

Learning sequence-structure affinity using neural networks and a stochastic representation of protein folding motifs

Daniel St-Arnaud and François Major

*Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal,
C.P. 6128, succ. centre-ville,
Montréal, Québec, Canada H3C 3J7
{starnaud|major}@iro.umontreal.ca*

Abstract

A stochastic framework for the representation of arbitrary protein structural motifs is introduced. Based on a sound mathematical theory, this formalism allowed us to define a measure of sequence-motif affinity. This measure was learnt directly from data using artificial neural networks (ANN).

A structural motif – for instance the "common core" of the immunoglobulins – is defined as a set of amino acid positions and inter-position relations. Each position is also characterized by a set of properties such as solvent accessibility and secondary structure type. Using this definition, one can construct a Markov random field (MRF) representation for a motif. The probability density function (PDF) encoded by this MRF can be seen as a measure of sequence-motif affinity and can be used for protein fold recognition by optimal sequence-structure alignment, or "threading" (see White et al [1]). Under suitable conditions, the PDF associated with a structural motif M can be factorized into a simple product of local probability distributions by approximating the MRF with an equivalent Bayesian network structure. These local distributions may then be learnt from sequence and structure data using ANNs. The resulting assembly of ANNs defines a stochastic model for all sequences sharing the common structural motif M .

Unlike usual pairwise interaction (threading) potentials, our NN-based approach can learn higher order dependencies between adjacent amino acid positions. The NN approach eliminates the need to discretize environmental attributes and allows us to test different encodings of amino acids. This might provide insights on properties that are important for protein folding. Also, the probabilistic nature of our sequence-motif affinity measure allows for easy alternative model comparison within the Bayesian inference framework.

We are currently trying to apply this methodology to the problem of beta-fold recognition and, although we have yet to complete the implementation, we expect that it will compare favorably to existing methods.

References

- [1] White, J. Muchnik, I. and Smith, T.F. (1994). Modeling protein cores with Markov random fields. *Math. Biosci.* 124, 149-179.