

456d Characterization of Protein Sequence Landscapes Using Flat-Histogram Monte Carlo Algorithms

M. Scott Shell, Pablo G. Debenedetti, and Athanassios Z. Panagiotopoulos

The protein design problem, the determination of amino acid sequences which fold into target backbone structures, depends strongly on the number of amino acids to be considered for a particular design problem, and on the nature of the solvent-mediated interactions between them (i.e., the amino acid alphabet). Therefore, to better understand protein designability, we seek to quantify the dependence of the collection of all possible sequences--the sequence landscape--upon the particular amino acid alphabet used. We characterize sequence landscapes in several heteropolymer models of proteins using an efficient flat-histogram Monte Carlo search method. Our approach involves determining the distribution along various fitness parameters of all sequences of a given length, when threaded through a common backbone. These calculations are performed for a number of Protein Data Bank structures using three well-studied contact potentials. Our results indicate significant differences among the studied potentials in terms of the "smoothness" of their landscapes. In particular, one model reveals unusual cooperative behavior among its species' interactions, resulting in what is essentially a set of phase transitions in sequence space. Such phase transitions may possess evolutionary significance, and can have a profound effect on the performance of protein design algorithms. Moreover, our calculations permit a quantitative determination of designability by performing a statistical "counting" of the number of sequences which target a given configuration; importantly, our approach works for chain lengths far exceeding those for which an exhaustive enumeration of sequences can be performed.