PROFcon: prediction of internal protein contacts

Punta M.(1,2),Rost B.(1,2,3)

(1) CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York (2) Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York (3) NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York

Motivation

Novel and more efficient structural prediction methods are needed in order to reduce the ever-increasing gap between the number of known protein sequences and structures. Better predictions of non-local contacts between residues could help improve fold recognition and de novo modeling. In the present study, we introduce PROFcon, a novel contact prediction method based on a neural network that combines information from various sources. PROFcon predictions are then used to estimate experimental folding rates of two-state proteins (i.e. proteins that fold without intermediate states).

Methods

Datasets: the protein dataset used to train and test our neural network was extracted from the December 2003 EVA release (Koh et al. (2003), Nucleic Acids Research, 31: 3311-3315): a set of 3201 protein chains of known structure. We removed all non-X-ray structures, all membrane and coiled-coil proteins. We divided the proteins into three data sets: training, validation and test set, containing 748, 466 and 633 proteins, respectively.

Predictions of experimental folding rates were based on the set of 37 two-state folders introduced by Ivankov and Finkelstein (Ivankov & Finkelstein (2004) Proc. Natl. Acad. Sci. U.S.A. 101, 8942-8944).

Neural Network: we trained standard feed-forward neural networks with back-propagation and momentum term (Rost & Sander (1993) J. Mol. Biol., 232, 584-599). In total, we used 738 input, 100 hidden, and 2 output units (contact, non-contact). The input features corresponded to three different ways of describing each pair of residues. In particular, we used information from the local environment of the residues, information from the segment connecting the residues, and global information from the entire protein.

LROpred: the observed number of long-range contacts in two-state folders (known as long range order, LRO) has been shown to correlate with the proteins experimental folding rates (Gromiha & Selvaraj (2001) J. Mol. Biol. 310, 27-32). In the attempt to apply PROFcon predictions of non-local contacts to the prediction of experimental folding rates, we defined the following quantity: LROpred = N(S,T)/L*2, where N(S,T) is the number of pairs predicted by PROFcon with a score $\geq$T (the score is the raw neural network output) and separated by at least S sequence positions. L is the protein length. We used T=0.45 and S=14.

Results

Contact Predictions: we confirmed (Gorodkin, et al. (1999) Ismb, 95-105) that the information from the segment connecting two residues i and j improves the prediction of contacts and that this improvement is much more relevant for residues separated by fewer residues (sequence separation between 6 and 24). In general, PROFcon performance depended on the protein

length, structural class and family size. Shorter proteins and proteins with a larger number of available homologous sequences were better predicted. Alpha/beta proteins enjoyed the highest accuracy in long-range contacts predictions (sequence separation≥24). PROFcon's good performances were confirmed at the recent (2004) CASP6 experiment where our neural network was assessed as one of the best contact prediction methods.
Folding rates predictions: LRO_{pred} as calculated from PROFcon contact predictions correlated with LRO obtained from proteins structures. We then used LRO_{pred} to predict folding rates of two-state folders. Our estimates obtained using information from sequence alone were almost as accurate as estimates obtained from protein structure.

Contact email: mp2215@columbia.edu
URL: http://www.predictprotein.org/