

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Eigenvector centrality for characterization of protein allosteric pathways

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Negre, C.F., Morzan, U.N., Hendrickson, H.P., Pal, R., Lisi, G.P., Patrick Loria, J., et al. (2018). Eigenvector centrality for characterization of protein allosteric pathways. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 115(52), E12201-E12208 [10.1073/pnas.1810452115].

Availability:

This version is available at: <https://hdl.handle.net/11585/669278> since: 2019-02-20

Published:

DOI: <http://doi.org/10.1073/pnas.1810452115>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Christian F. A. Negre, Uriel N. Morzan, Heidi P. Hendrickson, Rhitankar Pal, George P. Lisi, J. Patrick Loria, Ivan Rivalta, Junming Ho, Victor S. Batista, Proceedings of the National Academy of Sciences Dec 2018, 115 (52) E12201-E12208.

The final published version is available online at :
<https://doi.org/10.1073/pnas.1810452115>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

LA-UR-17-22957 (Accepted Manuscript)

Eigenvector centrality for characterization of protein allosteric pathways

Negre, Christian Francisco Andres
Hendrickson, Heidi
Pal, Rhitankar
Rivalta, Ivan
Ho, Junming
Batista, Victor

Provided by the author(s) and the Los Alamos National Laboratory (2019-05-09).

To be published in: Proceedings of the National Academy of Sciences

DOI to publisher's version: 10.1073/pnas.1810452115

Permalink to record: <http://permalink.lanl.gov/object/view?what=info:lanl-repo/lareport/LA-UR-17-22957>

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Eigenvector Centrality Distribution for Characterization of Protein Allosteric Pathways

Christian F. A. Negre,^{1,*} Uriel N. Morzan,^{2,†} Heidi P. Hendrickson,² Rhitankar Pal,²
George P. Lisi,² J. Patrick Loria,^{2,3} Ivan Rivalta,^{4,‡} Junming Ho,⁵ and Victor S. Batista^{2,§}

¹*Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545*

²*Department of Chemistry, Yale University, P.O. Box 208107,
New Haven, CT 06520-8107, and Energy Sciences Institute,
Yale University, P.O. Box 27394, West Haven, CT 06516-7394*

³*Department of Molecular Biophysics and Biochemistry,
Yale University, New Haven, CT, United States.*

⁴*École Normale Supérieure de Lyon, CNRS, Université Lyon 1,
Laboratoire de Chimie UMR 5182, 46, Allée d'Italie, 69364 Lyon Cedex 07, France.*

⁵*Agency for Science, Technology and Research, Institute of High Performance Computing,
1 Fusionopolis Way #16-16 Connexis North, Singapore 138632.*

Determining the principal energy pathways for allosteric communication in biomolecules, that occur as a result of thermal motion, remains challenging due to the intrinsic complexity of the systems involved. Graph theory provides an approach for making sense of such complexity, where allosteric proteins can be represented as networks of amino acids. In this work, we establish the eigenvector centrality metric in terms of the mutual information, as a mean of elucidating the allosteric mechanism that regulates the enzymatic activity of proteins. Moreover, we propose a strategy to characterize the range of the physical interactions that underlie the allosteric process. In particular, the well known enzyme, imidazol glycerol phosphate synthase (IGPS), is utilized to test the proposed methodology. The eigenvector centrality measurement successfully describes the allosteric pathways of IGPS, and allows to pinpoint key amino acids in terms of their relevance in the momentum transfer process. The resulting insight can be utilized for refining the control of IGPS activity, widening the scope for its engineering. Furthermore, we propose a new centrality metric quantifying the relevance of the surroundings of each residue. In addition, the proposed technique is validated against experimental solution NMR measurements yielding fully consistent results. Overall, the methodologies proposed in the present work constitute a powerful and cost effective strategy to gain insight on the allosteric mechanism of proteins.

Allostery is a ubiquitous process of physico-chemical regulation in biological macromolecules such as enzymes. The fundamental step in the allosteric regulation is the binding of a ligand at a particular enzymatic site affecting the activity at a different and often very distant position of the protein. While allosteric processes have long been of interest, especially due to their relevance in developing potent and selective therapeutics, the mechanism for energy transfer between allosteric sites remains poorly understood. Thus, establishing a molecular level understanding of communication pathways between the physically distant enzymatic sites is crucial for the design of innovative drug therapies[1, 2] and protein engineering[3–5].

Recently, there have been significant efforts toward the development of computational tools to support, interpret and/or predict experimental evidences for elucidation of allosteric pathways in proteins [2, 6–12]. Network analysis has been extensively used in this context, by incorporating concepts and methodologies from graph theory into the realm of molecular dynamics simulations [13–19] 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 and to detect possible allosteric

pathways [20–26].

In these network theory-based approaches, a protein is represented as a network consisting of a set of nodes, n connected by edges, m . Usually, each amino acid is associated to a node (typically positioned on the alpha carbon or the center of mass of the residue side chain). Depending on the physical property of interest, there are multiple quantities that can characterize the edges (i.e. the connections between nodes), such as the magnitude of the dynamical correlations [9, 27, 28], the energetic coupling [29], or the spatial distance between residues [30]. Given a network of N nodes, its graph can be represented with an $N \times N$ adjacency matrix \mathbf{A} whose elements \mathbf{A}_{ij} are related with the strength of the physical interaction under consideration.

