A Combination of Support Vector Machines and Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction - (session: Structural Genomics)

Alessio Ceroni, Alessandro Vullo, Paolo Frasconi

Università di Firenze

We present a number of algorithms for improving the prediction of protein secondary structure. Our baseline predictor is based on support vector machines (SVM). Three separate classifiers are trained to discriminate helices, beta sheets, and coils from the rest, respectively. The margins of the three classifiers are subsequently
transformed into conditional probabilities by a multivariate normalized exponential function (softmax) whose parameters are estimated by maximum likelihood. We propose two different approaches to improve prediction accuracy. First, we train a bidirectional recurrent neural network (BRNN) as a structure-structure filter, i.e. we use the probabilistic SVM predictions as inputs and let the recurrent network to exploit upstream and downstream contextual information to refine predictions. Second, we use a Viterbi decoder (VD) controlled by a finite automaton that encodes prior knowledge on the minimum length of helices and beta sheets.

We validated our methods on a set of 979 proteins from PDB Select, defining a random split with 490 sequences for training and 326 for accuracy estimation. Multiple alignment profiles were generated by running psi-blast on a non-redundant set of proteins' chains. The SVM classifiers used a gaussian kernel and kernel hyperparameters were estimated using the remaining 163 sequences as a validation set.

Using VD on the top of SVM yields a 1.7% relative error reduction on Q3 and a 13% relative error reduction on segment overlap (SOV). Using the BRNN on the top of the SVM yields 5.8% relative error reduction on Q3 and 13.4% relative error reduction on SOV. Finally, the VD on the top of the combination of SVM and BRNN yields a further 3.4% relative error reduction on SOV, while Q3 remains similar. The advantage of the BRNN in the combination is particularly evident for the prediction of beta sheets that improves from 60.4% to 65.7% (a 13.3% relative error reduction). Our current best system achieves Q3=78% and SOV=74%.