) use cutoffs of 4.0 and 4.5 Å, respectively, but disregard hydrogen atoms when calculating edges and centers of mass. Finally, and as done in the previous study,[33] each edge was weighted using the correlation coefficients as defined in eq 1, by setting internode distances as $-\log(|r_{ij}|)$.

With the protein structure networks determined, we have calculated shortest paths between residues T104f and C84h for each of the methods by using the Floyd−Warshall algorithm.[24] Residue T104f is part of the effector binding site in HisF, directly interacting with bound effector molecules. Cys84h is a member of the catalytic triad that constitutes the HisH active site. These two residues are therefore reasonable candidates for determination of a signaling pathway characteristic of IGPS.

Table 1 shows the determined shortest paths for the different methods, with interdomain edges highlighted in bold letters. It can be seen that the determined pathways do not agree.

Table 1. Shortest Paths between Residues T104f and C84h as Calculated by the Floyd−Warshall Algorithm[a]

method	pathway
PA-3.0	T104f-A106f-I113f-A117f-**S122f-P119h**-H120h-H141h-Q176h-C84h
PA-4.5	T104f-N103f-I102f-A89f-**I93f-N12h**-V51h-C84h
PH-4.0	T104f-N103f-I102f-T78f-F77f-**P76f-Y138h**-H178h-C84h
PH-4.5	T104f-N103f-I102f-S101f-**K99f-Y138h**-F177h-C84h

[a]The topology and internode distances were defined in different ways: the first approach used the all-atom trajectory to calculate contacts and correlation coefficients, employing different cutoff values for topology determination (first two rows). The second approach used only heavy (non-hydrogen) atomic coordinates to calculate contacts and correlation coefficients, also employing different cutoffs (last two rows). Interdomain edges are highlighted in bold letters.

Furthermore, the interdomain edges, which are responsible for signal propagation through the critical monomer−monomer interface, do not coincide in any case. It is interesting to compare the pathways for methods PH-4.0 and PH-4.5, where a small (0.5 Å) change in the cutoff value is responsible for changing the HisF interface signaling residue from Pro76f to Lys99f. We further note that the pathway determined in the previous study[33] also identified Pro76f as the HisF-interface signaling residue, even though the employed criteria were equivalent to our method PH-4.5.[33,57] These observations indicate that even small changes in the way the Pearson correlation coefficients are calculated and analyzed may change qualitative features of the network representation.

Figure 3 shows the pathways obtained by methods PA-3.0 (red pathway) and PA-4.5 (yellow pathway) highlighted in the

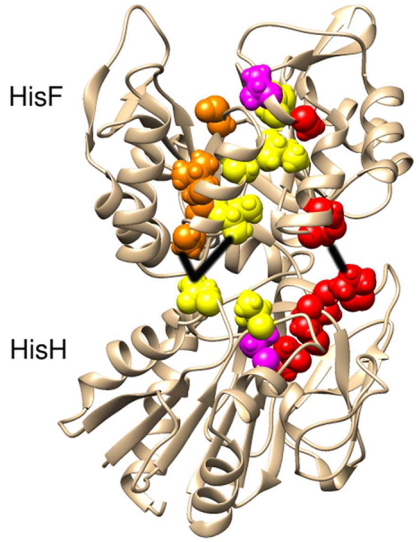


Figure 3. Shortest paths as determined by the Floyd−Warshall algorithm. Shortest paths between residues T104f and C84h were determined with different distance cutoffs. A cutoff of 3.0 Å between any atoms resulted in the path shown in red, while a cutoff of 4.5 Å between any atoms resulted in the path shown in yellow. The shortest path between residues G81f and C84h with a cutoff of 4.5 Å between any atoms is shown in orange. Residues T104f and C84h are shown in magenta and black lines are drawn for the interdomain edges in all paths.

structure of IGPS. It can be seen that, even though the end points are the same (T104f and C84h, shown as magenta spheres), the pathways go through substantially different regions of the protein structure. In addition, when changing the starting point to G81f (a residue less than 6 Å apart from T104f), the resulting pathway (orange in Figure 3) overlaps with the T104f-C84h pathway in the HisH domain but not in the HisF domain. This shows that different starting points can lead to different conclusions about signaling propensities.