One of the corner stones of network analysis is the concept of centrality, i.e., the relative importance of a node or clusters of nodes. Measures of centrality are crucial to identify the most influential nodes in a network. The importance is usually quantified by a real-valued function, related to a type of flow or transfer across the network (e.g., the amount of momenta transported by a given atom in a protein). There are many measures of centrality characterizing slightly different aspects of the network. Probably the simplest of all is the degree, k_i , of

each node, i , which is defined by the number and strength of the connections attached to it,

$$k_i = \sum_{j=1}^n \mathbf{A}_{ij}. \quad (1)$$

The degree centrality (DC), provides a measure of the relative connectivity of each node within a network. A node that is well connected is expected to have a large ‘‘influence’’ on the graph. While the DC can provide useful information, it is not a true ‘‘node-centrality’’ as defined by Ruhnau,[31] 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,

$$b_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \quad (2)$$

where n_{st}^i is 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, and hence, the removal of such nodes may disrupt the communication in the network. However, communication do not always take the shortest path, and hence, the BC may provide a misleading interpretation of the real 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 sum of the centralities of all nodes that are connected to it by an edge, \mathbf{A}_{ij} ,

$$c_i = \epsilon^{-1} \sum_{j=1}^n \mathbf{A}_{ij} c_j, \quad (3)$$

therefore, \mathbf{c} is the eigenvector associated to the eigenvalue ϵ of \mathbf{A} . The EC is, hence, a measure of how well connected a node is to other well connected nodes in the network. Noteworthy, the EC serves as a measure of the connectivity against a fixed scale when normalized, and so it can be used to reliably compare different networks.[31] For example, the normalization becomes essential when analyzing differences between graphs, e.g., 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.[32] Structurally, IGPS is a tightly associated heterodimer (see 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* unit and reacts with the effector PRFAR to form imidazole glycerol phosphate (IGP). While Gln binding is unaffected by the presence of PRFAR, the hydrolysis of Gln is accelerated 5000-fold upon PRFAR binding through a mechanism that, for many years, has remained elusive [33]. IGPS is thus a V-type enzyme and a model system to study non-cooperative allostery involving conformational changes.

In a recent study [9], we have carried out a BC-based community network analysis 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 as compared to the CNA based on optimal modularity (the connection between CNA and the EC is analyzed in detail in the SI). The results presented here are both complementary and fully consistent with our previous findings. Additionally, at variance with our previous CNA approach, the strategy proposed in this work allows to capture the long range contribution to the correlation pattern evidencing fundamental aspects of the allosteric behavior of IGPS. Therefore, the methodology presented here represents an ideal technique for the identification of mutation targets to inhibit or enhance the IGPS catalytic activity, 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.

The present paper is organized as follows: We first summarize the method of CNA and results for reference [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 reference [9] and [34].

COMMUNITY NETWORK ANALYSIS

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 using the generalized correlation coef-

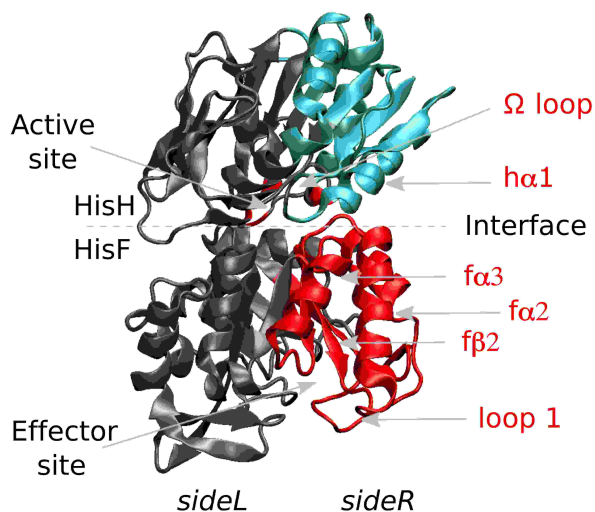


FIG. 1. Molecular representation of IGPS. We have added labels for some key molecular features that are directly involved in the allosteric regulation. Communities **h2** (cyan) and **f3** (red) are also depicted.

ficients, based on the mutual information (MI) between two residues $\mathbf{r}_{MI}[\mathbf{x}_i, \mathbf{x}_j]$ [27]:

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

where the fluctuation or atomic displacements vectors \mathbf{x}_k are computed from a molecular dynamics (MD) simulations. For clarity, we have kept the original notation used in references [9, 27, 34], where a detailed explanation on the calculation of the generalized correlation coefficients can be found.

The mutual information between the two residues is computed as:

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

where

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

$$H[\mathbf{x}_i, \mathbf{x}_j] = - \int \int p([\mathbf{x}_i, \mathbf{x}_j]) \ln(p([\mathbf{x}_i, \mathbf{x}_j])) d\mathbf{x}_i d\mathbf{x}_j, \quad (7)$$

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 molecular dynamics 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 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:

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

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 [35] is then used to determine the matrix of minimum distance (maximum correlation), $\mathbf{w}_{ij}^{(M)}$ 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 [36, 37]. Details of the computation of the communities structure based in the maximum modularity from the generalized correlation matrix can be found in references [9, 34].

