Identifying Structure Transitions for Protein Secondary Structure Prediction

Armano G[1], Manconi A[1]

Motivation

Predicting protein secondary structures is still one of the open challenges in bioinformatics. To simplify, we can consider last generation predictors as a pipeline of three functional abstractions: i) encoding, ii) primaryto- secondary structure prediction, and iii) secondary-to-secondary structure prediction (also called refinement). We deem that interesting margins of improvement can be found in the refinement phase. In particular, since a significant number of prediction errors is associated with transitions, algorithms aimed at correcting secondary structure transitions might improve the performance of a predictor. In a previous work we assessed the behavior of existing predictors in predicting secondary structure transitions. The experiments performed have shown that only a small part of transitions are properly predicted, whereas most of these prediction are slightly anticipated or postponed. Experimental results performed to clarify this issue highlighted that repairing all errors within a window of 5 amino acids would give a significant improvement of Q3 (order of magnitude 3-5%, depending on the specific predictor). A first step towards this goal is to identify correct vs. non correct transitions. In so doing, it will be easier to put into practice suitable refinement procedures, which could be focused only on non correct transition predictions – leaving unchanged the correct ones.

Methods

The information about the structure of a protein being embedded into its amino acid sequence, studying the relationship between amino acid sequences and the corresponding structures can be useful to understand the principles that govern the folding of protein chains. In particular, the analysis of interactions that bring about secondary structure transitions can be useful to decide whether or not a transition is correct. To this end, we devised an approach based on decision trees, aimed at classifying correct vs. non correct transitions. Experiments have been carried out using the WHAT IF dataset of 9077 structures with resolution < 2.5 and R-factor < 0.25. Two thirds of the sequences have been used for training the system, whereas one third for testing and validation. A decision tree has been trained with data extracted by a moving window of fixed length run along each amino acids sequence. As for positive samples, the moving window has been centered on

[1] Department of Electrical and Electronic Engineering - University of Cagliari, Cagliari, Italy

actual transitions (i.e.. from alpha helices to coils, from beta strands to coils, from coils to alpha helices, and from coils to beta strands). Negative samples have been extracted by centering the moving window on splices with no transitions. The resulting decision tree has been assessed with predicted secondary structures. It is worth pointing out that the class distribution of the problem being imbalanced, accuracy is an inappropriate performance metric. Hence, we defined two metrics, say g+ and g- able to take into account the imbalance. The first is the geometric mean between positive predictive value and true negative rate, whereas the second is the geometric mean between negative predictive value and true positive rate. These metrics permit to assess separately the behavior of a decision tree in the task of identifying actual transitions. Let s be a threshold defined for g+ and g-. A value of g+ or g- greater than s shows that the system performs well in the task of discriminating between correct vs. non correct transitions. On the other hand, a value of g+ or g- lower than s shows that the decision tree does not perform well, for the predictor currently under analysis, in the task of discriminating between correct vs. non correct transitions.

Results

Experiments have been performed on six secondary structure predictors: GOR IV, Prof, Predator, SSPRO, JNET, and PSI-PRED. Preliminary results are promising. For some predictors the decision tree works very well, whereas for others the metrics cited above do not allow to obtain real advantages. For instance, analyzing predicted transitions from coils to beta strand in GOR IV, with a window size of 5 results a value of g+= 0.94.

Contact e-mail

manconi@diee.unica.it