Using Current-flow Scheme to Capture Protein Allosteric Regulation

Wesley M. Botello-Smith, Yun Luo*

Department of Pharmaceutical Sciences, Western University of Health Sciences, Pomona CA 91766
Telephone: (909) 469 8377
Email: luoy@westernu.edu

Abstract

Network analysis provides a powerful set of tools for investigating protein allosteric regulation. Commonly utilized network construction methods based upon inter-residue contact and pairwise correlated motion often suffer from instabilities, thus could lead to inconclusive or misleading results. Here, we have shown that this instability can be mitigated by utilizing current-flow betweenness scoring, which includes contributions from all possible paths connecting a given set of nodes. By calculating the network shift between a wild-type serine/threonine kinase and three gain-of-function mutants, we have shown that current-flow betweenness method allows for identification of statistically relevant changes in edge or node path usage with greater sensitivity than when correlation is used directly as network weights. Therefore, current-flow betweenness can be seen to have great potential utility for future works in which network analysis of biomolecular simulations is employed.

Introduction

Protein allosteric regulation is ubiquitous inside the cell and serves to regulate cell functions in a delicate way. Nevertheless, basic questions about the nature of the allosteric coupling between the protein-protein complex and the molecular mechanism underlying the regulation of protein function upon binding remained difficult tasks to tackle. Inparticular, protein-protein interactions usually encompass several communication sub-optimal paths that are often difficult to capture using conventional geometric based network analysis based solely on the pair-wise atomic

motion correlations. Here we developed a novel score function in order to capture the complex protein-protein communication upon binding and how the communication is transmitted from binding interface to the protein functional domain.

Network analysis has gained increasing attention in the biology over the past decade. Application to molecular modeling and simulations is particularly promising. One application of network analysis in the context structural analysis of biomolecular systems, is determination of which residues of a molecule are important for propagation of signals along the molecule. For instance, it is well known that many proteins will bind to an inhibitor in one region of a protein, resulting in a conformational change that inactivates a distant binding region. By constructing a network of interactions between nearby residues, one may then employ techniques from graph theory (i.e. network analysis) to determine which residues in the protein play an important role in the propagation of conformational changes from one site to the other, such as protein or small ligand binding site and catalytic site.

Current methods make use of betweenness (a.k.a. geodesic) centrality, which measures the number of times that an edge or vertex appears on a shortest path connecting two other nodes. This measurement is quite straight forward to compute and it provides a useful metric to determine which nodes in a graph (e.g. residues in a protein) are important with respect to general propagation of signals (e.g. energy, motion, conformational change, etc) across the network. While this may be useful in the case of an exploratory study, often, one is concerned with propagation of changes between a known pair or set of residues or domains. When applied under that context, betweenness centrality may yield erroneous estimates because it will contain contributions from paths which are not relevant to transmitting a signal between the domains of interest. On a related note, betweenness centrality considers only the shortest paths between nodes in a network. However, edges and nodes that lie 'near to' but not exactly on a shortest path may provide relevant contributions, will be ignored by a standard betweenness centrality computation. These two drawbacks can lead to artifacts such as undervaluing the importance of nodes and edges that are near but part of shortest paths, and also to overvaluing of edges that lie on non-relevant paths. Moreover, since edges lying near to shortest paths may never be included

at all, this can lead to significant instability when applying betweenness centrality to molecular simulations, wherein the network topology may fluctuate over time.

When only transmission between a pair of known residues or domains is of interest, a modified approach to betweenness centrality may be employed. Here one computes only shortest paths connecting residues in the two domains of interest. This reduces overestimating contributions from nodes and edges that do not lie between regions of interest. Next, contributions of edges and nodes that lie 'near to' but not on the shortest path may be included by computing 'suboptimal' paths. This method, dubbed 'Weighted Implementation of Suboptimal Paths' (WISP)[1] helps address the aforementioned drawbacks of standard betweenness centrality when applied to this specific, but common, sub-network analysis problem.

One drawback to this approach, however, is that one must either choose a priori what amount of additional path length will be considered, or else scan iteratively until satisfactory results are yielded. This can be quite time consuming. Moreover, when applied to the common method of constructing networks based on inter-residue contacts and correlation based edge weights computed from a molecular simulation e.g. as in the NetworkView[2] plugin available under the VMD[3] modeling and simulation package, the results can be unstable with respect to the subset of the simulation trajectory that is chosen. This is discussed at length, along with alternative energy based network construction and weighting in a pair of papers by Ribeiro and Ortiz[4, 5]. These papers make it clear that such instabilities provide a complication when assessing the relative importance of nodes or edges based on contact topology and correlation edge weighting. Here, we show that there is an alternative betweenness metric that can alleviate these instabilities. Furthermore, this betweenness metric is generalizable to other network topology and weighting schemes and thus could be employed for analysis of energy based network topology and weighting schemes as well to further improve stability.

One cause of instability of betweenness centrality (and even WISP based betweenness measurements) when applied to networks generated using contact based topologies and correlation weighting is that the shortest path, and even paths near to the shortest path, may fluctuate quite significantly over the course of a molecular simulation. The WISP method

provides additional stability here because edges and nodes in a shortest path that has disappeared may still be on a 'suboptimal' path. Thus, one may compensate, at least in part, for these fluctuations by choosing a sufficiently large path length cutoff. However, it is often difficult to know a priori what additional length to consider and computing sets of suboptimal paths for extremely long cutoff lengths can become quite cumbersome.

In current work, we investigate an alternative method for computing betweenness scores, known as current-flow betweenness, that takes its roots in analysis of electrical circuits[6]. Interestingly, this method can be shown to be analogous to a metric commonly employed in analysis of information propagation networks, known as Information Centrality[7], which was derived in a much different manner (yet, surprisingly, is mathematically equivalent). It can be shown that information centrality and current-flow centrality yield essentially equivalent results mathematically[6]. Moreover, it has been shown that these metrics are upper or asymptotic limits for certain random walk based methods[6, 8]. Here, we will focus on the current-flow betweenness since the derivation is conceptually simpler and more intuitive than the derivation of the analogous Information Centrality measurement.

The primary advantage of current-flow betweenness over geodesic betweenness is much the same as the advantages derived from calculating betweenness scores based on the WISP method. Specifically, it may be applied (and in fact is most easily applied) when one is concerned only with paths connecting a specific set of nodes. Thus it does not suffer from the draw back of standard betweenness centrality where paths connecting extraneous pairs of nodes may lead to overvaluing the importance of a node or edge. Like the WISP method, current-flow betweenness includes contributions from edges and nodes that are 'near to' but do not lie on a shortest path. However, current-flow betweenness includes contributions from all possible paths connecting the set of nodes of interest, not just those falling within a specified cutoff as in the WISP method. Thus, in some sense, current-flow betweenness can be conceived of as a more robust version of the WISP method with respect to computing betweenness.

Current-flow betweenness is framed in terms of a network of electrical resistor connecting a pair (or set) of source nodes and sink nodes across which a sufficient voltage is applied to induce a

unit of current to flow. In the case of a system of electrical resistors, it is known that current will flow through all possible paths between a given source and ground simultaneously. If one were to measure the current-flowing through a given edge (resistor) in such a network, given that sufficient current has been applied to induce a unit of current-flow between the supply and ground, this would then yield the "current-flow betweenness", i.e. as the current-flow through that node or edge in the electrical network. For simple networks, this may be computed quite easily by hand by employing Kirkoff's laws. Figure 1 shows an example of using Kirkoff's Laws to compute current-flow through each edge of network of resistors. In protein dynamics, the negative log of pairwise correlation can be treated as resistance between each edge.

