Machine learning to predict gene and protein function

Andrew Garrow, James Bradford, Matthew Care 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 the start of a major new effort in Bayesian network learning, which will provide a new avenue of attack to predict protein interactions and the effects of mutations, and open the new problem of predicting the relatedness of gene function from ‘-omics’ data using the Gene Ontology.

Protein function prediction and classification using uncertainty

The aim of this project is to investigate 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. This aim will be met through the following objectives:

• to demonstrate the use of Bayesian networks for prediction of functional effects of single nucleotide polymorphisms (SNPs) and other mutations;
• to construct a Bayesian network for prediction of protein-protein interaction interfaces;
• to research 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;
• and, in all cases, to compare Bayesian networks with other methods, including support vector machines (SVMs), decision trees and standard neural networks in terms of prediction performance and usability issues.

The most advanced aspect of this project currently is the prediction of the effects of non-synonymous single nucleotide polymorphisms (nsSNPs). This has been studied by various research groups using a variety of probabilistic and machine learning tools. All of the methods use a range of structural and sequence attributes to try and predict deleterious mutations, those affecting protein function.

This project aims to use Bayesian networks for SNP prediction problems, using a diverse set of attributes to predict the effects of mutations. Bayesian networks have several advantages over other machine learning techniques. Firstly the structure of the Bayesian network is a pictorial representation of the assumptions made for the prediction process, showing the causal links between the attributes and the possible effect. In addition Bayesian networks can function with reasonable accuracy even if there are missing attributes in the sample for which a prediction is to be made. Thus they can be trained using both protein structural and sequence attributes without any punishment for proteins for which structural information has not been derived.

An additional project goal is to use the final Bayesian networks to produce a SNP prediction web server. This would allow SNP queries to be submitted to a server, with the predictions and derived attributes returned by email. This would then be followed up by a
comprehensive SNP study using a variety of the best Bayesian network models on a dataset derived from the OMIM database.

**Searching genomes for transmembrane barrel proteins**

Transmembrane 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 transmembrane 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 are being developed using a number of machine learning techniques (e.g. support vector machines (SVMs) and genetic algorithms). Another related project is focusing on development of Hidden Markov Models for detection of transmembrane 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