3.2. Identifying Critical Network Nodes. The observations presented in the previous section demonstrate the need for analysis that focuses less on individual pathways, and more on the entire network, by determining which nodes are present in the greatest number of pathways between different parts of the network. Calculation of the shortest paths between all different pairs of nodes provides a way to identify critical network nodes. The number of shortest paths that pass through a node is taken as a measure of its "bottleneckness"[58] (also named *node betweenness*), and is one of several measures of network centrality.[13] We have applied this analysis to IGPS by calculating the normalized node betweenness η, defined as the number of shortest paths that go through a node divided by the maximum number of shortest paths that go through any node.

Figure 4 shows the structure of IGPS with residues that have a value of η greater than 0.5 colored according to their

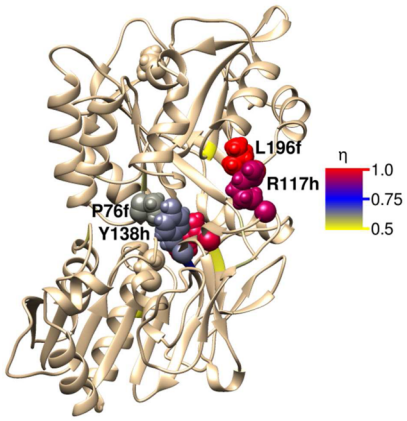


Figure 4. Normalized node betweenness (η) for the analysis based on heavy-atom correlation coefficients with a cutoff of 4.0 Å (method PH-4.0). Residues for which $\eta \geq 0.5$ are colored according to the above scale. Residues T104f and C84h are represented as spheres.

respective values, obtained from the analysis based on method PH-4.0. It can be seen that the residues that have the highest values of η are mostly within the monomer−monomer interface. Furthermore, the critical Asp98f-Lys181h salt-bridge is not highlighted, evidencing that this method is not able to correctly identify residues that are important for allosteric function. In addition, the residue with the highest η value is Leu196f, which is in close proximity to Arg117h (another residue exhibiting a high η value). This method, therefore, points to a relatively weak interaction between a leucine and an arginine as a major agent for signal propagation. Finally, it is noted that the correlation coefficients obtained with this method for D98f-K181h and L196f-R117h are virtually the same (0.22 and 0.20, respectively). For these reasons, it is concluded that Pearson correlation coefficients do not yield sufficiently specific weights to overcome the loose topological definitions dictated by arbitrary cutoffs.

3.3. Generalized Correlation Coefficients. Pearson correlation coefficients are a well-established way to quantify correlated motions in MD simulations of biomolecular systems. However, they fail to identify correlated motions in several different situations (see Methods section). To investigate whether a more robust assessment of correlated motions can successfully identify residues important for allosteric signal propagation in IGPS, we calculated generalized correlation coefficients based on the mutual information concept[48] for the MD simulation and evaluated the signaling properties of the resulting network. The criteria used to define the topology were identical to method PH-4.0, that is, a cutoff of 4 Å with a heavy-atom representation of the protein. We will refer to this method as MIH-4.0.

Figure 5 shows the structure of IGPS with residues that have a value of η greater than 0.5 colored according to their respective values. It can be seen that the critical residues are spatially distant from the corresponding residues of method

dx.doi.org/10.1021/ct400977r | J. Chem. Theory Comput. 2014, 10, 1762−1769

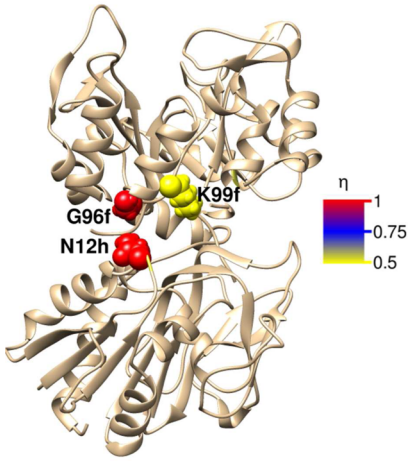


Figure 5. Normalized node betweenness (η) for the analysis based on generalized correlation coefficients. Residues for which η > = 0.5 are colored according to the above scale.

