Machine learning to predict gene and protein function
James Bradford, Matthew Care, Andrew Garrow, and David Westhead

Introduction
Machine learning techniques are being applied to several biological problems, in collaboration with groups in computer science and statistics. Projects employ a variety of learning methods including support vector machines, decision trees and Bayesian networks, and the applications range through protein structure prediction, the prediction of gene function and the effects of mutations, and the prediction of protein interactions. Following our earlier work in these areas, this year has seen a major new effort in Bayesian network learning, which has provided a successful avenue of attack to predict protein interactions and the effects of mutations. This next year, we will focus on the new problem of predicting the relatedness of gene function from ‘-omics’ data using the Gene Ontology.

Protein-protein binding site prediction
Identifying the interface between two interacting proteins provides important clues to the function of a protein, and is becoming increasing relevant to drug discovery. This last year we have focused on predicting both protein-protein binding site location and interaction type using Bayesian networks in combination with surface patch analysis. In doing so, insights have been gained into the properties that characterise a binding site and drive complex formation. Our method predicts protein-protein binding sites with a high success rate of 82% on a benchmark dataset of 180 proteins, improving on previous work by 6% (see Bradford & Westhead 2005). The method was also able to handle incomplete datasets automatically. With this in mind, we also carried out a study on the Mog1p family for which evolutionary information was sparse and were able to suggest binding sites for Ran and other signalling proteins on Mog1p itself. Our results on other members of the family suggest that proteins can still bind to different proteins and probably have different functions even though they share the same overall fold. We also demonstrated the applicability of our method to drug discovery efforts by successfully locating a number of binding sites involved in the protein-protein interaction network of papilloma virus infection. In a separate study of obligate and non-obligate interfaces, we found that such was the similarity between the two types, we were able to use obligate binding site properties to predict the location of non-obligate binding sites and vice versa.

Modelling the effect of missense mutations on protein function
Prediction of the effects of non-synonymous single nucleotide polymorphisms (nsSNPs) has been studied by various research groups using a variety of probabilistic and machine learning tools. Most methods use a range of structural and sequence attributes to try and predict deleterious or missense mutations that affect protein function.

Bayesian networks have successfully been applied to two protein mutagenesis datasets (lac repressor and T4 lysozyme) yielding results that are comparable with those produced by other machine learning techniques. In addition, the results showed that Bayesian networks generalise well to new data, are robust to training from incomplete data, and handle missing data such as structural or evolutionary information. Having discovered the most important contributors to prediction, we reduced our Bayesian network from 15 to only four nodes. This simpler model, even though no evolutionary information was used, maintained similar classification performance to the full network.
Current work has involved producing a larger dataset of SNPs to more accurately predict their effects. The Swiss-Prot Variant database of Human protein variants was parsed to generate ~12,000 disease SNPs (from ~1000 proteins) and ~8,000 polymorphic SNPs (from ~3000 proteins). It is hoped that this diverse "real world" dataset can be used to train machine learning algorithms to analyse existing un-annotated SNP databases.

**Searching genomes for trans-membrane barrel proteins**

Trans-membrane barrel (TMB) proteins are a functionally important and diverse group of molecules found spanning the outer membranes of Gram negative and acid fast Gram positive bacteria, mitochondria and chloroplasts. Structurally they are well understood with entries from over 23 families in the protein databank (PDB). However, unlike with alpha helical trans-membrane proteins, development of TMB computational screening techniques has proven difficult with TM strands composed of a short and aliphatic, inside-outside dyad repeat motif.

In this project high accuracy composition based discrimination algorithms have been developed using a number of machine learning techniques (e.g. support vector machines (SVMs) and genetic algorithms; see Garrow *et al.* 2005). Another related project has focused on development of Hidden Markov Models for detection of trans-membrane strands.

**Collaborations**

Drs. Andy Bulpitt and Chris Needham in the School of Computing, University of Leeds. Dr Alison Agnew in the School of Biology, University of Leeds

**Publications**


**Funding**

This work is funded by the MRC, BBSRC, and the BBSRC E-Science Initiative