Dynamic Feature Extraction from Molecular Mechanics Trajectories

Han Altae-Tran, Muneeb Sultan[*], Vijay Pande[†]

December 14, 2013

Introduction

With the recent advancements in distributed computing, it is now possible to simulate molecular dynamics of large systems over unprecedented time scales. However, due to the sheer size of the generated trajectories, it is no longer possible to reliably extract relevant chemical and biological information from a simulation through visual inspection alone - a single trajectory alone may consist of well over a thousand of atoms, charted over hundreds of time points. Currently, Markov state models are being used to compute a discretization of the conformational space of large proteins. However, one current challenge lies in understanding the transitions between the states. To address this problem, we consider restoring the dynamic systems perspective and look for methods capable of detecting transition mechanisms in this context.

Since physical interactions between atoms are typically proportional to distances between corresponding atoms, we are particularly interested in analyzing pairwise atomic distances. Unfortunately, the space of pairwise distances scales with the number of atoms squared. It is therefore desirable to develop methods that reduce the dimensionality of the space into readily interpretable coordinates that largely account for variations in the data.

While offering a promise of fewer dimensions, conventional methods, such as Principal Components Analysis (PCA), often produce difficult to interpret coordinates, as each of the coordinates mixes a significant number of atoms that makes it difficult to definitely attribute motion along a principal component to motion in real space. Furthermore, the assumption of linearity is weak in the context molecular dynamics. Similar problems arise for factor analysis. Finally, many of these techniques are not readily applicable to time series data because measurements taken close in time cannot be considered independent samples from the distribution of atomic configurations.

Here we seek to address this issue by proposing an unsupervised feature selection method for a system of time series that incorporates basic time dependent relationships within the system. For the application, we will be working with pairwise distances from single molecular dynamics trajectories; however, this method should be applicable to other feature spaces and possibly other fields.

[*]Department of Chemistry, Stanford

[†]Department of Chemistry, Computer Science, and Structural Biology, Stanford

Methods

Consider a system of n possibly related processes, such as the set of pairwise distances in a protein, where the relationships between the processes are undetermined, and could possibly depend on time and/or the state of the system. However, for a narrow time range, and small set of system states, it may be possible to uncover basic relationships between the processes in this limited domain of operation. In the context of large scale transitions between states in a system, we are therefore interested in detecting processes that exhibit anomalous behavior in the beginning or even possibly before a large scale transition of the system. One possibility for detecting this type of relationship involves lagging one process relative to another and computing a heuristic that determines whether the first process exhibits anomalous behavior that is followed by anomalous behavior in the second process, essentially providing a notion of directionality between the two processes.

Arguably there are other methods for detecting such variation within a system, with principal components analysis being one of the more famous examples. However, many of these methods are not catered toward time series data and do not necessarily address the issue of detecting anomalous behavior that proceeds transitions. We will save this discussion for the end.

Directionality Measure

Consider a sample of the n processes, $\left\{X^i_{t_k}\right\}_{i=1,\ldots,n}$, over the equally spaced times $t_1, \ldots, t_{T_f}$. Because we are interested in anomalous behavior (deviations from standard behavior), we mean subtract each of the sampled processes using a moving average of global parameter σ. The moving average is essentially a causal smoothing filter applied to the data set (causal meaning that only time points behind a given time are included in the smoothing). Denote the moving average subtracted sampled processes as $Y^i_{t_k} = X^i_{t_k} - \left\langle X^i_\sigma \right\rangle_{t_k}$, where $\left\langle X^i_\sigma \right\rangle_{t_k}$ is the moving average of $X^i_{t_k}$ with parameter σ. A moving average is preferred over a complete average because despite being constrained to a limited range of operations, the processes might posses time dependent average values. For an example, consider two atoms that suddenly move close together at some time t_k and remain fixed around this new distance for the remainder of the sampled times. Furthermore, this allows a more natural comparison between rapid switch like motion and step like motion (e.g. the previous

example).