PH-4.0 (Figure 4). More importantly, residues G96f and N12h exhibit the highest η values. These residues are close to the important salt-bridge D98f-K181h. Residue N12h forms hydrogen bonds with N15h and interacts with the salt-bridge, contributing to its stabilization.[56] Usage of the generalized correlation coefficients as weights thus seems to correctly identify regions important for signal propagation; however, specific interactions are still not correctly identified.

3.4. Energy Based Method. The employment of correlation coefficients or number of atom−atom contacts to specify internode distances relies on the fact that interatomic interactions are ultimately responsible for any process that occurs on the molecular level. To determine whether a more accurate account of signaling pathways in proteins can be obtained by the employment of protein structure networks, we have chosen to use an energy based method of defining network topologies and internode distances. Our approach is based on two assumptions: first, chemical bonds provide an optimal way for propagation of signals; second, propensities for signal propagation through noncovalent interactions depend on their strength, with these propensities not being higher than the ones for covalent bonds.

The proposed method involved calculating residue−residue nonbonded interaction energies ε_{ij}, averaged over the entire MD trajectory. Weights ω_{ij} are assigned to different residue pairs as follows:

$$\omega_{ij} = \begin{cases} \omega_b, & \text{if } i \text{ and } j \text{ are covalently bound} \\ \chi_{ij}, & \text{otherwise} \end{cases} \quad (4)$$

The parameter ω_b was set with the value 0.99. The function χ_{ij} is evaluated by calculating the average interaction energy between all pairs of noncovalently bound residues ε_{av} (excluding residue pairs for which $|\varepsilon_{ij}| < 1$ kJ/mol). We then define $\chi_{ij} \equiv 0.5\{1 - (\varepsilon_{ij} - \varepsilon_{av})/5\varepsilon_{rmsd}\}$, where ε_{rmsd} is the root-mean-square deviation of the distribution of interaction energies. Weights greater than 0.99 were reassigned to this value. Node pairs for which $\omega_{ij} < 0.01$ were considered disconnected, and internode distances were then taken as $-\log(\omega_{ij})$. We note that the above definition gives higher weights for strong, attractive interactions. The average and root-mean-square deviation of the interaction energies were −9.5 and 14.6 kJ/mol, respectively. The normalization range of 5 standard deviations for the edge weights thus led to relatively few nonbonded edges receiving the maximum weight. We also note that strong repulsive interactions between nonbonded residues are extremely unlikely since the system will naturally evolve to increase the distance between the two residues and/or use solvent molecules to reduce the strength of the interaction. In practice, the proposed method resulted in no residue pairs that exhibited a nonzero interaction energy being considered disconnected. Therefore, it can be said that the method as formulated is free of cut-offs.

This method was used to calculate η values as defined previously, with the results shown in Figure 6. It can be seen

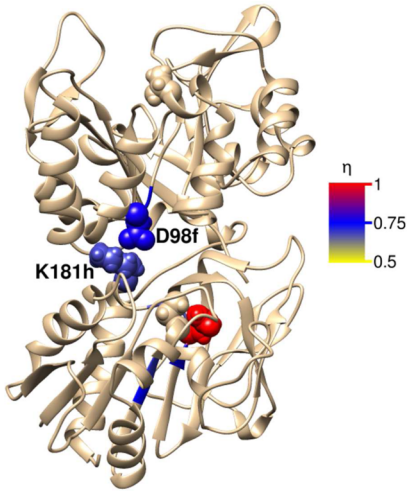


Figure 6. Normalized node betweenness (η) for the analysis based on energy evaluation. Residues for which $\eta \geq 0.5$ are colored according to the above scale. Residues T104f and C84h are represented as spheres. The residue with η = 1 is Q176h (red), which is close to the HisH active site.

that this method correctly identifies the salt-bridge D98f-K181h as a major agent in signal propagation for IGPS, being in fact the only such edge in the monomer−monomer interface. It is also interesting to note that the residue with the highest η value is Glu176h, which is in close proximity to the HisH active site.

