Homework Assignment #9

Due in class Tuesday, Dec 5

One of the goals of the course is to provide you with the ability to understand scientific discussions of protein dynamics. We have looked at different ways to characterize the ensemble of protein conformations available from molecular simulations. In fact, this is an active area of research. I have attached an edited version of a paper from one of the leading labs in molecular dynamics. The lab is that of Vijay Pande at Stanford University and he runs the Folding@home project (http://folding.stanford.edu/). The paper was published in 2009 but is still very relevant today. It outlines an approach to identify the important states in a molecular ensemble. The methods have become more complicated since this paper but the concepts are the same.

Your assignment is to read the attached paper and answer the following question. Please hand in a hard copy of your answer at the beginning of class.

What is the key conceptual addition that this group has added to allow one to identify meaningfully important states in a molecular simulation?

Note: I have edited the paper to remove jargon, unfamiliar topics and the citations to make it easier to read. If you would like to read the original it is available on the course website at http://pages.jh.edu/pfleming/compbio/files/bowman_methods_2009.pdf.

.

Using generalized ensemble simulations and Markov state models to identify conformational states

Gregory R. Bowman, Xuhui Huangb and Vijay S. Pande
Standford University

[Extracted from Methods 49(2) 197-201, 2009]

Abstract
Part of understanding a molecule's conformational dynamics is mapping out the dominant metastable, or long lived, states that it occupies. Once identified, the rates for transitioning between these states may then be determined in order to create a complete model of the system's conformational dynamics. Here we describe the use of the MSMBuilder package to identify the metastable states from simulations.

1. Introduction

Molecular Dynamics (MD) and Monte Carlo (MC) computer simulations have the potential to complement experiments by elucidating the chemical details underlying the conformational dynamics of biological macromolecules like proteins and RNA. Such simulations sample a system's free energy landscape, which is characterized by long lived, or metastable, states separated by large free energy barriers. Thus, understanding a system's conformational dynamics can be broken down into two steps: (1) identifying the long lived, or metastable, states visited by the system and (2) determining the rates of transitioning between these states. Unfortunately, it is extremely difficult to adequately sample the conformational space accessible to biomolecules. Furthermore, even if adequate sampling can be achieved, the resulting datasets are often quite large and, therefore, difficult to analyze and interpret.

A popular approach to the first step is to use [methods that] achieve broad sampling at the temperature of interest by performing a random walk in temperature space. Thus, they are a suitable way to sample the accessible space.

Clustering methods [may be used to identify metastable states]. However, most clustering algorithms group conformations together based solely on their structural similarity, so they may fail to capture important kinetic properties. To illustrate the importance of integrating kinetic information into the clustering of simulation trajectories, one can imagine two people standing on either side of a wall. Geometrically these two individuals may be very close but kinetically speaking it could be extremely difficult for one to get to the other. Similarly, two conformations from a simulation dataset may be geometrically close but kinetically distant and, therefore, a clustering based solely on a geometric criterion would be inadequate for describing the system's dynamics.

[What we need is] a form of clustering that incorporates kinetic information by grouping conformations that can interconvert rapidly into the same state and conformations that cannot interconvert rapidly into different states. Thus, conformations in the same

metastable state, which may be thought of as a large free energy basin, will be grouped together while conformations separated by large free energy barriers will not.

A biomolecular folding free energy landscape may be thought of as a hierarchy of basins. The number of metastable states to be constructed controls the resolution of the model by determining how large a barrier must be in order to divide phase space into multiple states.

There are four major steps in the procedure: (1) dividing the data into small sets called microstates based on their structural similarity, (2) lumping kinetically related microstates together into metastable states (also called macrostates), (3) extracting representative conformations for each state, and optionally (4) calculating populations of each state to judge convergence. Steps 1-3 are depicted schematically in Fig. 1. The conformations extracted with this method represent the space explored by the system and thus give insights into its dynamics.

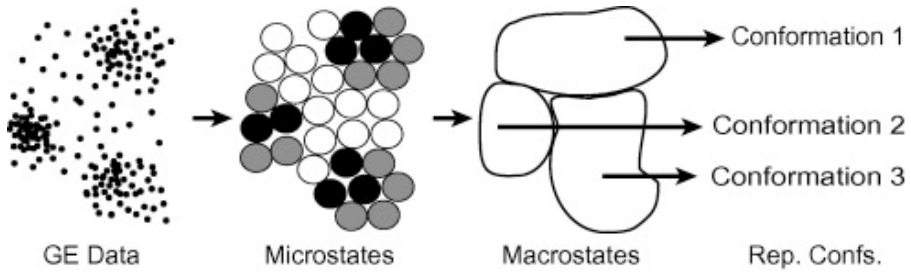

Figure 1. Schematic of the steps required for obtaining representative conformations for each state. First, conformational data represented by points are grouped into microstates represented by circles, with darker circles for more highly populated microstates. Kinetically related microstates are then lumped together into macrostates, or metastable states, represented by amorphous shapes. Finally, representative conformations are obtained by extracting the most probable conformation from each macrostate.