With mean subtraction in place, we can now consider the cross correlation between two sampled processes, X^i and X^j:

$$C_{ij}(\ell) = Y^i \star Y^j(\ell) \sum_{k=-\infty}^{\infty} Y^i_{t_k} Y^j_{t_{k+\ell}}$$

where we have defined $Y^i_{t_k} = 0$ when $k \notin \{1, \ldots, T_f\}$. In the idealized continuous setting, the cross correlations amounts to the integral over t of $Y^i_t Y^j_{t+\ell}$, which computes the overlap between Y^i_t and Y^i_t when the latter is lagged in time by the amount ℓ. We create a normalized cross correlation function, $C'_{ij}(\ell)$ so that $\left|C'_{ij}(\ell)\right| \leq 1$ for all ℓ and $C'_{ii}(0) = 1$ for all i. This ensures that cross correlations of different pairs i_1, j_1 and i_2, j_2 are more naturally comparable. Now, given two processes (i, j), we would like to define a notion of directionality, z_{ij}, between the pair. We will base the idea on the max one sided cross correlation,

$$c_{ij} = \max_{\ell \leq 0} \left|C'_{ij}(\ell)\right|$$

which is maximum possible correlation between $Y^i_{t_k}$ and $Y^j_{t_k}$ when the former is lagged in time. The core idea is that if both $Y^i_{t_k}$ and $Y^j_{t_k}$ display anomalous behavior, a lagged $Y^i_{t_k}$ should substantially overlap $Y^j_{t_k}$ if $Y^j_{t_k}$ displays this behavior behavior before $Y^j_{t_k}$ does. However, using c_{ij} to describe the relationship structure between the sampled processes will lead to many problems. The first is the lack of distinguishable directionality between sampled processes - the difference between c_{ij} and c_{ji} may be small even if c_{ij} is large. This reflects an overall inability to distinguish between the two notions $Y^i_{t_k}$ proceeds $Y^j_{t_k}$ (denoted $Y^i_{t_k} \to Y^j_{t_k}$) and $Y^j_{t_k} \to Y^i_{t_k}$. Furthermore, there is the added problem of spurious noise alignment, whereby noise from two series line up in one lag direction, but not the other. Lastly, there is the problem of correlated motion giving rise to large c_{ij}, c_{ji}, and $|c_{ij} - c_{ji}|$. To address these issues we introduce two multiplicative factors for c_{ij},

$$\Delta_{ij} = |c_{ij} - c_{ji}|$$

and

$$\gamma_{ij} = \left[1 - \mathrm{corr}_\ell \left(C'_{ij}(\ell),\, C'_{ji}(\ell)\right)\right]^2$$

where corr_ℓ indicates the standard linear correlation over ℓ. The first factor quantifies how lopsided the difference between the two notions, $Y^i_{t_k} \to Y^j_{t_k}$ and $Y^j_{t_k} \to Y^i_{t_k}$, are. Therefore, if $\Delta_{ij} \gg 0$, then the c_{ij}, c_{ji} pair is a non trivial comparison. Note that this factor substantially reduces c_{ij} between two processes that are highly correlated at low lag time. In the context of molecular mechanics, this amounts to reducing the effects of atoms that consistently move together. γ_{ij}, on the other hand, determines how dissimilar $Y^i_{t_k} \to Y^j_{t_k}$ and $Y^j_{t_k} \to Y^i_{t_k}$ are in the context of all possible lag times. If $C'_{ij}(\ell)$, and $C'_{ji}(\ell)$ are highly correlated, then increasing lag times produces the same type of change in $C'_{ij}(\ell)$ and $C'_{ji}(\ell)$, which imply that the two notions $Y^i_{t_k} \to Y^j_{t_k}$ and $Y^j_{t_k} \to Y^i_{t_k}$ are reversible. Therefore $\gamma_{ij} \gg 0$ when the the $C'_{ij}(\ell)$ and $C'_{ji}(\ell)$ are uncorrelated (or even anti correlated which implies strong irreversibility). Together, these two factors help ensure that the c_{ij} are meaningful.

