Eigenvector centrality for characterization of protein allosteric pathways

Christian F. A. Negre[a,b,c,1,2], Uriel N. Morzan[b,c,1,2], Heidi P. Hendrickson[b,c,d], Rhitankar Pal[b,c], George P. Lisi[b,e], J. Patrick Loria[b,f], Ivan Rivalta[g,h,2], Junming Ho[i], and Victor S. Batista[b,c,2]

[a]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545; [b]Department of Chemistry, Yale University, New Haven, CT 06520-8107; [c]Energy Sciences Institute, Yale University, West Haven, CT 06516-7394; [d]Department of Chemistry, Lafayette College, Easton, PA 18042; [e]Department of Molecular Biology, Cell Biology & Biochemistry, Brown University, Providence, RI 02903; [f]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520; [g]Université de Lyon, École Normale Supérieure de Lyon, CNRS, Université Claude Bernard Lyon 1, Laboratoire de Chimie UMR 5182, Lyon, France; [h]Dipartimento di Chimica Industriale "Toso Montanari," Università degli Studi di Bologna, Viale del Risorgimento, 4I-40136 Bologna, Italy; and [i]School of Chemistry, University of New South Wales, Sydney NSW 2052, Australia

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved November 7, 2018 (received for review June 18, 2018)

Determining the principal energy-transfer pathways responsible for allosteric communication in biomolecules remains challenging, partially due to the intrinsic complexity of the systems and the lack of effective characterization methods. In this work, we introduce the eigenvector centrality metric based on mutual information to elucidate allosteric mechanisms that regulate enzymatic activity. Moreover, we propose a strategy to characterize the range of correlations that underlie the allosteric processes. We use the V-type allosteric enzyme imidazole glycerol phosphate synthase (IGPS) to test the proposed methodology. The eigenvector centrality method identifies key amino acid residues of IGPS with high susceptibility to effector binding. The findings are validated by solution NMR measurements yielding important biological insights, including direct experimental evidence for interdomain motion, the central role played by helix hα1, and the short-range nature of correlations responsible for the allosteric mechanism. Beyond insights on IGPS allosteric pathways and the nature of residues that could be targeted by therapeutic drugs or site-directed mutagenesis, the reported findings demonstrate the eigenvector centrality analysis as a general cost-effective methodology to gain fundamental understanding of allosteric mechanisms at the molecular level.

allostery | graph theory | eigenvector centrality | information theory | IGPS

Allostery establishes a wide range of regulatory processes in biological macromolecules. The primary step in the allosteric regulation often involves binding of a ligand effector that regulates catalytic activity far away from its biding site. The mechanisms of energy transfer between the allosteric and catalytic sites are essential for design of selective therapeutic methods. However, they are typically poorly understood due to the intrinsic complexity of the systems and the lack of effective characterization methods. Thus, establishing methodologies for understanding communication pathways between physically distant sites in allosteric enzymes remains an important outstanding challenge. Such methods could expedite the design of innovative drug therapies (1, 2) as well as protein engineering strategies (3–5).

Significant efforts have been recently reported in the development of computational tools to support, interpret, and/or predict experiments focused on the elucidation of allosteric pathways (2, 6–12). Network analysis has been extensively used in this context by incorporating concepts and approaches from graph theory in the realm of molecular dynamics (MD) simulations (9, 13–22). For instance, community network analysis (CNA) has emerged as a powerful and increasingly popular approach to analyze the dynamics of enzymes and protein/DNA (and/or RNA) complexes in studies of allosteric mechanisms (23–29).

Graph theory represents proteins as networks of nodes corresponding to amino acid residues or DNA/RNA bases, linked by edges. The length of the edges corresponds to the magnitude of a physical property correlating the nodes, such as the dynamical correlation (9, 30, 31), coupling strength (32), or distance between residues (33). For a network of N nodes, the corresponding graph is described by an $N \times N$ adjacency matrix $\mathbf{A}$ with elements $\mathbf{A}_{ij}$ defining the strength of the physical correlation between nodes i and j.

One of the cornerstones of network analysis is the concept of centrality—that is, the relative importance of an individual member in a group. Measures of centrality are crucial to identify the more influential nodes in a network. There are many measures of centrality characterizing slightly different aspects of the network. Probably the simplest of all is the degree centrality (DC), k_i, providing a measure of the relative connectivity of node i in the network, as follows:

$$k_i = \sum_{j=1}^{n} \mathbf{A}_{ij}, \qquad [1]$$

where $\mathbf{A}_{ij}$ defines the strength of the physical correlation between nodes i and j. A node that is well connected is expected to have a large "influence" on the graph. While the DC can

Significance

Allosteric processes are ubiquitous in macromolecules and regulate biochemical information transfer between spatially distant sites. Despite decades of study, allosteric processes remain generally poorly understood at the molecular level. Here, we introduce the eigenvector centrality measure of mutual information to disentangle the complex interplay of amino acid interactions giving rise to allosteric signaling. The analysis of eigenvector centrality is tested in imidazole glycerol phosphate synthase (IGPS), a prototypical V-type allosteric enzyme. The resulting insights allow us to pinpoint key amino acids in terms of their relevance in the allosteric process, suggesting protein-engineering strategies for control of enzymatic activity.

Author contributions: C.F.A.N., U.N.M., I.R., and V.S.B. designed research; C.F.A.N., U.N.M., H.P.H., R.P., G.P.L., J.P.L., and J.H. performed research; C.F.A.N. and U.N.M. contributed new reagents/analytic tools; C.F.A.N., U.N.M., H.P.H., G.P.L., and I.R. analyzed data; and C.F.A.N., U.N.M., I.R., and V.S.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] C.F.A.N. and U.N.M. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: uriel.morzan@yale.edu, cnegre@lanl.gov, ivan.rivalta@ens-lyon.fr, or victor.batista@yale.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810452115/-/DCSupplemental.

Published online December 10, 2018.

provide useful information, it is not a true “node centrality” as defined by Ruhnau (34) and thus does not give a measure of centrality based on a fixed scale that allows comparisons between different graphs.

An alternative definition is the betweenness centrality (BC), b_i, which provides a measure of how information can flow between nodes (or edges) in a network. The BC can be quantified as the number of times a node acts as a bridge along the geodesic (shortest) path between two other nodes,

$$b_i = \sum_{st} \frac{n^i_{st}}{g_{st}}, \qquad [2]$$

where n^i_{st} is the number of shortest paths between nodes s and t that pass through node i, and g_{st} is the total number of shortest paths between nodes s and t. The nodes with high BC have a large influence on the overall information passing by flow, and, hence, the removal of such nodes may disrupt the communication in the network. However, communication does not always take the shortest path, and, hence, the BC may provide only partial information on the relevance of each amino acid in the functional dynamics of a protein.

