Quantifying Conformational Ensemble Changes in Proteins Using Inverse Machine Learning

Mohsen Botlani, Ahnaf Siddiqui and Sameer Varma

Department of Cell Biology, Microbiolgy and Molecular Biology
University of South Florida, FL-33620

Abstract

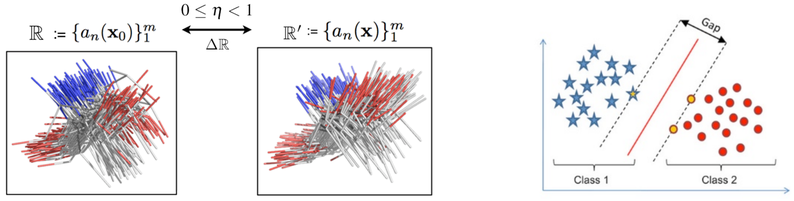
Background: Protein activities are regulated tightly in biological environments. An understanding of their regulatory mechanisms entails assessment of their various states, including active and inactive states. For many proteins, their states can be distinguished based on their minimum-energy conformations since, the magnitudes of thermal fluctuations, or dynamics, are negligible compared to the differences in minimum-energy structures. This approximation, however, breaks down for several other proteins. The states of these proteins can only be distinguished categorically from each other when their finite-temperature conformational ensembles are considered alongside their minimum-energy structures. The list of such proteins has grown rapidly in the last decade, which now includes GPCRs, PDZ domains, nuclear transcription factors, heat shock proteins, T-cell receptors and viral attachment proteins. Applicability of molecular simulations toward understanding mechanisms in this latter category of proteins requires development of new methods that can deal with high-dimensional conformational ensemble data.

Description: The traditional approach to compare protein conformational ensembles is to compare their respective summary statistics. However, if a subset of the summary statistics from the two ensembles is found to be identical, it does not imply that the remaining summary statistics will also be identical. The general problem of finding and choosing a feature that appropriately distinguishes ensembles can be overcome by comparing ensembles directly against each other and prior to any dimensionality reduction. We have developed a method to accomplish just that – it performs excellently for both Gaussian and non-Gaussian distributions. The difference between ensembles is computed by solving the inverse machine learning problem and in terms of a metric that satisfies the conditions set forth by the zeroth law of thermodynamics.

Conclusions: Such a quantification permits statistical analyses and quantitative data mining necessary for establishing causality in protein functional regulation. We have applied this method to (a) quantitatively understand the effect of ligand binding on the structure and dynamics of a viral protein whose function is controlled by dynamic allostery; (b) understand the role of water in the inception of allosteric signals; (c) determine intersecting signaling pathways. This method is available under standard GNU license on SimTk.**(https://simtk.org/projects/conf_ensembles).**

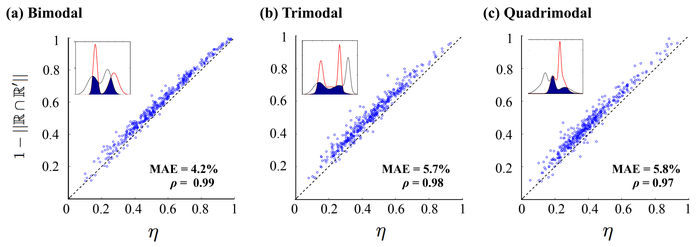
Inverse machine learning

$\mathbb{R} := \{a_n(\mathbf{x}_0)\}_1^m \quad \overset{0 \le \eta < 1}{\underset{\Delta\mathbb{R}}{\longleftrightarrow}} \quad \mathbb{R}' := \{a_n(\mathbf{x})\}_1^m$