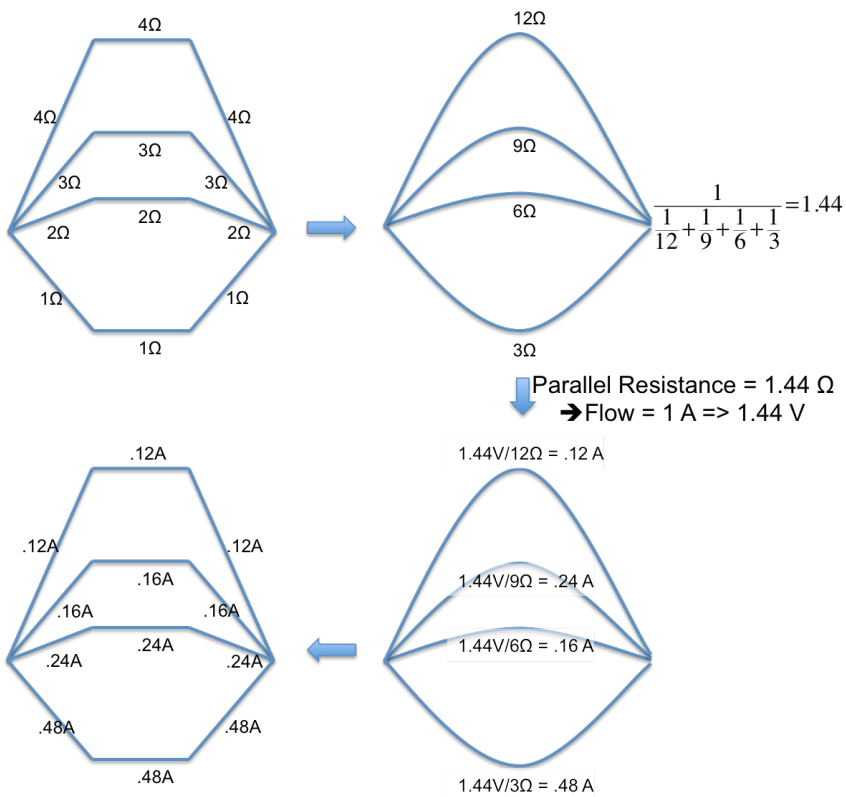


Figure 1: Example of using Kirkoff's Laws to compute current-flow through each edge of network of resistors.

As mentioned earlier, it can be shown that this 'current-flow betweenness' is equivalent to another betweenness measure, which is well known in other branches of network theory, known as information centrality. Similarly, several betweenness scoring metrics have been proposed based on random walk approaches. In many cases, such scores can also be shown to be equivalent or approximate of information centrality. These metrics have been shown to provide better results as compared with shortest path betweenness in other network analysis fields. In this paper we investigate the utility of current-flow betweenness to study of allosteric networks in proteins. For the interested reader, a more detailed in depth derivation of flow betweenness can be found in the paper by Brandes and Daniel[6], cited earlier in the introduction, along with discussions of how it can be shown to be equivalent to information centrality. A summary of the relevant equations and considerations are described in Methods section.

Here we use three kinase systems as examples to illustrate how to capture the protein-protein communications that are critical for protein function using current-flow betweenness score in the network analysis. Protein kinases are signaling enzymes that serve to regulate cell growth, proliferation, metabolism, differentiation, and migrations. The human kinome encodes over 500 protein kinases. Among them, serine/threonine kinase receptor I (STKR1) represent an important family that controls the TGF-beta and bone morphogenetic protein signaling. The malfunction of these family kinases is often associated with cancer and bone formation diseases. In the absence of ligand, STKR1 kinase activity is physiologically inhibited by binding of a negative regulator protein FKBP12 to the GS domain. Gain-of-function mutations in the GS and kinase domains are somehow able to bypass FKBP12-mediated inhibition and produce aberrant STKR1 signaling. For instance, Activin-Like Receptor 2 (ALK2) is one of the seven STKR1. Several gain-of-function mutations, namely, R206H, R207E, and R207D near the binding interface of ALK2 and FKBP12 are associated with various diseases such as heterotopic ossifications. Surprisingly, those mutations do not necessarily reduce the protein-protein binding affinity between ALK2 and FKBP12, but somehow allosterically alter the kinase activation. We hereby performed betweenness analysis for all three disease-related mutations using four different betweenness metrics. Our results consistently show how a single mutation at the protein-protein binding interface change the protein-protein communication landscape and allosterically shift the kinase towards catalytic competent configuration. We compared with the

suboptimal path identified by conventional correlation scores and showed quantitatively how current-flow betweenness score improve the convergence of the suboptimal path search. Statistical analysis of network path is also included in this study.

Methods

System Preparation

The system preparation has been described in our previous work[9]. Briefly, wild-type ALK2 (wtALK2) with FKBP bound structure complex was taken from the crystal structure (PDB ID: 3H9R). The missing A-loop residues 362 to 374 and the β-turn residues 273 to 275 were transplanted from PDB ID 3Q4U. Each mutant system (R206H, R207E, and R207D) was prepared from wtALK2-FKBP using MOE software. The pKa values of each amino acid were calculated using PROPKA. The pKa value the histidine in R206H was previously calculated to be 6.3 using constant-pH simulation and the dominant protonation state of 206H is the uncharged state with hydrogen on the Nδ atom[9]. CHARMM-GUI [10] was used to read in the PDB file and generate solvated systems. All simulations employed the all-atom CHARMM C36 force field for proteins and ions, and the CHARMM TIP3P force field [11] for water.

Molecular Dynamics Simulations

All molecular dynamics simulations were performed with NAMD2.9b using periodic boundary conditions under constant temperature and pressure (NPT ensemble). A temperature of 300 K was applied via Langevin thermostat and 1 atm of pressure was maintained via Andersen-Hoover barostat. Long-range electrostatic interactions were treated using the particle-mesh Ewald (PME) method. A smoothing function was applied to both electorstatic and VDW forces over the range of 10 Å to 12 Å. The non-bonded interaction list was updated on every integration step using a cutoff of 13.5 Å. The SHAKE algorithm was applied to all hydrogen atoms. A molecular dynamics equilibrium was set to relax the atomic system by releasing the harmonic constraints (force constant of 50 kcal/mol/$Å^2$) stepwise (every 200 ps) on water and ion molecules, protein side chains, and eventually the protein backbone. Additionally, analysis methods subdivided each 300 ns production run into five 60 ns sub-windows with the first window serving as an additional 60 ns of unrestrained equilibration time. The last four windows (4x60 ns) of each trajectory were considered for network analysis.

