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:

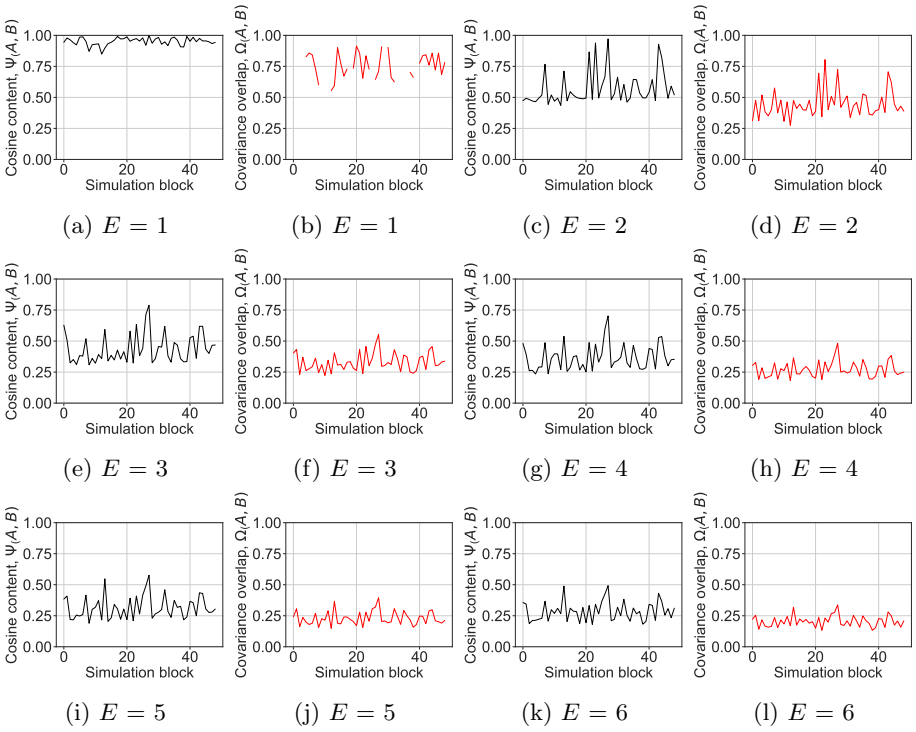
$$\Omega_{A:B} = 1 - \{\frac{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B) - 2\sum_{i=1}^{3N}\sum_{j=1}^{3N}(\lambda_i^A\lambda_i^B)^{0.5}(\mathbf{v}_i^A \cdot \mathbf{v}_i^B)^2}{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B)}\}^{0.5} \tag{3}$$

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

A 10 ns simulation of a deca-alanine peptide in vacuum was used as a test system for examining the behaviour of the time-evolution of mutual information matrices derived from a structural-alphabet representation of the trajectory. In particular, we wanted to test the numerical behviour of the cosine content and covariance overlap with different eigenmodes used in the calculation. The time evolution of both the cosine content and the cov-



(a) $E = 1$ (b) $E = 1$ (c) $E = 2$ (d) $E = 2$

(e) $E = 3$ (f) $E = 3$ (g) $E = 4$ (h) $E = 4$

(i) $E = 5$ (j) $E = 5$ (k) $E = 6$ (l) $E = 6$

Figure 1: The numerical behaviour of spectral subspace overlap with different eigenmodes

airance overlap decreases with increased numbers of eigenmodes included in the calculation. This is because a higher spectral variance is explicitly considered when more eigenmodes are summed over in equations (2) and (3). Also of note, the decreased subspace overlap converged after the first three eigenmodes. This convergent behaviour is consistent with the rapid decay in the eigenvalues, shown below. In this particular example of deca-alanine in vacuum for 10 ns at 300 K there is very little conformational motion in the peptide. This is reflected by the high degree of overlap between the

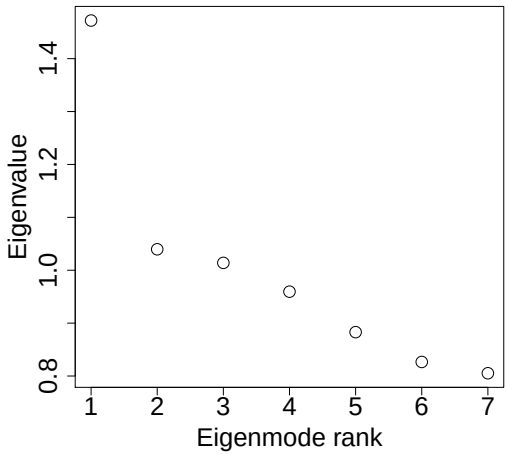


Figure 2: Scree plot of eigenvalue distribution of mutual information matrix

trajectory blocks, shown in figure 1. Furthermore, figure 2 indicates that the first eigenmode represents a significant amount of the positional variance in the trajectory and that subsequent eigenmodes contain negligable variance. To this point, the convergence in the subspace overlap measures ω and Ψ result from increased variance included in the measure. This raises an important statistical question: which eigenmodes are statistically signficant?