3.5. Error Analysis. We have shown that the energy based method, as opposed to the methods based on correlation coefficients/geometrical cutoffs, yields an accurate description of signal propagation in IGPS. However, the precision of these methods remains to be established. The dependence of the shortest paths between specific residue pairs on the employed correlation coefficient method discussed earlier indicates that the associated statistical error should be significant. Another source of error that needs to be characterized relates to atomic fluctuations that can lead to topological changes, as well as changes in the employed weights, when different trajectories or trajectory segments are used.

To evaluate how the signaling properties depend on these issues, we have chosen to focus on the overall node signaling propensities, as measured by the η values. Four steps were followed to evaluate this error. First, for each of the networks in an analyzed set, we ranked the residues according to the respective η values. This means that, for each network, the residue with an η value of 1 gets a rank of 1, the residue with the second highest η value gets a rank of 2, and so forth. Second, the average rank and root mean squared deviation in

rank (ε) were calculated for each residue. For example, if the analysis of three networks returns the ranks of a residue to be (2,3,2), its average rank will be $\langle r \rangle = (2 + 3 + 2)/3 = 2.33$, and its root mean squared deviation in rank will be $\varepsilon = \{\{(2 - 2.33)^2 + (3 - 2.33)^2 + (2 - 2.33)^2\}/3\}^{1/2}$. Third, we rank residues according to their $\langle r \rangle$ values. Finally, under the assumption that higher ranking nodes are more relevant for precision determination, we plotted the average root mean squared deviation in rank as a function of nodes included in the average. In other words, we evaluate the precision of the different methods by calculating the average rank error of the n leading positions: $E_n = \sum_{i=1}^{n} \varepsilon_i / n$, with ε_1 being the rank error of the node with the lowest $\langle r \rangle$ value.

Figure 7 shows the results from this error analysis when applied to the methods outlined in this paper. Four different

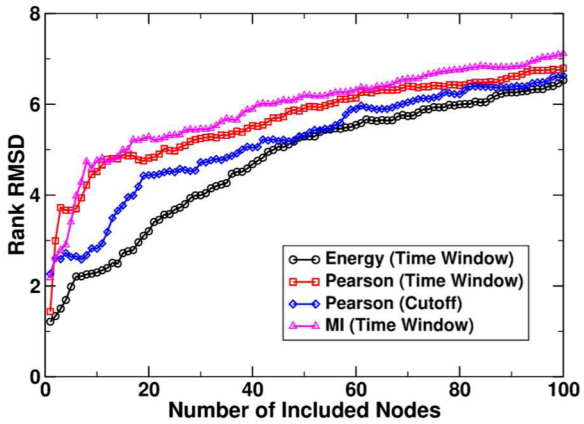


Figure 7. Average rank RMSD (E_n) as a function of included nodes (see text for details). The error analysis was evaluated for different time windows of the MD trajectory, when analyzed by methods PH-4.0, MIH-4.0, and Energy. The error analysis for the employment of methods PA-3.0, PA-4.5, PH-4.0, and PH-4.5 to analyze the entire MD trajectory is also presented.

sets of results are presented. The Energy (black) data set in Figure 7 was constructed by applying the error analysis described in the previous paragraph to a set of five networks constructed by applying the energy-based method on five different 10-ns segments extracted from the total 50-ns MD trajectory (0−10 ns, 10−20 ns, 20−30 ns, 30−40 ns, 40−50 ns). The red and magenta data sets were constructed in the same way but using the PH-4.0 and MIH-4.0 methods, respectively. Finally, the blue data set was constructed by applying the error analysis described above to a set of four networks constructed by applying the four Pearson correlation coefficient methods (PA-3.0, PA-4.5, PH-4.0, and PH-4.5), each on the entire 50-ns MD trajectory. It can be seen from Figure 7 that substantial reductions in error are achieved with the energy based method, especially when considering the leading positions.

