Global analysis of SNPs, proteins and protein-protein interactions: approaches for the prioritisation of candidate disease genes.

Dobson, Richard James Butler

Global analysis of SNPs, proteins and protein-protein interactions: approaches for the prioritisation of candidate disease genes.

Richard James Butler Dobson, BSc

Clinical Pharmacology and the Genome Centre, William Harvey Research Institute,

Bart's and The London School of Medicine and Dentistry,

Queen Mary University of London

PhD thesis

2009

Abstract

Understanding the etiology of complex disease remains a challenge in biology. In recent years there has been an explosion in biological data, this study investigates machine learning and network analysis methods as tools to aid candidate disease gene prioritisation, specifically relating to hypertension and cardiovascular disease.

This thesis comprises four sets of analyses: Firstly, non synonymous single nucleotide polymorphisms (nsSNPs) were analysed in terms of sequence and structure based properties using a classifier to provide a model for predicting deleterious nsSNPs. The degree of sequence conservation at the nsSNP position was found to be the single best attribute but other sequence and structural attributes in combination were also useful. Predictions for nsSNPs within Ensembl have been made publicly available.

Secondly, predicting protein function for proteins with an absence of experimental data or lack of clear similarity to a sequence of known function was addressed. Protein domain attributes based on physicochemical and predicted structural characteristics of the sequence were used as input to classifiers for predicting membership of large and diverse protein superfamiles from the SCOP database. An enrichment method was investigated that involved adding domains to the training dataset that are currently absent from SCOP. This analysis resulted in improved classifier accuracy, optimised classifiers achieved 66.3% for single domain proteins and 55.6% when including domains from multi domain proteins. The domains from superfamilies with low sequence similarity, share global sequence properties enabling applications to be developed which compliment profile methods for detecting distant sequence relationships.

Thirdly, a topological analysis of the human protein interactome was performed. The results were combined with functional annotation and sequence based properties to build models for predicting hypertension associated proteins. The study found that predicted hypertension related proteins are not generally associated with network hubs and do not exhibit high clustering coefficients. Despite this, they tend to be closer and better connected to other hypertension proteins on the interaction network than would be expected by chance. Classifiers that combined PPI network, amino acid sequence and functional properties produced a range of precision and recall scores according to the applied

weights.

Finally, interactome properties of proteins implicated in cardiovascular disease and cancer were studied. The analysis quantified the influential (central) nature of each protein and defined characteristics of functional modules and pathways in which the disease proteins reside. Such proteins were found to be enriched 2 fold within proteins that are influential ($p<0.05$) in the interactome. Additionally, they cluster in large, complex, highly connected communities, acting as interfaces between multiple processes more often than expected. An approach to prioritising disease candidates based on this analysis was proposed.

Each analyses can provide some new insights into the effort to identify novel disease related proteins for cardiovascular disease.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The science of biology describes the organisation and processes of organisms at each level ranging from the molecular up to the ecosystem (Roberts & King, 1987). Each level has connected, complex systems and understanding the relationships and connections between the component parts is an important challenge. Key to understanding each level has been the move away from reductionist approaches to wholist approaches (Katagiri, 2003). Reductionists focus on one element of a system with the aim to learn everything about that element. Reductionist approaches have been rather successful and continue to be important in understanding the details of the component parts which will facilitate a systems level understanding since we know more about the individual building blocks. In contrast, wholist approaches observe all the components at a specific level together.

Wholist approaches have led to a range of terms with the suffix *-ome* (eg. genome, transcriptome, proteome), all are used to describe all the components of a system at a particular level. Terms ending in *-omics* (eg. genomics, transcriptomics, proteomics) are used to describe the approaches and technologies for studying each level, and allow us to get a snapshot of these whole systems at a particular level (-omes and -omics glossary taxonomy, 2009). Genomics technologies include genome wide linkage screens where variable Simple Sequence Repeats (SSRs) are used as molecular markers for a range of applications including mapping disese genes and forensics. Genome wide association (GWA) studies are a second approach, in this case the variation in genotype frequen-