Current-flow Betweenness

To compute current-flow betweenness, one must first construct the associated adjacency and laplacian matrices. The adjacency matrix, **A**, has as its elements, $\mathbf{a_{i,j}}$, the absolute values of the corresponding network weights $\mathbf{w_{i,j}}$. In the case of a network of electrical resistors, these weights would correspond to the conductance of each resistor. Here, we are interested in transfer of motion as described by a contact map, weighted by correlation of atomic motions. Thus the entries of the absolute values of the entries in the correlation matrix will serve as the weights of the corresponding edges in the contact map. I.e. given the atomic correlation matrix **C**, we assign $\mathbf{a_{i,j}} = |\mathbf{c_{i,j}}|$ if i, j corresponds to an edge in the contact map, and $\mathbf{a_{i,j}} = 0$ if not.

$$\mathbf{d_{i,i}} = \sum_{j=1}^{N} \mathbf{a_{i,j}}$$

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

Where $\mathbf{d_{i,I}}$ are the elements of the diagonal matrix **D**, constructed by summing over each row of the adjacency matrix **A**. Next, the inverse of this network laplacian **L** must then be computed and can be used to compute the betweeness of each edge (Figure 2).

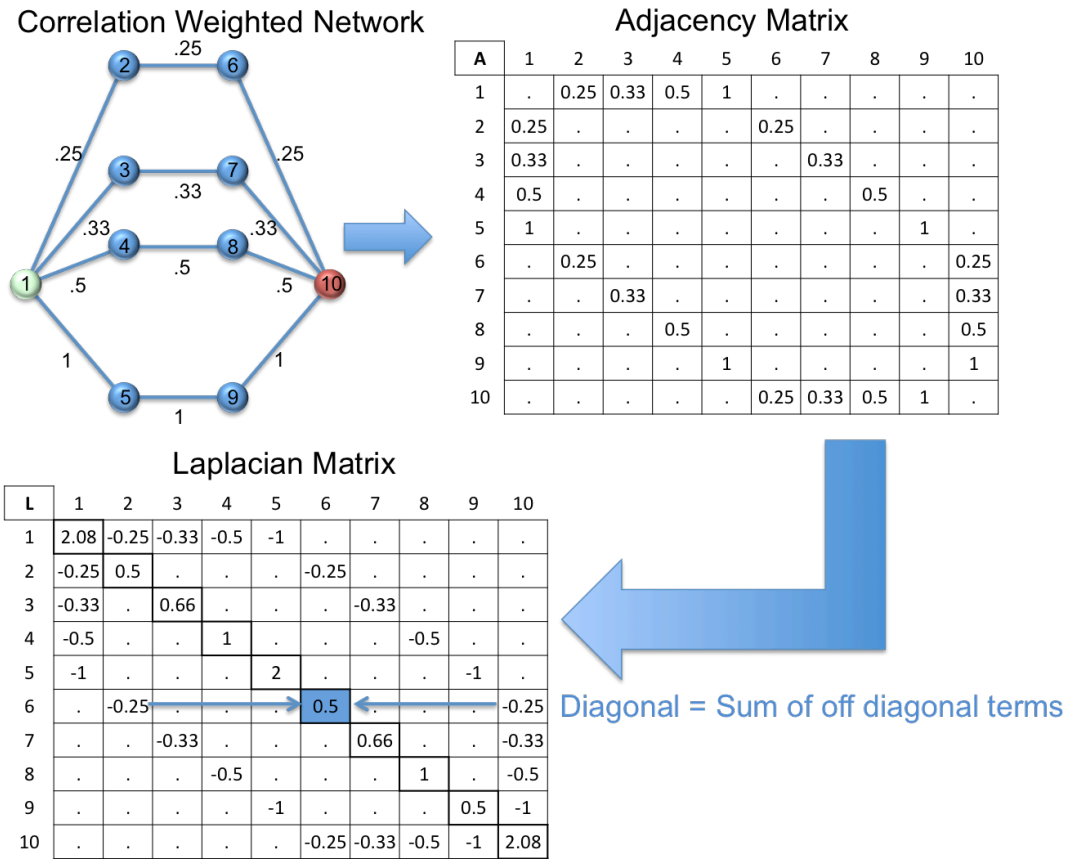


Figure 2: Example of a weighted network and corresponding adjacency and laplacian matricies. This network is the correlation analog to the electrical resistor network from Figure 1.

Betweenness of each node can then be computed by summing the betweenness of each connected edge. Unfortunately, by its construction, the laplacian matrix is guaranteed to be 'singular', meaning that no unique inverse exists. Fortunately, an appropriate 'pseudo-inverse' can serve just as well. In particular, the Moore-Penrose pseudo inverse has been used (although other methods such as single value decomposition may also work). It should be noted that computation of matrix inverses and pseudo-inverses can become quite computationally taxing for very large networks. However, the process is reasonably fast for networks consisting of thousands or even tens of thousands of nodes. In the case of allosteric network applications, network nodes are typically taken to represent individual residues within a protein and are computed from molecular dynamics or other simulation techniques. These techniques seem to have similar size limitations and so, direct computation of pseudo-inverses should be tractable.

In the event where they are not, approximation methods are available, such as the eigenvalue decomposition methods given in the paper by Bozzo and Franceschet[13].

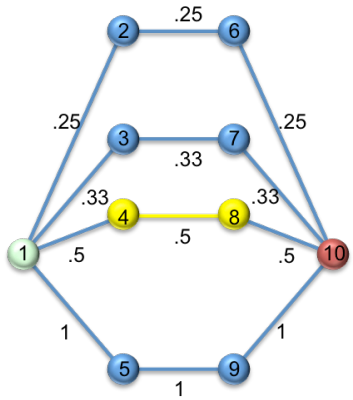
Given a network's adjacency matrix, **A**, and a suitable inverse (or approximation thereof) to its Laplacian, $\mathbf{L}^+$, the current-flow betweenness between nodes i and j, $\mathbf{B}_{\mathbf{i,j}}$, for a given source node, s, and target node, t, is given as:

$$v_i^{s,t} = \mathbf{L}_{i,s}^+ - \mathbf{L}_{i,t}^+$$
$$\mathbf{B}_{i,j} = \mathbf{A}_{i,j} * | v_i^{s,t} - v_j^{s,t} |$$

where $v_i^{s,t}$ is the potential at a given node i. In the case of multiple nodes as source and target, one may attain the current-flow as the sum over all combinations of source and target nodes divided by the number of combinations. This is a relatively trivial double sum when the two sets are disjoint, e.g:

$$\mathbf{B}_{i,j} = \frac{1}{|S \cup T|} \sum_{s \in S} \sum_{t \in T} \mathbf{A}_{i,j} * | v_i^{s,t} - v_j^{s,t} |$$

Where S, T are disjoint sets of source and target nodes respectively. When S and T are not disjoint (e.g. as in computing a generalized current betweenness over the entire network) one must take care to avoid / remove contributions from double summations. Here, however, we consider only the former case where the sources are disjoint from the targets. The interested reader may consult the paper by Brades and Fleischer[6] for details on how to handle the latter. See Figure 3 for an illustrative example of flow betweenness calculation for a simple 10 node network with a single source and sink. The corresponding 'by hand' calculation using Kirkoff's Laws was shown previously in Figure 1 for the analogous network of electrical resistors, where each edge's resistance is equal to the reciprocal of the correlation shown in Figure 2 and 3. One can see that the result for edge 4-8 ($\mathbf{B}_{4,8}$) using the above matrix formalism is identical to the result from electrical network analog.



