Inferring Kinetic Information from Unordered Images of Proteins

Min Cheol Kim (mincheoly@stanford.edu)
Christian Choe (cachoe@stanford.edu)
TJ Lane (tjlane@stanford.edu)

Introduction and Background

Our previous worked focused on reconstructing linear protein dynamics from unordered images of protein conformations. From this project, we learned that using the Isomap algorithm is effective in estimating a low dimensional representation of the protein conformations, in which the true trajectory of the protein dynamics is more obvious. We were able to reconstruct these linear trajectories by first performing dimensionality reduction via the Isomap algorithm, then by applying a relatively simple graph algorithm to find a linear path through the Isomap space.

However, proteins in nature do not simply behave in a linear fashion; their behavior is more of a graph, where the nodes represent the different protein conformations and the edges probability measures of transitioning from the source node to the sink node on either ends of the edge. In this report, we are attempting to recover the kinetic information from scrambled trajectories; in other words, given a set of unordered protein conformations, we try to 1) identify the representative, stable, states in our data and 2) infer the kinetic relationship between these states.

However, this linear approach makes the assumption that there is only a single behavior of the protein that we may find interesting. If there are two distinct dynamical processes occurring, we would want to identify both of these directions. In this report, we continue to recover kinetic information from scrambled image of proteins. Specifically, we would like to find functions that identify these distinct dynamical processes when applied to raw protein structures.

The current state-of-the-art technique for analyzing large sets of molecular dynamics (MD) simulations (where the images are ordered and time information is available) finds such functions using time structured Independent Component Analysis (tICA) and Markov State Models. In this report, we compare the performance of our techniques on tICA computed from the true trajectories.

Techniques

Principal Components Analysis

Principal Component Analysis (PCA) is a technique to find the direction in the given input space that maximizes the variance of the data upon projection. The use of PCA in studying protein dynamics has been extensively studied, although not specifically for our purpose. In general, when PCA is used to analyze MD datasets, PCA is applied to neighboring regions of the produced trajectory, and the use of PCA on the global dataset does not match the performance of tICA.

However, in our situation, any technique we use must involve the entire dataset. The particular order of our dataset is meaningless, since the frames are scrambled snapshots of proteins.

Isomap

Isomap is a procedure for finding a low dimensional embedding of high dimensional data points. Isomap computes the lower dimensional coordinates by first computing the geodesic distances between each point, or "hops" it requires to get one point to the other based on a k-nearest-neighbors metric. Then, it finds a smaller set of descriptors that still preserves these geodesic distances as best as possible.

This technique has shown promise before in both linear reconstruction and finding the folding coordinate of FsPeptide, as previously discussed in our Winter Quarter Report. Isomap requires a relatively dense sample of protein structures to find a meaningful underlying manifold, since it does not make any prior assumptions except that nearby neighbors in the original space must be nearby in the reduced dimension.

Spectral Embedding

Spectral Embedding is a technique that performs dimensionality reduction by first computing some type of normalized affinity matrix (graph Laplacian, nearest neighbors, Euclidean distance), and computing the eigenvectors of that matrix. It then looks for a spectral gap, where eigenvalues jump by a significant amount. It then uses the first few eigenvectors as-

sociated with the largest eigenvalues to reconstruct the lower dimensional representation of the original coordinates.

This lower dimensional representation is also called Diffusion Maps, a technique that has previously been studied for inferring protein kinetics.

Nonlinear Dimensionality Reduction + PCA

Our proposed technique combines a nonlinear dimensionality reduction technique with PCA. These nonlinear dimensionality reduction technique is affective in finding the few points that are representative of the data, and the PCA at the last step may help "refine" the identified dynamical direction by finding the direction of the maximum direction in this smaller space.

Simple Examples

We first present some simple toy examples that may capture the nature of the the underlying manifold on which protein structures may lie.