Figure 1 shows the two most important communities **h2** (cyan) and **f3** (red) projected into the residue space of the IGPS in the *apo* state as determined in [9]. Secondary structural element of **h2** involves $h\beta1$, $h\beta2$, $h\beta3$, $h\beta4$, $h\beta11$, $h\alpha1$, $h\alpha2'$ and Ω -loop. Secondary structural element of **f3** instead involves $f\beta1$, $f\beta2$, $f\beta3$, $h\beta7$, $h\beta8$, $f\alpha1$, $f\alpha2$, $f\alpha3$, $h\alpha4$ and Loop1.

We have previously showed that the correlation between communities **h2** and **f3** is enhanced (with larger inter-betweenness) 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 the binding of PRFAR. This was described as a low-frequency inter-domain 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 [38].

The recognition of the local interactions that deter-

mine variations in the breathing motion (and, thus, in the **h2-f3** inter-communities 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 f β 2 (in *HisF*), altering salt-bridge formations at the surface exposed fa2, fa3 and ha1 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, f β 2, fa2, fa3, ha1 and Ω -loop (indicated with red labels in Figure 1). The active allosteric role of some of these residues has been recently proved by single-site mutation experiments [39].

The CNA provides an introspection tool for visualizing the most important transformation induced by the allosteric effector in a coarse-grained fashion, allowing easy detection of effector-driven changes in the overall inter-communities information flows. However, we have showed 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 pathways 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.

EIGENVECTOR CENTRALITY ANALYSIS

Lets define the adjacency matrix as follows:

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

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 \mathbf{A} are the weights associated to every edge. Additionally, an exponential damping factor with a length parameter λ has been introduced to expression 9. This parameter can be adjusted to control the locality of the correlations under consideration based on the average distance between residues (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. $\lambda \rightarrow \infty$, $\mathbf{A}_{ij} = \mathbf{r}_{MI}[\mathbf{x}_i, \mathbf{x}_j] \forall i \neq j$). By adopting such damping factor, we obtain a two-fold benefit for the EC analysis: i) by setting reasonably small damping values we could mimic the distance cutoff employed in the CNA and we can then fairly compare EC and CNA results; 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 last section.

As mentioned in the introduction, the eigenvector centrality (EC) measurement arises from an eigendecomposition of the adjacency matrix, $\mathbf{A}\mathbf{c} = \epsilon\mathbf{c}$, 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, being N 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 to 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 equation 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 eigenvector centrality is ensured by the Perron-Frobenius theorem which states that any symmetric matrix (property i) with non-negative entries (property ii) has a unique largest real eigenvalue (see SI). To illustrate the practical consequence of this theorem in the case of *apo* and PRFAR bound IGPS, Figure 2 shows that there is almost two orders of magnitude separating the highest eigenvalues from the remaining ones.

Based on this definition, the EC values c_i can be computed by diagonalizing matrix \mathbf{A} and keeping the eigenvector \mathbf{c} that corresponds to the maximum eigenvalue ϵ . The power method [40] is an alternative to matrix diagonalization which is computationally more efficient and would be more appropriated 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 [41].

As the set of eigenvectors of \mathbf{A} is orthonormal, the sum of all the squared centralities is one ($\sum_i c_i^2 = 1$). The latter plus the fact that the centralities are positive suggests that the squared centralities c_i^2 could be interpreted as the probability for a signal to pass through node i [41]. The eigenvalue ϵ , in turn, gives a measure of the net-

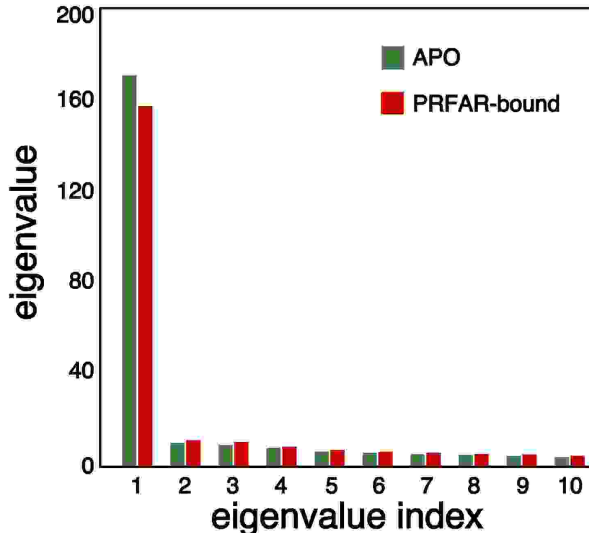


FIG. 2. Largest 10 eigenvalues obtained from the adjacency matrix defined by 9 (in the limit of $\lambda \rightarrow \infty$) for the *apo* (green) and PRFAR bound (red) IGPS. **(CHRIS: isn't this part of the SI?)**

work degree of connectivity. At $\lambda \rightarrow \infty$ (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 previously suggested by inspecting the correlation matrix [9]. Moreover, our solution NMR spectroscopic measures characterizing the conformational exchange (k_{ex}) for numerous amino acids in *HisF* domain, indicate that nearly every residue increases its flexibility upon PRFAR binding [42]. 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 (Figure 3), displaying the c_i coefficients for each amino acid with a color scale from white (zero centrality) to red (maximum centrality). In all the cases, a renormalization of the centrality values was applied for plotting purposes (See SI). Figure 3 shows the values of \mathbf{c} for both *apo* and PRFAR-bound IGPS proteins, as computed by setting the damping distance to infinity. Noteworthy, 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 Figure 3, the highest EC values shift collectively from *sideL* to *sideR* of the IGPS 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 con-

sistent with the enhancement in the betweenness of **h2-f3** pair of communities [9].

