MULTISEQUENCE ALGORITHM FOR COARSE-GRAINED BIOMOLECULAR SIMULATIONS: EXPLORING THE SEQUENCE-STRUCTURE RELATIONSHIP OF PROTEINS

Adekunle Aina, Memorial University of Newfoundland, Canada
akaina@mun.ca
Stefan Wallin, Memorial University of Newfoundland, Canada

Key Words: Monte Carlo, generalized ensembles, protein folding, protein fold switching.

Many biologically motivated problems naturally call for the investigation and comparison of molecular variants, such as determining the mechanisms of specificity in biomolecular interactions or the mechanisms of molecular evolution. We consider a generalized ensemble algorithm for coarse-grained simulations of biomolecules which allows the thermodynamic behavior of two or more sequences to be determined in a single multisequence run. By carrying out a random walk in sequence space, the method also enhances conformational sampling. Escape from local energy minima is accelerated by visiting sequences for which the minima are shallower or absent. We test the method on an intermediate-resolution coarse-grained model for protein folding with 3 amino acid types and explore the potential for large-scale coverage of sequence space by applying it to sets of more than 1,000 sequences each. The resulting thermodynamic data is used to analyze the structures and stability properties of sequences covering the space between folds with different secondary structures. Besides demonstrating that the method can be applied to large number of sequences, the results allow us to carry out a more systematic analysis of the biophysical properties of sequences along mutational pathways connecting pairs of different folds than has been previously possible.

Figure 1 – (A) Example of a multisequence simulation of 144 sequences in a single run. The plot shows the Monte Carlo evolution of the sequence s, the total potential energy E and the root-mean-square deviation (RMSD) calculated against the representative fold IA (light blue) and fold IB (dark red) structures in (B). Representative structures of folds (B) IA, IB, (C) IIA and IIB, chosen to be the minimum-energy conformations found for 4 reference sequences