cies of markers such as single nucleotide polymorphisms (SNPs) are compared between cases and controls or are tested for association with a quantitative trait. SNPs are genetic variations representing a simple single base pair difference (allele) between individuals at a particular position within the DNA sequence, they are expected to contribute to the causes of many complex traits. SNP arrays enable identification of SNP variation across the genome using 1000's of SNPs in 1000's of people and potentially identifying associations with disease. Transcriptomics technologies include expression microarrays for identifying genes that are over or under expressed in disease affected tissue relative to normal tissues. Other techniques include the analysis of epigenetic changes leading to phenotypic variation through mechanisms such as DNA methylation (epigenomics) and the study of protein-protein interactions (PPI) through approaches such as yeast two-hybrid screens, potentially leading to improved understanding of biological functions (interactomics). At all levels, the relationships between components of a system are of interest and can be considered as networks where the components are vertices and the relationships are edges (Junker & Schreiber, 2008). These large emerging datasets from each network provide scientists with a wealth of data that has to be explored, described and understood using approaches that include data mining and the application of machine learning and network analysis.

A very important application is in the study of diseases, which often involves the disruption of a functional pathway involving multiple genes and their products. In genome wide linkage screens regions as large as 30 million bases (30cM) have been identified. In GWA studies the associated regions tend to vary in size from a few kilobases (Kb) to 1000's of Kb (McCarthy *et al.*, 2008). With both types of study, investigators are left with huge expanses of DNA which contain many hypothetical genes. There is a need to devise strategies to aid in the identification and prioritisation of genes within these regions. The function of hypothetical genes must also be considered. It is also important to be able to isolate the functional SNPs from the multiple SNPs that are inherited together in a linkage disequilibrium (LD) block. Candidate genes may also be prioritised and knowledge of disease etiology may be acquired by biological network analysis of protein-protein interactions contained within the human interactome.

The work described in this thesis investigates approaches for predicting deleterious SNPs and protein function and performing topological analysis of protein-protein interaction (PPI) networks for the identification and prioritisation of candidate genes for complex diseases. The studies are focused on datasets from cardiovascular disease (cvd), hypertension and cancer, but the methods can potentially be applied to any disease phenotype.

The organisation of the thesis is as follows:

- Chapter 1 introduces the rationale for the studies and the focus areas and discusses the applied computational principles of machine learning and graph theoretic approaches.

- Chapter 2 describes the methods and results obtained from analysing SNPs, using supervised machine learning classifiers for predicting deleterious non-synonymous SNPs (nsSNPs).

- Chapter 3 describes the methods and results for predicting protein superfamily using classifiers with a set of sequence based attributes. This analysis focuses on large diverse superfamilies where it is difficult to assign function using traditional sequence homology based methods.

- Chapter 4 describes the methods and results obtained from surveying topological properties of hypertension related proteins within the human interactome. The resultant hypertension protein network properties are combined with sequence and functional based information to build a model for predicting novel candidate hypertension related proteins.

- Chapter 5 describes interactome analysis of proteins implicated in cvd and cancer. The influential nature of these proteins is quantified and community structures are analysed. An approach for prioritising cvd candidate genes is shown.

- Chapter 6 highlights the primary results, compares with previous work and describes possible future work and directions within the studied fields.

1.2 An overview of Machine Learning

The increasing amount of information from numerous genomes and the easy access on the world wide web has led to many opportunities for bioinformatics research. Algorithms are required to extract information, knowledge and patterns within this data. Such algorithms can be used to search the genomic space to determine a hypothesis that fits the space. In this chapter we provide an overview of machine learning and describe the main methods used in this thesis.

Artificial intelligence (AI) falls within the field of computer science and engineering, it aims to produce computer programs that can cope with problems requiring intelligent behaviour, learning and adaptation. Machine learning is a branch of AI concerned with the development of algorithms for learning (Michalski *et al.*, 1983). Deductive learning is where a conclusion is arrived at using previously known facts or fulfilling conditions. The conclusion is always true on condition of the facts being true. In contrast, inductive learning is where the facts may predict a conclusion with a probability, but there is no guarantee of the conclusion being true. Machine learning methods use inductive learning techniques to create programs by producing rules based on patterns within data sets. Simple pattern discovery alone may be more accurately classified as data mining. Machine learning has many uses within the field of bioinformatics where patterns and rules are used for well characterised examples to classify instances that are less well understood.

1.2.1 Supervised vs unsupervised algorithms

The three most common types of machine learning algorithms are: *supervised learning*, *unsupervised learning* and *semi-supervised learning*. *Supervised learning* or *classification learning* takes a set of examples that are classified, and creates a set of rules to classify samples where the status is unknown. In contrast, *unsupervised* learning models a set of inputs where labelled examples are not available. The algorithm finds a way of clustering the data based upon the known features and then provides descriptions for these clusters. *Semi-supervised learning* utilises both labelled and unlabelled instances in order to create a classifier.

