Protein function prediction and classification using uncertainty

James Bradford*[1], Chris Needham[2], Andy Bulpitt[2] and David Westhead[1]

[1] School of Biochemistry and Microbiology, University of Leeds
[2] School of Computing, University of Leeds

1 Introduction

We are currently investigating the use of Bayesian networks to integrate information, express relationships and make inferences or predictions on biological problems, motivated by data generation in genomics and proteomics. A Bayesian network is a directed acyclic graph, which encodes probabilistic relationships among the variables of interest. Nodes of the graph represent random variables and edges represent conditional dependencies between these variables. There are several benefits of using Bayesian networks:

- They can handle missing data, as they encode dependencies between input variables
- They can be used to learn causal relationships.
- The combination of causal and probabilistic semantics allows the combination of data and prior knowledge.
- In conjunction with Bayesian methods they offer an efficient and principled way of avoiding over fitting of data. All available data may therefore be used in training and useful information may be drawn from limited data.

Our initial objective is to demonstrate the use of Bayesian networks for prediction of protein-protein interfaces and the functional effects of single nucleotide polymorphisms (SNPs; a type of genetic mutation) using attribute and training sets from previous work (Krishnan and Westhead, 2003; Bradford and Westhead, 2005). This will lead to research into the implementation of an existing classification ontology such as the Gene Ontology (GO) as a Bayesian network to handle uncertain data and relate functional categories. In all cases, Bayesian networks will be compared with other methods, including support vector machines (SVMs), decision trees and standard neural networks in terms of prediction performance and usability issues.

We will be presenting preliminary results of Bayesian network construction and prediction on the functional effects of SNPs, and protein-protein interfaces. We will also present some statistical analysis into the dependencies between the input variables (attributes such as sequence conservation) in order to ascertain the most predictive properties of a protein-protein interface or SNP functional effect. Any differences between these results and those from our previous work using SVMs (Krishnan and Westhead, 2003; Bradford and Westhead, 2005) and decision trees (Krishnan and Westhead, 2003) will also be highlighted.

References

Bradford, J.R. and Westhead, D.R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487-1494.

Krishnan, V.G. and Westhead, D.R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199-2209.