Structural determination of protein through the extended continuous similarity

Lexin Chen, Arup Mondal, Alberto Perez and Ramón Alain Miranda-Quintana

Department of Chemistr and Quantum Theory Project, University of Florida, Gainesville, FL 32611

Molecular dynamics (MD) is a computer simulation method for analyzing systems' dynamics by integrating Newton's law of motion. Although the advent of graphical processing units has made microsecond timescale MD simulations a routine, simulation post-processing analysis has not caught up to speed with the increasing size of simulation datasets. Clustering, which groups objects based on structural similarity, is pertinent for finding an ensemble of representative structures, which is key to finding the protein folding pathway. Traditionally, clustering algorithms, such as agglomerative clustering, scale quadratically, which is unfavorable with the increasing dataset size. Recently, we have proposed clustering algorithms and frame-selection methods based on extended continuous similarity indices, leading to linear-scaling workflows, which can predict the number of clusters, and create a diverse selection of representative structures. In this study, thirteen proteins from the sparsely labeled NMR-assisted prediction of the CASP 13 database were simulated using MD simulations and the extended continuous similarity was applied to identify the ensemble representative structures and elucidate the protein folding pathways.