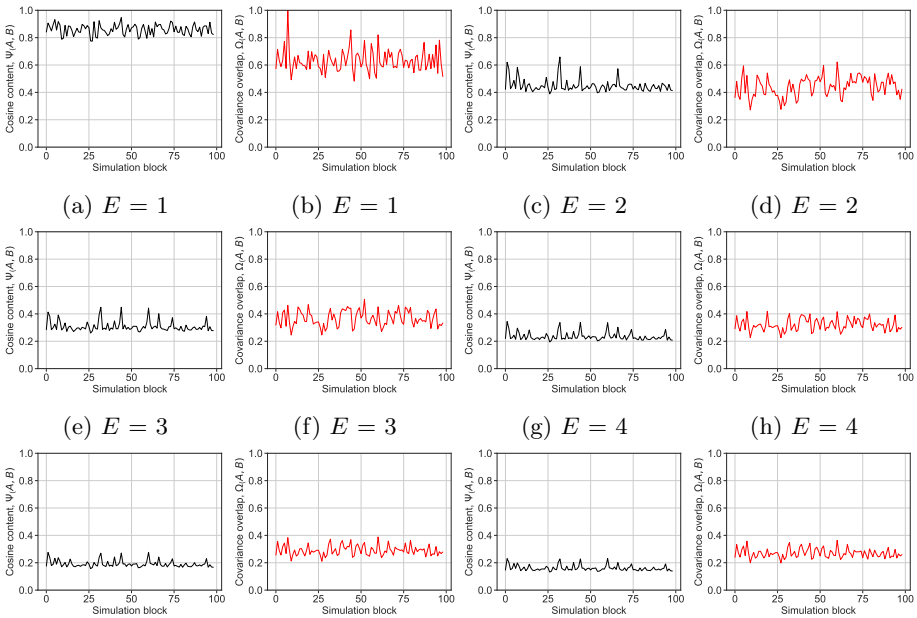
Going back to the initial problem of detecting subspace sampling of MD trajectories it became obvious that the deca peptide, which a good test case for assessing the numerical behaviour of ω and Ψ, was not very useful for interrogating whether or not the mutual information was sensitive to global conformational changes. To investigate this problem, we measured the subspace overlap of a more complex system - monomeric PKM2 bound to Tepp-46.

Test case 2: 1 μs MD simulation of monomeric PKM2 bound to Tepp-46

We have seen that monomeric PKM2 undergoes a significant ligand-dependent conformational rearrangement upon binding Tepp-46. Particularly, the B-domain cap closes over the catalytic pocket when the allosteric activator is bound and remains in an 'open' conformation when in the 'apo' state.

When Tepp-46 is bound, the B-domain cap closes at around 150 ns (not shown). Here, given that the protien undergoes a significant conformational rearrangement during the simulation we hypothesised that it would be a good test system to determine whether global conformational changes were accounted for my local correlations in the fragment scheme.

We started by measuring the subspace overlap for different numbers of eigenmodes, to investigate whether local correlations are sensitive to global conformational changes. Similar to the deca-alanine peptide, the subspace



(a) $E = 1$ (b) $E = 1$ (c) $E = 2$ (d) $E = 2$

(e) $E = 3$ (f) $E = 3$ (g) $E = 4$ (h) $E = 4$

(i) $E = 5$ (j) $E = 5$ (k) $E = 6$ (l) $E = 6$

Figure 3: The numerical behaviour of spectral subspace overlap with different eigenmodes for monomeric PKM2 + Tepp-46. The MD trajectory was split into 100 blocks of 10 ns. The subspace overlap was calculated for each blocks i_n and i_{n-1}

overlap explored by the mutual information matrices of monomeric PKM2 + Tepp-46 was normally distributed about an average values of $\Omega_{A:B}$ and $\Psi_{A:B}$ and did not deviate as a result of the closure of the B-domain cap (shown below). Similarly, $\mu(\omega_{A:B})$ and $\mu(\psi_{A:B})$ decay to a convergent value, with increasing number of eigenmodes. From this, it appeared that the time evolution of backbone correlations calculated using the structural alphabet