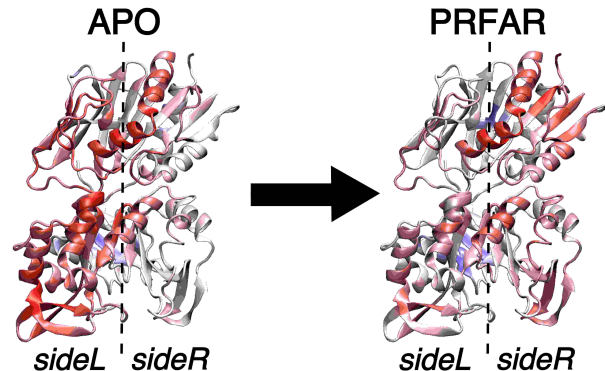


FIG. 3. Computed centrality values for both APO and PRFAR-bound IGPS. The color scale goes from blue ($c = 0.0$) to red (max values of c).

The methodology introduced above somehow resembles the well known essential dynamics (ED) scheme in which the global trajectory of a system 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

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

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 the upper panel of Figure 4 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 mutual information 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 non-colinear correlations. The first observation is consistent with the fact that the behavior of the essential mode coefficients (orange line, upper panel) is quite similar to the root mean square fluctuation per residue (blue curve, upper panel). Therefore, this analysis illustrates that the ED and the EC extracted from the mutual information are two complementary methodologies that provide a different insight on the systems dynamics. In particular the technique presented in this work constitutes a powerful alternative to analyze allostery because it isolates the

principal component in terms of the correlation and not in terms of flexibility as in the case of essential dynamics.

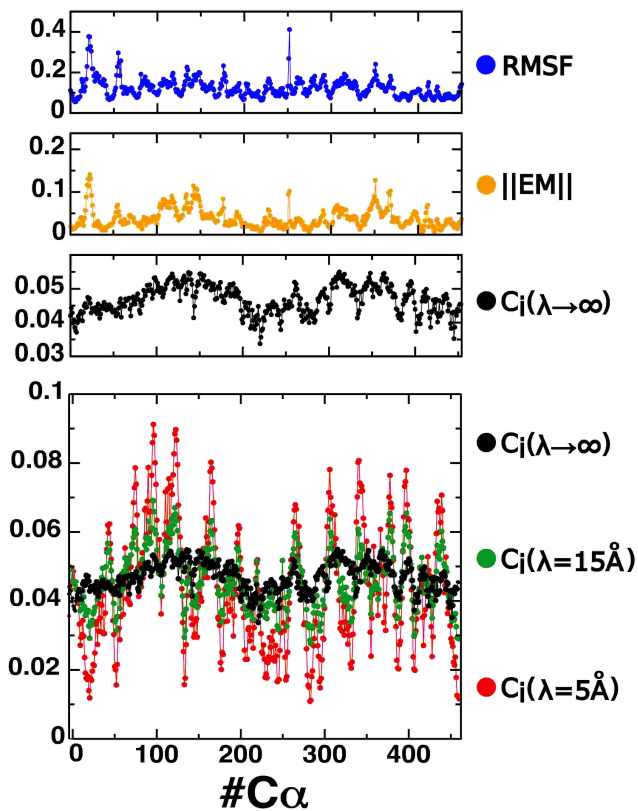


FIG. 4. (Upper Panel) Comparison between the Euclidean norm of the elements of the first essential mode associated with each C α (orange line), the centrality coefficients obtained from the first eigenvector of the adjacency matrix defined in equation 9 with $\lambda \rightarrow \infty$ (black line), and root mean square fluctuation per residue (RMSF) (blue line). (Lower Panel) Effect of the length parameter in the exponential damping factor of the adjacency matrix defined in equation 9. Values of $\lambda = 5 \text{ \AA}$, 15 \AA and $\lambda \rightarrow \infty$ are depicted in red, green and black respectively.

The lower panel of Figure 4 shows the effect of the length parameter λ defined in expression 9. In the limit of $\lambda \rightarrow \infty$ 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 aminoacids. In this context, the exponential damping appears as a strategy to elucidate the *correlation paths* triggered by short range molecular interactions, thus providing a physically relevant description of the momentum transfer within the protein residue network.

CENTRALITY VARIATION TRIGGERED BY EFFECTOR BINDING

In order to highlight the changes in the EC distribution caused by the binding of the effector PRFAR (see Figure 3), we have examined the EC differences associated to PRFAR binding ($c_i^{PRFAR} - c_i^{APO}$) for each residue i . Figure 5 shows that there is significant redistribution of the EC values upon PRFAR binding. Two protein regions feature increased centralities, namely residues around 5-100 (in HisF) and around 254-330 (1-46 in HisH), involving the $\alpha 1$, $\alpha 2$, loop1 and $h\alpha 1$ fragments. Connections between the loop1 and Ω -loop are hence established after PRFAR is bound to IGPS as suggested in Ref. [9] and as clearly depicted in the centrality differences analysis presented in Figure 5.