Somehow, in between these two definitions of centrality (i.e., degree and betweenness centralities), the eigenvector centrality (EC) emerges as an alternative that takes into account both the number of connections of a given node and its relevance in terms of information flow. The EC of a node, c_i, is defined as the weighted sum of the centralities of all nodes that are connected to it by an edge, $\mathbf{A}_{ij}$,

$$c_i = \epsilon^{-1} \sum_{j=1}^{n} \mathbf{A}_{ij}\, c_j, \qquad [3]$$

where $\mathbf{c}$ is the eigenvector associated to the eigenvalue ϵ of $\mathbf{A}$. The EC is a measure of how well connected a node is to other well-connected nodes in the network. Importantly, the EC serves as a measure of the connectivity against a fixed scale when normalized, so it can be used to reliably compare different networks (34). For example, the normalization becomes essential when analyzing differences between graphs, for example, to study the pattern of centrality variation between the *apo* and *holo* states of a protein.

In the present work, we illustrate the potential of the EC measure to provide a molecular-level characterization of the allosteric mechanism of enzymes. In particular, we focus on the prototypical case of the imidazole glycerol phosphate synthase (IGPS), a bacterial enzyme present in the amino acid and purine biosynthetic pathways of most microorganisms, making it an attractive target for antibiotic, pesticide, and herbicide development (35). Structurally, IGPS is a tightly associated heterodimer (Fig. 1) in which each monomer catalyzes a different reaction: The *HisH* enzyme promotes the hydrolysis of glutamine (Gln) to produce ammonia, which diffuses to the *HisF* subunit and reacts with the effector *N*-[(5-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR) to form imidazole glycerol phosphate and AICAR. While Gln binding is unaffected by the presence of PRFAR, the hydrolysis of Gln is accelerated 5,000-fold upon PRFAR binding through a mechanism that, for many years, has remained elusive (36). IGPS is thus a V-type enzyme and a model system to study noncooperative allostery involving conformational changes.

In a recent study (9), we carried out a BC-based CNA by optimizing the modularity function to explore the underlying allosteric mechanism of this enzyme. We now present an alternative strategy, exploring the description of allostery provided by the EC compared with the CNA based on optimal modularity (the connection between CNA and the EC is analyzed in detail in SI Appendix). This approach identifies the most important amino acids for the allosteric signaling, providing an ideal route for the identification of mutation targets to inhibit or enhance the IGPS catalytic activity and opening the doors to a plethora of combined theoretical–experimental studies oriented to increase the control of its function and develop new alternatives for drug discovery. Additionally, the strategy introduced in this work allows us to capture long-range contributions to the correlation pattern beyond our previous CNA study and fundamental aspects of the allosteric behavior of IGPS. In particular, we show that while the correlation between residues is enhanced by a conformational breathing motion, the allosteric pathway is dominated by short-range contacts (9).

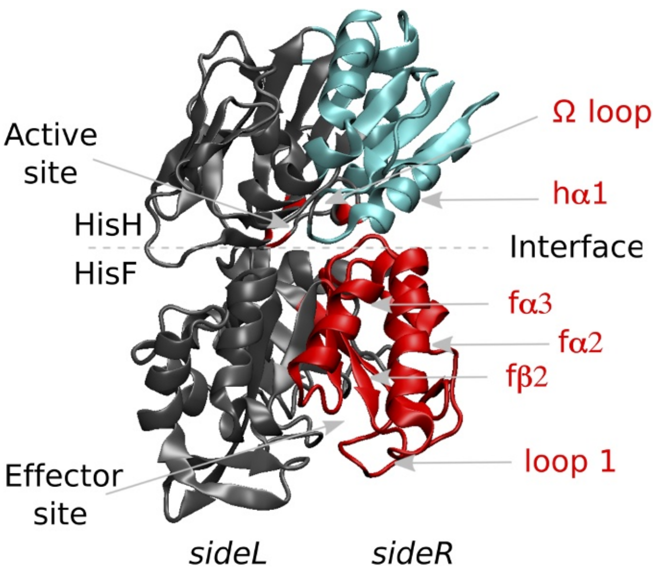

Fig. 1. Molecular representation of IGPS. Red labels indicate secondary structure elements that are directly involved in the allosteric regulation. Communities **h2** (cyan) and **f3** (red) in the sideR of IGPS are also depicted.

The present paper is organized as follows: We first summarize the method of CNA and results for ref. 9. Next, the method of EC is introduced and applied to the IGPS systems. Results are discussed and compared with CNA. Correlation matrices are obtained from the same trajectories and following the same protocol as in ref. 9.

CNA

Consider a protein residue network where each node represents the α-carbon of an amino acid in the protein, and each edge represents the dynamical correlation between the two residues (nodes) it connects. The latter can be quantified by using the generalized correlation coefficients, based on the mutual information (MI) between two residues $\mathbf{r}_{MI}[\mathbf{x}_i, \mathbf{x}_j]$ (30):

$$\mathbf{r}_{MI}[\mathbf{x}_i, \mathbf{x}_j] = \left(1 - \exp\left(-\frac{2}{3}\mathbf{I}[\mathbf{x}_i, \mathbf{x}_j]\right)\right)^{1/2}, \qquad [4]$$

where the fluctuation or atomic displacements vectors $\mathbf{x}_k$ are computed from MD simulations. For clarity, we have kept the original notation used in refs. 9 and 30, where a detailed explanation on the calculation of the generalized correlation coefficients can be found.

The MI between the two residues is computed as:

$$\mathbf{I}[\mathbf{x}_i, \mathbf{x}_j] = H[\mathbf{x}_i] + H[\mathbf{x}_j] - H[\mathbf{x}_i, \mathbf{x}_j], \qquad [5]$$

where

$$H[\mathbf{x}_i] = -\int p[\mathbf{x}_i]\ln(p[\mathbf{x}_i])\,d\mathbf{x}_i, \qquad [6]$$

$$H[\mathbf{x}_i,\mathbf{x}_j] = -\iint p([\mathbf{x}_i,\mathbf{x}_j])\ln\left(p([\mathbf{x}_i,\mathbf{x}_j]\right)\,d\mathbf{x}_i\,d\mathbf{x}_j, \qquad [7]$$

are the marginal and joint Shannon entropies, respectively, obtained as ensemble averages over the atomic displacements $(\mathbf{x}_i, \mathbf{x}_j)$, with marginal and joint probability distributions $p[\mathbf{x}_i]$ and $p[\mathbf{x}_i, \mathbf{x}_j]$ computed over thermal fluctuations sampled by MD simulations of the system at equilibrium. The coefficient $\mathbf{r}_{MI}$ ranges from zero for uncorrelated variables to 1 for fully correlated variables.