We can now proceed to define the directionality between $Y^i_{t_k}$ and $Y^j_{t_k}$ as

$$z_{ij} = c_{ij} \Delta_{ij} \gamma_{ij}$$

which sets up a relationship structure between the n processes that we can proceed to analyse. However, a potential problems might arise at this stage, and we address a few of them below.

Noisy Data

If the sampled processes are especially noisy, the cross correlation will be as well, it might be necessary to smooth $C'_{ij}(\ell)$ with a moving average of window h. This will prevent spurious maxima from appearing, the idea being that $\ell^*_{ij} = \arg\max_{\ell \leq 0} \left|C'_{ij}(\ell)\right|$ represents the location of true maximum of $\left|C'_{ij}(\ell)\right|$ if neighboring values of ℓ also demonstrate large values of $\left|C'_{ij}(\ell)\right|$. Note that we are not necessarily interested in determining a *lag time*, τ_{ij}, that gives a definitive time delay in the relationship between the two processes. First off, a constant time delay between the two processes may not be adequate in describing the relationship. Moreover, we are more interested in the question of whether or not $Y^i_{t_k}$ proceeds $Y^j_{t_k}$, which can be dealt with here without explicit reference to lag times. As a result, smoothing the cross correlation function poses no significant difficulty to the task at hand.

High Dimensionality

Another issue might arise when we n is very large, as computing z_{ij} requires $O(n^2 T_f)$ calculations. If we have a prior notion of when some of the $Y^i_{t_k}$ do not contribute to system in the domain of operation, then we can remove these sampled processes from the analysis. In the context of molecular mechanics, such a notion would be encapsulated in the idea of a stationary, low variance sampled process, as such a process would not be changing throughout the domain of operation (and is hence a constant along the domain). This is not to say it is not useful in the system as a whole, just that it does not contribute in the context of dynamics in the queried domain. For example, distances between covalently bonded atoms usually do not contribute significantly to a conformational change in a protein, since this is structurally fixed. One way to quantify this notion is the use of a signal energy. High signal energies of mean subtracted processes typically correspond to high amplitude oscillations, and hence signify a non constant signal. One last point should be made that in the context of molecular mechanics, distances between atoms matter. An oscillation of 1 angstrom between atoms that are 100 on average angstroms apart is not nearly as important as an oscillation of 1 angstrom between atoms that are 5 angstroms apart. To account for this, we define a *relative*

fluctuation, given by

$$\tilde{Y}^i_{t_k} = \frac{Y^i_{t_k}}{\langle X^i_\sigma \rangle_{t_k}}$$

which weights the mean subtracted sampled process at a time t_k, $Y^i_{t_k}$, using the sampled process mean estimate at time t_k, $\langle X^i_\sigma \rangle_{t_k}$. Sampled processes with large deviations relative to the mean estimate have large relative fluctuations. Small relative fluctuations amounts to little relative motion. In the context of molecular mechanics, the latter is seen in the distances between structurally stable components. To quantify the presence of large relative fluctuations in a given process, $Y^i_{t_k}$, we compute the signal energy of the relative fluctuations, which we call the relative fluctuations energies,

$$\begin{aligned} E_i &= \langle \tilde{Y}^i_{t_k}, \tilde{Y}^i_{t_k} \rangle \\ &= \sum_{k=-\infty}^{\infty} |\tilde{Y}^i_{t_k}|^2 \end{aligned}$$

and then place a cutoff on these energies depending on which sampled processes one is interested in keeping. once can then curate a training set of sampled processes to keep and sampled processes to discard and use a logistic regression trained on this set to automatically keep or discard each of the remaining sampled processes.