S-shaped curve

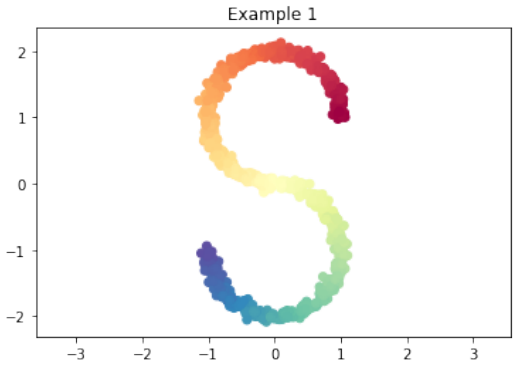

Figure 1: Simple S-curve

This S-shaped curve represents a situation where related protein conformations lie next to each other, with the trajectory following the S curve. The results for our proposed techniques above (PCA, Isomap + PCA, and Spectral Embedding + PCA) are shown above. Because

this curve is highly nonlinear and have a high density of points, Isomap + PCA projects the dynamical process the most accurately.

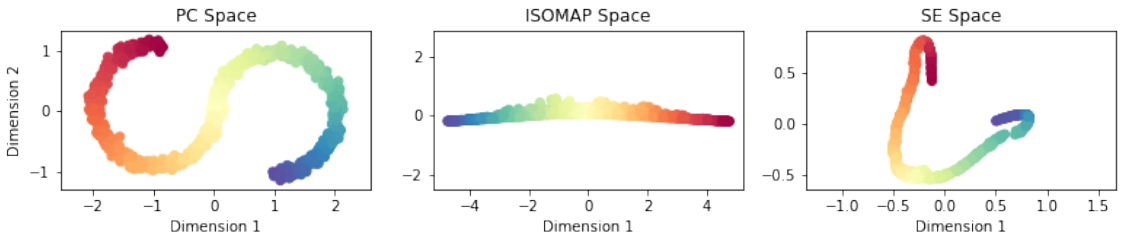

Figure 2: Space found by various techniques

Moving bivariate Gaussian distribution

Suppose we have a process moving in the y-direction, perhaps an interchange between conformations in the purple ellipse and those in the magenta ellipse. This is a toy example used by Frank Noe in one of his lectures, to show that tICA performs much better than PCA when the direction of the dynamical process is different from the direction of maximum variance.

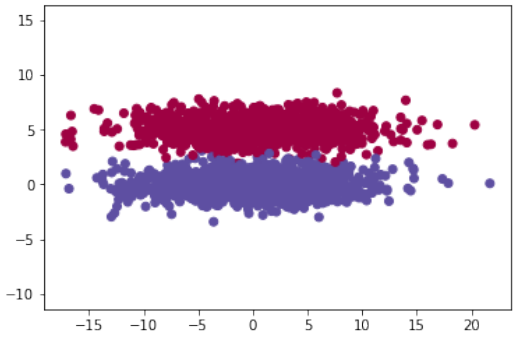

Figure 3: Moving Gaussian distribution

Our method of Spectral Embedding + PCA performs best in separating these two distinct groups (Fig. 4). As expected, PCA fails to distinguish these two groups with the direction of maximum variance.

