A Methodology to Optimally Extract Local Structural Information in Proteins

Armando D. Solis **S. Rackovsky**
armando.solis@mssm.edu shelly@msvax.mssm.edu
Department of Biomathematical Sciences, Mount Sinai Medical Center Box 1023,
One Gustave Levy Place, New York, NY 10029, USA

Keywords: protein sequence-structure relationship, optimized representations

1 Introduction

Protein bioinformatics examines the relationship between protein sequence and native conformation by rationalizing the structural information found in the growing, but still limited, mass of experimentally solved protein structures. Derived from experimental databases, empirical potentials, which link key sequence patterns to structural attributes, have proven useful in (a) understanding the coding properties of amino acids and sequences [3,8], (b) identifying similarities in folds from alignments of protein chains [9], and (c) providing a means to bias the search for the native conformation in computational predictions [2].

Using basic ideas from information theory, our methodology, built on our previous work [8], extracts the optimum amount of structural information available in the local sequence. From a non-redundant set of protein X-ray structures, we construct local sequence-dependent structural distributions or potentials which summarize the influence of local sequence on backbone conformation. These empirical potentials (a) utilize multi-residue information, (b) exploit the similarities in the local coding properties of amino acids to streamline the local sequence description, (c) assist in extracting maximal information from local sequence, (d) work from the currently existing data set of protein structures, and (e) prove to be readily accessible and useful for structural prediction and other explorations.

2 Methods and Results

Two simplification strategies are employed: we use discretized structural representations (phi-psi dihedral angles [5], 3-state secondary structural classes [1], or virtual backbone path [4]), and we group amino acids that show similarities in local coding properties [8]. The relationship between local sequence and the corresponding range of conformations is expressed as a structural distribution formed by two components: (a) the set of all instances of the particular local sequence in the non-redundant data set of solved protein chains, and (b) a background distribution, or the local sequence-independent distribution, which guards against complete memorization [7]. The combination of these two sets is facilitated by a mixing coefficient, which, along with the collapse of the amino acid alphabet, is optimized by our automatic procedure.

The fitness of a sequence/structure representation scheme is measured by the information gain, an information-theoretic quantity derived from the Shannon entropy [6]. Local sequence-dependent distributions (or potentials) are constructed by first collapsing the amino acid alphabet into a given number of groups. Then, for a given length of local sequence and a choice of backbone conformation, the instances are compiled into discrete sets, one for each unique local sequence. These are then combined with a given proportion of the background distribution to form the sequence-dependent distribution. The information gain is measured for the specified mixing coefficient and amino acid grouping. A direct Monte-Carlo search across different amino acid groupings, and then across different mixing coefficients, locates the scheme that results in the maximum information gain, and consequently in the most optimal set of distributions.

Major results of this study include: (a) a viable strategy for extracting and measuring structural information from local sequence, (b) a catalogue of local sequence-dependent potentials for different structural descriptors (phi-psi dihedral angles, 3-state secondary structural classes, and virtual bond geometry), and (c) an elucidation of the local coding properties of amino acids and short sequences.

References

[1] Kabsch, W. and Sander, C., Dictionary of protein secondary structure: Pattern recognition of Hydrogen-bonded and geometrical features, *Biopolymers*, 22(12):2577–2673, 1983.

[2] Lee, B., Kurochkina, N., and Kang, H.S., Protein folding by a biased Monte Carlo procedure in the dihedral angle space, *FASEB J.*, 10(1):119–125, 1996.

[3] Rackovsky, S., On the nature of the protein folding code, *Proc. Natl. Acad. Sci.*, 90(2):644–648, 1993.

[4] Rackovsky, S., Quantitative organization of the known protein X-ray structures, *Prot. Struct. Funct. Genet.*, 7:378–402, 1990.

[5] Ramachandran, G.N. and Sasisekharan, V., Conformation of polypeptides and proteins, *Adv. Prot. Chem.*, 23:283–437, 1968.

[6] Shannon, C.E., A mathematical theory of communication, *Bell Syst. Tech. J.*, 27:379–423. 1948.

[7] Sippl, M.J., Calculation of conformational ensembles from potentials of mean force, *J. Mol. Biol.*, 213(4):859–883, 1990.

[8] Solis, A.D. and Rackovsky, S., Optimized representations and maximal information in proteins, *Prot. Struct. Funct. Genet.*, accepted.

[9] Wako, H. and Blundell, T.L., Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins, *J. Mol. Biol.*, 238(5):693–708, 1994.