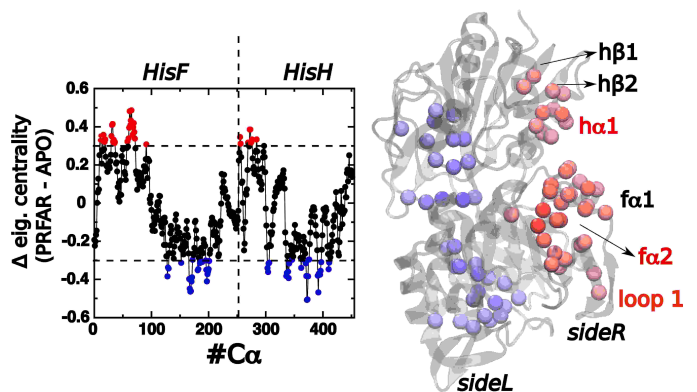


FIG. 5. Centrality differences (PRFAR-bound - APO) for an exponential damping $\lambda = 5 \text{ \AA}$ as a function of the residue index (a) and plotted on top of the protein representation (b). Red and blue values are regions that respectively gain and lose centrality upon PRFAR binding.

Previous studies have suggested the existence of two dynamically differentiated sides in IGPS, i.e. left and right or *sideL* and *sideR* respectively [9, 38] (Figure 5). Detailed inspection of MD trajectories have suggested that the allosteric signal propagates through *sideR*. Noteworthy, in agreement with that observation, Figure 5 shows that the 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 to clearly identify the two sides of IGPS, confirming our previous hypothesis.

Importantly, the residues identified by this analysis are perfect targets for mutations that may have a major impact on IGPS catalytic behavior. In particular, helix $h\alpha 1$ appears as a specially promising and unexplored fragment for site directed mutagenesis experiments oriented to refine the control of IGPS activity.

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 “neighbor-

hood” of those nodes. This sort of modification may play a more subtle role on altering the proteins 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. A strategy to obtain this *neighborhood centrality* measure is to subtract the degree centrality (DC) coefficients from the original EC values:

$$c'_i = \epsilon^{-1} \sum_{j=1}^n A_{ij} c_j - \sum_{k=1}^n A_{ik}. \quad (11)$$

Figure 6 illustrates the measurement of the c'_i coefficients associated to the transition between the APO and PRFAR bound states (i.e. $c'_i = c'_i(\text{PRFAR}) - c'_i(\text{APO})$). This analysis highlights residues fN14, fV48, fR59, fT61, fL65, fQ67, fV69, fR95, fG96 and hN14 as the ones neighboring the aminoacids with a large increase of centrality upon PRFAR binding. With the exception of residues fT61, fL65 and fV69, all the aminoacids pointed out by this measurement coincide with the ones that have the larger PRFAR induced EC variation. Remarkably, single-point mutation on residue fV48 and fN98 (which is in the vicinity fG96) have shown to have a dramatic effect on the PRFAR-induced activation of IGPS catalytic activity [39]. On the other hand, the relevance of fV48 as part of the hydrophobic cluster in f β 2 and fE67 and fR95 as part of the surface salt-bridge network at f α 2/f α 3 has been indicated by detailed MD trajectories inspections while here it is rapidly detected by the comparative EC analysis.

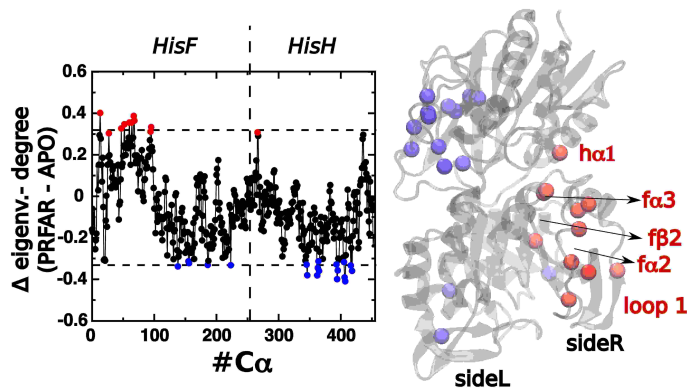


FIG. 6. Difference between EC and DC, c'_i , for the PRFAR binding process (PRFAR-bound - APO) for an exponential damping of $\lambda = 5 \text{ \AA}$ as a function of the residue index (a) and plotted on top of the protein representation (b). Red and blue values are regions that respectively gain and lose of centrality with central aminoacids upon PRFAR binding.

THE LOCALITY FACTOR

In order to further analyze the impact of the locality factor in the overall centrality distribution, Figure 7 shows the calculated EC coefficients at different values of λ . Importantly, adjusting the damping parameter down to $\lambda = 3.3 \text{ \AA}$ does not seem to have a significant effect on the overall trend of the 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. This implies that short range correlations dominate the protein dynamics, and hence the residue-to-residue effect is the main mechanism that underlies the momentum transmission in IGPS. Another important point to note is the fact that the disruptions of the centrality values disappear upon the application of the locality factor recovering the smoothness of a residue-to-residue short range transmitted signal.