2. Description of method

2.1. Dividing the data into microstates

The first step is to divide the [trajectory conformations] into thousands of microstates based on their structural similarity. For conformational dynamics we measure structural similarity by the RMSD for some subset of the atoms. While the RMSD may not be very meaningful for large distances, it does have a kinetic interpretation for small distances. That is, conformations with very small RMSDs should be able to interconvert rapidly. Thus, if a microstate is small enough that every member has a very small RMSD to every other member then one may assume that their structural similarity implies a kinetic similarity.

However, one must also take care not to generate microstates that are too small because it is important to see a sufficient number of transitions between them. For example, if every conformation were put into its own microstate no pair of trajectories would ever visit the same microstate. Preliminary work in our lab shows that RMSD radii on the order of 2-2.5 Å seem appropriate for protein systems.

A k-centers clustering algorithm is chosen that yields clusters of approximately equal volume (as judged by using the maximal RMSD distance between the cluster center and any other point in the cluster as the radius of a sphere). This property is of value because it means that the population of a cluster is approximately proportional to its density in phase space. (Instructor's note: In k-centers clustering you divide up the available space into a known number of clusters and insure that all possible spaces are represented. Other clustering methods, e.g. hierarchical, let the data determine the number and size of clusters.)

2.2. Lumping microstates into macrostates

Conceivably, one could extract a representative conformation for each microstate to get an idea of the conformational space explored by the system of interest. However, this would only be a slight improvement upon examining the raw data itself. Instead, it is valuable to lump kinetically related microstates together into metastable states, also called macrostates.

The first step in generating a set of macrostates is to determine how many of them to create. [This task may be accomplished by building a transition probability matrix which contains the probability that a simulation will be in microstate j at time $t + \Delta t$ given that it was in state i at time t.] A series of implied timescales are then calculated. These implied timescales correspond to the timescales for transitioning between different sets of microstates. An appropriate number of macrostates to build can be determined based on the locations of the major gaps in the implied timescales.

2.3. Extracting representative conformations

There are a number of ways of extracting representative conformations for each macrostate. A simple way of getting a single conformation is to use [an algorithm that generates the geometric center of the coordinates of the macrostate.] The geometric center [is, however] not necessarily the most probable member of each macrostate.

To understand the distribution of conformations in each macrostate one may identify the geometrically central conformation [as a first step]. [Then] the conformations for a given macrostate may then be overlaid in a viewer for visual analysis. Such an approach may be cumbersome if there are too many microstates in each macrostate. One alternative is to randomly select a reduced number of conformations from each macrostate. A major shortcoming of these methods is that they select conformations with a more or less uniform distribution across the macrostate.

Probably the best way of extracting representative conformations is to [calculate] a list of the microstates in each macrostate ordered from densest to sparsest. That is, the most probable to the least probable. Any number of the most probable structures in a given macrostate may then be selected and overlaid in a viewer to get an idea of the distribution of conformations within the state.

2.4. Judging convergence

Unfortunately there is no analytic way of checking that a single set of simulations has explored the entire accessible space for a given system and, therefore, yielded representative conformations that accurately describe the conformational dynamics. [Instructor's note: This is a way of restating the that the ergodic hypothesis can not be proven in this case.] To the best of our knowledge, the most effective way to ensure that the entire space has been explored is to run two distinct sets of simulations started from very different initial configurations. The populations for each state may then be calculated for each dataset. If they agree then one can be relatively sure that the entire space has been explored because the thermodynamics found are independent of the starting conformation.

Of course, due to the stochastic nature of conformational dynamics the two sets of populations are unlikely to agree exactly. To make a valid comparison error bars on the populations from each dataset [may be calculated]. If the populations agree within error then the two simulations may be considered to have converged to the true equilibrium distribution and one may be relatively sure that the entire accessible space has been explored. Thus, the conformations extracted in step 3 will provide an accurate depiction of the conformational dynamics of the system.