Predicting whether a non-synonymous SNP (nsSNP) is disease related or whether

a protein belongs to a certain functional group are questions that can be addressed via machine learning methods. The method of *supervised learning* is appropriate as the aim is to assign an instance, either a nsSNP or an unannotated protein to one of a number of classes. In the case of nsSNP classification it is possible to use a set of nsSNPs where the disease status is known as a training set to form a set of rules that could be used to make a prediction for nsSNPs where the function is unknown. The work performed in this thesis utilises *supervised learning* exclusively, so the focus from this point will be on this approach. The supervised learning classifiers, support vector machines (SVMs) and decision trees (described below) are amongst the most commonly used classifiers within the field of bioinformatics.

1.2.2 Support Vector Machines

Support vector machines (SVMs) are a kernel based *supervised learning* classifier developed by Corte & Vapnik (1995). They have been shown to be very accurate in many disciplines including bioinformatics, benefitting from the ability to handle high dimensional data with a small number of instances, finding a good balance between training set accuracy and test data error. For a given set of training vectors labelled with two classes, a SVM can find the optimal linear hyperplane that maximally separates instances of the classes by maximizing the margin between the two classes (Figure 1.1).

1.2.2.1 Non-linear classification

Very often problems are not immediately linearly separable, and so the vectors must be transformed into some higher dimensional space and the optimal hyperplane found in this transformed feature space. Non-linear discrimination can be achieved through the application of a range of *kernel functions*. The performance of the SVM is controlled by this function and the regularization of the C parameter. The C parameter is used to trade between training errors and larger hyperplane margins.

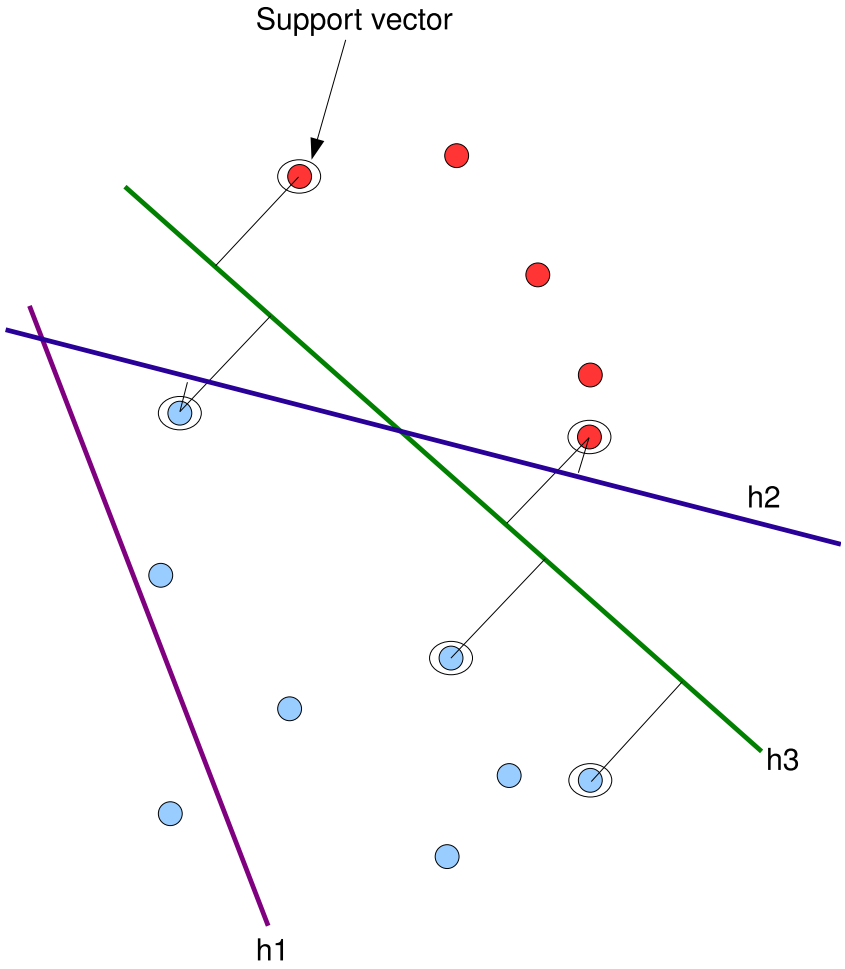


Figure 1.1: Support vector machine (SVM) hyperplanes. Three hyperplanes are displayed; h1 does not separate the two classes (orange and blue circles), h2 separates the classes with a small margin, h3 with the maximum margin. The support vectors are circled

1.2.2.2 MultiClass SVM

If there are more than two classes, various SVM techniques have been designed to overcome the problem.