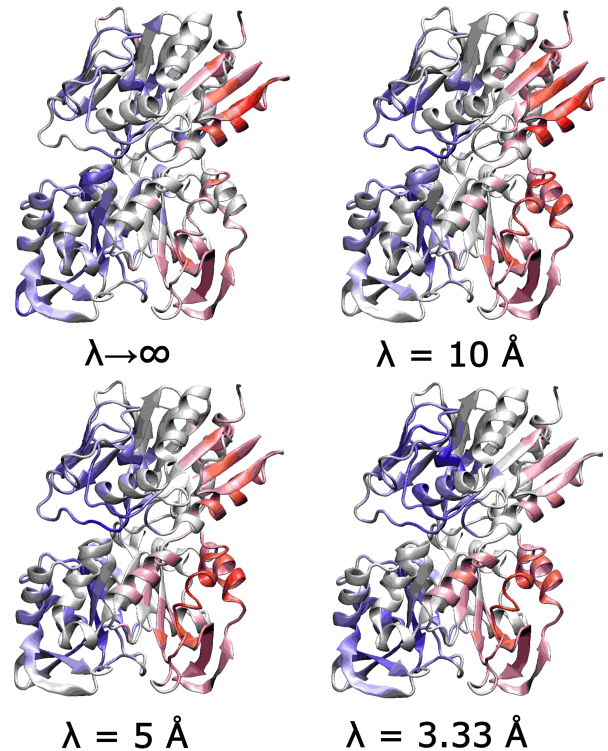


FIG. 7. Centrality differences (PRFAR-bound - APO) for different values of λ . Regions in red and blue correspond to gains and loss of centrality respectively.

The average C_α - C_α distance is around 3.8 \AA , therefore when the value of $\lambda < 4 \text{ \AA}$, the correlation matrix becomes almost diagonal (see SI), and the key EC trend is most likely masked by numerical errors.

As discussed above, by introducing the locality factor λ it is possible to select from the overall motion of the

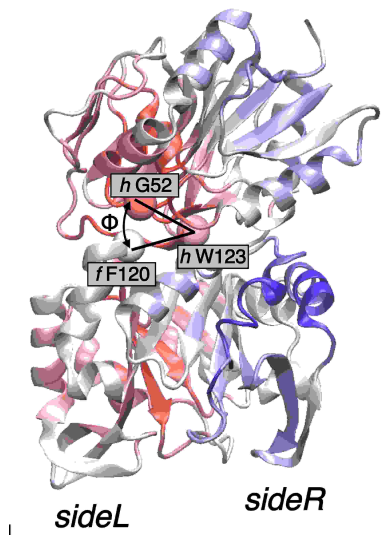


FIG. 8. Variation in the PRFAR-induced centrality coefficients caused by the application of the locality factor ($\lambda = 5$ Å). Red to blue scale characterizes a gain or loss of centrality respectively upon the application of the locality factor.

system the correlations arising exclusively from physical interactions whose range are below certain distance threshold. On the other hand, despite having shown that the resulting short range component is the one that dominates the overall correlation pattern, it is possible to analyze the nature of the long range contribution. Figure 8 introduces a measurement of the long range component of the PRFAR induced EC coefficients computed as:

$$d_i^{\lambda_0} = [c_i^{\text{PRFAR}} - c_i^{\text{APO}}]_{\lambda \rightarrow \infty} - [c_i^{\text{PRFAR}} - c_i^{\text{APO}}]_{\lambda = \lambda_0} \\ = [c_i^{\lambda \rightarrow \infty} - c_i^{\lambda = \lambda_0}]_{\text{PRFAR}} - [c_i^{\lambda \rightarrow \infty} - c_i^{\lambda = \lambda_0}]_{\text{APO}}, \quad (12)$$

for $\lambda_0 = 5, 10$ and 20 Å (panels A, B and C respectively) (**CHRIS: Panels are missing?**). Remarkably, the long range d_i distribution also preserves qualitatively the *sideL*/*sideR* structure, but in this case 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*. This is consistent with the presence of an interdomain “breathing” motion, as previously reported [9, 38] (Figure 8.A, dashed black lines forming an angle ϕ). The large structural (long range) rearrangement associated to this motion increases its frequency upon PRFAR binding almost four times [38]. 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* (Figure 9), which supports the existence of a large motion with maximum amplitude on

sideL as determined by the long range centrality analysis. Furthermore, *sideL* of subunit *HisF* appears more static with weaker effectors than PRFAR [42], suggesting that this breathing motion might be determining the allosteric activation of IGPS in some extent. But more generally, the NMR study presented in Figure 9 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.

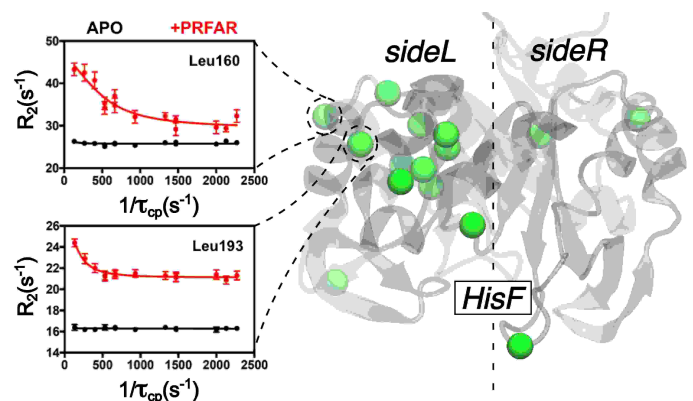


FIG. 9. NMR relaxation dispersion experiments characterizing the PRFAR-induced millisecond motions in *HisF* subunit of IGPS. The right panel highlights the residues that show the highest variation on their relaxation-dispersion profile upon IGPS binding. The left panel shows two representative relaxation dispersion curves for residues Leu160 (upper panel) and Leu193 (lower panel) in the APO and PRFAR bound states (black and red respectively).

