

Hidden Markov Models in cascade with neural networks generate a better predictor of segments of protein secondary structures

P. Martelli e Rita Casadio

Universita' di Bologna - via Irnerio 42, 40126 Bologna

Neural networks have been proved to be the most efficient methods for the protein secondary structure prediction (1). However one of the problems in using neural networks for sequence analysis is the independence of every prediction from the others. This is introducing noise in the results. For example the most evident effect is the presence of helices and strands one-residue long in the prediction of protein secondary structure. Since the shortest helical stretch is 3-residue long (if the 310 helices are included in this class) and the shortest strand is 2-residue long, these predictions can be regarded as 'syntax errors'. As a consequence, the length distributions of the segments of secondary structure predicted by Neural Networks is quite different from those of helices, strands and coil extracted from the atomic-resolved structures of the PDB database. In this work we propose a new kind of filter, based on Hidden Markov Models (2). We take advantage of their capabilities in capturing the duration of phenomena. For instance it is possible to include in HMMs a 'minimum length' constraint, in order to avoid the single residue predictions of helices and beta-strands. Our tool consists of a 6-states HMM: 3 states are labeled as helical states, 2 as strand and 1 as coil, following the minimal length observed in the database for these structural types. The transition probabilities among the states are then computed from the databases of known structures. Every residue along the sequence is emitted by each state with probability values equal to the outputs of the Neural Network predictor. The prediction of the HMM is given by the Viterbi-decoding, namely by the computation of the most probable path through the states of the model, given the Neural Network outputs. We prove that the filter doesn't affect the prediction efficiency of the Neural Network (72 % of overall accuracy when three structural states are discriminated). However its application improves considerably the length distributions of the predicted structures as compared to a Neural Network based filter previously adopted (3).

1. Rost B., Malyutina S., Marryng structure and genomics, *Structure* 6:259-263 (1998)
2. Durbin R., Eddy S., Krogh A., Mitchinson G.: *Biological Sequence Analysis*. Cambridge University Press, Cambridge (1998).
3. Vivarelli F., Fariselli P., Casadio R. - The prediction of protein secondary structure with a cascade correlation learning architecture of neural networks- *Neural Computing and Applications* 6:57-62 (1997)