The protein graph connectivity is then built, excluding direct connections of first neighbors (in amino acid sequence) and according to two cutoffs: Two nodes are considered connected if the distance between their α-carbons is within a distance cutoff (generally 4–6 Å) for a certain percentage of the MD trajectories (percentage cutoff, usually 65–85%). The distances between all of the connected nodes (i, j) in the graph topology define a matrix of elements $\mathbf{w}_{ij}^{(0)}$ obtained from $\mathbf{r}_{MI}[\mathbf{x}_i, \mathbf{x}_j]$, according to:

$$\mathbf{w}_{ij}^{(0)} = -\log[\mathbf{r}_{MI}[\mathbf{x}_i,\mathbf{x}_j]], \qquad [8]$$

setting the $\mathbf{w}_{ij}$ distance to infinity (in practice to extremely large values) when two nodes are not connected, as defined by the connectivity rules. The Floyd–Warshall algorithm (37) is then used to determine the matrix of minimum distance (maximum correlation), $\mathbf{w}_{ij}^{(\mathrm{M})}$, considering direct distances as well as up to N possible intermediate residues mediating indirect communication pathways (where N is the total number of residues in the system). The total number of residues for the IGPS case is $N = 454$.

The edge-betweenness matrix with elements $\mathbf{b}_{ij}$ is defined as the number of shortest paths that include edge (m_{ij}) as one of its communication segments. In other words, the edge-betweenness matrix is an estimation of the information "traffic" passing through the edge connecting residues i and j in the network. The edge-betweenness matrix is then used for partitioning the network into communities according to the Girvan–Newman algorithm, which is based on maximizing the modularity Q measure (38, 39). Details of the computation of the community structure based in the maximum modularity from the generalized correlation matrix can be found in ref. 9.

Fig. 1 shows the two most important communities **h2** (cyan) and **f3** (red) projected onto the residue space of IGPS in the *apo* state as determined in ref. 9. Secondary structural elements of **h2** involve hβ1, hβ2, hβ3, hβ4, hβ11, hα1, hα2′, and Ω-loop. Secondary structural elements of **f3** instead involve fβ1, fβ2, fβ3, hβ7, hβ8, fα1, fα2, fα3, hα4, and Loop1.

We have previously shown that the correlation between communities **h2** and **f3** is enhanced (with larger interbetweenness) after PRFAR binding. Furthermore, it was shown that the explanation for this enhancement relies on the increase in the frequency of an interdomain motion at the dimeric interface (*HisH–HisF*) upon binding of PRFAR. This was described as a low-frequency interdomain breathing motion that allows for fluctuations between two states (open and closed IGPS heterodimer) that are accessible at thermal equilibrium in both the *apo* and PRFAR complexes. Disruption of this breathing mode with drug-like compounds was recently suggested as a method for inhibiting the allosteric mechanism (20).

The recognition of the local interactions that determine variations in the breathing motion (and, thus, in the **h2–f3** intercommunities correlations) has been performed by detailed comparative analysis of chemical interactions along the MD trajectories of *apo* and PRFAR-bound IGPS complexes (9). In particular, it was observed that PRFAR binding affects specific hydrophobic interactions in Loop1 and fβ2 (in *HisF*), altering salt-bridge formations at the surface-exposed fα2, fα3, and hα1 helices (at the *HisF*/*HisH* interface) that, in turn, determine modification of the breathing motion and of the hydrogen-bonding network between the Omega loop and the oxyanion strand nearby the *HisH* active site. Thus, among the secondary structure elements of communities **h2** and **f3**, the following elements have been retained as allosteric pathways: Loop1, fβ2, fα2, fα3, hα1, and Ω-loop (indicated with red labels in Fig. 1). The active allosteric role of some of these residues has been recently proved by single-site mutation experiments (40).

The CNA provides an introspection tool for visualizing the most important transformations induced by the allosteric effector in a coarse-grained fashion, allowing easy detection of effector-driven changes in the overall intercommunities information flows. However, we have shown that to recover direct information on allosteric pathways, a detailed analysis of the MD trajectory is still necessary (9). Therefore, CNA can successfully assist the tedious allosteric pathway detection by indicating major network changes due to the effector binding, but it cannot provide an easy detection and immediate visualization of the sequence of amino acids involved in the allosteric-to-active-site signal propagation. Here, we show that a comparative EC approach, on the other hand, can provide fast detection of allosteric nodes and easy interpretation of the signal pathways "activated" by the effector binding.

EC Analysis

Let us define the adjacency matrix as follows:

$$\mathbf{A}_{ij} = \begin{cases} 0, & \text{if} \quad i = j \\ \mathbf{r}_{MI}[\mathbf{x}_i,\mathbf{x}_j]\exp(-\frac{d_{ij}}{\lambda}) & \text{if} \quad i \neq j. \end{cases} \qquad [9]$$

Just as in the CNA approach, here, each node of the graph corresponds to the α-carbon of an amino acid residue, and the off-diagonal elements of **A** are the weights associated with every edge. Additionally, an exponential damping factor with a length parameter λ has been introduced to Eq. **9**. This parameter can be adjusted to control the locality of the correlations under consideration based on the average distance between residues

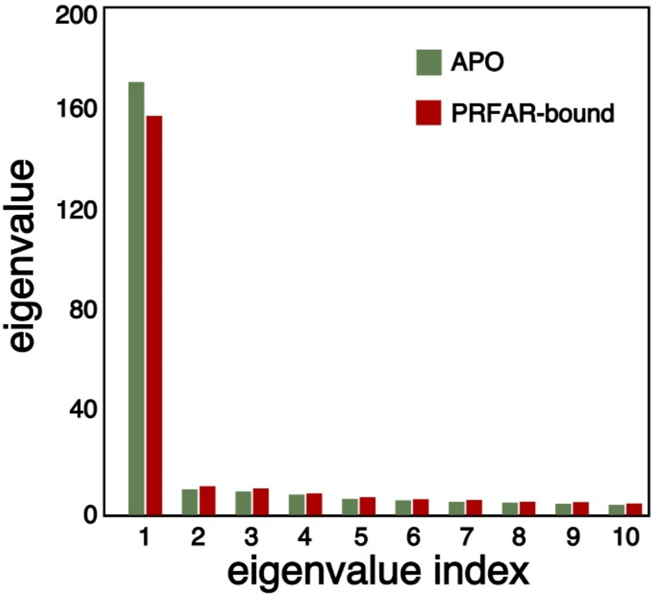

Fig. 2. Largest 10 eigenvalues obtained from the adjacency matrix (as defined by **9** in the limit of $\lambda \to \infty$) for the *apo* (green) and PRFAR-bound (red) IGPS.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

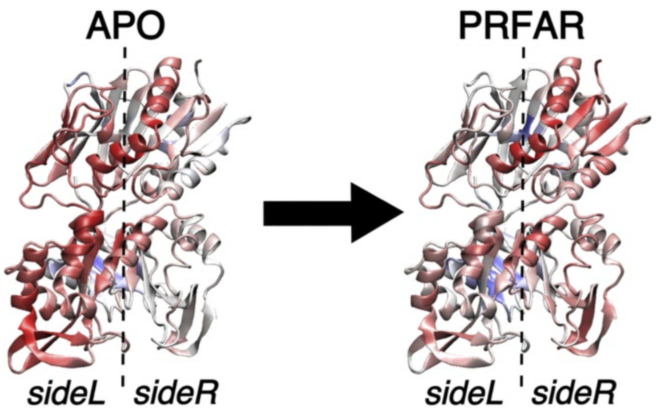

Fig. 3. Computed centrality values for both *apo* and PRFAR-bound IGPS. The color scale goes from blue ($c = 0.0$) to red (maximum values of c).