Adjacency Matrix

A	1	2	3	4	5	6	7	8	9	10
1	.	0.25	0.33	0.5	1	.	.	.	.	.
2	0.25	.	.	.	.	0.25	.	.	.	.
3	0.33	.	.	.	.	.	0.33	.	.	.
4	0.5	.	.	.	.	.	.	0.5	.	.
5	1	.	.	.	.	.	.	.	1	.
6	.	0.25	.	.	.	.	.	.	.	0.25
7	.	.	0.33	.	.	.	.	.	.	0.33
8	.	.	.	0.5	.	.	.	.	.	0.5
9	.	.	.	.	1	.	.	.	.	1
10	.	.	.	.	.	0.25	0.33	0.5	1	.

$\mathbf{L}^+$ (Moore-Penrose Pseudo Inverse of $\mathbf{L}$)

$\mathbf{L}^+$	1	2	3	4	5	6	7	8	9	10
1	0.64	0.32	0.32	0.32	-0.01	0.00	0.00	0.00	-0.67	-0.32
2	0.32	2.83	0.16	0.16	-0.17	1.33	0.00	0.00	-0.67	-0.16
3	0.32	0.16	2.18	0.16	-0.17	0.00	1.01	0.00	-0.67	-0.16
4	0.32	0.16	0.16	1.49	-0.17	0.00	0.00	0.67	-0.67	-0.16
5	-0.01	-0.17	-0.17	-0.17	0.16	-0.33	-0.33	-0.33	-0.67	-0.49
6	0.00	1.33	0.00	0.00	-0.33	2.67	0.00	0.00	-0.67	0.00
7	0.00	0.00	1.01	0.00	-0.33	0.00	2.02	0.00	-0.67	0.00
8	0.00	0.00	0.00	0.67	-0.33	0.00	0.00	1.33	-0.67	0.00
9	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67	-0.67
10	-0.32	-0.16	-0.16	-0.16	-0.49	0.00	0.00	0.00	-0.67	0.16

$$v_4^{1,10} = \mathbf{L}^+_{4,1} - \mathbf{L}^+_{4,10} = .32 - (-.16) = .48$$

$$v_8^{1,10} = \mathbf{L}^+_{8,1} - \mathbf{L}^+_{8,10} = 0 - 0 = 0$$

$$\mathbf{B}_{4,8} = \mathbf{A}_{4,8} * | v_4^{1,10} - v_8^{1,10} | = .5 * | .48 - 0 | = .24$$

Figure 3: Calculation of flow betweenness for edge 4-8 of an example network (top left corner).

Network Analysis

As discussed in the simulation protocol above, trajectories for ALK2 wild-type and each of the investigated mutants, R206H, Q207D, and Q207E, were split into five 60 ns windows and the last 4 windows were then analyzed. The conformational change of R375 on A-loop has been previous reported as a critical step for kinase activation[12]. The goal here is to investigate whether and how each activating mutation alters the communication between the inhibitory binding protein FKBP and R375. A visual overview of the analysis procedure can be seen in Figure 4. First, contact correlation maps were computed for each window using the program CARMA[14] . The resulting data matrices were then loaded into R to compute associated current-flow betweenness scores using nodes (alpha carbon of each amino acid) from FKBP in contact with the kinase domain (within 5 Å from any atoms on kinase domain) as sources and the residue R375 of the salt bridge as a sink. The resulting current-flow betweenness weighted networks for

each window were then exported back into matrix data files to perform windowed ranking error analysis (see section below). Finally, the window-averaged correlation and betweenness weighted networks for each mutant were compared against that of the wild type using ANOVA followed by Tukey's Honest Significant Differences analysis to determine which edges exhibited significant changes as a result of mutation.

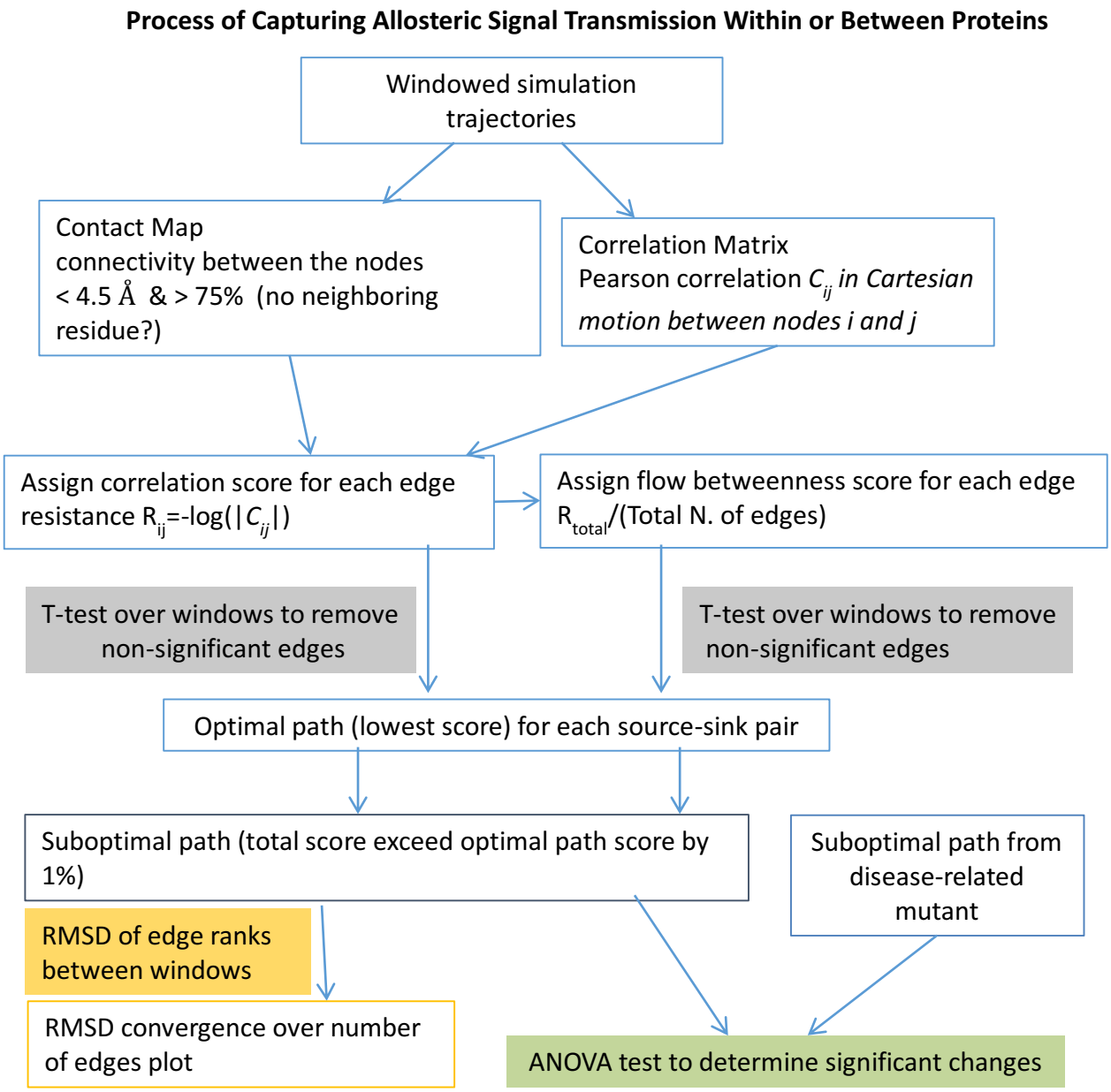


Figure 4: Flow chart of network analysis protocol using current-flow betweenness and ANOVA analysis.

