A data base of minimally frustrated alpha-helical segments extracted from proteins according to an entropy criterion

Rita Casadio (a, b), Mario Compiani (a, c), Piero Fariselli (a) and Pier Luigi Martelli (a)

(a) CIRB, Biocomputing unity, University of Bologna;
(b) Laboratory of Biophysics, Dept.of Biology, Via Irnerio 42, I-40126-Bologna, Italy;
(c) Dept of Chemical Sciences, University of Camerino;

Supervised Neural Networks have been proved to be some of the most efficient tools to predict secondary structure of proteins from their aminoacid sequences. We developed a method that is able to evaluate the reliability of the predictions and the stability of helical structural motifs. A neural network with a 13 residue-long input window and a 2 neuron output is trained to recognize 2 classes: residues that have or not have a native a-helical structure in the protein data base. The two activations of the output neurons are interpreted as the probabilities for the central residue of the input fragment to be or not to be in helical structure and the Shannon entropy of the output is used as a measure of the prediction reliability [1].
A data base of minimally frustrated alpha helical segments is then defined by filtering a set comprising 822 non redundant proteins, which contain 4783 alpha helical structures. The data base definition is performed using the neural network-based alpha-helix predictor, whose outputs are rated according to an entropy criterion. A comparison with the presently available experimental results indicates that a subset of the data base contains the initiation sites of protein folding experimentally detected and also protein fragments which fold into stable isolated alpha-helices. This suggests the usage of the data base (and/or of the predictor) to highlight patterns which govern the stability of alpha helices in proteins and the helical behavior of isolated protein fragments.

[1] Compiani M., Fariselli P., Martelli P.L. and Casadio R., An entropy criterion to detect minimally frustrated intermediates in native proteins , Proc. Natl. Acad. Sci. USA 95(16): 9290-9294 (1998)