Analysing the Directionality Structure

The structure imposed by the z_{ij} can be thought of in the context of directed graphs, where z_{ij} and z_{ji} provide the two asymmetric weights between the i^{th} and j^{th} nodes (sampled series). The usefulness in this interpretation of the z_{ij} lies in graph partitioning, whereby we split the directed graph into representative subgraphs, each of which represents some distinct source or destination of motion within the system under the limited domain of operation. To make this explicit, let Z be the adjacency matrix formed by $Z_{ij} = z_{ij}$. If we are interested in partitioning the graph described by Z into K representative subgraphs, roughly, we would attempt to pick K subsets of the n nodes that maximize the flow (directionality) across these subsets. In the more usual language, we seek K clusters of the n points (sampled series). Unfortunately much of the modern clustering work addresses symmetric directed graphs, where the weight between two nodes can be likened to a distance. While there are methods for clustering asymmetric weighted graphs, we will instead take the approach of graph symmetrization which will allow us to use the more familiar methods on some symmetrization of Z. The most common symmetrization is $Z + Z^T$, which essentially ignores the directed structure of the original graph by averaging z_{ij} and z_{ji} to form the symmetric weights between the i^{th} and j^{th} nodes. In the context of the above directionality measure, this is not a particularly helpful symmetrization, though it is widely used for other problems. A less commonly found symmetrization is the bibliometric symmetrization,

$$Z^{\text{sym}} = ZZ^T + Z^TZ$$

. Since the i^{th} row of Z represents the outflow of the i^{th} node, $(ZZ^T)_{ij}$ is the shared outflow of the i^{th} and j^{th} nodes. Similarly, $(Z^TZ)_{ij}$ is the shared inflow of the i^{th} and j^{th} nodes. The symmetrization Z^{sym} therefore relates two nodes, i and j, if they exhibit highly shared flow to the rest of the graph, even if they do not exhibit flow between them.

The resulting symmetrization Z^{sym}_{ij} can be used as a similarity matrix for a symmetric weighted graph clustering problem. Various clustering techniques exist, such as the ubiquitous k-means clustering. However, here we are interested in finding connected components in the Z^{sym} induced structure, rather than centroids because not all nodes in a cluster need to be closely related to one another for the nodes as a whole to represent a distinct concerted motion within the system. To achieve this, we make use of a technique called spectral clustering, which projects the data into a lower dimensional subspace using the eigenvectors of the normalized graph Laplacian matrix, and then clusters the nodes in the reduced dimensional space using k-means. If the graph is split into many disconnected components, the graph Laplacian will be roughly of block diagonal form, and consequently its eigenvectors will have nonzero entries corresponding to combinations of the blocks. The nodes in the lower dimensional space are then separated according to the disconnected component they belong in. Therefore, disconnected components within the graph may be detected using spectral clustering, resulting in K clusters $Q_1, \ldots, Q_K$.

Cluster Aided Feature Selection

Finally, we would like to select representatives from each of the clusters $Q_1, \ldots, Q_K$ to constitute a reduced set of features that explains the different types of directed motion within the system in the domain of operation. Here we seek to optimize some function $f(i)$, $i \in Q_q$ for each cluster Q_q, with $\arg\max_{i \in Q_q} f(i)$ corresponding to the representative node i^*_q of Q_q. Depending on the goals (e.g. detecting sources vs destinations etc.), different choices of f may be made (such as letting $f(i)$ be the eigenvector centrality of the i^{th} node in Z^{sym}). For our purposes, where we are looking to detect sources of motion, we take f to be the influence of node i on the graph, that is the difference of squared outflows and inflows, that is

$$f(i) = \sum_{j=1}^{n} \left(Z^2_{ij} - Z^2_{ji} \right)$$

which rewards large outflow and penalizes large inflow, while neglecting small inflow/outflow that might not substantially affect the relationship of the i^{th} node to the system as a whole. Finally, we select the representative nodes, i^*_q, using by maximizing f over the clusters Q_q, and order the i^*_q in the order of decreasing $f(i^*_q)$. The top representative nodes are therefore those that represent the most substantial sources of motion within the system. We now proceed by testing the effectiveness of this method on molecular mechanics data.