Statistical Analysis of Betweenness

The windowed correlation and flow betweenness weighted networks were then analyzed to filter out edges with weights that were not statistically meaningful. This was accomplished by applying a t-test to the distribution of weight values from the four windows in order compute p-values for each edge individually. Edges with p-values greater than .10 were discarded, i.e. those edges for which the average weight was sufficiently small relative to the variance such that there would be a greater than a 10% chance that the same average weight could be attained by sampling a random distribution with a mean of zero. The averages weights for the remaining edges were then used to construct the FKBP-R375 path sub-networks from the window averaged networks for each system.

Construction of sub-networks

The sub-network of paths leading from FKBP to the regulatory salt bridge residue R375 for each system was constructed using the 'subopt' program from the networkView plugin for VMD (ref). This was accomplished by running the program for each node in FKBP in contact with a node from the kinase domain as the starting point and the node corresponding to R375 as the end point. The optimal path between each source node on FKBP12 and the sink node R375 is defined by the shortest path, in which the distance is the sum of –ln(betweenness core) or –ln(correlation scores) for all edges in a given path. A path dilation of roughly 1% of the distance was used to construct the suboptimal paths. 15 for betweenness weighted paths and 4 for correlation weighted paths was found within 1% dilation.

Sub-network edge ranking stability

To assess the utility of flow betweenness weights in ranking residue-residue interactions, FKBP-R375 sub-networks were constructed for each of the four trajectory windows of each system. Following a procedure analogous to the one described in Ribeiro and Ortiz[4] for node based rank error analysis, η values were assigned to each edge by counting the number of paths the edge occurred in from the output of each FKBP – R375 'subopt' computation, divided by the total number of paths computed. These values were then used to rank each edge in each window with a rank of 1 assigned to the edges with the highest η value, rank of 2 to the next highest η value,

etc. The average of each edge's rank <r> was computed over all 4 windows for each system. Next a second round of ranking was performed based on the average ranking <r>. Finally, the average root mean square deviation of edge ranks was plotted versus the number of nodes included was plotted for each system to compare ranking based on correlation weights versus betweenness weights. E.g. for each system, the top n edges were selected based on average rank score and the RMSD of ranking over the 4 windows was computed using those top n edges.

Statistical analysis and visualization of wild-type vs. mutant sub-networks

The FKBP-R375 sub-networks were computed for each window of the correlation and betweenness weighted contact networks for each system. In order to assess changes in the allosteric network induced by these mutations, each edge of the sub-networks for each mutation was compared against the corresponding edge in the wild type matrix by applying an ANOVA over the given edges value among all windows, followed by Tukey's Honest Significant Differences test. For clarity, any edge that did not appear in the FKBP-R375 sub-network for at least one window of either the wild type or a given mutant was omitted. Any edges which exhibited a p-value of greater than 25% (E.g. 75% confidence level) were then filtered out as well. The remaining network edges were then exported to a pair of matrix data files, with one file containing the network topology with edge thickness based upon the p-value of each edge, and the other file containing the corresponding difference in edge weights, computed as difference between the wild type value, minus the given mutant value. This process was repeated using the sub-network occurrence frequency (η) that was computed previously for each edge using correlation or flow betweenness weights as discussed under the assessment of ranking stability methodology section.

The resulting networks were then visualized using the NetworkView plugin of the VMD molecular graphics program. The p-value matrices served as network topology and edge thickness definitions while edges were colored according to the difference in either edge value or occurrence frequency accordingly. The resulting network depictions were then compiled into a set of three four panel figures (one figure for each wild type versus mutant comparison). These depictions are shown in Figure 6-8 for WT versus Q207D, WT versus Q207E), and WT versus R206H. It should be noted that while most edges have been labeled, the number of resulting

edges in some figures was simply too large to include labels for all edges directly on the network plot. For the interested reader, the corresponding data is available in table form in the supplementary materials.

Results and Discussion

Rank RMSD analysis

As discussed in the papers by Ribeiro and Ortiz[4, 5] network constructed based on inter-residue contact maps may exhibit different network topologies as a function of simulation length. Choice of weighting scheme can also impact this phenomenon. We follow a similar protocol here to compare the stability of flow betweenness weights versus correlation weights.

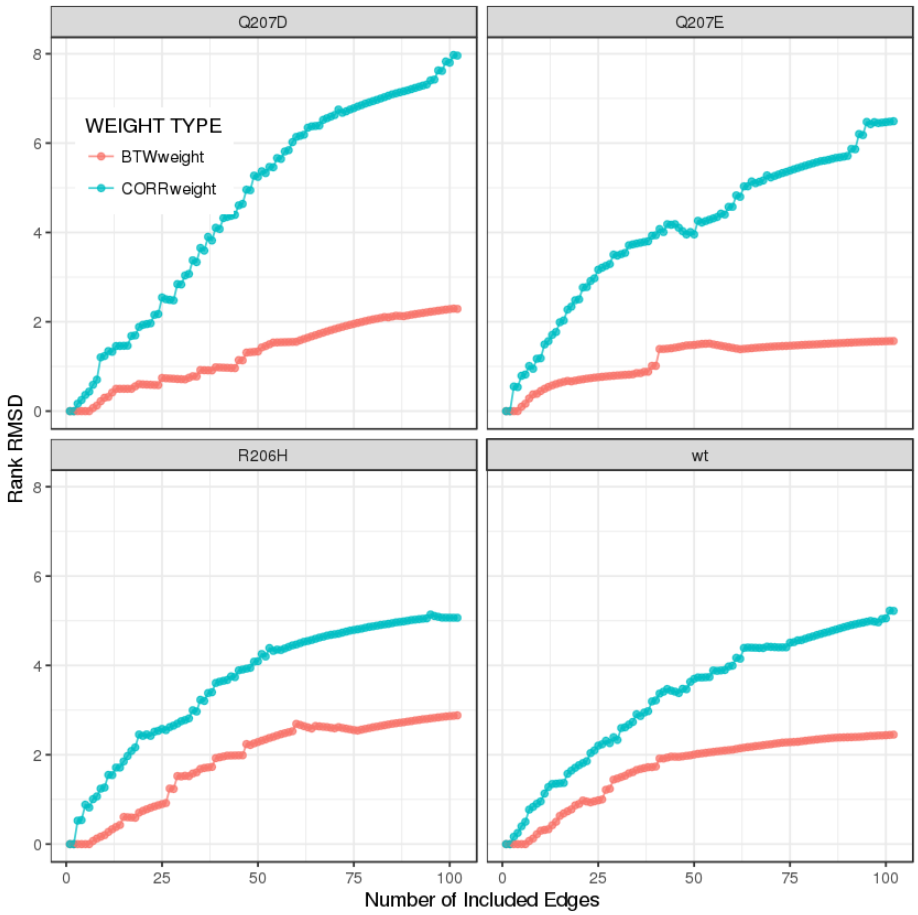


Figure 5: Comparison of edge ranking stability for correlation and flow betweenness weighted contact networks for each system studied.