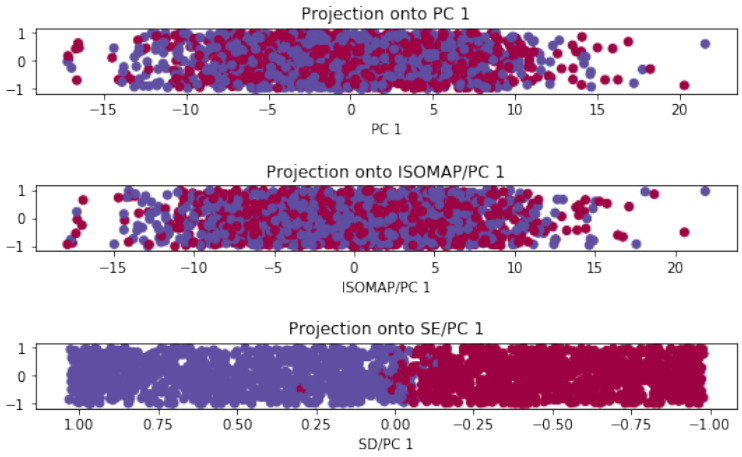

Figure 4: First dynamical direction found by various techniques

FsPeptide

The first system we studied was FsPeptide, a small polypeptide that contains 21 alanines. FsPeptide is a simple model that explores folding of a single alpha helix. From the MSM-Builder analysis it is known that the FsPeptide only has a single dominant reaction coordinate which corresponds to the folding coordinate. In this section we try the techniques mentioned above to find the dynamical process in FsPeptide and compare the results to tICA. (Fig. 5)

Distance in tICA space vs. Distance in inferred space

In our initial analysis, the first metric of evaluating the accuracy of the found space was comparing the distances between the frames in this space against the distances between the same conformations in the space found by tICA. As seen in Fig. 5 the distances in PCA, Isomap/PCA, and the raw dihedral angle space correlate well with the distances in the tICA space.

Mean first passage time vs. Distance in inferred space

In addition to using tICA distances, another metric for evaluating accuracy was comparing the distances between frames in this space against the known mean first passage time (MFPT) between these conformations. We performed this analysis in our report last quarter.

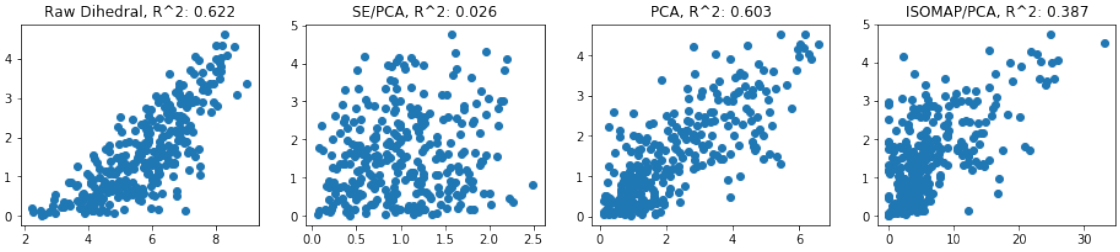

Figure 5: Distance in tICA space vs distance in inferred space

Similar to our plots for distances in the tICA space, we observed that distances in PCA, Isomap/PCA and the raw dihedral angle space correlate quite well with the MFPT (Fig. 6).

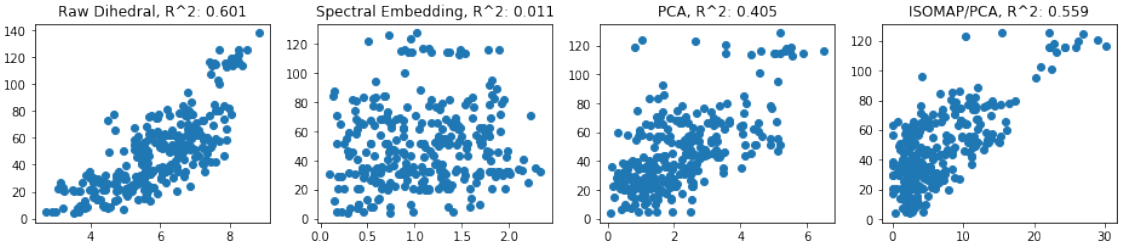

Figure 6: MFPTe vs distance in inferred space

Energy landscapes

Since we sampled a collection of protein structures at equilibrium found by MSMBuilder, we can assume the sampled data is representative of the various protein conformations. As a result, we can estimate the density maps of our structures in various two dimensional space estimate the true energy landscape. For FsPeptide, we observe one large clust with a single hot spot. Only PCA shows a similar pattern to tICA, our gold standard. All the hot spots in the different dimensional space estimates correspond to the structure of FsPeptide that is folded in an alpha helix (Fig. 7).

