Poster I-33

Information-Theoretic Analysis of Protein Fold Potentials

Authors:
Armando D. Solis (*Mount Sinai School of Medicine*)
S. Rackovsky (*Mount Sinai School of Medicine*)

Short Abstract: We examine fundamental information-theoretic roots of protein fold potentials to explore effective ways to build better empirical functions. We estimate parameters from finite data for potential functions chosen to minimize information loss while also avoiding information redundancy. The effect of these improvements on performance in fold recognition is assessed.

Long Abstract:
We examine the fundamental roots in information theory of protein fold potentials, with the ultimate goal of finding improved ways to build such empirical functions. An information theory based view of protein folding potentials involves measuring mutual information shared between sequence and conformation. We show that details of the sequence-conformation mutual information equation reveal the informatic nature of requisite sequence-dependent structural probabilities, and provide another justification of log-odds scores in structure prediction. More importantly, one can, in principle, construct log-odds informatic functions from the set of known folds to evaluate possible conformations for a given sequence, and identify the native conformation as that which gives the largest score. Certainly, the difficulty of empirically constructing such log-odds functions stems from the disproportionate number of degrees of freedom in the sequence and structure spaces as compared to the amount of data available currently. Therefore, a parallel issue is how best to estimate each probability term from finite data, especially since estimating detailed conditional probabilities can easily become intractable.

Four simplifications of the mutual information equation may be taken: (a) positional independence of conformation; (b) independence and additivity of conformational terms; (c) reduced seqence effect; (d) reasonable discretization of sequence and structure spaces. These strategies allow partition the mutual information into discrete terms, each referring to an aspect of the sequence-structure relationship. Therefore, the information equation can be written as a summation, where each term specifies information (the log-odds “energy” score) contained in particular pairings of subsections of sequence and structure descriptions. Such an equation underlies the basic informatic nature of empirical energy functions widely used in protein folding. The advantage of the informatic formalism springs from the notion that parameter optimization is possible, desirable, and straightforward. The challenge is to find the right descriptors and levels of discretization that will minimize information loss, or alternatively, maximize mutual information.

Our recent [1-3] and current work are built on the basis that such empirical "energies" are fundamentally informatic functions which can be maximized, and which have optimizable parameters. Understanding that many potential functions are simplified forms of the complete

information equation could also lead to better scoring functions that can be tailored to the current data set size. Some issues we are tackling presently are (a) sequence alphabet reduction, (b) the informatic nature of the reference state, (c) information content of widely used potentials like the backbone dihedral, contact, pairwise distance-dependent side chain, and solvent accessibility potentials, (d) parameter optimization for those potentials, and (e) better functional forms of the scoring function, in the interest of maximizing information extraction from limited data.

References

[1] Solis, A.D. & Rackovsky, S. 2000. Optimized Representations and Maximal Information in Proteins. Prot. Struct. Funct Genet. 38:149-164.
[2] Solis, A.D. & Rackovsky, S. 2002. Optimally Informatic Backbone Structural Propensities in Proteins. Prot. Struct. Funct Genet. 48:463-486.
[3] Solis, A.D. & Rackovsky, S. 2006. Improvement of Statistical Potentials and Threading Score Functions Using Information Maximization. Prot. Struct. Funct. Bioinform. 62:892-908.