3.6. Importance of Chemical Bonds. The proposed energy based method is able to identify residues that are important for allosteric function of IGPS. This is done by considering that chemical bonds and strong nonbonded interactions provide optimal ways for signal propagation. It must be noted that a similar energy based method to construct protein structure networks has recently been proposed.[59,60] The authors have calculated and analyzed average interaction energies between nonbonded residues in an analogous manner to the present study. However, the authors have disregarded chemically connected residues in the construction of the network. It must be noted that this approach inherently assumes that the contribution of covalent interactions to signal propagation is negligible.

In order to assess the effects of disregarding covalent interactions in the network representation, we analyzed how the signaling properties of the network depend on the ω_b parameter. Figure 8 presents the η values for the network that

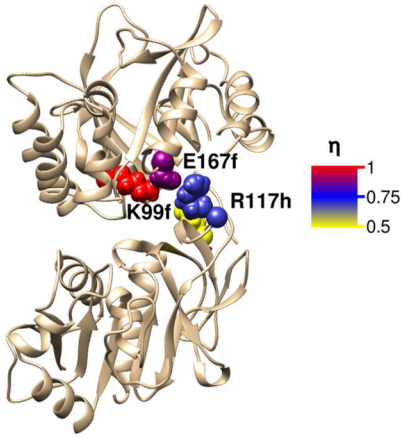


Figure 8. Normalized node betweenness (η) for the analysis based on energy evaluation with chemically bound residues disregarded. Residues for which $\eta \geq 0.5$ are colored according to the above scale.

completely disregards covalent interactions ($\omega_b = 0$). The results are significantly different from Figure 6, with the network highlighting the interaction between Glu167f and Arg117h as the major agent of signal propagation between monomers.

The obtained results show that completely disregarding covalent interactions has significant effects on the network representation, with the ability to identify the salt bridge D98f-K181h as a critical edge for signal propagation lost. To further analyze the importance of chemical bonds for signal propagation, we have constructed networks with intermediate values of ω_b. We use the ranks of residues D98f, K181h, and R117h as a measure of network quality. Figure 9 gives the results for different values of ω_b. It can be seen that the ranks of

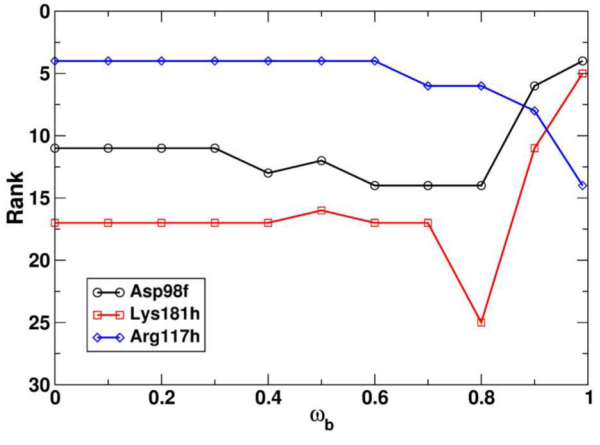


Figure 9. Ranks of residues Asp98f, Lys181h, and Arg117h as a function of the parameter ω_b.

dx.doi.org/10.1021/ct400977r | J. Chem. Theory Comput. 2014, 10, 1762−1769

residues D98f and K181h significantly increase for the highest values of ω_b, while the opposite is true for R117h. This indicates that the inclusion of covalently bound residues in the topology is essential for an adequate assessment of the signaling properties of IGPS.

