Gerstein Lab Contributions to the Prediction of Allosteric Residues

Using models of protein conformational change, we have developed a comprehensive mathematical framework that incorporates protein structure and dynamics for predicting allosteric residues both on the surface and in the interior (stress.molmovdb.org). Computational efficiency has been a priority in the development of this framework, thereby enabling high-throughput analysis for large protein datasets, as well as elucidating properties that are general to allosteric residues.

Given that knowledge of protein dynamics is so integral to this framework, we have also developed a pipeline for identifying alternative conformations of proteins throughout the PDB. The identification of likely allosteric residues within this set of dynamic proteins allows us to examine the biophysical and evolutionary features of the identified allosteric hotspots in a straightforward fashion. We have utilized this framework to generate and distribute a tool that enables users to submit their own structures for analysis. Several of the unique features include the fact that it is easy-to-use, computationally tractable, and capable of simultaneously identifying residues both at the surface and within the protein interior.

1. Our Work on Identifying Allosteric Residues on Protein Surfaces

To identify likely ligand binding sites, we use a modified version of the binding leverage method introduced by Mitternacht and Berezovsky.[1] This approach identifies cavities whose occlusion would interfere with large-scale motions. Once candidate sites for each protein are generated, we use both anisotropic network models (ANMs) and alternative crystal structures to generate models of conformational change. We then score each site based on the degree to which deformations at the site couple to the modeled conformational changes. High-scoring sites (i.e., sites at which occlusion strongly interferes with conformational change) constitute the predicted set of surface allosteric residues.

Our approach differs from previous ones in several key ways. First, our highly efficient implementation of this method enables more exhaustive Monte Carlo searches. In contrast to other techniques, we also use the heavy atoms of the protein when evaluating a ligand's affinity for each location, thereby generating a more selective set of candidate sites. In addition, we use principles from protein folding (specifically, the concept of energy gaps) in order to sensibly threshold the list of predicted sites. As a validation, we have implemented this method in order to predict known-ligand binding sites in well-studied systems.

2. Our Work on Identifying Critical Interior Residues via Dynamic Network Analysis

The framework described above captures hotspot regions at the protein surface, but residues in the interior may also play allosteric roles. These interior residues often act by functioning as essential 'bottlenecks' in the communication pathways between distal regions. Therefore, we use principles from network theory, in conjunction with our models of conformational change, to predict allosteric residues within the interior.

We model proteins as networks, wherein residues represent nodes and edges represent contacts between residues. Using this model, the problem of identifying interior-critical residues is thus reduced to a problem of identifying nodes that participate in network bottlenecks. We weigh edges according to the correlated motions of contacting residues; a strong correlation in the motion between contacting residues implies that knowing how one residue moves better enables one to predict the motion of the other, suggesting a strong information flow between the two residues. Then, using the motion-weighted network, we identify "communities" of nodes using the well-established Girvan-Newman formalism.[2] Finally, we calculate the betweenness of each edge, where the betweenness of an edge is the number of shortest paths between all pairs of residues that pass through that edge, with each path representing the sum of node-node 'distances' assigned in the weighting scheme above. Those residues that are involved in the highest-betweenness edges between pairs of interacting communities are identified as the interior-critical residues.

3. Our Public Software Tool for the Identification of Allosteric Residues (stress.molmovdb.org)

The implementations for finding both surface- and interior-critical residues have been made available to the scientific community through a new software tool, STRESS (for STRucturally-identified ESSential residues). This tool allows users to specify a PDB to be analyzed, and the output provided constitutes the set of identified critical residues. To magnify the impact of this work and to obviate the need for long wait times, we host this service on the Amazon cloud and use an extremely efficient algorithmic implementation.

4. Our Work on Identification of Alternative Conformations in Large Protein Datasets

Our framework for identifying potential allosteric residues assumes that these proteins undergo pronounced conformational changes. Therefore, to better ensure that the proteins studied exhibit well-characterized distinct conformations, we systematically identify instances of alternative conformations within the PDB. We perform multiple structure alignments for thousands of structures, with each alignment consisting of sequence-identical structures. Within each alignment, we cluster the structures using structural similarity to determine the distinct conformational states. This is accomplished through a combination of multidimensional scaling and a means of identifying the optimal number of groups in K-means clustering (i.e., the "K" value).[3] We then use information regarding protein motions to identify potential allosteric sites on the surface and within the interior.

References

1) Mitternacht, Simon, and Igor N. Berezovsky. "Binding leverage as a molecular basis for allosteric regulation." PLoS computational biology 7.9 (2011): e1002148.
2) Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12 (2002): 7821-7826.
3) N Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.