One-vs-Others method The one-vs-others method is a simple method for dealing with multi-class problems containing *n* classes (Brown *et al.*, 2000). The problem is transformed into *n* 2 way classifiers. Each classifier contains a single class as 'class 1' and all of the others classes combined as 'class 2'.

For a query instance where the class is unknown the system tests against each of the 2 class models to see whether it belongs to 'class 1' or 'class 2'. This leads to *n* scores from the *n* classifiers. Ideally there will only be one case where the query is assigned to 'class 1'. In reality there may be false positives, whereby more than one of the models assigns the query to 'class 1'. The complexity of 'class 2' may lead to the false positives.

Unique One-vs-Others method The unique one-vs-others method adds a second step to the one-vs-others method for dealing with instances where there are false positives (Ding & Dubchak, 2001). This step involves the creation of 2 class classifiers for each of the false positives. The final assigned class is the class that was selected most in these models built from the false positives. In this step false positives should be eliminated.

One-vs-One method In the unique one-vs-others method 2-way classifiers are built to break the ties between the false positives. In the one-vs-one or pair wise coupling method, the first step is abandoned altogether so the process is composed solely of the second step of one against one classifiers (Hastie & Tibshirani, 1998). The final chosen class is the one that receives the most votes from each of these pair wise classifiers. This approach is used by implementations such as SVM SMO (Platt, 1998) and LibSVM (Chang & Lin, 2001).

1.2.3 Decision Trees

Decision trees are supervised classifiers composed of a graph (tree structure) of decisions (Quinlan JR, 1993). Each interior node of the tree relates to a variable where a decision is

made on which branch to take based on the value of the variable. The decisions are usually simple single attribute tests to divide the data. A leaf represents the predicted class based on values at the nodes on the path from the root. Decision trees have the advantage over many classifiers in that they produce interpretable rules. Once a tree has been built new instances can be classified by starting at the root and following a path down to a leaf. An example of a decision tree can be seen in Figure 1.2 where an activity for the day is chosen based on a number of attributes.