4. CONCLUSIONS

We have calculated the signaling properties of protein structure networks of the allosteric enzyme IGPS using different criteria for determination of topologies and internode distances. It is shown that Pearson correlation coefficients are not able to correctly identify residues or regions important for allosteric function. The employment of generalized correlation coefficients improves the obtained results (as previously noted in Lange and Grubmüller[49] and Rivalta et al.[34]) but still cannot correctly identify important residues for allosteric signal propagation.

We have introduced a new energy based approach to characterize the signaling pathways of IGPS. This method is based on the assumption that signal propagation occurs preferentially through covalently bound residues or residues that have a strong noncovalent interaction energy. The obtained results showed that this approach is able to correctly identify important residues for allosteric signal propagation in IGPS. We have also analyzed the importance of chemical bonds for signal propagation. The neglect of covalent interactions in the determination of network topologies leads to a significant deterioration of the description of signal propagation.

Protein structure networks may be defined in several different manners and error estimates are essential for meaningful comparisons to be made. We have calculated average rank errors for the different calculated networks and the results show that the energy based method is able to determine residues important for allosteric function both accurately and reliably. We also note that these results were obtained for the apo state of the enzyme (effector molecule unbound), indicating that possible communication pathways may be determined without knowledge of allosteric binding sites or effector molecules. The fact that the energy based method yields low statistical errors when different trajectory segments are analyzed is also significant, as shorter MD simulations may be used to obtain equivalent results.

AUTHOR INFORMATION

Corresponding Author

*E-mail: vortiz@columbia.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1125698.