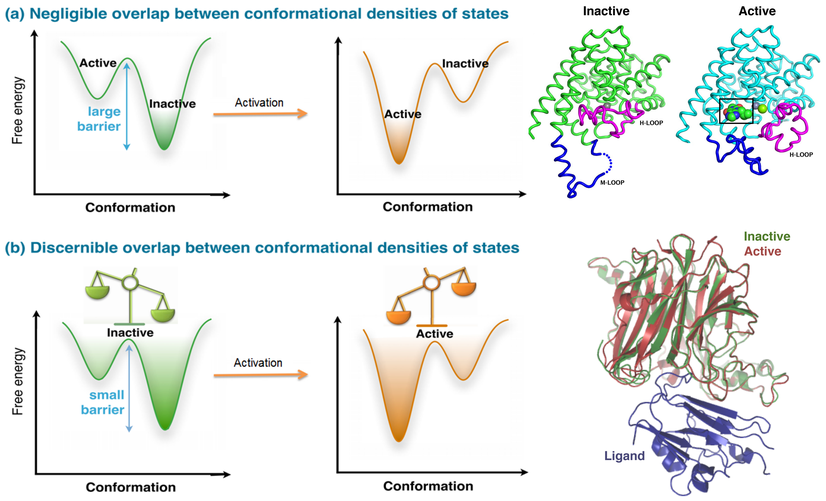


Traditionally, a support vector machine (SVM) is used for binary classification. It is first trained on a set of instances for which their group identities are known, and then used for predicting the group identities of unclassified instances. In our approach, we train the SVM to recognize the difference of two n-particle conformational ensembles, but instead of using the trained SVM for predictive purposes, we utilize the mathematical properties of the underlying classification function to obtain a physically meaningful quantitative estimate for the difference between the ensembles. The method is trained on Gaussian distributions, and works excellently without need for any data fitting. From a theoretical standpoint, the method should also work for multi-Gaussian distributions, and by extension, for any distribution, because the overlap between two multi-Gaussian distributions is essentially a sum of overlaps between Gaussian distributions,

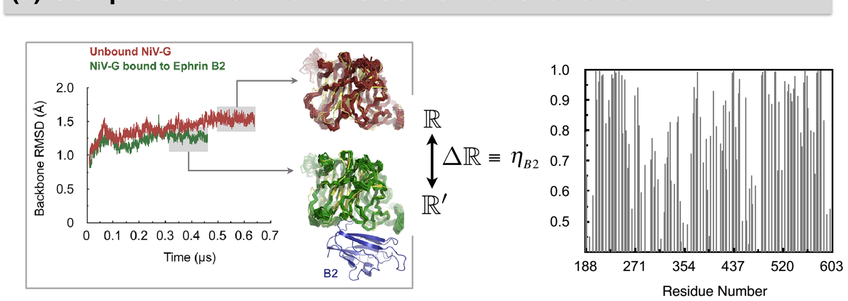
$$\eta = 1 - \|\sum_{i=1}^{n} c_i f_i \cap \sum_{j=1}^{n} c'_j f'_j\| = 1 - \|\sum_{i,j=1}^{n} c_i f_i \cap c'_j f'_j\|$$

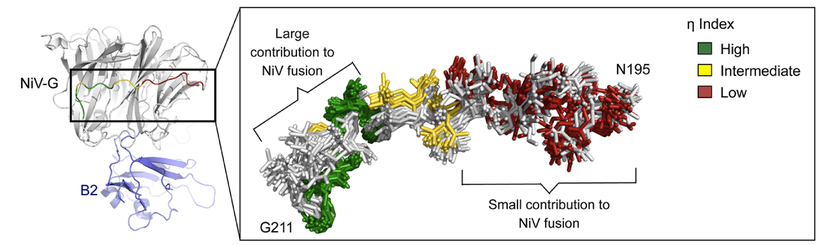


Functional Regulation via small structural changes



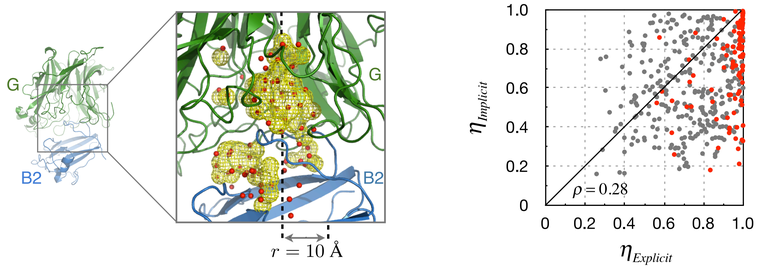
(a) Comparison between two conformational ensembles





(b) Comparison between two conformational ensemble shifts

Example: Effect of force field on ligand-induced conformational ensemble shifts. $\eta_{Explicit}$ and $\eta_{Implicit}$ are computed, respectively, from stochastic dynamics simulations in explicit and implicit solvent.