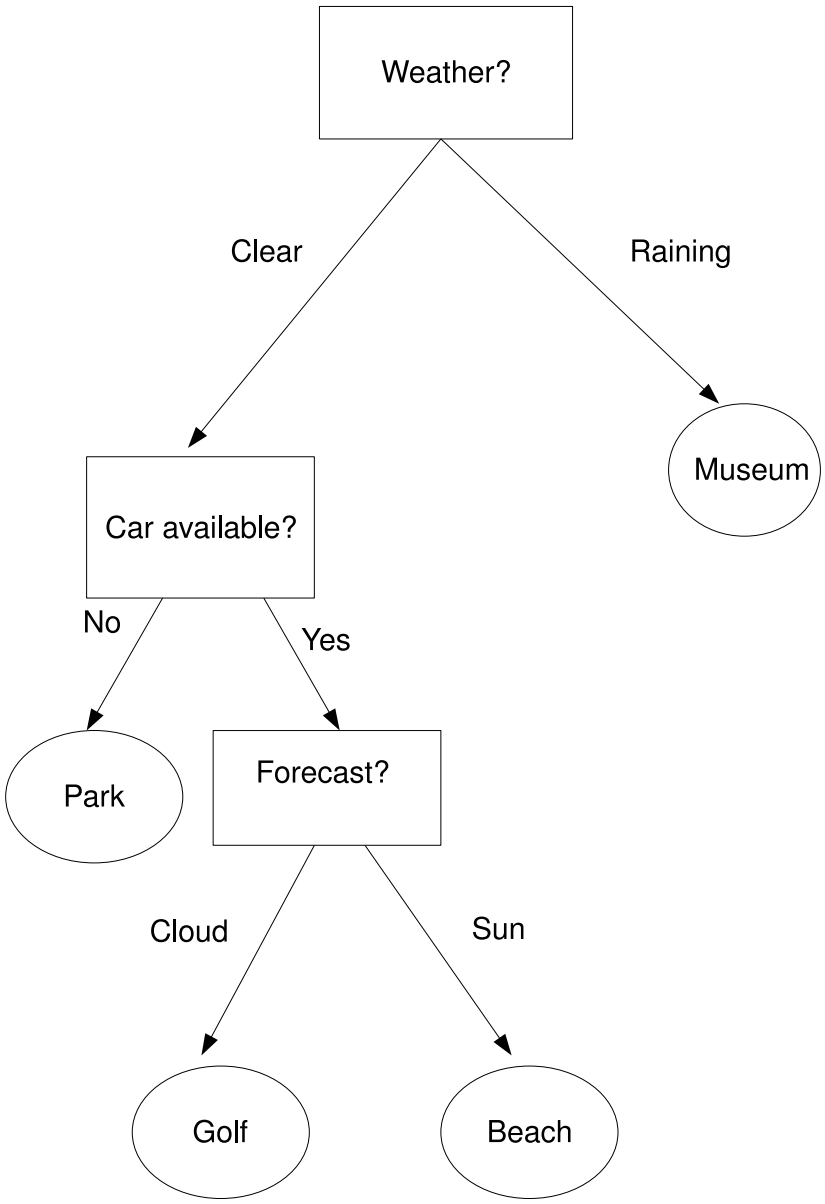


Figure 1.2: An example of a decision tree for choosing a weekend activity, showing decisions at the nodes, and final classification at the leaves.

When the attribute at a node is nominal, there will be one branch for each attribute value. If the attribute is continuous then it will usually be split into 2 and a decision based on whether the instance is above or below a threshold cut-off. There are a number of methods for deciding which attribute should be used at each node. Information gain of the split is a commonly used measure, which measures the information required after using the attribute as a classifier at a node subtracted from the information required before using the attribute as a classifier. The Gini measure calculates statistical dispersion defined as a ratio between 0 and 1 with lower values representing equal distribution.

Decision trees apply varied criteria for halting tree growth and then pruning it back. This is done to prevent trees being produced that are too specific to the training dataset. The aim is to produce a tree that is general enough to be applied to any new instances that require clasification, avoiding overfitting. The algorithms are efficient and therefore able to handle large volumes of data due to the simple partitioning approach taken by the algorithm. However, one drawback to this divide and conquer approach is that the divisive partitioning can mean that interesting relationships between attributes within the data can be separated early on.

A very popular decision tree algorithm and one used in this thesis is C4.5 (Quinlan JR, 1993). It is a very easy to use algorithm and is commonly used within bioinformatics. Performance of this classifier is often used as a benchmark to which other classifiers are compared. This algorithm uses information gain to partition the data at each node. The algorithm is capable of handling many types of attributes: empty nominal attributes, nominal attributes, numeric attributes, unary attributes, missing values, binary attributes, and date attributes.

Random forest (RF) is a supervised classifier consisting of multiple decision trees (Breiman, 2001) whereby the final class selected for an instance is the mode class selected by the multiple decision trees. RF combines two machine learning methods of ‘bagging’ and ‘random feature selection’. Each tree is created from a bootstrap sample of the training data where about one-third of the cases are left out (out-of-bag (OOB) data). OOB data is used to obtain an unbiased estimate of the error during the training. This is known as bagging. RF extends bagging because rather than using all features, RF ran-

domly selects a subset of input variables to decide what decision should be made at each node of the tree. Advantages of random forest classifiers include the fact that the error can be balanced when the class population sizes are imbalanced and there are good methods for handling missing data and overfitting can be avoided. The algorithm can handle the same array of attribute types as C4.5 (Quinlan, 1993).

Another decision tree based classifier used in this thesis is the PART decision list which uses a separate-and-conquer approach. The algorithm builds a partial C4.5 decision tree in each iteration and makes the 'best' leaf into a rule (Frank & Witten, 1998a). Again, the algorithm can handle the same array of attribute types as C4.5.

1.2.4 Weka workbench

Weka is a freely available collection of machine learning algorithms for data mining tasks (Witten & Frank, 1999) available from the web site `http://www.cs.waikato.ac.nz/ml/weka/`. The work in this thesis extensively used this workbench and its implementations of the various machine learning algorithms. The workbench has its own implementation of the C4.5 algorithm called J48. The algorithms can be applied to a dataset through a graphical user interface (GUI), using a command line interface (CLI) or called from Java code directly. Tools are included for data pre-processing, classification, regression, clustering, association rules, and visualization. The weka workbench is a commonly used package within bioinformatics, with an extension library called BioWeka, created specifically for many common bioinformatics related tasks (Frank *et al.*, 2004).

