Prediction of protein coarse contact maps

P. Frasconi, A.Vullo

Università di Firenze

Prediction of protein three-dimensional structure from sequence is today one of the most challenging unsolved problems in computational molecular biology. Although for some classes of proteins (or domains) prediction is possible by means of homology modeling or fold recognition, a definitive solution is still lacking for the general case involving arbitrary sequences. In principle, machine learning methods for sequences could be used to map amino acid sequences into sequences of 3D atom coordinates. Unfortunately, translational and rotational invariances make the task very difficult to formulate in a straightforward way.

Another step towards ab initio prediction of protein structure is the prediction of useful geometrical features, such as secondary structure and contact maps. A protein contact map is an undirected graph having vertices associated with amino acids (or secondary structure segments) and arcs representing a spatial neighborhood relation. Its knowledge is very important in this context because the associated spatial constraints allow one to reconstruct the compatible spatial conformation of residues using distance geometry or other efficient stochastic energy minimization algorithms [1].

The specific problem we consider in this work is the prediction of coarse contact maps, i.e. maps whose nodes are associated with secondary structure elements (helices, beta sheets, and coils). Our methods are based on recursive neural networks, a machine learning model capable of dealing with structured data representations [2]. Training maps are derived from a nonredundant set of proteins in PDP.

The first algorithm we propose is a graph search method. For each protein in the training set, given the true map G and a set of hypothetical maps {G_i}, the network is trained to predict a measure of closeness between each pair (G_i, G). After training, a greedy search algorithm (such as hill-climbing or beam-search) is used to iteratively refine the predicted conformation, using the neural network output as a heuristic evaluation function. The second algorithm is based on a 2-dimensional generalization of noncausal recurrent neural networks like those used in [3] for the prediction of protein secondary structure. The network architecture relies in this case on a factorial state space representation, with states arranged according to a NxN dimensional grid (being N the sequence length). Four different layers are used in the factorization, each associated with one of the four possible directions of acyclic grid propagation (NW->SE, NE->SW, SW->NE, SE->NW). The output grid has units associated with graph edges. The network is trained to produce high outputs on edge units and low outputs on non-edge units --- work on this second architecture is related to ongoing collaboration with P. Baldi and A. Pollastri (UC Irvine). Our preliminary results on PDB data show the viability of the approach.

[1] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. Folding and Design, 2:295--306, 1997.
[2] M. Gori, P. Frasconi and A. Sperduti. A general framework for adaptive processing of data structures, IEEE Transactions on Neural Networks, 9(5):768-786, 1998.
[3] S. Brunak, P. Baldi, P. Frasconi, G. Pollastri and G. Soda. Exploiting the past and the future in protein secondary structure prediction, Bioinformatics, 15(11), 1999.