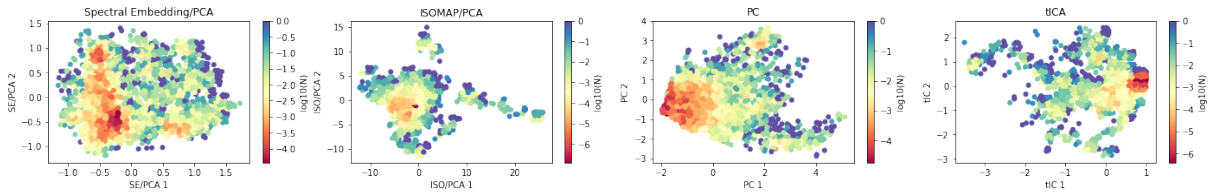

Figure 7: Inferred energy landscapes for FsPeptide

Tracking the folding coordinate with percent helix

In the study of FsPeptide, percent helix can be used to monitor how folded a given protein conformation is. Percent helix is suitable for FsPeptide since the fully folded structure of FsPeptide is a single, continuous alpha helix. By checking if the phi and psi angles along the protein backbone fall in the range of alpha helix in the Ramachandran plot a, percent helix can be calculated. As the structure transitions along the true dynamical coordinate, we expect the percent helix to continuously increase. in Fig. 8 we see that only Isomap/PCA does not follow a continuous increase but instead levels off and drops.

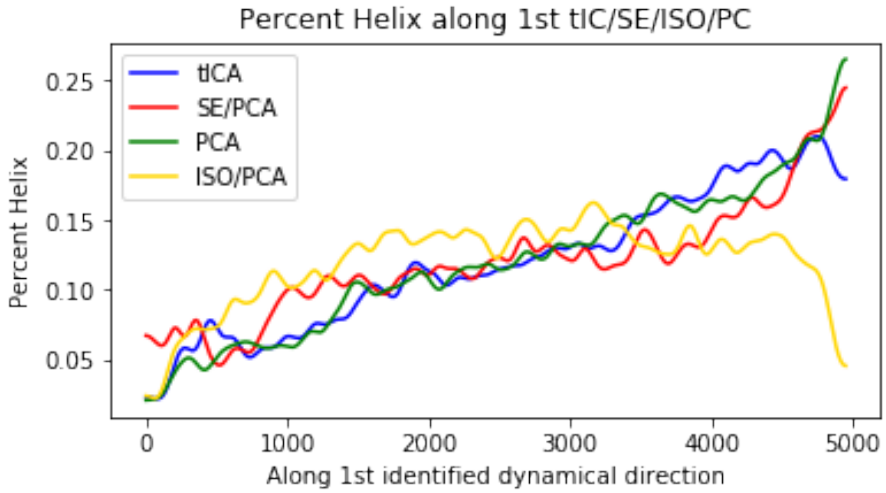

Figure 8: Percent helix along first dynamical direction for FsPeptide

Calmodulin

The second system we studied was calmodulin, a calcium binding protein that has two distinct forms: the *apo* and the *holo* forms. We know from previous studies that the two most distinct dynamical procsses are the transition between the *apo* and the *holo* forms, as well as the partial unfolding of the protein. In this section, we try the abovementioned techniques to find these dynamical processes and compare our results to tICA (Fig. 9).

Distance in tICA space vs. distance in inferred space

One metric for evaluating the accuracy of the found space was comparing the distance between conformations in this space against the distance between the same conformations in the space found by tICA. We observed that distances in PCA, Isomap/PCA and the raw dihedral angle space correlate quite well with the distance in the tICA space.