(d_{ij}). This means that if λ is short enough, the correlation between residues that are far away from one another will be disregarded, and the effect of the locality in the allosteric pathway will be revealed. On the other hand, if λ is set to a very large value, all correlations, including those between residues separated by long distances, will be accounted for (i.e., $\lambda \to \infty$, $\mathbf{A}_{ij} = \mathbf{r}_{MI}[\mathbf{x}_i, \mathbf{x}_j]\ \forall\, i \neq j$). By adopting such damping factor, we obtain a twofold benefit for the EC analysis: (*i*) By setting reasonably small damping values, we could mimic the distance cutoff used in the CNA, and we can then fairly compare EC and CNA results; and (*ii*) comparison of EC values at various damping distances provides direct information on the role of long-range correlations in allosteric pathways. This will be discussed in further detail in *The Locality Factor*.

As mentioned in the introduction, the EC arises from an eigendecomposition of the adjacency matrix, $\mathbf{Ac} = \epsilon\mathbf{c}$, where $\mathbf{c}$ is the vector containing the centralities c_i for each node i and ϵ is the associated eigenvalue. Therefore, there is a set of N solutions to this eigenvalue problem, with N being the number of α-carbon atoms in the protein. However, we will rely here on the assumption that the functional dynamics of the protein can be assigned to the major collective mode of correlation. Consequently, the eigenvectors associated with the remaining eigenvalues will be neglected. The election of this leading eigenvector as the principal component of the correlation pattern can be formally justified, considering that the adjacency matrix $\mathbf{A}$ defined by Eq. **9** has the following mathematical properties: (*i*) $\mathbf{A}_{ij} = \mathbf{A}_{ji}\ \forall\ i, j$; and (*ii*) $0 \leq \mathbf{A}_{ij} \leq 1\ \forall\ i, j$. Hence, uniqueness of the definition of the EC is ensured by the Perron–Frobenius theorem, which states that any symmetric matrix (property *i*) with nonnegative entries (property *ii*) has a unique largest real eigenvalue. Fig. 2 shows that the highest eigenvalue exceeds the others by almost two orders of magnitude, illustrating the Frobenius theorem in practice for *apo* and PRFAR-bound IGPS.

The EC values c_i are computed by diagonalizing A and keeping the eigenvector c corresponding to the maximum eigenvalue. The power method (41) is an alternative to matrix diagonalization that is computationally more efficient and would be more appropriate for large systems. The information encoded on the resulting eigenvector $\mathbf{c}$ reveals the importance of the nodes for the whole connectivity of the network. The nodes with the highest centralities will act as the principal "channels" for momentum transmission across the protein. This strategy has been applied as a means of visualizing dynamical phenomena in other domains of science (42). The eigenvalue ϵ, in turn, gives a measure of the network degree of connectivity. At $\lambda \to \infty$ (no exponential damping), the values of ϵ are 166.8 and 154.0 for *apo* and PRFAR-bound, respectively. This indicates that the system experiences an overall decrease of correlation as a consequence of PRFAR binding as suggested by inspecting the correlation matrix (9). Moreover, our solution NMR spectroscopic measures characterizing the conformational exchange (k_{ex}) for numerous amino acids in the *HisF* domain indicate that nearly every residue increases its flexibility upon PRFAR binding (21). This increase in flexibility is translated into an effective reduction of the intermolecular connectivities and, hence, results fully consistent with the predicted drop in the overall correlation.

The EC values for each node can be easily visualized in the protein structure (Fig. 3), displaying the c_i coefficients for each amino acid with a color scale from blue (zero centrality) to red (maximum centrality). In all of the cases, a renormalization of the centrality values was applied for plotting purposes (SI Appendix). Fig. 3 shows the values of $\mathbf{c}$ for both *apo* and PRFAR-bound IGPS proteins, as computed by setting the damping distance to infinity. Importantly, the subgraph composed by the most important nodes in the network changes dramatically with the effector binding, highlighting the connection between the EC distribution and the momentum transport pathway. As indicated in Fig. 3, the highest EC values shift collectively from *sideL* to *sideR* in IGPS upon PRFAR binding. This variation of the relative EC distribution evidences a change in the correlation pattern that is in agreement with our previous analysis and

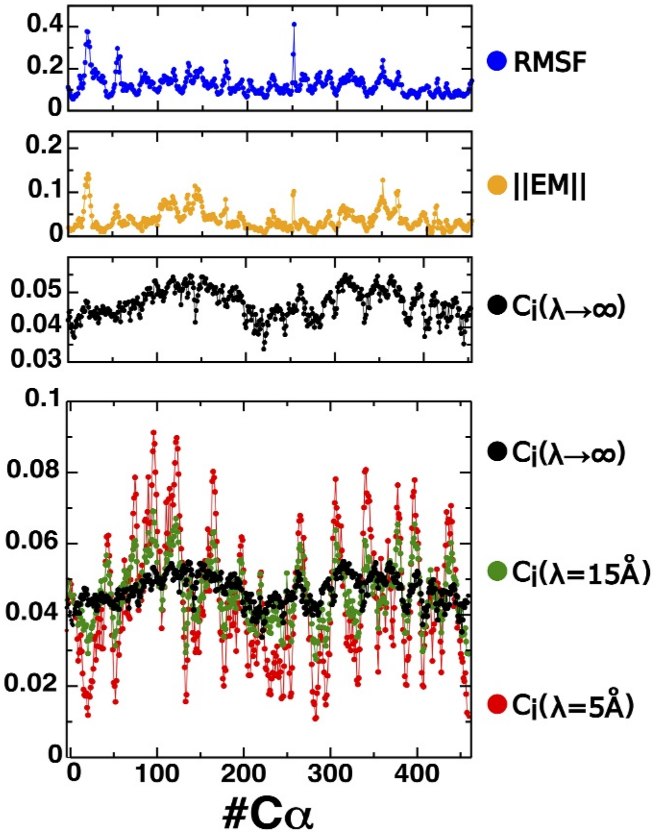

Fig. 4. (*Upper*) Comparison between the Euclidean norm of the elements of the first essential mode associated with each C_α (orange line), the centrality coefficients obtained from the first eigenvector of the adjacency matrix defined in Eq. **9** with $\lambda \to \infty$ (black line), and root-mean-square fluctuation per residue (RMSF; blue line). (*Lower*) Effect of the length parameter in the exponential damping factor of the adjacency matrix defined in Eq. **9**. Values of $\lambda = 5$ Å, 15 Å, and $\lambda \to \infty$ are depicted in red, green, and black, respectively.

