A DECISION GRAPH EXAMPLATION OF PROTEIN SECONDARY STRUCTURE PREDICTION

D.L. Dowe, J. Oliver, L. Allison, C.S. Wallace & T.I. Dix

ABSTRACT

Oliver and Wallace recently introduced the machine-learning technique of decision graphs, a generalisation of decision trees. Here it is applied to the prediction of protein secondary structure to infer a theory for this problem. The resulting decision graphs provides both a prediction method and, perhaps more interestingly, an explanation for the problem. Many decision graphs are possible for the problem; a particular graph is just one theory or hypothesis of secondary structure formation. Minimum message length encoding is used to judge the quality of different theories and, for example, prevents learning the *noise* in the training data. Minimum message length encoding is a general technique in inductive inference.

The predictive accuracy for 3 states (Extended, Helix, Other) is in the range achieved by current methods. Many believe this is close to the limit for methods based on only local information. In addition, a decision graph is an explanation of a phenomenon. This contrasts with some other predictive techniques which are more "opaque". The reason for a prediction can be of more value than the prediction itself. Rules containing both conjunction and disjunction can be extracted from a decision graph in a natural way which makes for good expressive power. For example, they can express pair-wise interactions and conditional rules. Groups of related amino acids can be identified from the graph. The rules are probabilistic and indicate a degree of confidence in a prediction. The more confident predictions are generally more accurate. For this initial application of decision graphs to protein structure problems, the data comprised 10K amino acids from 50 proteins. The data was divided into various training and test sets.