PNAS
www.pnas.org

Supplementary Information for

Eigenvector Centrality for Characterization of Protein Allosteric Pathways

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

Christian F. A. Negre E-mail: cnegre@lanl.gov,
Uriel N. Morzan E-mail: uriel.morzan@yale.edu,
Ivan Rivalta E-mail: ivan.rivalta@ens-lyon.fr,
Victor S. Batista E-mail: victor.batista@yale.edu

This PDF file includes:

Supporting Information Text

1. Relationship between modularity and eigenvector centrality

It is possible to define an eigenvector centrality-based modularity matrix that could offer a clear connection between the community structure and the plots of the centrality. Lets consider a weighted graph $\tilde{G} = (V, E)$ with nodes in V edges in E where the weights w_{ij} between nodes i and j are given by the elements of the adjacency matrix, this is $w_{ij} = A_{ij}$. Following the definition of modularity matrix B of reference (1) we have:

$$B_{ij} = A_{ij} - \frac{g_i g_j}{2m} \quad [1]$$

Where $m = 1/2 \sum_l g_l = 1/2 \sum_{lk} A_{lk}$ is the total number of edges and $g_{i/j}$ is the node degree (total connection) for node i/j. From the definition of the eigenvector centrality, we know that:

$$\sum_l c_l^2 = 1 \Leftrightarrow \sum_l 2mc_l^2 = 2m = \sum_l g_l \quad [2]$$

Where c_l is the eigenvector centrality value for node l.

With this new definition of the node degree as a scaled node centrality, we can rewrite the modularity matrix of reference (1) as follows:

$$B_{ij} = A_{ij} - 2mc_i^2 c_j^2 \quad [3]$$

$$\tilde{B}_{ij} = \frac{A_{ij}}{2m} - c_i^2 c_j^2 \quad [4]$$

The two terms of equation 4 can be interpreted as a sort of probability of signal flow. The first term can be viewed as the probability of a signal flowing from i to j or vice-versa through a direct connection whereas the second term can be interpreted as the probability of flowing from i to j or vice-versa through an indirect connection. This stems form the fact that the square of the centrality value has the property of a probability measure since $\sum_i c_i^2 = 1$ and $0 \leq c_i^2 \leq 1 \, \forall \, i$.

Let us now have a vector of labels $\mathbf{s}$ such that s_i = -1 or +1 depending if node i belongs to one or another community. To get the optimal community partitioning we want $max_{\mathbf{s}}(Q(\mathbf{s})) = max_{\mathbf{s}}(\mathbf{s}^t \tilde{B} \mathbf{s})$ and the ij cotribution to the argument of the "min" function will be:

$$s_i \frac{A_{ij}}{2m} s_j - s_i c_i^2 c_j^2 s_j \quad [5]$$

From the probability interpretation we note that if nodes i and j have a strong indirect connection but with $A_{ij} \simeq 0$ then we need to classify them as belonging to different communities ($s_i s_j = -1$) in order to maximize Q. In the opposite case, if $A_{ij} >> 0$ and the indirect connection is low, we need to classify them as belonging to the same communities ($s_i s_j = 1$) in order to maximize Q. The situations in between will be decided by the balance between the first and second term of equation 4.

2. Renormalization of centrality values for plotting purposes

We have performed a plot of the centrality values with a color scale that can be added in temperature factor field (beta) (2) of a PDB file. In order to do this, the following transformation is applied to the centrality values:

$$c_i \leftarrow 2 \frac{c_i - c_{min}}{(c_{max} - c_{min})} - 1 \quad [6]$$

With this transformation we ensure that if $c_i = c_{min} \Rightarrow c_i \leftarrow -1$ and if $c_i = c_{max} \Rightarrow c_i \leftarrow 1$

3. Correlation matrix as a function of the damping

Figure S1 shows the correlation matrices plotted for different values of λ. We can see a clear transition to a diagonal dominant matrix when $\lambda = 1.66$.

4. Centrality-Degree comparisson

The EC (Eq. 3) measure is also intimately related to the degree centrality (DC) (Eq. 1). However, in Section 3 it is shown that the difference between these two measurements encodes very fundamental aspects of the system's behavior. In order to gauge this difference, the upper panel of Figure S2 represents the normalized PRFAR - *apo* values of EC plotted against the corresponding DC distribution. As expected, the two measurements are very similar. Interestingly, the decrease of the locality factor λ enhances the difference between EC and DC. This shows that the nature of the EC - DC difference is manifested mainly in the local component of the correlations, which is fully consistent with the interpretation of the distribution defined in eq. 11 as a *neighborhood centrality*.