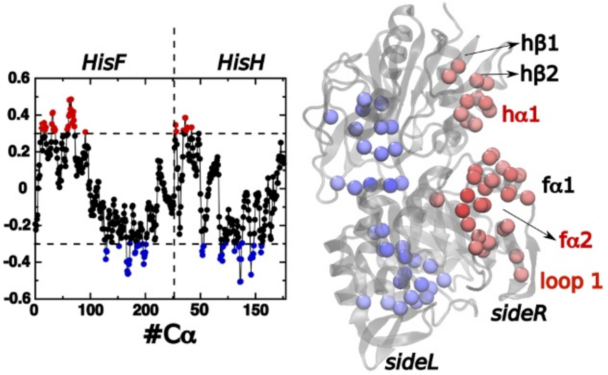

Fig. 5. Centrality differences (PRFAR-bound − APO) for an exponential damping $\lambda = 5$ Å as a function of the residue index (*Left*) and plotted on top of the protein representation (*Right*). Red and blue values are regions that, respectively, gain and lose centrality upon PRFAR binding. The domains with higher PRFAR-induced centrality increase are loop1 (*HisF*: 16–31), fα1 (*HisF*: 31–43), fα2 (*HisF*: 59–72), hβ1 (*HisH*: 1–5), hα1 (*HisH*: 12–25), and hβ2 (*HisH*: 30–35).

consistent with the enhancement in the betweenness of **h2–f3** pair of communities (9).

The methodology introduced above resembles the well-known essential dynamics (ED) scheme in which the global trajectory of a system is analyzed in terms of its major collective modes of fluctuation. (43–46) These modes—usually called essential modes—are obtained by diagonalizing the covariance matrix, defined as

$$\mathbf{C}_{ij} = \langle(\mathbf{x}_i(t) - \langle\mathbf{x}_i(t)\rangle)(\mathbf{x}_j(t) - \langle\mathbf{x}_j(t)\rangle)\rangle. \qquad \textbf{[10]}$$

Normally, despite not being formally guaranteed, it is observed that the protein dynamics is dominated by a few essential modes. Therefore, this scheme also provides a way to obtain eigenvector coefficients that reveal the relevance of each node in the overall behavior of the network. Nevertheless, the measure of relevance can have several meanings; in particular, Fig. 4, *Upper* shows that the nature of the eigenvector coefficients obtained from the first essential mode (the one associated to the highest eigenvalue) is qualitatively different from that of the EC coefficients. There are two main reasons that justify this difference: (*i*) While in the latter case, the generalized MI matrix is only a measure of the dynamical correlation between pairs of nodes, in the former case, the covariance matrix is both a measure of correlation and the amount of fluctuation. (*ii*) On the other hand, the covariance measure fails to account for noncolinear correlations. The first observation is consistent with the fact that the behavior of the essential mode coefficients (orange line, Fig. 4, *Upper*) is quite similar to the root-mean-square fluctuation per residue (blue curve, Fig. 4, *Upper*). Therefore, this analysis illustrates that the ED and the EC extracted from the MI are two complementary methodologies that provide different insight on the system's dynamics. In particular, the technique presented in this work constitutes a powerful alternative to analyze allosterism because it isolates the principal component in terms of the correlation and not in terms of flexibility, as in the case of ED.

Fig. 4, *Lower* shows the effect of the length parameter λ defined in Eq. **9**. In the limit of $\lambda \to \infty$, the off-diagonal elements of the adjacency matrix become equivalent to the generalized correlation function for each pair of nodes. The centrality coefficients obtained in this way exhibit a smooth variation. In contrast, when λ is short enough, only the local components of the correlations survive, and the centrality coefficients reveal the relevance of each residue in terms of its dynamical correlation with neighboring amino acids. In this context, the exponential damping filters out long-range correlations, thus providing a strategy to elucidate the allosteric paths triggered by short-range molecular correlations.

Centrality Variation Triggered by Effector Binding

We have examined the EC differences associated with PRFAR binding ($c_i^{PRFAR} - c_i^{APO}$) for each residue i to analyze changes in the EC distribution caused by binding of the effector PRFAR (Fig. 3). Fig. 5 shows that there is significant redistribution of the EC values upon PRFAR binding. Two protein regions feature increased centralities, namely, residues around fL10–fG80: loop1 (*HisF*: 16–31), fα1 (*HisF*: 31–43), fα2 (*HisF*: 59–72), and hM1-hQ36: hβ1 (*HisH*: 1–5), hα1 (*HisH*: 12–25), and hβ2 in *HisH*. Connections between the loop1 and Ω-loop are hence established after PRFAR is bound to IGPS, as depicted in the centrality-differences analysis presented in Fig. 5.

Previous studies have postulated the existence of two dynamically differentiated sides in IGPS—that is, left and right or *sideL* and *sideR*, respectively (9, 20) (Fig. 5). Detailed inspection of MD trajectories have suggested that the allosteric signal propagates through *sideR*. Importantly, in agreement with that observation, Fig. 5 shows that binding of the effector PRFAR causes an increase in the centrality values of *sideR* amino acids. Moreover, the pattern shown by the centrality distribution allows clear identification of the two sides of IGPS, confirming our previous hypothesis.

The identified residues, including 10–80 (in *HisF*) and 1–36 (in *HisH*) (Fig. 5, highlighted in red), represent promising targets for site-directed mutagenesis studies since they exhibit the highest increase in centrality upon PRFAR binding. Importantly, we identify helix hα1 as one of the domains with higher centrality increase upon PRFAR binding. We anticipate that these findings should stimulate significant interest for site-directed mutagenesis studies or the use of small allosteric drugs targeting helix hα1. Therefore, the reported results provide biological insights that are potentially useful for therapeutic applications that could aim at disrupting IGPS functionality by targeting the hα1 dynamics.

In addition, instead of focusing on the nodes that are important per se, another criteria that can be relevant to guide mutagenesis efforts is to focus on the "neighborhood" of those nodes. This sort of modification may play a more subtle role in altering

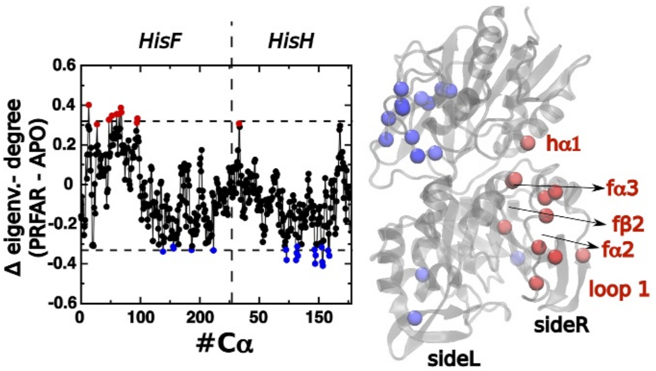

Fig. 6. Difference between EC and DC, c_i', for the PRFAR-binding process (PRFAR-bound − *apo*) for an exponential damping of $\lambda = 5$ Å as a function or the residue index (*Left*) and plotted on top of the protein representation (*Right*). Red and blue values are regions that, respectively, gain and lose correlation with central amino acids upon PRFAR binding. The domains with higher PRFAR-induced c_i' increase are labeled.