From inspection of

Figure 5 it is quite clear that flow betweenness weighting yields much more stable estimates of edge ranking as compared with direct use of correlation as weights. Both the correlation and flow betweenness weighted networks shared identical underlying contact maps, thus the effect of weighting choice can be compared without being confounded by changes in network connectivity construction. Finally, It was noted in the papers by Ribeiro and Ortiz[4, 5] that contact based topologies exhibit inherent fluctuation regardless of the weight type used and that energy based connectivity schemes provide a better choice. While we did not investigate such connectivity map construction and weighting schemes here, the flow betweenness metric should be readily transferable to energy based connectivity and weighting schemes as well and may provide additional improvement.

Comparison of correlation, flow betweenness, and occurrence frequency

The utility flow betweenness based network weights versus correlation based weights, or the corresponding sub-network occurrence frequency based weighting schemes, can be qualitatively assessed by examining the differences in corresponding sub-network paths. In this case, we are concerned with the sub-networks connecting the FKBP binding protein to residue R375 of the known salt-bridge lock. While the most straightforward approach would be to compare the sub-networks for each weighting scheme side by side, the resulting sub-networks, are quite large. This is, in part, due to the fact that the FKBP domain portion of the FKDBP- kinase interface, which serves as the source of the sub-networks (while R375 serves as the 'sink') is itself quite large. Thus, even if a relatively small dilation is used to generate additional 'suboptimal' paths (e.g. selecting paths exceeding the shortest path by no more than 1% path length), the resulting sub-network contained hundreds of edges. The networks resulting from the direct application of this suboptimal path methodology can be found in the supplementary material, but are omitted here for clarity.

We present here the results from statistical analysis of differences between the mutant and wild type sub-networks generated under each edge-weighting scheme. As discussed in the methods section, we include only edges satisfying a 75% minimum confidence level. The results are

shown below in Figure 6-8. As can be seen, there are still a large number of significant differences. Noticeably, there are only a handful for which the difference is both substantial in magnitude (as indicated by edge coloring) and also high in statistical significance (as indicated by edge thickness). Similarly, there are a handful of interactions which are common among different weight types and mutants. Of key interest here, is determination of which changes are detected under flow betweenness weighting but not correlation weighting, or vice versa.

One such difference can be seen when considering the edges connecting residues 206 and 207 to the binding protein. These interactions are present in most of the displayed plots with R(H)206-Q54 present when Q207 is mutated to D or E, or K53-Q207 present in the panels for R206H. The detection of these is encouraging since these residues 206 and 207 are the locations of the mutated residues and this result shows that mutation induces a loss of interaction between residue 206 and FKBP is modulated by the R206H, Q207D, and Q207E mutations. Furthermore, while the changes in interaction weight are detected under both correlation and flow betweenness weighting schemes for R206H and Q207E, however, for Q207D, only the flow betweenness weighting scheme was successful in detecting a change (albeit, at the low end of the confidence level limit).

Similarly, the interaction between N373 and K243 between A-loop and αC-helix shows increased strength in all three mutants when flow betweenness scoring is used for weighting, however, if correlation weighting is used, it is only visible in the R206H mutant. This again provides an example of where betweenness weighting can detect changes that may otherwise be overlooked if correlation is directly used for weighting.

Another feature of interest is the interaction G356-R375 between DLG motif and A-loop. Change in this interaction was detected for every system. Under flow betweenness weighting scheme, the edge exhibits a reduction in strength for mutant systems. However, under correlation weighting schemes, this edge shows a gain of interaction for Q207D and Q207E and is only diminished in R206H. When frequency based weighting is used, flow betweenness based usage frequency agrees with betweenness weighting, however, the correlation frequency shows this edge as either absent (no statistically significant change) or diminished in mutant instead of

strengthened. This observation is consistent with the nature of flow betweenness which seeks to estimate edge usage, averaged harmonically over all possible paths while the frequency metrics here include only the optimal path and paths exceeding optimal length by no more than 1%.

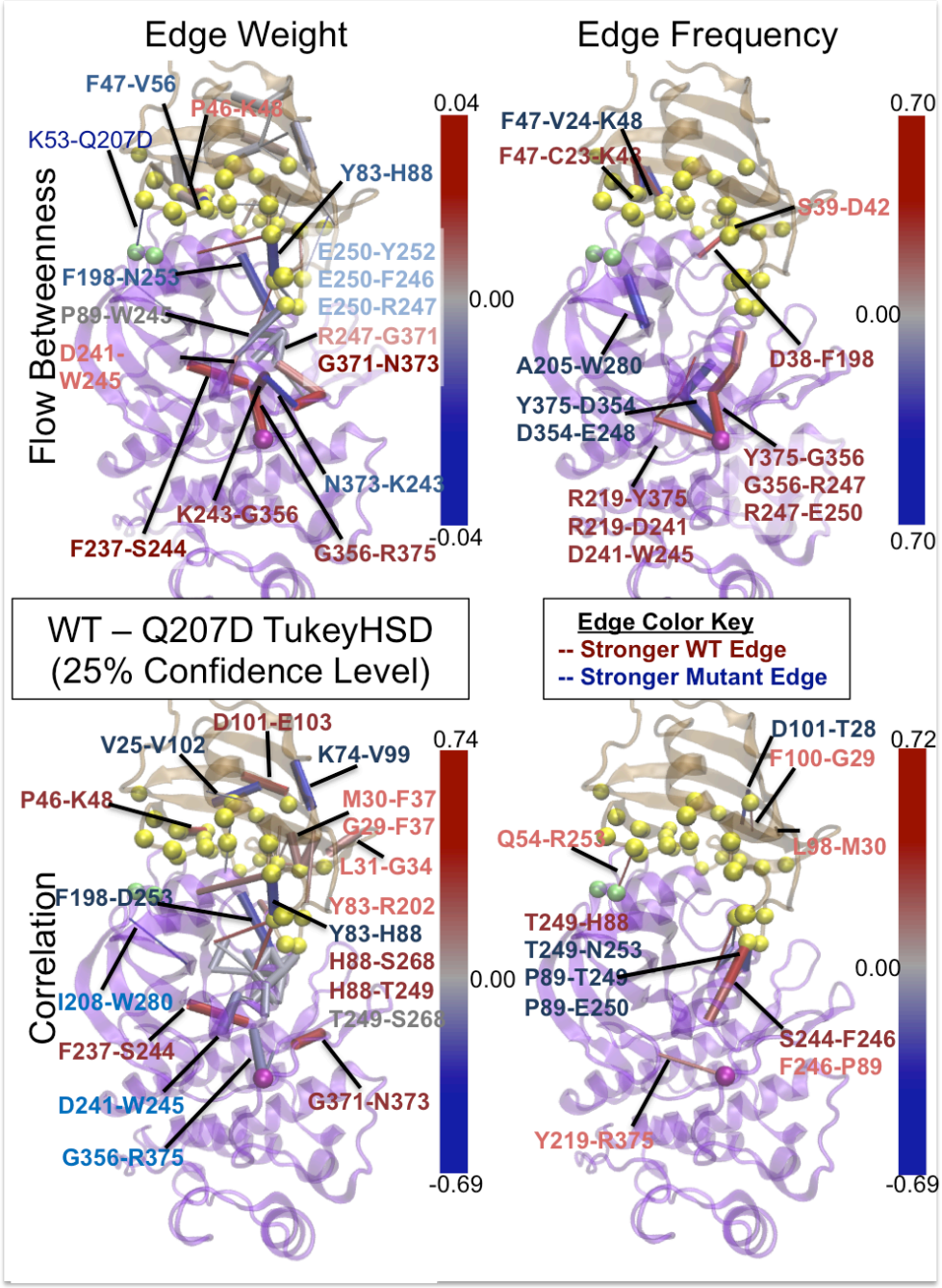


Figure 6: Comparison of WT versus Q207D networks. Each panel depicts the results of statistical analysis of mutant and wild type structure over all windows for each edge weighting scheme: Direct Flow Betweenness (top left), Direct Correlation (bottom left), and Suboptimal path occurrence frequency using Flow Betweenness (top right), or Correlation (bottom right) weights. Edge thickness depicts confidence level with the thinnest edges representing the 75%