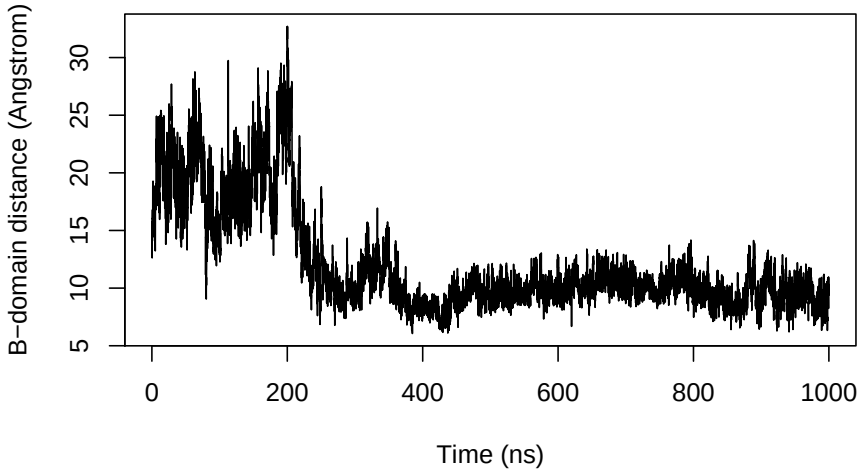


Figure 4: The B-domain closed over the catalytic pocket of monomeric PKM2 when Tepp-46 is bound to its N-terminal allosteric pocket.

approach was not sensitive enough to pick up global structural changes. This is perhaps to be expected, because the mutual information matrices are a statistical measure of distal local correlations. Although an eigenvalue transformation of the mutual information matrices did not explicitly give information on global structural changes, we returned to the original problem of extracting meaningful correlations from the mutual information matrices. A critical question regarding the use of correlation signals is whether the signal strength observed. To filter out residual couplings from those reflecting an allosteric signal we used a random matrix theoretical framework. The fundamental aim is to detect those fragment couplings, which represents a true allosteric signal (ie. non-random). There is significant noise in the fragment couplings, because only some of the couplings are truely informative in terms of their contribution to allosteric signal transmission. A fundamental result from random matrix theory describes the eigenvalue distribution of a correlation matrix analytically - under certain assumptions, if entries are drawn from a Gaussian distribution with zero mean and unit variance, the probability of having an eigenvalue λ is given by the canonical

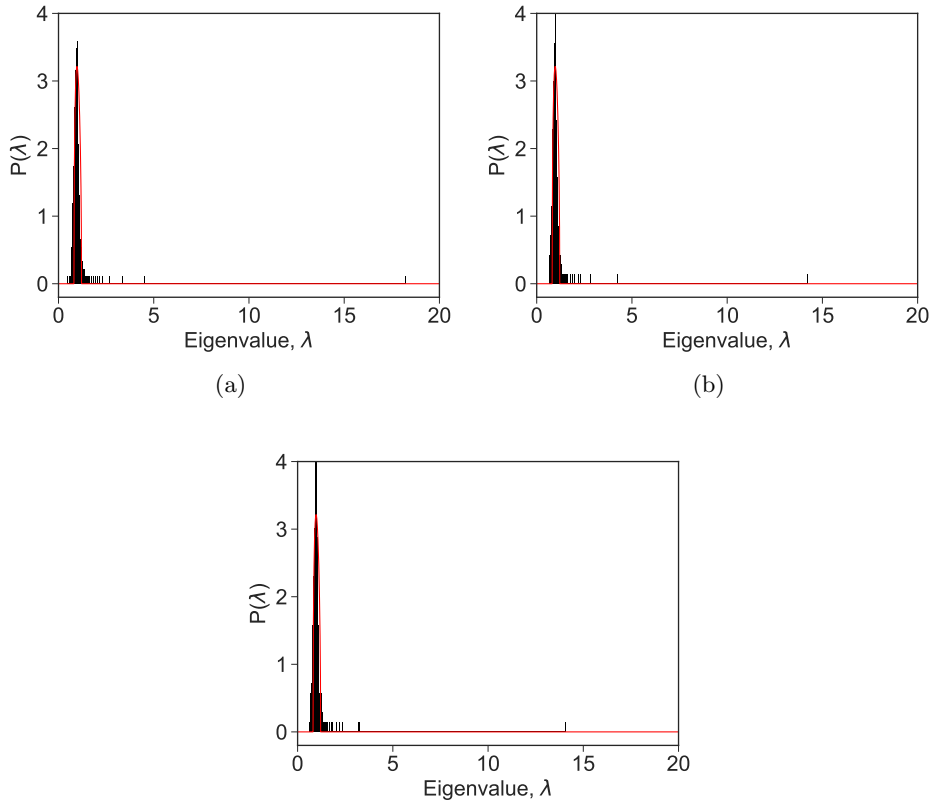


Figure 5: Marcenko-Pastur distribution plots for spectral decompositions of mutual information signals for three independent MD trajectories of monomeric PKM2 + Tepp-46.

Marcenko-Pastur distribution:

$$P(\lambda) = \frac{\sqrt{((1+\sqrt{\gamma})^2 - \lambda)(\lambda - (1-\sqrt{\gamma})^2)}}{2\pi\gamma\lambda} \tag{4}$$

where γ describes how well-sampled the dataset is. The probability that a random matrix has eigenvalues larger than $1 + \sqrt{\gamma}^2$ in the absence of any signal is very low. As a result, those eigenvalues above $1 + \sqrt{\gamma}^2$ correspond to statistically significant signals.