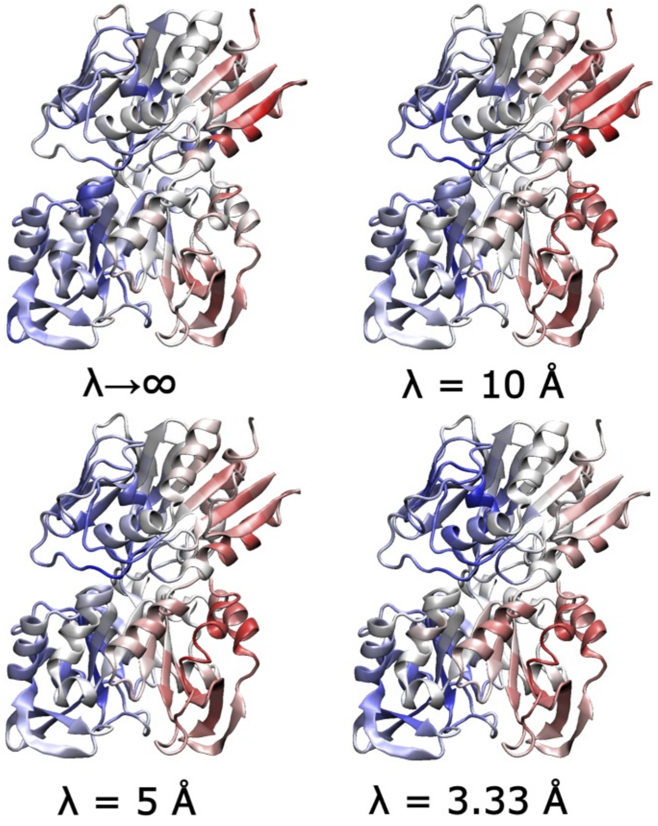

Fig. 7. Centrality differences (PRFAR-bound − APO) for different values of λ. Regions in red and blue correspond to gains and lose of centrality, respectively.

the protein activity, which can be potentially relevant for applications like drug discovery in which the desired effect comes from disrupting the environment of key residues in the protein. Given that the difference between DC (Eq. **1**) and the EC is the fact that the former weights the correlation by the centrality of the neighbors, a strategy to obtain this neighborhood-centrality measure is to subtract the DC coefficients from the original EC values:

$$c_i' = \epsilon^{-1} \sum_{j=1}^{n} A_{ij}\, c_j - \sum_{k=1}^{n} A_{ik}. \qquad \textbf{[11]}$$

Fig. 6 illustrates the c_i' coefficients associated with the transition between the *apo* and PRFAR-bound states [i.e., $c_i' = c_i'(PRFAR) - c_i'(APO)$]. This analysis highlights residues fN14, fV48, fR59, fT61, fL65, fQ67, fV69, fR95, fG96, and hN14 as the ones neighboring the amino acids with a large increase of centrality upon PRFAR binding. With the exception of residues fT61, fL65, and fV69, all of the amino acids pointed out by this analysis coincide with those that have large PRFAR-induced EC variation. Remarkably, single-point mutation of residues fV48 and fN98 (in the vicinity fG96) have a dramatic effect on the PRFAR-induced activation of IGPS catalytic activity (40). On the other hand, the relevance of fV48 as part of the hydrophobic cluster in fβ2 and fE67 and fR95 as part of the surface salt-bridge network at fα2/fα3 have been suggested by tedious inspection of MD trajectories, while here they are rapidly detected by the comparative EC analysis.

Interestingly, the amplitude of the distribution $c' = EC - DC$ increases with the reduction of the locality factor λ (*SI Appendix*, Fig. S2, *Upper*). This result shows that the difference between EC and DC arise mainly from short-range correlations, which is fully consistent with the neighborhood-centrality interpretation (Eq. **11**).

The Locality Factor

Fig. 7 shows the calculated EC coefficients at different values of λ to further analyze the impact of the locality factor in the overall centrality distribution. We note that reducing the damping parameter down to $\lambda = 3.3$ Å does not significantly affect the overall EC differences between *apo* and PRFAR-bound IGPS. The same allosteric pathway for IGPS is revealed whether or not we include the correlations between residues separated by long distances. Moreover, the *sideL*/*sideR* structure is maintained at all λ's. These results imply that the allosteric pathway is dominated by short-range correlations. We note that the locality factor decays with the average distance between residues along the entire MD trajectory. Thus, the locality factor filters long-range correlations and also infrequent short-range correlations (i.e., short-lived local interactions). Since no qualitative changes are observed for a broad range of damping factors (Fig. 7), we conclude that the flow of allosteric communication does not include infrequent contacts or long-range conformational motions. These findings point to a very fundamental aspect of IGPS allosterism with implications for design of therapeutic agents.

The average $C_\alpha - C_\alpha$ distance is ~3.8 Å. Therefore, the correlation matrix becomes almost diagonal (*SI Appendix*) when $\lambda < 4$ Å, and the key EC trend is most likely masked by numerical errors.

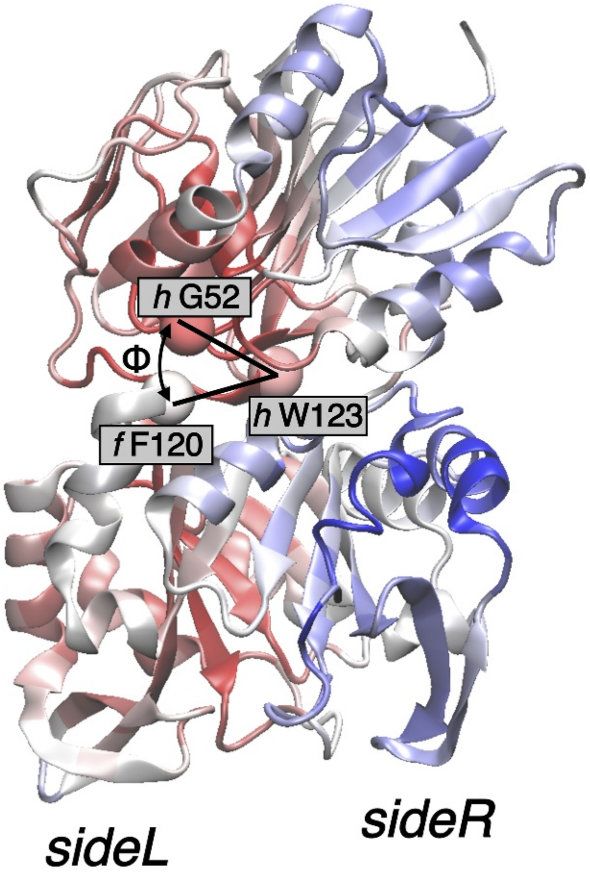

Fig. 8. Variation in the PRFAR-induced centrality coefficients caused by the application of the locality factor ($\lambda = 5$ Å). Red to blue scale characterizes a gain or loss of centrality, respectively, upon the application of the locality factor.

