

Additional Paper for Biological Data Mining: “Predicting Post-synaptic activity in proteins with data mining”

Paper reference:

“Genome Scale Prediction of Protein Functional Class from Sequence using Data Mining”. Ross D. King, Andreas Karwath, Amanda Clare & Luc Dehaspe. *Conference on Knowledge Discovery in Data, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. Pages: 384 – 389.

Abstract

This paper tackles the problem of predicting protein function from amino acid sequence which has always been an area of active research. The goal is also to come up with rules that are easily interpretable by biologists. The authors apply a hybrid data mining algorithm to predict protein function from sequence using the *Mycobacterium tuberculosis* genome as an example. They use a data set for training which consists of a hierarchically arranged assignment of functions to proteins. The actual data mining is done using a hybrid of clustering and rule learning. The use of clustering is suggested since it improves the representation for learning using the expressive power of inductive logic programming (ILP). WARMR, an ILP data mining algorithm was used to identify frequent patterns in the sequence descriptions. Nearly, 18,000 frequent queries were identified which were then encoded into binary presence/absence features. The rule learning was done using the familiar C4.5 and C5 algorithms. The rules were selected based on performance on a validation set and the unbiased accuracy of these rules was estimated on a test set. The authors take efforts to balance accuracy with unidentified gene coverage. The tradeoff between commission and omission (making incorrect predictions Vs missing genes) is identified and dealt with. The authors observe that it is possible to come up with good rules that predict function from sequence at all levels of the functional hierarchy. The focus is on coming up with previously unknown, biologically interpretable rules that can predict protein function even in the absence of identifiable sequence homology.

Discussion

In the paper discussed in class, the focus was on applying a data mining algorithm, namely C4.5 to the task of predicting whether or not a protein has post-synaptic activity. This was essentially a binary classification problem that made use of various features of the protein like its PROSITE patterns, molecular weight and sequence length. In the current paper, the authors employ a hybrid of clustering and rule learning algorithms to solve a multi-class classification problem. The rule learning algorithms used are C4.5 and C5. Here the objective is, given a training set, to be able to predict protein class from function. This method is an improvement on using methods based on direct sequence similarity to learn the mapping between sequence and function. Such methods can be considered as approximations to nearest neighbor type functions. The more complicated homology recognition methods can be considered to be analogous to case based learning functions. The training set here is a hierarchical, functional classification of proteins. The objective is then to learn discriminatory functions to map sequence to biological function. The focus in both papers is primarily twofold. Firstly, to obtain good predictive accuracy and secondly to come up with rules that are comprehensible and ‘interesting’ to biologists. In addition, the learnt rules also provide evolutionary insight. Even though the actual function of a gene can only be determined by ‘wet’ methods, methods like the one suggested in this paper make this experimental validation much more efficient by restricting it to just high probability hypotheses. The authors also raise questions about the evolutionary causation of these general rules, outlining several possibilities to explain their existence. Both papers address the issue of automatic knowledge discovery with respect to proteins using similar data mining algorithms, the emphasis being as much on interpretability of rules learnt as it is about the accuracy of prediction. The authors foresee that tapping into bio-informatics data from various other sources like expression profiles, pathway analysis, structural studies etc. and combining them would be likely to produce more powerful predictions than using any single one in isolation.