MAR15-2014-020464

Abstract for an Invited Paper
for the MAR15 Meeting of
the American Physical Society

Using evolutionary sequence variation to make inferences about protein structure and function

LUCY COLWELL, Cambridge University

The evolutionary trajectory of a protein through sequence space is constrained by its function. Collections of sequence homologs record the outcomes of millions of evolutionary experiments in which the protein evolves according to these constraints. The explosive growth in the number of available protein sequences raises the possibility of using the natural variation present in homologous protein sequences to infer these constraints and thus identify residues that control different protein phenotypes. Because in many cases phenotypic changes are controlled by more than one amino acid, the mutations that separate one phenotype from another may not be independent, requiring us to understand the correlation structure of the data. To address this we build a maximum entropy probability model for the protein sequence. The parameters of the inferred model are constrained by the statistics of a large sequence alignment. Pairs of sequence positions with the strongest interactions accurately predict contacts in protein tertiary structure, enabling all atom structural models to be constructed. We describe development of a theoretical inference framework that enables the relationship between the amount of available input data and the reliability of structural predictions to be better understood.