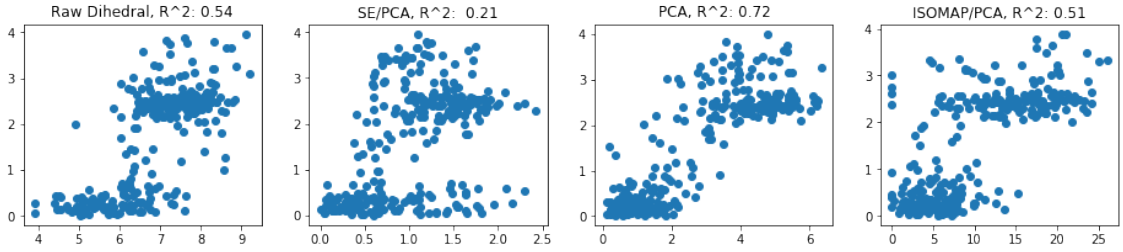

Figure 9: Distance in tICA space vs distance in inferred space

Mean first passage time vs. distance in inferred space

Another metric for evaluating the accuracy of the found space was comparing the distance between conformations in this space against the known mean first passage time (MFPT) between these conformations. We performed this analysis in our report last quarter for FsPeptide. Similar to our plots for distances in the tICA space, we observed that distances in PCA, Isomap/PCA and the raw dihedral angle space correlate quite well with the MFPT (Fig. 10).

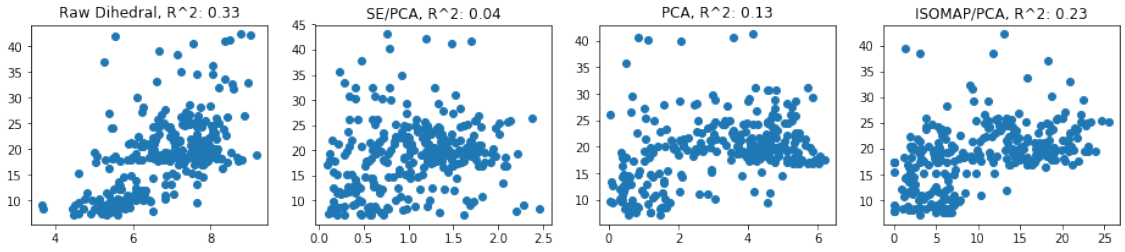

Figure 10: MFPTe vs distance in inferred space

Phe92-Phe141 distance as a proxy for *Apo-Holo* transition

For calmodulin, one metric for monitoring the *apo* and the *holo* forms is measuring the Euclidean distance between the phenylalanine amino acid at position 92 and the phenylalanine

amino acid at position 141. This interaction is significant since the overlap of these aromatic groups reveal a second calcium binding site, a hallmark of the *holo* form. As the structures change from the *apo* to the *holo* form, we expect to see a decrease in this Phe92-Phe141 distance.

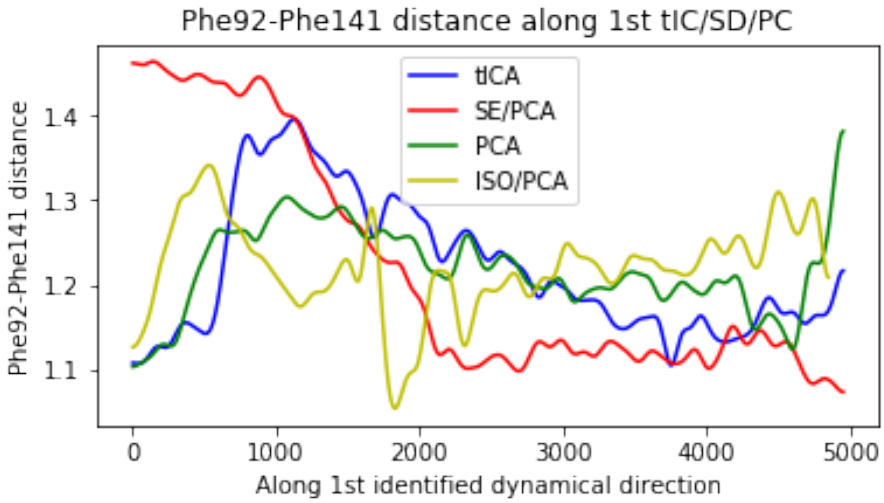

Figure 11: Observing the Phe92-Phe141 distance

Here, we created a sequence of protein conformations by sorting by their values found for our estimate of the first dynamical process (for example, the value of the first coordinate in the PCA space). This is our best guess at how the very first dynamical process proceeds. We compared our results to the ordering found by the first tIC, shown in blue. We observed that Spectral Embedding + PCA is quite good at picking up sharp jumps in the dynamical process, while Isomap + PCA and PCA alone did not find such a transition (Fig. 11).