REFERENCES

(1) Cui, Q.; Karplus, M. *Protein Sci.* **2008**, *17*, 1295−1307.
(2) Changeux, J.-P.; Edelstein, S. J. *Science* **2005**, *308*, 1424−1428.
(3) Changeux, J.-P. *Protein Sci.* **2011**, *20*, 1119−1124.
(4) Lindsley, J. E.; Rutter, J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10533−10535.
(5) Berezovsky, I. N. *Biochim. Biophys. Acta* **2013**, *1834*, 830−835.
(6) Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 321−323.
(7) Kamp, M. W. V. D.; Shaw, K. E.; Woods, C. J.; Mulholland, A. J. *J. R. Soc. Interface* **2008**, *5*, 173−190.
(8) Levitt, M. *J. Mol. Biol.* **1983**, *168*, 595−620.
(9) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. In *Computational Molecular Dynamics: Challenges, Methods, Ideas*; Deuflhard, P., Hermans, J., Leimkuhler, B., Mark, A. E., Reich, S., Skeel, R., Eds.; Springer-Verlag: Berlin, 1998; Vol. *4*; pp 39−65.
(10) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.
(11) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435−439.
(12) Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. *J. Mol. Biol.* **2005**, *354*, 173−83.
(13) Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. *Phys. Rep.* **2006**, *424*, 175−308.
(14) Alves, N. A.; Martinez, A. S. *Phys. A* **2007**, *375*, 336−344.
(15) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. *J. Mol. Biol.* **2004**, *344*, 1135−1146.
(16) Bagler, G.; Sinha, S. *Phys. A* **2005**, *346*, 27−33.
(17) Csermely, P. *Tren. Biochem. Sci.* **2008**, *33*, 569−576.
(18) Böde, C.; Kovács, I. A.; Szalay, M. S.; Palotai, R.; Korcsmáros, T.; Csermely, P. *FEBS Lett.* **2007**, *581*, 2776−2782.
(19) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. *Proteins* **2001**, *44*, 150−165.
(20) Atilgan, A. R.; Akan, P.; Baysal, C. *Biophys. J.* **2004**, *86*, 85−91.
(21) Bhattacharyya, M.; Bhat, C. R.; Vishveshwara, S. *Protein Sci.* **2013**, *22*, 1399−1416.
(22) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. *J. Mol. Biol.* **1989**, *212*, 151−166.
(23) Lockless, S. W.; Ranganathan, R. *Science* **1999**, *286*, 295−299.
(24) Floyd, R. W. *Commun. ACM* **1962**, *5*, 345−345.
(25) Dijkstra, E. *Numer. Math.* **1959**, *1*, 269−271.
(26) Johnson, D. B. *J. ACM* **1977**, *24*, 1−13.
(27) Papaleo, E.; Lindorff-Larsen, K.; De Gioia, L. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12515−12525.
(28) Tehver, R.; Chen, J.; Thirumalai, D. *J. Mol. Biol.* **2009**, *387*, 390−406.
(29) Gasper, P. M.; Fuglestad, B.; Komives, E. A.; Markwick, P. R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 21216−21222.
(30) Freddolino, P. L.; Gardner, K. H.; Schulten, K. *Photochem. Photobiol. Sci.* **2013**, *12*, 1158−1170.
(31) Sethi, A.; Tian, J.; Derdeyn, C. A.; Korber, B.; Gnanakaran, S. *PLoS Comp. Biol.* **2013**, *9*, e1003046.
(32) Ghosh, A.; Vishveshwara, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15711−15716.
(33) Vanwart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. *J. Chem. Theor. Comp.* **2012**, *8*, 2949−2961.
(34) Rivalta, I.; ASultan, M. M.; Lee, N.; Manley, G. A.; Loria, J. P.; Batista, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1428−E1436.
(35) Bahar, I.; Lezon, T. R.; Yang, L.; Eyal, E. *Ann. Rev. Biophys.* **2010**, *39*, 23−42.
(36) Chennubhotla, C.; Yang, Z.; Bahar, I. *Mol. Biosys.* **2008**, *4*, 287−292.
(37) Yang, Z.; Májek, P.; Bahar, I. *PLoS Comp. Biol.* **2009**, *5*, e1000360.
(38) Chennubhotla, C.; Bahar, I. *PLoS Comp. Biol.* **2007**, *3*, 1716−1726.
(39) Chennubhotla, C.; Bahar, I. *Mol. Sys. Biol.* **2006**, *2*, 36.
(40) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V. J.; Smith, J. L. *Struct.* **2001**, *9*, 987−997.
(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(42) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W. E. I.; Yang, R.; Cieplak, P.; Luo, R. A. Y.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(43) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinf.* **2013**, *29*, 845−854.
(44) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(45) Parrinello, M.; Rahman, A. *Phys. Rev. Lett.* **1980**, *45*, 1196−1199.
(46) Hess, B. *J. Chem. Theor. Comp.* **2008**, *4*, 116−122.
(47) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952−962.
(48) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; John Wiley & Sons: New York, 2006; p 19.
(49) Lange, O. F.; Grubmüller, H. *Proteins* **2006**, *62*, 1053−1061.
(50) Bowman, G. R.; Geissler, P. L. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 11681−11686.
(51) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. *J. Chem. Theor. Comp.* **2009**, *5*, 2486−2502.
(52) Kraskov, A.; Stögbauer, H.; Grassberger, P. *Phys. Rev. E* **2004**, *69*, 066138.
(53) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(54) Forman-Kay, J. D. *Nat. Struc. Biol.* **1999**, *6*, 1086−1087.
(55) Myers, R. S.; Amaro, R. E.; Luthey-Schulten, Z. a.; Davisson, V. J. *Biochemistry* **2005**, *44*, 11974−11985.
(56) Amaro, R. E.; Sethi, A.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. *Biochem.* **2007**, *46*, 2156−2173.
(57) Sethi, A.; Eargle, J.; Black, A.; Luthey-Schulten, Z. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 6620−6625.
(58) Yu, H.; Kim, P. M.; Sprecher, E.; Trifonov, V.; Gerstein, M. *PLoS Comp. Biol.* **2007**, *3*, e59.
(59) Vijayabaskar, M. S.; Vishveshwara, S. *Biophys. J.* **2010**, *99*, 3704−3715.
(60) Vijayabaskar, M. S.; Vishveshwara, S. *PLoS Comp. Biol.* **2012**, *8*, e1002505.