minimum confidence level cutoff. Color indicates the change in measured edge value computed as the wild type value minus the mutant value (blue color indicates the edge is stronger in mutant, red indicates the edge is stronger in wild-type).

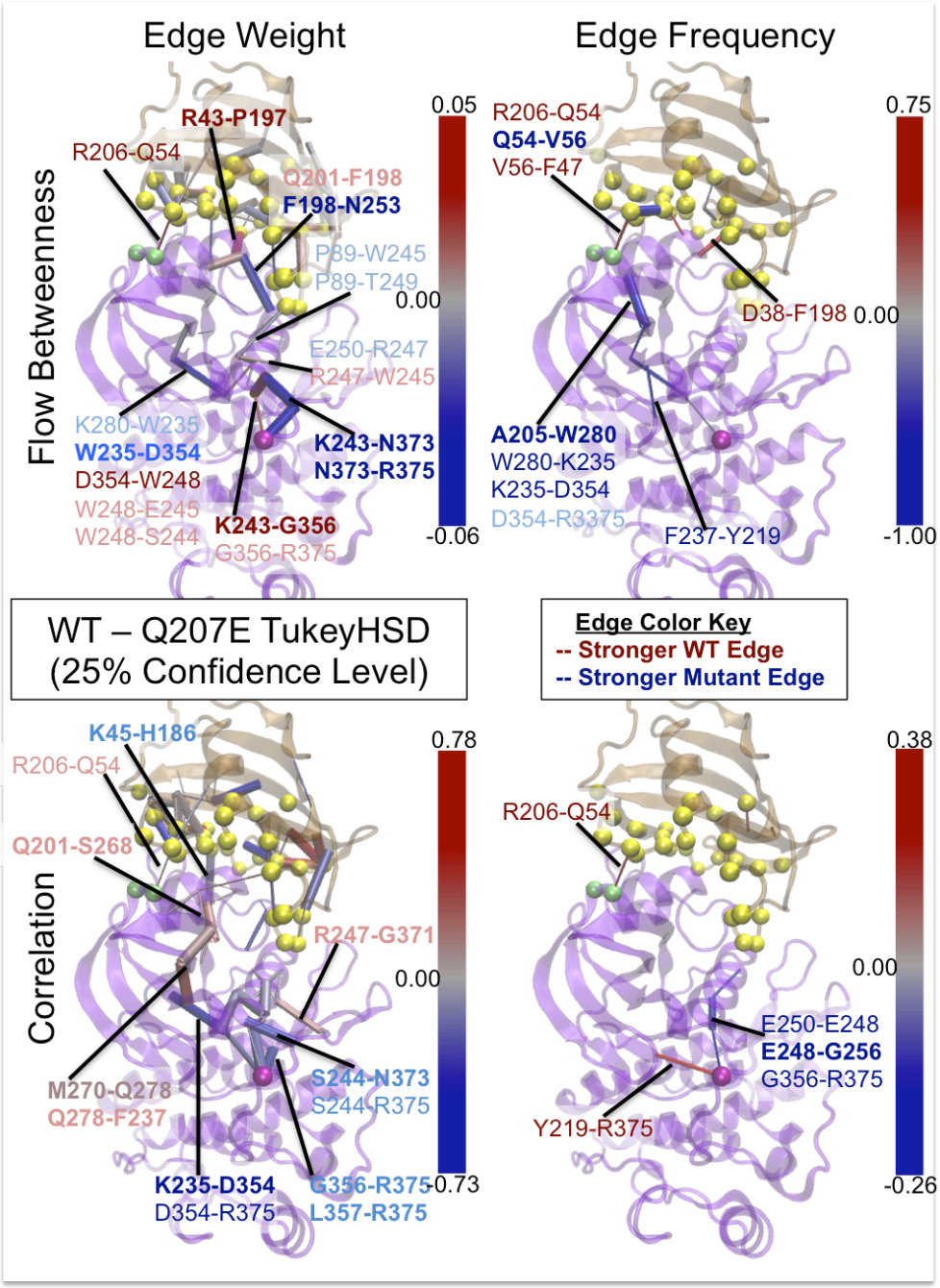


Figure 7: Comparison of WT versus Q207E networks. Each panel depicts the results of statistical analysis of mutant and wild type structure over all windows for each edge weighting scheme: Direct Flow Betweenness (top left), Direct Correlation (bottom left), and Suboptimal path occurrence frequency using Flow Betweenness (top right), or Correlation (bottom right)

weights. Edge thickness depicts confidence level with the thinnest edges representing the 75% minimum confidence level cutoff. Color indicates the change in measured edge value computed as the wild type value minus the mutant value.