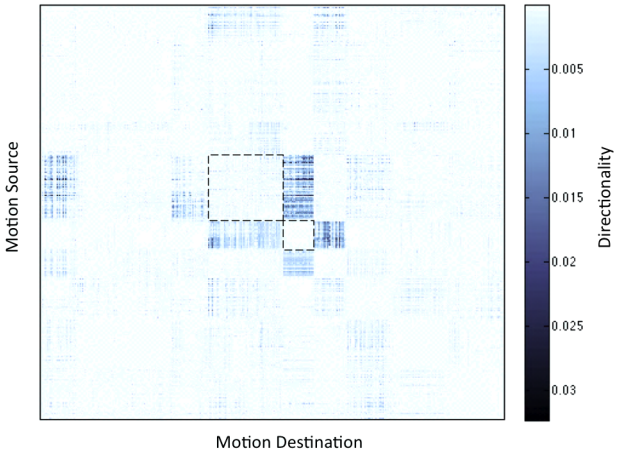

Figure 1: Heatmap of the spectral clustering results. The $(I, J)^{\text{th}}$ entry corresponds to the directionality measure $z_{I,J}$ between the I^{th} pairwise distance and the J^{th} pairwise distance. The two clusters with high outflow relative to inflow are boxed.

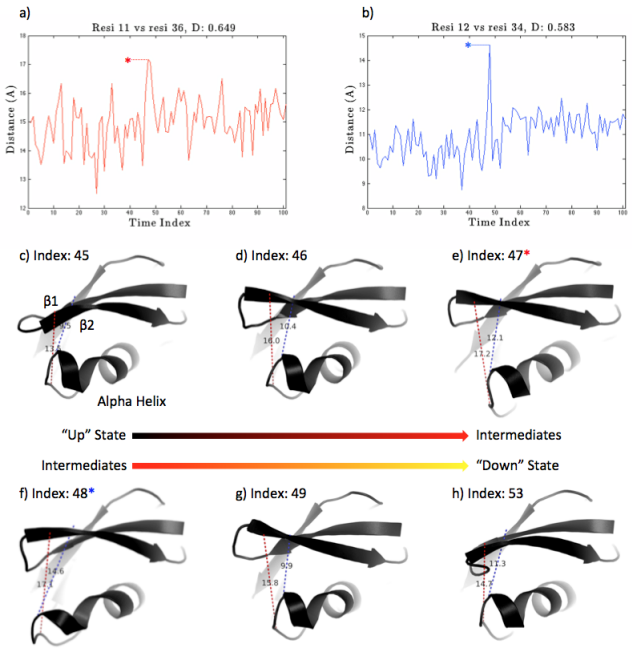

Figure 2: Results of cluster based feature selection. a,b) Time series plots of the top two pairwise distances scored by $D = f(I)$. c-h) Top two pairwise distances are mapped onto the ubiquitin structure at the given time indices. The intial ubiquitin up state is shown in c), with the final ubiquitin down state shown in h). The red and blue asterisks shown in a) and b) occur at the times represented by e) and f) respectively.

