The Challenge of Predicting Gene Function

Ross D. King

Department of Computer Science, University of Wales, Aberystwyth

Abstract. The biological sciences are undergoing an explosion in the amount of available data. New data analysis methods are needed to deal with the data. A central problem in bioinformatics is the assignment of function to sequenced open reading frames (ORFs). The most common approach is based on inferred homology using a statistically based sequence similarity (SIM) method e.g. PSI-BLAST. Alternative non-SIM based bioinformatic methods are now becoming popular. Application of machine learning methods to gene function prediction presents a number of challenges: the data is inherently relational, has multi-class labels, contains a large number of sparsely populated classes, the requirement to learn a set of accurate rules (not a complete classification), and the existence of a very large amount of missing values. We have developed a method we term Data Mining Prediction (DMP) to deal with these problems. DMP is based on combining evidence from sequence, predicted secondary structure, predicted structural domain, InterPro patterns, sequence similarity profile and expressions data; and uses a combination of Multi-relational (Inductive Logic Programming) data mining, and decision trees to produce prediction rules for functional class. DMP predictions are more general than is possible using homology. Our original work on DMP was applied to bacterial genomes. In 2000/1 DMP was used to make public predictions of the function of 1309 E. coli ORFs. Since then biological knowledge has advanced allowing us to test our predictions. We have therefore examined the updated annotations our predicted ORFs, and examined the scientific literature for direct experimental derivations of function. Both tests confirmed the DMP predictions. Accuracy varied between rules, and with the detail of prediction, but they were generally significantly better than random. For voting rules, accuracies of 75-100one of these DMP predictions have been confirmed by direct experimentation. The DMP rules also have interesting biological explanations. DMP is, to the best of our knowledge, was the first non-SIM based prediction method to have been tested directly on new data. We have since extended to deal with large eukaryotic genomes, first yeast (S. cerevisiae), and then Arabidopsis thaliana. Accurate rules were learned and predictions made for many of the ORFs whose function is currently unknown. These rules are informative, agree with known biology and enable for scientific discovery.