Interestingly, the overall difference between *sideR* and *sideL* d_i values is considerably reduced when going from $\lambda = 5$ 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 to 20 Å (see SI).

CONCLUSIONS

In the present work we have introduced a strategy based on the eigenvector centrality (EC) and mutual information metrics as a way of elucidating the allosteric pathways at an atomistic level and disentangle the local and non-local components of the characteristic distances that determine the allosteric mechanism. Furthermore, we have introduced a new perspective to measure centrality in terms of the environment relevance, allowing to interpret recent site directed mutagenesis experiments [39].

As opposed to other principal component analysis of widespread use in the literature, the EC scheme presented in this work provides a way to obtain the major

collective correlation mode, independent from the magnitude of the fluctuations. As a consequence, this methodology constitutes a powerful strategy to quantify the relevance of each amino acid in the overall pathways of momentum transfer. In addition, the correlation measure is based on the generalized mutual information, which correctly captures the non-collinear correlation, overcoming the well known limitation of the Pearson correlation coefficients.

We have used the IGPS protein as a test case to show that our approach successfully predicts the most important residues involved in the allosteric mechanism upon effector binding. The identified amino acids are localized around *sideR* of the *HisH-HisF* interface connecting the effector and the active sites. These residues belong to the same allosteric pathways detected in our previous community network analysis [9] further corroborated by recent experimental evidences [39]. The outcome indicates how the comparative EC analysis here developed can predict allosteric pathways and estimate the role of long-range correlations in allostery through a robust and cost effective protocol.

ACKNOWLEDGEMENT

We acknowledge support by the Office of Basic Energy Sciences of the U.S. Department of Energy (DE-FG02-07ER15909) for supercomputer time from NERSC, XSEDE and the Yale University Faculty of Arts and Sciences High Performance Computing Center partially funded by the National Science Foundation grant CNS 08-21132.

* Equally contributing authors; cnegre@lanl.gov

† Equally contributing authors; uriel.morzan@yale.edu

‡ ivan.rivalta@ens-lyon.fr

§ victor.batista@yale.edu

- [1] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, *Pharmacology and Therapeutics* **138**, 333 (2013), arXiv:1210.0330.
- [2] J. R. Wagner, C. T. Lee, J. D. Durrant, R. D. Malmstrom, V. A. Feher, and R. E. Amaro, *Chemical Reviews* **116**, 6370 (2016).
- [3] N. M. Goodey and S. J. Benkovic, *Nat. Chem. Biol.* **4**, 478 (2008).
- [4] M. T. Reetz, P. Soni, J. P. Acevedo, and J. Sanchis, *Angew. Chem.* **121**, 8418 (2009).
- [5] M. Ozbil, A. Barman, R. P. Bora, and R. Prabhakar, *J. Phys. Chem. Lett.* **3**, 3460 (2012).
- [6] R. J. Hawkins and T. C. B. McLeish, *Phys. Rev. Lett.* **93**, 098104 (2004).
- [7] D. Ming and M. E. Wall, *Phys. Rev. Lett.* **95**, 198103 (2005).
- [8] M. Palumbo, L. Farina, A. Colosimo, K. Tun, and P. K. Dhar, *Curr. Bioinf.* **1**, 219 (2006).
- [9] I. Rivalta, M. M. Sultan, N.-S. Lee, G. A. Manley, J. P. Loria, and V. S. Batista, *Proc. Natl. Acad. Sci. USA* **109**, E1428 (2012), <http://www.pnas.org/content/109/22/E1428.full.pdf+html>.
- [10] A. T. Vanwart, J. Eargle, Z. Luthey-Schulten, and R. E. Amaro, *Journal of Chemical Theory and Computation* **8**, 2949 (2012), arXiv:NIHMS150003.
- [11] A. A. S. T. Ribeiro and V. Ortiz, *Chemical Reviews* **116**, 6488 (2016).
- [12] K. Blacklock and G. M. Verkhivker, *PLOS Computational Biology* **10**, 1 (2014).
- [13] X. Sun, H. gren, and Y. Tu, *The Journal of Physical Chemistry B* **118**, 14737 (2014), pMID: 25453446, <http://dx.doi.org/10.1021/jp506579a>.
- [14] Y. Zhu, B. Ma, R. Qi, R. Nussinov, and Q. Zhang, *The Journal of Physical Chemistry B* **120**, 3551 (2016), pMID: 27007011, <http://dx.doi.org/10.1021/acs.jpcc.5b12299>.
- [15] R. Appadurai and S. Senapati, *Biochemistry* **55**, 1529 (2016), pMID: 26892689, <http://dx.doi.org/10.1021/acs.biochem.5b00946>.
- [16] L. Xu, W. Ye, C. Jiang, J. Yang, J. Zhang, Y. Feng, R. Luo, and H.-F. Chen, *The Journal of Physical Chemistry B* **119**, 2844 (2015), pMID: 25633018, <http://dx.doi.org/10.1021/jp510940w>.
- [17] A. T. VanWart, J. Eargle, Z. Luthey-Schulten, and R. E. Amaro, *Journal of Chemical Theory and Computation* **8**, 2949 (2012), pMID: 23139645, <http://dx.doi.org/10.1021/ct300377a>.
- [18] G. Palermo, C. G. Ricci, A. Fernando, R. Basak, M. Jinek, I. Rivalta, V. S. Batista, and J. A. McCammon, *Journal of the American Chemical Society* **0**, null (0), pMID: 28764328, <http://dx.doi.org/10.1021/jacs.7b05313>.
- [19] J. Guo and H.-X. Zhou, *Chemical Reviews* **116**, 6503 (2016), pMID: 26876046, <http://dx.doi.org/10.1021/acs.chemrev.5b00590>.
- [20] S. Li, J. Zhang, S. Lu, W. Huang, L. Geng, Q. Shen, and J. Zhang, *PLOS ONE* **9**, 1 (2014).
- [21] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten, *Proceedings of the National Academy of Sciences* **106**, 6620 (2009).
- [22] C. G. Ricci, R. L. Silveira, I. Rivalta, V. S. Batista, and M. S. Skaf, *Scientific Reports* **6**, 19940 (2016).
- [23] E. Papaleo, K. Lindorff-Larsen, and L. De Gioia, *Phys. Chem. Chem. Phys.* **14**, 12515 (2012).
- [24] H. David-Eden and Y. Mandel-Gufreund, *Nucleic Acid Res.* **36**, 4641 (2008).
- [25] X. Jiang, C. Chen, and Y. Xiao, *J. Comput. Chem.* **31**, 2502 (2010).
- [26] A. Szilagy, R. Nussinov, and P. Csermely, *Curr. Topics in Med. Chem.* **13**, 64 (2013).
- [27] O. Lange and H. Grubmüller, *Proteins: Structure, Function, and Bioinformatics* **62**, 1053 (2006).
- [28] O. F. Lange and H. Grubmüller, *Proteins: Structure, Function, and Bioinformatics* **70**, 1294 (2008).
- [29] B. M. Savoie, K. L. Kohlstedt, N. E. Jackson, L. X. Chen, M. Olvera de la Cruz, G. C. Schatz, T. J. Marks, and M. A. Ratner, *Proceedings of the National Academy of Sciences* **111**, 10055 (2014), <http://www.pnas.org/content/111/28/10055.full.pdf>.
- [30] U. Doshi, M. J. Holliday, E. Z. Eisenmesser, and D. Hamelberg, *Proceedings of the National Academy of Sciences* **113**, 4735 (2016),

