Evaluating protein secondary structure predictions

Motivation *Secondary structure predictions* have by tradition been evaluated by their confusion matrices. A confusion matrix indicates, for each of the structural categories, the percentage of correct predictions (true positives) and the percentage of incorrect predictions (false positives) by category. The confusion matrix gives an overview of how good the prediction is, but lacks the information which is contained in the sequence, namely the extent to which the categories are predicted in segments. By the nature of secondary structure, the categories of helix and beta strand most often occurs in long stretches of the sequence, i.e. consecutive segments of helix and consecutive segments of beta strand. This means that a predictor that predicts interrupted segments of helix and strand is much less useful than a predictor that correctly predicts areas of consecutive structure that overlaps with the actual structural segments.

In 1994, the SOV measure was introduced as a supplement to the confusion matrix. The segment overlap measure gives a single score, which is based not only on the percentage of true positive predictions, but also on the amount of consecutive segments predicted. However, while using a single score is useful for ranking predictions by their quality, it is not of much use in identifying the strong and weak points of the predictions.

Method We propose *a new method of evaluating segment predictions.* The comparison statistics is an extension to the traditional confusion matrix. In addition to the confusion categories for each type of structure, we add sequence dependent subcategories for each structural category: Helix Overprediction categories - Predicted helix segment beyond the limit of the true helix segment (extension) - Predicted helix segment extended across a gap in the true helix segment (bridge) Helix Underprediction categories - Predicted helix segment shorter than the true helix segment (shortage) - Predicted interrupted helix segment within a non-interrupted true helix segment (gap)

An overview of the comparison statistics will then reveal the quality of the segment prediction, and can also be used to identify strengths and weaknesses of specific predictions. The segment comparison can be generalised and applied, not only to secondary structure predictions, but also to other sequence comparisons containing predicted segments to compare.

Project The project goal is to *implement the category statistics* for three structural states, so it can be used to evaluate secondary structure predictions. This calculation of the category statistics can be implemented as a *state machine* that simultaneously reads the predicted and true sequence. If time allows the project will also include a comparison of the segment prediction statistics with the SOV measure on *existing datasets* of secondary structure predictions.

Contact The project is a collaboration between IMADA (Lene Favrholdt, lenem@imada.sdu.dk) and Centre for Molecular and Biomolecular Informatics (*Fiona Nielsen*, fnielsen@cmbi.ru.nl)

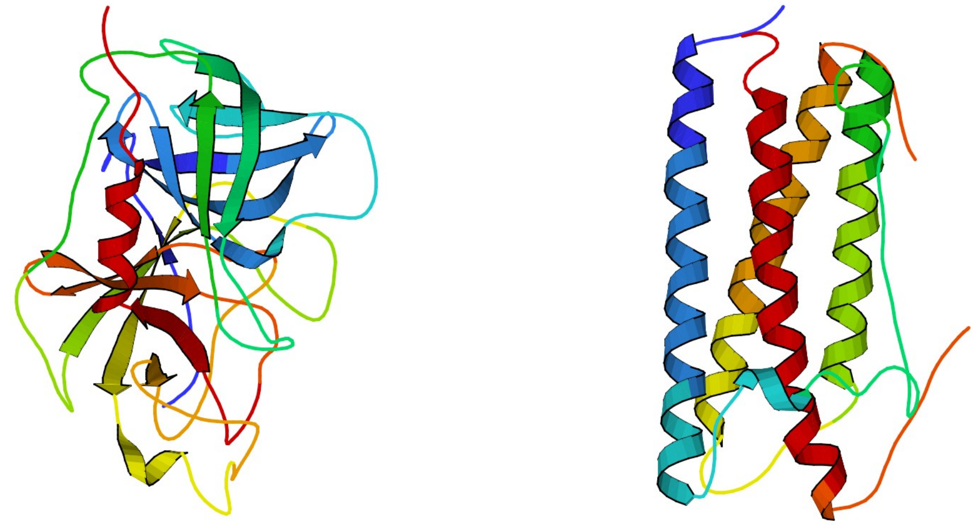

Figure 1: To protein-strukturer