To be presented as a poster at ISMB '99, Heidelberg, August 6 - 10, 1999

CoDe (*Co*nsensus formation by *De*cision tree learning) A new method for consensus protein secondary structure prediction

Theo Mevissen, Joachim Selbig and Thomas Lengauer

Institute for Algorithms and Scientific Computing (SCAI)
GMD - German National Research Center for Information Technology
Schloß Birlinghoven, D-53754 Sankt Augustin, Germany

{mevissen,selbig,lengauer}@gmd.de

Predicting secondary structure as exactly as possible for protein sequences of unknown structure provides information that is useful for other protein structure prediction methods like fold recognition and *ab initio* 3D prediction. A consensus prediction constructed from the output of several methods should yield more reliable results than each of the individual methods. Consensus formation by combining the results of several prediction methods may be considered as applying consensus rules of the form "*IF* . . . *THEN* . . ." where the condition consists of a conjunctive combination of predicted secondary structure states and the conclusion is the predicted consensus state. Such rules are directly coded by paths inside decision trees. Therefore, decision tree learning is our machine learning method of choice in order to find the best set of consensus rules.

We tested our approach described in more detail in (1) on the new dataset CB396 of 396 protein domains proposed in (2) and on 11 CASP3 prediction targets (3). The individual secondary structure predictions were taken over from the JPRED (4) and the NPS (5) servers.

For evaluating the results we used the Q_3 value (fraction of residues predicted correctly in the secondary structure states helix, strand, and coil), its standard deviation, and the segment overlap value (SOV).

First results achieved by our system show that consensus formation based on decision tree learning improves the accuracy of the prediction by more than 10 % in some cases. The CoDe system will be integrated into the *ToPLign* system (*To*olbox for *P*rotein a*Lign*ment) which can be accessed at http://cartan.gmd.de/ToPLign.html (6).

1. J. Selbig, T. Mevissen, and T. Lengauer: Decision tree-based formation of consensus protein secondary structure prediction. Submitted to *Bioinformatics*.
2. J.A. Cuff and G.J. Barton,G.J.: Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins Struct. Funct. Genet.*, 34 (1999) 508–519.
3. http://moult.carb.nist.gov/casp/
4. http://circinus.ebi.ac.uk:8081/
5. http://pbil.ibcp.fr/
6. H. Mevissen, R. Thiele, R. Zimmer, and T. Lengauer: The ToPLign software environment - Toolbox for protein alignment. *Bioinformatik '94*, Institut für molekulare Biotechnologie, Jena, 1994.