Chapter 12
Using Data Mining Techniques to Probe the Role of Hydrophobic Residues in Protein Folding and Unfolding Simulations

Cândida G. Silva
University of Coimbra, Portugal

Pedro Gabriel Ferreira
Center for Genomic Regulation, Spain

Paulo J. Azevedo
University of Minho, Portugal

Rui M. M. Brito
University of Coimbra, Portugal

ABSTRACT

The protein folding problem, i.e. the identification of the rules that determine the acquisition of the native, functional, three-dimensional structure of a protein from its linear sequence of amino-acids, still is a major challenge in structural molecular biology. Moreover, the identification of a series of neurodegenerative diseases as protein unfolding/misfolding disorders highlights the importance of a detailed characterisation of the molecular events driving the unfolding and misfolding processes in proteins. One way of exploring these processes is through the use of molecular dynamics simulations. The analysis and comparison of the enormous amount of data generated by multiple protein folding or unfolding simulations is not a trivial task, presenting many interesting challenges to the data mining community. Considering the central role of the hydrophobic effect in protein folding, we show here the application of two data mining methods – hierarchical clustering and association rules – for the analysis and comparison of the solvent accessible surface area (SASA) variation profiles of each one of the 127 amino-acid residues in the amyloidogenic protein Transthyretin, across multiple molecular dynamics protein unfolding simulations.

DOI: 10.4018/978-1-60566-816-1.ch012

INTRODUCTION

Molecular dynamics (MD) is one of the most realistic simulation techniques available to study protein folding *in silico*. In MD simulations, the structural fluctuations of a single protein can be tracked over time by numerically solving Newton's equations of motion (Adcock & McCammon, 2006). When using molecular dynamics simulations to study protein folding and unfolding processes, multiple simulations need to be considered to probe the large conformational space and multidimensional potential energy surface available to the polypeptide chain, and obtain significant statistical mechanical averages of the system properties (Brito, 2004; Kazmirski, 1999; Scheraga, 2007). Even though the computational power available keeps increasing, it is still a major challenge to simulate protein folding or unfolding processes in its real time scale (hundreds of µs to seconds or more). However, it has been suggested that performing multiple short simulations (usually 5 to 10) provides better sampling of the conformational space than having a single long simulation (Caves, 1998). Thus, performing multiple simulations on the 10 to 100 ns time scale is becoming routine, which generates huge amounts of data to be analysed and compared. Furthermore, a large set of structural and physical properties (such as root mean square deviation, radius of gyration, secondary structure content, native contacts, and solvent accessible surface area) is usually calculated from the MD trajectories to characterize the conformational space explored.

Most of the structural and physical properties calculated from the MD trajectories are easy to extract. However, the next challenge for data analysis in multiple MD simulations is to identify, among the properties, those that are essential in describing the protein unfolding or folding processes. Additionally, it is important to define the relative importance of each property along the folding/unfolding pathway. It is expected that some of the properties best describe initial stages of the processes under study, while others may be more sensitive to later stages. Looking at a wide range of properties and experimental conditions further increases the amount of data generated by such simulation models. Analyzing and interpreting these data requires automated methods such as data mining. These issues have been addressed before by Kazmirski *et al* (1999) and Brito *et al* (2004). While Kazmirski *et al* (1999) presented several methods based on structure and property data to compare different MD trajectories, Brito *et al* (2004) discussed the usefulness of data mining techniques, which include machine learning, artificial intelligence, and visualization, to address the data analysis problem arising from multiple computational simulations, including protein folding and unfolding simulations. Figure 1 depicts a general overview of this process, from the initial system under study to the interpretation of the results using data mining tools. The researcher begins by performing multiple MD simulations, starting from the same experimental structure (same atom coordinates) but different initial atom velocities. For each simulation, a set of varying atom coordinates and velocities over time (a trajectory) is obtained. At the end of each simulation, a collection of molecular properties may be calculated to characterize the structural variation of the protein during the process. Finally, the molecular property variation profiles may be subjected to analysis using data mining tools.

The solvent accessible surface area (SASA) is one of the molecular properties that might be calculated for each MD trajectory. SASA reports on an important parameter from the protein conformational stability point of view: solvent exposure and protein compactness. Its value may be calculated for the entire protein, but also for subsets of amino-acid residues, accounting for example for the polar or non-polar contributions. Furthermore, the study of the SASA variation of each individual amino-acid residue provides a greater level of detail on the individual contributions for the folding or unfolding processes.