Met124-Ala128 distance as a proxy for partial unfolding

Similar to our analysis above, one metric for monitoring the partial unfolding of the protein is the distance between he methionine amino acid at position 124 and the alanine amino acid at position 128.

Here, we created a sequence of protein conformations by sorting by their values found for our estimate of the second dynamical process (for example, the value of the second coordinate in the PCA space). This is our best guess at how the second dynamical process proceeds. We compared our results to the ordering found by the second tIC, shown in blue. We observed that the signal is more noisy than our estimate of the first dynamical process, in consistent

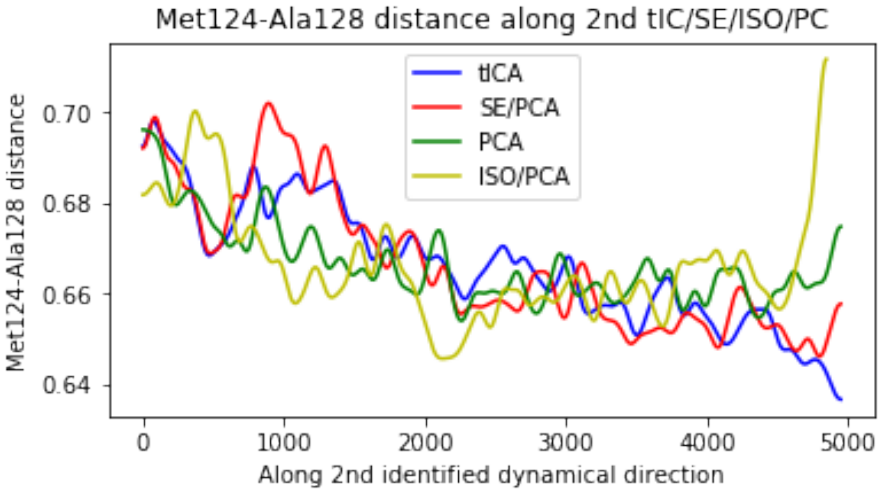

Figure 12: Observing the Met124-Ala128 distance

with our intuition that more subtle processes are harder to infer. All techniques except Isomap + PCA is good at sensing the general decrease in this distance as the protein goes from a more unfolded to a more folded state. Spectral Embedding + PCA seems to track tICA best (Fig. 12).

Energy landscapes

Because we sampled our collection of protein structures at equilibrium, we can assume that the density maps of our structures in various two dimensional spaces estimate the true energy landscape. For calmodulin, we observe a relatively clear clusters in each of the estimates found by various techniques, with Isomap + PCA and PCA showing a pattern most similar to tICA, our gold standard. These clusters refer to the *apo* and the *holo* forms (Fig. 13).

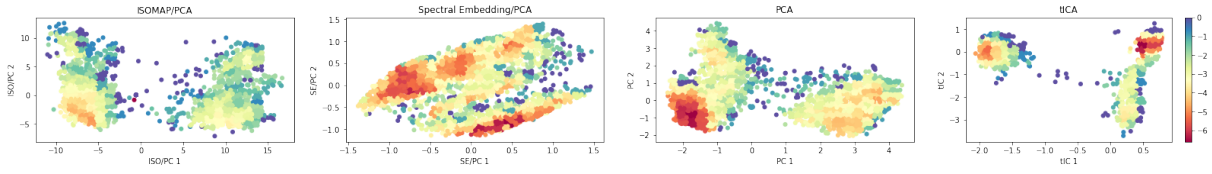

Figure 13: Inferred energy landscapes for calmodulin