- <http://www.pnas.org/content/113/17/4735.full.pdf>.
- [31] B. Ruhnau, *Social Networks* **22**, 357 (2000).
- [32] B. N. Chaudhuri, S. C. Lange, R. S. Myers, S. V. Chittur, V. Davisson, and J. L. Smith, *Structure* **9**, 987 (2001).
- [33] R. S. Myers, J. R. Jensen, I. L. Deras, J. L. Smith, and V. J. Davisson, *Biochem.* **42**, 7013 (2003).
- [34] I. Rivalta, M. M. Sultan, N.-S. Lee, G. A. Manley, J. P. Loria, and V. S. Batista, *Proc. Natl. Acad. Sci. USA* **109**, E1428 (2012), supporting information document, <http://www.pnas.org/content/109/22/E1428.full.pdf+html>.
- [35] R. W. Floyd, *Commun. Acm.* **5**, 345 (1962).
- [36] M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 7821 (2002), <http://www.pnas.org/content/99/12/7821.full.pdf+html>.
- [37] M. E. J. Newman, *Proceedings of the National Academy of Sciences of the United States of America* **103**, 8577 (2006), arXiv:0602124 [physics].
- [38] I. Rivalta, G. P. Lisi, N.-S. Snoeberger, G. Manley, J. P. Loria, and V. S. Batista, *Biochem.* **55**, 6484 (2016), PMID: 27797506, <http://dx.doi.org/10.1021/acs.biochem.6b00859>.
- [39] G. P. Lisi, K. W. East, V. S. Batista, and J. P. Loria, *Proceedings of the National Academy of Sciences* **114**, E3414 (2017), <http://www.pnas.org/content/114/17/E3414.full.pdf>.
- [40] D. S. Watkins, *Fundamentals of matrix computations, third edition* (John Wiley & Sons, 2010).
- [41] J. Jimenez-Martinez and C. F. A. Negre, *Phys. Rev. E* **96**, 013310 (2017).
- [42] G. Lisi, G. Manley, H. Hendrickson, I. Rivalta, V. S. Batista, and J. Loria, *Structure* **24**, 1155 (2016).
- [43] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Proteins: Structure, Function, and Bioinformatics* **17**, 412 (1993).
- [44] S. Hayward and B. L. de Groot, "Normal modes and essential dynamics," in *Molecular Modeling of Proteins*, edited by A. Kukol (Humana Press, Totowa, NJ, 2008) pp. 89–106.
- [45] T. Meyer, C. Ferrer-Costa, A. Prez, M. Rueda, A. Bidon-Chanal, F. J. Luque, C. A. Laughton, and M. Orozco, *Journal of Chemical Theory and Computation* **2**, 251 (2006).
- [46] U. N. Morzan, L. Capece, M. A. Marti, and D. A. Estrin, *Proteins: Structure, Function, and Bioinformatics* **81**, 863 (2013).