Application to Molecular Mechanics

Here we present an application of the above unsupervised learning method to a simulated transition between the β_1-β_2 region (residues 6-10) up state and down state of ubiquitin (see fig 2c, fig 2h for the up state and down state respectively). To reduce the initial dimensionality, we only consider the alpha carbons (backbone carbons) within the protein. The space of all pairwise distances between alpha carbons is then considered, where a sampled process, $Y_{t_k}^{[i,j]}$, is now the pairwise distance between the alpha carbon of residue i and the alpha carbon of residue j computed at time t_k. Relative fluctuations energies are calculated and used to filter out pairwise distances that are relatively constant, and the noisy data protocol of smoothing of the cross correlation function by a window of 10 frames is employed to reduce the effects of thermal noise on the cross correlation function. Z_{ij} is then computed and symmetrized to form Z^{sym}, which is then split into $K = 10$ groups using spectral clustering. The results of spectral clustering can be seen in figure 1, which is a heatmap of Z_{ij} arranged by cluster. Notice that there are only two clusters with high outflow compared to inflow, and two notable clusters with higher inflow than outflow, while the remaining 6 clusters are relatively inactive. Therefore, in the context of this clustering, we consider the top two pairwise distances, as these correspond to the two clusters that are, in the network sense, sources of motion. We use $f(i)$ to select the representative pairwise distances, $Y_{t_k}^{[i,j]_q^*}$ from each cluster Q_q, $q = 1, \ldots, K$. The top two representatives are the distance between Lysine11 and Leucine36 ($Y_{t_k}^{[11,36]}$), and the distance between Threonine12 and Threonine34 ($Y_{t_k}^{[12,34]}$), shown in figure 2 a,b. The implicitly sought after anomalous behavior for the two pairwise distances are shown with asterisks. The two distances are mapped to the cartoon representation of the structure at various times during the transition in figure 1 c-h. What $Y_{t_k}^{[11,36]}$ reveals is the preparatory rearrangement of the loop region below alpha helix - the loop moves down starting at index 46, and reaching a maximum distance from the β_1-β_2 region at time index 47. This is followed by an extremely rapid increase in $Y_{t_k}^{[12,34]}$ at index 48, which can be contextually understood as the rapid expansion of the alpha helix. This event is followed by a rapid contraction of the alpha helix at index 49, after which the system settles into the down state conformation. The top two selected pairwise distances therefore capture anomalous behavior that proceeds the system transition.

Regarding the number of clusters, K, when we use $K = 5, \ldots 15$, the single top pairwise distance is consistently $Y_{t_k}^{[11,36]}$ (which will follows from how we determine the i_q^*), while the second pairwise distance is either the distance from the β_2 sheet to the tip of the alpha helix, or the distance from the end of the alpha helix to the tip of the alpha helix (i.e. the alpha helix length). In either case, both of these latter two distances captures the expansion/contraction event of the alpha helix.

Retrospectively we may also investigate the effectiveness of z_{ij} in inducing a directionality structure between the pairwise

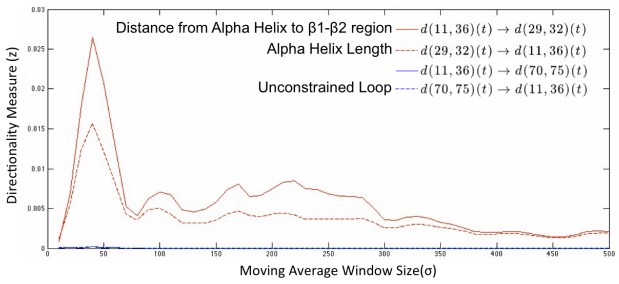

Figure 3: Directionality measure between pairwise distances expected to have high directionality (red) and between pairwise distances not expected to have high directionality (blue) computed for different moving average window sizes.

distances. Evidently, the loop below the alpha helix moves before the alpha helix does, which, truth be told, was not obvious by visual inspection alone. We should therefore expect the notion $Y_{t_k}^{[11,36]} \to Y_{t_k}^{[12,34]}$ to be stronger than the opposing notion $Y_{t_k}^{[12,34]} \to Y_{t_k}^{[11,36]}$. Furthermore, there is an unconstrained loop with very high positional variability present in the simulation, and we do not expect the tail end of this loop to contribute much to the transition, nor do we expect the transition to much affect the behavior of the loop. To quantify this we introduce $Y_{t_k}^{[70,75]}$, which measures the distance from the base of the loop to the end of the loop, and examine $Y_{t_k}^{[11,36]} \to Y_{t_k}^{[70,75]}$ and $Y_{t_k}^{[70,75]} \to Y_{t_k}^{[11,36]}$ (both of which should correspond to low directionality measures). To test that our directionality measure faithfully reproduces these relationships, we compute z_{ij} for these comparisons, for a variable moving average window size, σ. The results are plotted in figure 3. Notice that regardless of the window size, $z_{[11,36],[12,34]} > z_{[12,34],[11,36]}$, which tells us that we can accurately reproduce the notion that anomalous behavior of$Y_{t_k}^{[11,36]}$ proceeds that of $Y_{t_k}^{[12,34]}$. Furthermore, note that as expected, the unconstrained loop exhibits low outbound or inbound directionality.