(c) Comparison between multiple conformational ensembles

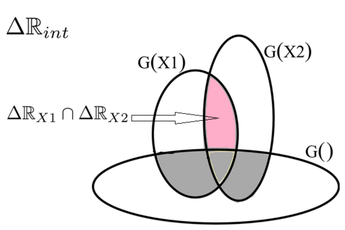
Intersecting signaling pathways: $\Delta\mathbb{R}_{signal} \subset \Delta\mathbb{R}_{int}$

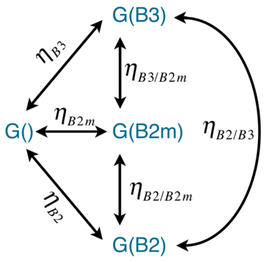
$\Delta\mathbb{R}_{int} := \Delta\mathbb{R}_{B2} \cap \Delta\mathbb{R}_{B3} \cap \Delta\mathbb{R}_{B2m}$

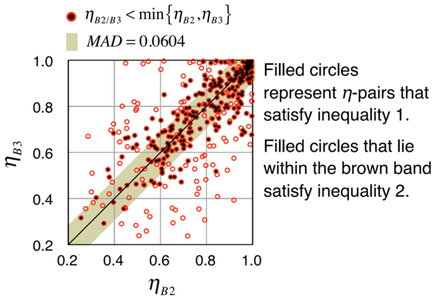
$\Delta\mathbb{R}_{X1} \cap \Delta\mathbb{R}_{X2}$ is defined by the inequalities:

1) $\min\{\eta_{X1}, \eta_{X2}\} > \eta_{X1/X2}$

2) $|\eta_{X1} - \eta_{X2}| < \langle \eta_{X1} - \eta_{X2} \rangle$



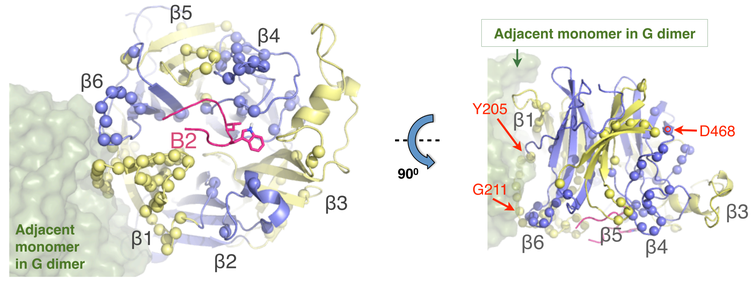




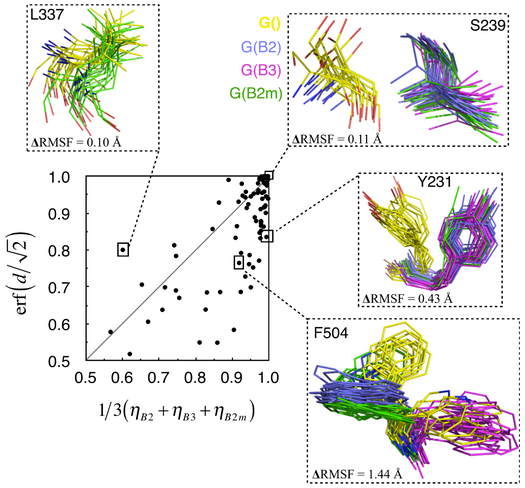
Filled circles represent η-pairs that satisfy inequality 1.

Filled circles that lie within the brown band satisfy inequality 2.

$\Delta\mathbb{R}_{int}$ **comprises of 106 residues,** which is ~25% of the residues in the G head domain.



Residues that are close to the diagonal undergo shifts primarily in backbone positions. Residues that lie below the diagonal undergo changes in side chain orientations and/or conformational entropy. Residues that lie above the diagonal represent cases where backbone deviations are swamped by smaller changes in whole residue deviations.



References

1. Leighty RE and Varma S. J Chem. Theory and Comput, 9: 868-875, 2013.
2. Varma S, Botlani M and Leighty RE. Proteins, 82: 3241-3254, 2014.
3. Dutta P, Botlani M and Varma S. J Phys. Chem. B, 118: 14795-14807, 2014.
4. Dutta P, Siddiqui A, Botlani M and Varma S. Biophys. J, 2016, Under revision.

Acknowledgments: All simulations were carried out at the Research Computing center of the University of South Florida.