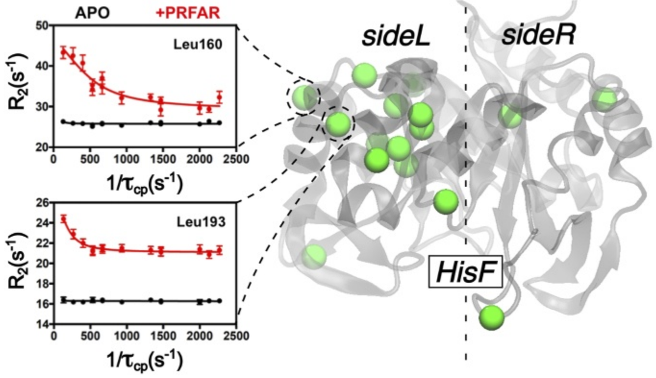

Fig. 9. NMR relaxation dispersion experiments characterizing the PRFAR-induced millisecond motions in the *HisF* subunit of IGPS. *Right* highlights the residues that show the highest variation on their relaxation-dispersion profile upon PRFAR binding. *Left* shows two representative relaxation dispersion curves for residues Leu160 (*Upper*) and Leu193 (*Lower*) in the *apo* and PRFAR-bound states (black and red, respectively).

As discussed above, it is possible to select the correlations whose range is below a certain distance threshold from the overall motion of the system simply by introducing the locality factor λ. On the other hand, it is possible to analyze the nature of long-range contributions, even though short-range components dominate the overall correlation pattern. Fig. 8 shows variations in the EC coefficients due to the long-range component of correlations, computed as follows:

$$\begin{aligned} d_i^{\lambda_0} &= [c_i^{\mathrm{PRFAR}} - c_i^{\mathrm{APO}}]_{\lambda\to\infty} - [c_i^{\mathrm{PRFAR}} - c_i^{\mathrm{APO}}]_{\lambda=\lambda_0} \\ &= [c_i^{\lambda\to\infty} - c_i^{\lambda=\lambda_0}]_{\mathrm{PRFAR}} - [c_i^{\lambda\to\infty} - c_i^{\lambda=\lambda_0}]_{\mathrm{APO}}, \end{aligned} \qquad \textbf{[12]}$$

for λ_0 = 5 Å. Remarkably, the long-range d_i distribution also preserves the qualitative *sideL*/*sideR* structure, although the trends are inverted with respect to the short-range picture, and the largest increase in the long-range centrality coefficients upon PRFAR binding is mainly located on *sideL*. These results are consistent with the presence of an interdomain "breathing" motion, as reported (9, 20) (Fig. 8, dashed black lines forming an angle ϕ). The large structural (long-range) rearrangement associated with this motion increases its frequency upon PRFAR binding almost fourfold (20). Consequently, the highest gain of long-range correlation that occurs mainly in *sideL* can be assigned to this low-frequency motion. In agreement with this, our solution NMR relaxation dispersion experiments show that the PRFAR-induced millisecond motions are primarily located on *sideL* (Fig. 9), which supports the existence of a large motion with maximum amplitude on *sideL*, as determined by the long-range centrality analysis. Furthermore, effectors weaker than PRFAR induce weaker perturbations on *sideL* of *HisF* (21), suggesting that the breathing motion influences the allosteric activation of IGPS. Remarkably, Fig. 9 shows experimental evidence of the suggested breathing motion (47).

The NMR study presented in Fig. 9 also provides an experimental proof for the presence of the *sideL*/*sideR* structure predicted by the EC analysis, in which the two sides of IGPS display clear differences in terms of their dynamical features. Interestingly, the overall difference between *sideR* and *sideL* d_i values is considerably reduced when going from λ= 5 to 10 Å, and for $\lambda = 20$ Å the d_i distribution becomes almost uniform. This indicates that the characteristic correlation distances involved in the breathing mode are within the range of 5–20 Å (*SI Appendix*).

Conclusions

We have introduced a methodology based on the EC of MI to elucidate allosteric pathways at an atomistic level. The method allows for identification of amino acid residues that are critical for allosteric signaling and characterization of the correlation distances that determine allosterism. Furthermore, the analysis of DC allows us to identify key residues neighboring amino acids with a large increase in centrality, consistent with recent site-directed mutagenesis experiments (40).

The EC scheme introduced in this work provides a valuable approach to obtain the main mode of collective correlation responsible for the allosteric signal, beyond the capabilities of standard principal component methods. The analysis is based on the generalized MI which correctly captures noncollinear correlations beyond the well-known limitations of methods based on the Pearson correlation coefficients.

We have applied the EC method to the IGPS enzyme to demonstrate the capabilities of our approach to identify the most important amino acid residues involved in the allosteric mechanism triggered upon effector binding. The EC results show excellent agreement with our solution NMR relaxation experiments, providing experimental evidence of the previously hypothesized interdomain breathing motion (9, 20, 40, 47).

The locality-based centrality analysis shows that the allosteric pathway is established by short-range correlations. Nevertheless, as observed (20), the resulting breathing motion enhances the allosteric signal. Furthermore, the EC method identifies helix hα1 (*HisH*: 12–25) as one of the domains with higher centrality increase upon PRFAR binding. We anticipate that site-directed mutagenesis or the use of allosteric drugs could target helix hα1 to control enzymatic activity. The reported results should motivate a wide range of studies to control IGPS activity by disrupting hα1 dynamics, considering that IGPS is a potential therapeutic target that is found in bacteria as well as in some plants and fungi, but not in mammals.

ACKNOWLEDGMENTS. J.P.L. and V.S.B. were supported by NIH Grant GM106121. V.S.B. also acknowledges supercomputer time from the National Energy Research Scientific Computing Center, XSEDE, and the Yale University Faculty of Arts and Sciences High Performance Computing Center, partially funded by National Science Foundation (NSF) Grant CNS 08–21132. I.R. was supported by École Normale Supérieure de Lyon (ENS-Lyon) "Fonds Recherche, MI-LOURD-FR15" and "Institut Rhônalpin des Systèmes Complexes (IXXI)" and the use of high-performance computing resources of the "Pôle Scientifique de Modélization Numérique" at the ENS-Lyon, France. J.P.L. was supported by NSF Grant MCB 1615415.

1. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R (2013) Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol Ther* 138:333–408.
2. Wagner JR, et al. (2016) Emerging computational methods for the rational discovery of allosteric drugs. *Chem Rev* 116:6370–6390.
3. Goodey NM, Benkovic SJ (2008) Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 4:478–482.
4. Reetz MT, Soni P, Acevedo JP, Sanchis J (2009) Creation of an amino acid network of structurally coupled residues in the directed evolution of a thermostable enzyme. *Angew Chem* 121:8268–8272.
5. Ozbil M, Barman A, Bora RP, Prabhakar R (2012) Computational insights into dynamics of protein aggregation and enzyme–substrate interactions. *J Phys Chem Lett* 3:3460–3469.
6. Hawkins RJ, McLeish TCB (2004) Coarse-grained model of entropic allostery. *Phys Rev Lett* 93:98104–98108.
7. Ming D, Wall ME (2005) Allostery in a coarse-grained model of protein dynamics. *Phys Rev Lett* 95:198103–198107.
8. Palumbo M, Farina L, Colosimo A, Tun K, Dhar PK (2006) Networks everywhere? Some general implications of an emergent metaphor. *Curr Bioinformatics* 1:219–234.

9. Rivalta I, et al. (2012) Allosteric pathways in imidazole glycerol phosphate synthase. *Proc Natl Acad Sci USA* 109:E1428–E 1436.
10. Vanwart AT, Eargle J, Luthey-Schulten Z, Amaro RE (2012) Exploring residue component contributions to dynamical network models of allostery. *J Chem Theor Comput* 8:2949–2961.
11. Ribeiro AAST, Ortiz V (2016) A chemical perspective on allostery. *Chem Rev* 116:6488–6502.
12. Blacklock K, Verkhivker GM (2014) Computational modeling of allosteric regulation in the Hsp90 chaperones: A statistical ensemble analysis of protein structure networks and allosteric communications. *PLoS Comput Biol* 10:1–21.
13. Sun X, Ågren H, Tu Y (2014) Microsecond molecular dynamics simulations provide insight into the allosteric mechanism of the Gs protein uncoupling from the $\beta 2$ adrenergic receptor. *J Phys Chem B* 118:14737–14744.
14. Zhu Y, Ma B, Qi R, Nussinov R, Zhang Q (2016) Temperature-dependent conformational properties of human neuronal calcium sensor-1 protein revealed by all-atom simulations. *J Phys Chem B* 120:3551–3559.
15. Appadurai R, Senapati S (2016) Dynamical network of HIV-1 protease mutants reveals the mechanism of drug resistance and unhindered activity. *Biochemistry* 55:1529–1540.
16. Xu L, et al. (2015) Recognition mechanism between lac repressor and DNA with correlation network analysis. *J Phys Chem B* 119:2844–2856.
17. VanWart AT, Eargle J, Luthey-Schulten Z, Amaro RE (2012) Exploring residue component contributions to dynamical network models of allostery. *J Chem Theor Comput* 8:2949–2961.
18. Palermo G, et al. (2017) Protospacer adjacent motif-induced allostery activates CRISPR-Cas9. *J Am Chem Soc* 139:16028–16031.
19. Guo J, Zhou HX (2016) Protein allostery and conformational dynamics. *Chem Rev* 116:6503–6515.
20. Rivalta I, et al. (2016) Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein–protein interface. *Biochemistry* 55:6484–6494.
21. Lisi G, et al. (2016) Dissecting dynamic allosteric pathways using chemically related small-molecule activators. *Structure* 24:1155–1166.
22. Palermo G, et al. (2018) Key role of the rec lobe during CRISPR–Cas9 activation by sensing, regulating, and locking the catalytic HNH domain. *Q Rev Biophys* 51:e9.
23. Li S, et al. (2014) The mechanism of allosteric inhibition of protein tyrosine phosphatase 1B. *PLoS ONE* 9:1–10.
24. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci USA* 106:6620–6625.
25. Ricci CG, Silveira RL, Rivalta I, Batista VS, Skaf MS (2016) Allosteric pathways in the ppar-rxr nuclear receptor complex. *Sci Rep* 6:19940.
26. Papaleo E, Lindorff-Larsen K, De Gioia L (2012) Paths of long-range communication in the e2 enzymes of family 3: A molecular dynamics investigation. *Phys Chem Chem Phys* 14:12515–12525.
27. David-Eden H, Mandel-Gufreund Y (2008) Revealing unique properties of the ribosome using a network based analysis. *Nucleic Acid Res* 36:4641–4652.
28. Jiang X, Chen C, Xiao Y (2010) Improvements of network approach for analysis of the folding free-energy surface of peptides and proteins. *J Comput Chem* 31:2502–2509.
29. Szilagyi A, Nussinov R, Csermely P (2013) Allo-network drugs: Extension of the allosteric drug concept to protein-protein interaction and signaling networks. *Curr Top Med Chem* 13:64–77.
30. Lange OF, Grubmüller H (2006) Generalized correlation for biomolecular dynamics. *Proteins: Struct Funct Bioinformatics* 62:1053–1061.
31. Lange OF, Grubmüller H (2008) Full correlation analysis of conformational protein dynamics. *Proteins: Struct Funct Bioinformatics* 70:1294–1312.
32. Savoie BM, et al. (2014) Mesoscale molecular network formation in amorphous organic materials. *Proc Natl Acad Sci USA* 111:10055–10060.
33. Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D (2016) Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proc Natl Acad Sci USA* 113:4735–4740.
34. Ruhnau B (2000) Eigenvector-centrality - A node-centrality?. *Soc Networks* 22:357–365.
35. Chaudhuri BN, et al. (2001) Crystal structure of imidazole glycerol phosphate synthase. *Structure* 9:987–997.
36. Myers RS, Jensen JR, Deras IL, Smith JL, Davisson VJ (2003) Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase. *Biochemistry* 42:7013–7022.
37. Floyd RW (1962) Algorithm 97: Shortest path. *Commun ACM* 5:345.
38. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826.
39. Newman MEJ (2006) Modularity and community structure in networks. *Proc Natl Acad Sci USA* 103:8577–8582.
40. Lisi GP, East KW, Batista VS, Loria JP (2017) Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity. *Proc Natl Acad Sci USA* 114:E3414–E3423.
41. Watkins DS (2010) *Fundamentals of Matrix Computations* (John Wiley & Sons, New York), 3rd Ed.
42. Jimenez-Martinez J, Negre CFA (2017) Eigenvector centrality for geometric and topological characterization of porous media. *Phys Rev E* 96:013310.
43. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins: Struct Funct Bioinformatics* 17:412–425.
44. Hayward S, de Groot BL (2008) *Normal Modes and Essential Dynamics* (Humana Press, Totowa, NJ).
45. Meyer T, et al. (2006) Essential dynamics: A tool for efficient trajectory compression and management. *J Chem Theor Comput* 2:251–258.
46. Morzan UN, Capece L, Marti MA, Estrin DA (2013) Quaternary structure effects on the hexacoordination equilibrium in rice hemoglobin rHb1: Insights from molecular dynamics simulations. *Proteins: Struct Funct Bioinformatics* 81:863–873.
47. Amaro RE, Sethi A, Myers RS, Davisson VJ, Luthey-Schulten ZA (2007) A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry* 46:2156–2173.