For future work, we will consider using functional atoms on amino acid side chains (rather than just alpha carbons), granting us a view into the mechanistic structures that underlie conformational changes. Furthermore, we might consider training our directionality measure on synthetic data of known lag times to obtain a more discriminative directionality measure of the form $z_{ij} = c_{ij}^{\kappa_1} \Delta_{ij}^{\kappa_2} \gamma_{ij}^{\kappa_3}$, where the κ_k are parameters to be trained.

We close our discussion with a brief comparison to PCA (which is actually used quite frequently in the context of molecular mechanics). Once again, we consider the set of processes, $Y_{t_k}^{[i,j]}$, treating each $[i,j]$ as a dimension, and each $Y_{t_k}^{[i,j]}$ as an observation for the dimension $[i,j]$. We then proceed by performing PCA on the time pooled data where each dimension $[i,j]$ is now populated by the sample points $\{Y_{t_k}^{[i,j]} : k = 1, \ldots, T_f\}$. The results are shown in figure 4. As expected, PCA is unable to even detect the transition, as the data points are not separated in time when projected

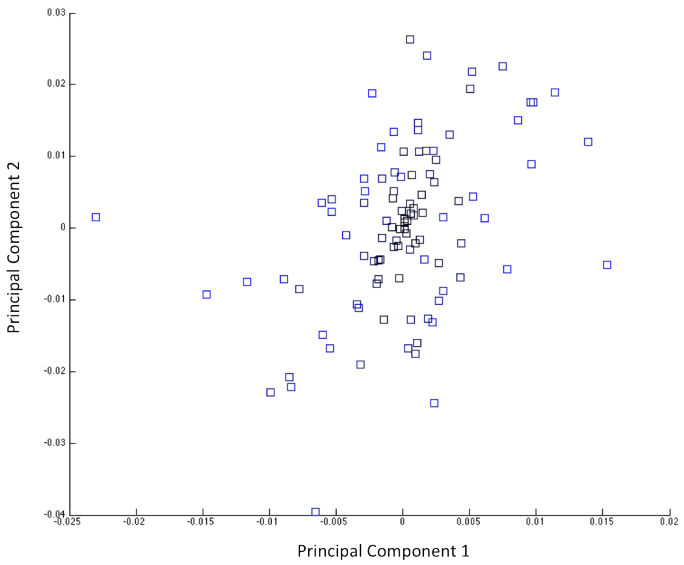

Figure 4: First two principal components of the simulated ubiquitin transition. The color of the points transitions from black to blue along the simulated protein trajectory

to this reduced dimensional space. Instead, PCA explains the difference between the up state and down state (without regard to intermediate states), as well as the motion of the unconstrained loop. The method we propose, on the other hand, is more explicitly equipped for time series analysis and irrelevant motion. Consequently, it is relatively successful in elucidating the types of motion that occur during system level transitions. Moreover, this method is at the heart a feature selection technique, and therefore does not mix any of the coordinates. This is important because it allows us to determine functionally relevant atomic pairs. In the context of other proteins, this might provide information regarding a hydrogen bond, or a salt bridge, that is critical to a particular conformational change. With additional refinements refinements, this method might even be helpful in understanding the transitions between protein conformations in a setting, such as drug development, where innumerable protein transitions need to be understood by a small number of researchers.

References

[1] Satuluri V, Parthasarathy S *Symmetrizations for clustering directed graphs.* Proceedings of the 14th International Conference on Extending Database Technology. 2013; 343-354.

[2] Zhang Y, Zhou L, Phillips AH, et al. *Conformational stabilization of ubiquitin yields potent and selective inhibitors USP7.* Nat Chem Biol. 2013; 9(1):51-8

[3] Teodoro ML, Phillips GN Jr, Kavraki LE *Understanding protein flexibility through dimensionality reduction.* J Comput Biol. 2003; 10(3-4):617-34