The lower panel of Figure S2 shows the *neighborhood centrality* values for $\lambda = 5$ Å as defined by Eq. 11 (red line). Alternatively, the EC/DC ratio yields qualitatively similar results, and hence, in principle this is an alternative way of defining

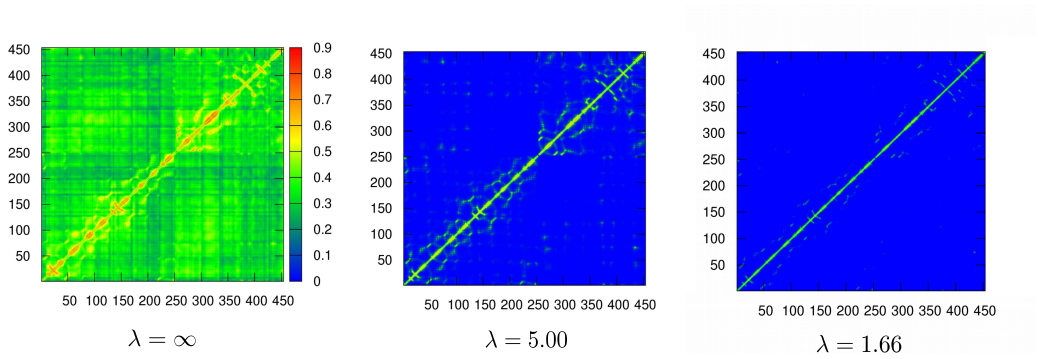

Fig. S1. Correlation matrix plots for APO for different values of λ. For $\lambda = 1.66$ the correlation matrix becomes almost diagonal.

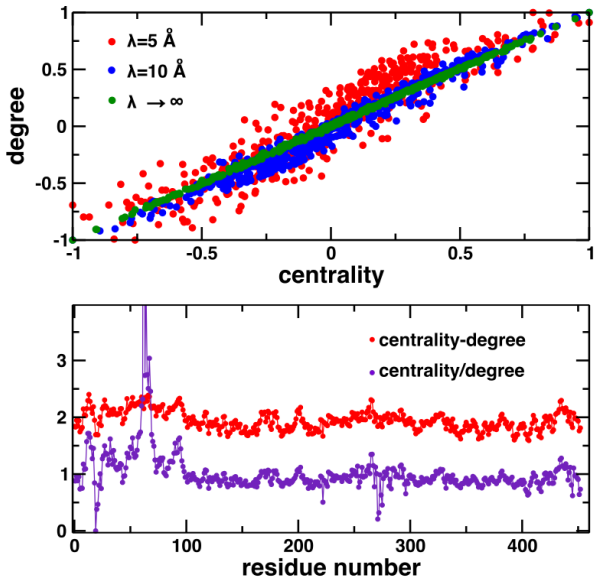

Fig. S2. (Upper panel) Normalized PRFAR - *apo* EC vs DC plots at $\lambda = 5$ Å, 15 Å and $\lambda \to \infty$. (Lower Panel) EC - DC (red) and EC/DC (purple) PRFAR - *apo* distributions. In order to avoid dividing by nearly zero values all the DC and EC values where shifted up by one.

the *neighborhood centrality*. However, one drawback associated to the EC/DC ratio is that the distribution presents locally large amplitude fluctuations that mask the global features of the system, hindering the clear interpretation of the colored representations presented in Figures 3, 5, 6, 7 and 8.

5. NMR Relaxation Dispersion Experiments and Data Processing

Multiple-quantum CPMG experiments probing ILV methyl groups ($^{1}3CH_3$) were performed on 14.1 T Varian Inova and 18.8 T Agilent NMR spectrometers at 30 ^{o}C. A constant relaxation delay period of 30 ms was used in the CPMG pulse sequence, with τcp delays of 0.0, 0.4167, 0.50, 0.625, 0.7682, 1.0, 1.4286, 2.0, 2.5, 3.333, 5.0, and 10.0 ms, and a recycle delay of 2.0 s. NMR spectra were processed with NMRPipe and analyzed in SPARKY. Transverse relaxation rates (R_2) were determined by measuring peak intensities of each ILV methyl resonance at multiple tcp delay points with a Perl-based exponential curve-fitting script. Relaxation dispersion curves were generated by plotting R2 versus $1/\tau_{cp}$ using in-house scripts. Relaxation dispersion data obtained at two static magnetic fields were fit simultaneously using the fast CPMG equation and uncertainty values were obtained from replicate spectra.

6. Protein Expression and Purification

The HisH and HisF proteins were expressed separately at 37 ^{o}C in M9 minimal medium containing $CaCl_2$, $MgSO_4$, and MEM vitamins. The HisF subunit was grown in 100% D_2O supplemented with $^{1}5NH_4Cl$ and $^{1}2C_6H_{12}O_6$ and isotopic labeling of isoleucine, leucine, and valine ($^{1}3CH_3$-ILV) methyl groups was achieved with 60 mg/L of alpha-ketobutyric acid [methyl-13C;

3,3-D2] and 100 mg/L of alpha-ketoisovaleric acid [3-methyl-13C; 3,4,4,4-D4] added 30min prior to induction. HisH was grown in deuterated M9 with naturally abundant nitrogen and carbon isotopes. Following induction at 37 ^{o}C with 1 mM IPTG, cells were harvested by centrifugation, resuspended in 10mM Tris, 10mM CAPS, 300mM NaCl, and 1 mM b-mercaptoethanol at pH 7.5 and co-lysed by ultrasonication. Cell debris was removed by centrifugation and the IGPS complex was purified by Ni-NTA affinity chromatography.

References

1. Newman MEJ (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* 103(23):8577–8582.
2. PDB (2017) Pdb file format.