Analysis of allosteric signals from MD simulations

It has long been known that molecular dynamics is subject to statistical erorrs introduced by temporal and physical approximations associated with the simulation. While empirical force-fields represent the best approximation of molecular interactions within a protein, a pressing issue which faces the simulation community are questions of (1) reproducibility and (2) convergence.

In order to estimate the sampling precision, it is commonplace to perform a number of independent replicates. Having a quantifiable measure of the similarity in conformational space sampled by multiple replicates of the sample system is crucial for understanding the reproducibility of measurables extracted from an MD simulation. Simply, one needs to answer the following: 'how likely is it that I will sample the same conformational space, if I repeat an MD simulation seeded from the same structure multiple times?'.

Secondly, ensuring convergence in biomolecular simulations is necessary to guarantee the quality of data. 'True convergence' is elusive to current computational capabilities given that to fully converge a protein would need to explore all possible conformational states, reversibly. This is difficult even for small peptide systems. As such, one can only hope to satisfy self-consistency checks. Several previous studies have shown that, even for a 1 ms simulation of BPTI, satisfiying self-consistency checks early in the simulation are misleading because they fail to anticipate conformational changes, which occur later on in the simulation.

In order to interpret allosteric signals emerging from a structural alphabet analysis of MD simulation of PKM2, it was necessary to first determine the level of noise associated with the mutual information score by averaging over a number of simulations. We also reasoned that it would not be sensical to average over multiple replicates of a simulation if they explored different unique conformational space or were uniquely exposed to conformational traps. This necessitated a quantitative comparison in order to determine which ergodic sectors were explored by the simulation replicates. Secondly, to determine the confidence associated with the allosteric signals extracted from the simulations, we needed a quantitative estimate of convergence, using similarity (or dissimilarity) of the mutual information matrix as an input.

Block covariance overlap method to measure egodicity of molecular dynamics simulation

The structural alphabet consists of 25 representative fragments of 4 consecutive Cα atoms, such that the protein conformation is reduced to a string of $n-3$ letters. A column of this alignment describes all of the conformational states sampled by a protein fragment along the simulation trajectory. Corrleation of conformational changes in a pair of protein fragments (i,j) was calculated as normalised mutual information:

$$I^n_{LL}(C_i : C_j) = \frac{I(C_i : C_j) - \epsilon(C_i : C_j)}{H(C_i : C_j)} \tag{1}$$

When the ensemble of protein conformations originates from an MD simulation, each of the eigenvectors resulting from the diagonalization of the mutual information matrix, describes a collective mode of motion that is not linearly correlated with any other in the system. The extent of this motion is given by the corresponding eigenvalue. This an eigenvalue decomposition of the mutual information matrix is a good way of assessing to what degree two simulations sample the same 'allosteric information space'.

One way to quantify the overlap between two matrices, is to perform an eigenvalue decomposition and then to compute the direction cosine between the eigenvector matrices:

$$\Psi_{A:B} = \frac{1}{n}\sum_{i}^{n}\sum_{j}^{n}(\mathbf{v}_i^A \cdot \mathbf{v}_i^B)^2 \tag{2}$$

Equation 2 is fairly permissive as a convergence criterium because the magnitudes along eigenvectors of the different modes are not considered, only the direction of the vectors. To expand on this, we can weight the direction cosine measure by the magnitude along the respective eigenvectors:

$$\Omega_{A:B} = 1 - \{\frac{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B) - 2\sum_{i=1}^{3N}\sum_{j=1}^{3N}(\lambda_i^A\lambda_i^B)^{0.5}(\mathbf{v}_i^A \cdot \mathbf{v}_i^B)^2}{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B)}\}^{0.5} \tag{3}$$

where λ^A and λ^B denote the eigenvalues of the mutual information matrices for simulations A and B, and N is the number of alpha carbon atoms in the calculations. We define this measure as the 'covariance overlap'.

The covariance overlap between two mutual information matrices

Test case 1: 10 ns MD simulation of deca-alanine in vacuum