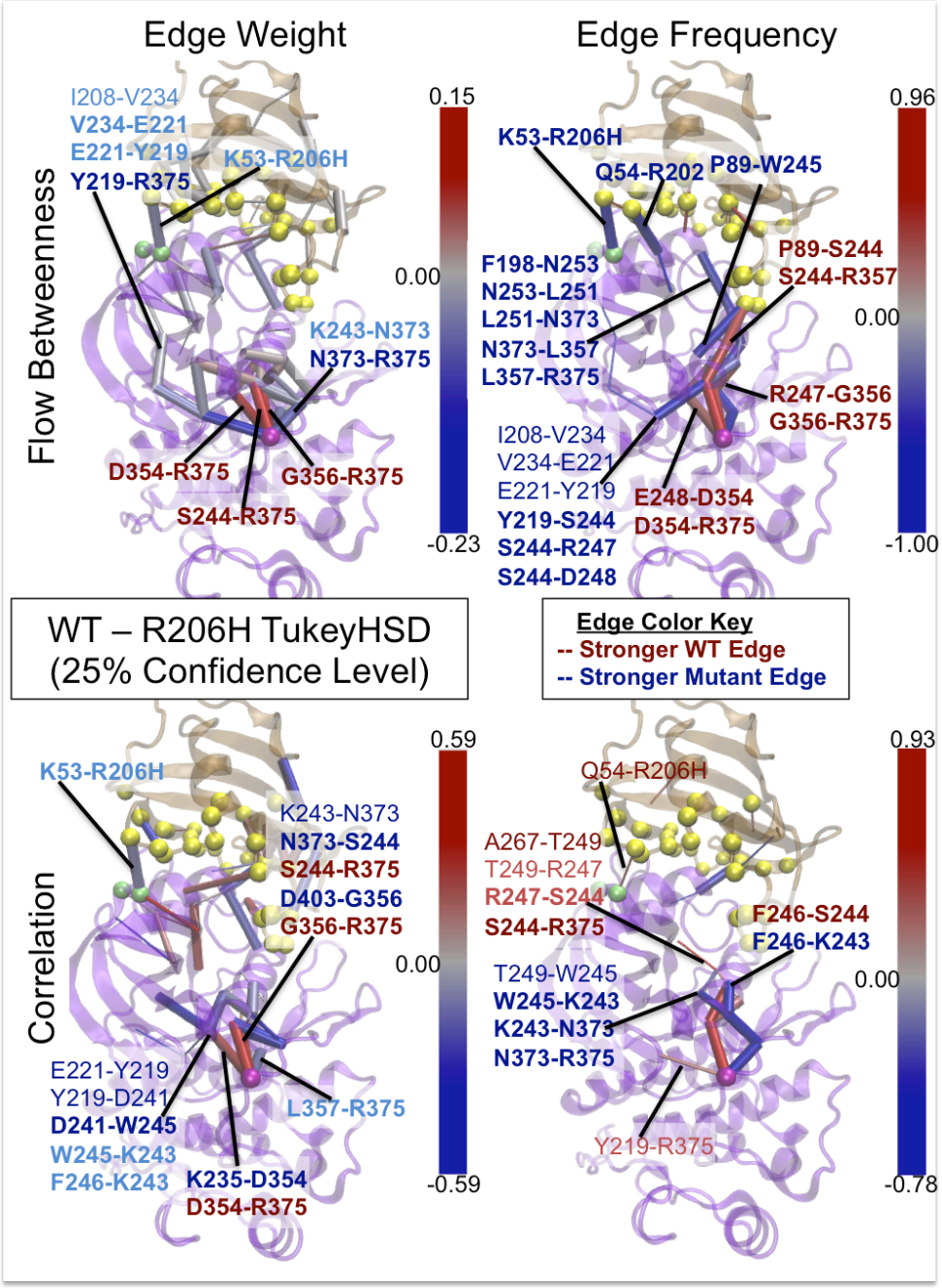


Figure 8: Comparison of WT versus R206H networks. Each panel depicts the results of statistical analysis of mutant and wild type structure over all windows for each edge weighting scheme: Direct Flow Betweenness (top left), Direct Correlation (bottom left), and Suboptimal path occurrence frequency using Flow Betweenness (top right), or Correlation (bottom right) weights. Edge thickness depicts confidence level with the thinnest edges representing the 75%

minimum confidence level cutoff. Color indicates the change in measured edge value computed as the wild type value minus the mutant value.

Conclusion

Network analysis provides a powerful set of tools for investigation of dynamical networks of signal propagation in proteins. It was shown previously that the commonly utilized methods for network construction based upon inter-residue contact topology and pairwise correlated motion based weighting schemes can suffer from instabilities when used in assessing the relative importance of residues and residue to residue interactions (e.g. ranking).

Here, we have shown that this instability can be, at least partially, mitigated by utilizing current-flow betweenness scoring rather than shortest or suboptimal path frequency based betweenness metrics. These methods are not specific to network topology or weighting schemes and could thus potentially be coupled with other methods, such as energy based topology and weighting to further improve stability and accuracy. Moreover, current-flow betweenness provides a more theoretically sound assessment of the relative importance of edges or nodes in a network since it implicitly includes contributions from all possible paths connecting a given set of nodes.

As was shown in the statistical analysis of differences between mutant and wild type systems for the ALK2 protein, changes in the correlation between a pair of nodes may not necessarily imply an increase in path usage. This is particularly relevant when considering systems over which the residue to residue contacts and correlations may fluctuate significantly. Since current-flow betweenness scores exhibit less pronounced fluctuations, they can allow for identification of statistically relevant changes in edge or node path usage with greater sensitivity than when correlation is used directly as network weights. Therefore, current-flow betweenness can be seen to have great potential utility for future works in which network analysis of biomolecular simulations is employed.

Lastly, it should be noted that while we employed network constructions based upon contact based topology and correlation weighting, current-flow betweenness could be employed upon any other topology and weighting scheme (provided that all edge weights are positive). Thus this

method could potentially yield improved results when coupled with interaction energy based network construction such as in the papers by Ribeiro and Ortiz[4, 5], which exhibit increased stability over contact based topology and correlation weighting schemes. Alternatively, smoothing functions may be applied to contact based construction methodologies to mitigate artifacts that arise due to the use of arbitrary cutoff distances and frequencies when constructing contact topologies. Indeed, a thorough and comprehensive comparison of current methodologies in network topology construction and weighting schemes could be extremely useful given the rapid rise in the application of network analysis methods to modeling and simulation investigations.

1. Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E., Weighted implementation of suboptimal paths (WISP): an optimized algorithm and tool for dynamical network analysis. *Journal of chemical theory and computation* **2014**, 10, 511-517.
2. Stone, J.; Developers, N. T.; Eargle, J.; Sethi, A.; Li, L.; Luthey-Schulten, Z., Dynamical Network Analysis. **2012**.
3. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, 14, 33-38.
4. Ribeiro, A. A.; Ortiz, V., Determination of signaling pathways in proteins through network theory: importance of the topology. *Journal of chemical theory and computation* **2014**, 10, 1762-1769.
5. Ribeiro, A. A.; Ortiz, V., Energy propagation and network energetic coupling in proteins. *The Journal of Physical Chemistry B* **2015**, 119, 1835-1846.
6. Brandes, U.; Fleischer, D. Centrality Measures Based on Current Flow. In STACS, 2005; Springer: 2005; Vol. 3404; pp 533-544.
7. Stephenson, K.; Zelen, M., Rethinking centrality: Methods and examples. *Social networks* **1989**, 11, 1-37.
8. Newman, M. E., A measure of betweenness centrality based on random walks. *Social networks* **2005**, 27, 39-54.
9. Botello-Smith, W. M.; Alsamarah, A.; Chatterjee, P.; Xie, C.; Lacroix, J. J.; Hao, J.; Luo, Y., Polymodal allosteric regulation of Type 1 Serine/Threonine Kinase Receptors via a conserved electrostatic lock. *PLoS Comput. Biol.* **2017**, 13, e1005711.
10. Jo, S.; Kim, T.; Iyer, V. G.; Im, W., CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, 29, 1859-65.
11. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. K., Comparison of simple potential functions for simulating liquid water *J. Chem. Phys.* **1983**, 79, 926-935.

12. Botello-Smith, W. M.; Alsamarah, A.; Chatterjee, P.; Xie, C.; Lacroix, J. J.; Hao, J.; Luo, Y., Polymodal allosteric regulation of Type 1 Serine/Threonine Kinase Receptors via a conserved electrostatic lock. *PLOS Computational Biology* **2017**, 13, e1005711.
13. Bozzo, E.; Franceschet, M., Effective and efficient approximations of the generalized inverse of the graph Laplacian matrix with an application to current-flow betweenness centrality. *arXiv preprint arXiv:1205.4894* **2012**.
14. Glykos, N. M., Software news and updates carma: A molecular dynamics analysis program. *Journal of computational chemistry* **2006**, 27, 1765-1768.