1.2.5 Generating a classifier

In performing successful pattern recognition using classifiers there are a number of general steps performed (Figure 1.3). Initially, the dataset of instances and features are collected, feature selection is then performed, the classifiers are trained, parameters are then tuned for the chosen algorithm and finally the performance evaluated. Feature selection removes redundancy and noise leaving only the most discriminatory features. The choice of algorithm is important, some can deal with a large number of instances and features better than others. For example, SVMs are good at coping with high dimensionality datasets

with a small number of samples, but these require a large amount of memory for large datasets (Witten & Frank, 1999). Some, such as SVMs and decision trees are sensitive to imbalance in the training dataset whereas others such as Naive Bayes are not (Witten & Frank, 1999). Where possible it is preferrable to test a number of classifiers to identify the most appropriate choice for the specific problem. Evaluation of the classifier aims to avoid overfitting by making sure the rules are not specific to the training dataset. The classifier will often be validated on a separate dataset after having been trained and tested.

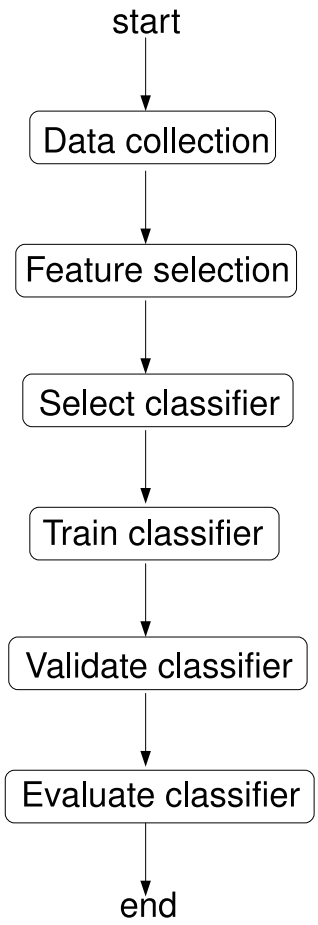


Figure 1.3: A typical machine learning approach (Al-Shahib, 2005)

1.2.5.1 Feature selection

Genomic data can be noisy, in that it is often extremely variable with some data even being incorrectly annotated. Also the data may not be complete and annotation may be missing for a number of training instances. Feature selection can help to remove or reduce the effect of noisy data.

1.2.5.2 Balanced vs unbalanced data

The number of instances belonging to each class in the training set may be imbalanced resulting in a danger that the classifier will have a preference for selecting the most populated class because the classifier assumes that there is a greater chance of an instance belonging to this class as it is more prevalent (Barandela *et al.*, 2003). The result is that performance is reduced for the minority dataset. However, it may be the case, such as when detecting fraudulent telephone calls for example, that detecting the minority case is of greater importance (Fawcett & Provost, 1997). This concern is addressed in the nsSNP analysis (Chapter 2) and when predicting hypertension related genes based on network topology (Chapter 4).

1.2.5.3 Evaluation of machine learning

A number of methods are available for evaluating machine learning results and showing the results are general enough to be applied to other data (Hand *et al.*, 2001). Some of the most common methods are described below. The various analyses in this study used all methods except for the bootstrap method. The choice was made depending on the size of the dataset, the employed classifier, and whether the classifier parameters were tuned.

Independent test data: If the training dataset was used to measure overall performance of the classifier, an over optimistic result would be obtained. Therefore it is important to evaluate performance on an independent test data set as it is a good way to gauge performance on future unseen datasets. Thus partitioning a dataset into training and independant test datasets is appropriate where a dataset is large.

Cross validation: Cross validation is especially useful for smaller datasets (Kohavi, 1995). The data is divided into (n) number of 'folds'. Each fold is treated as the test dataset in turn, with the remaining n-1 being used as training data. The performance of the classifier on each fold is measured and then a final accuracy is calculated based upon the average of all n folds. Stratified cross validation ensures that the distribution of class instances in the fold is similar to the distribution in the complete dataset. Leave-one-out cross validation is an extreme type of cross validation whereby each individual instance

is held out in turn meaning there are the same number of folds as there are instances. This maximises the amount of data available for training but is computationally expensive.

Validation datasets: When performing classifier parameter tuning steps with a large dataset, three independent data sets are required: a training set, a validation set, and a test set. The validation set is used to evaluate the effect of changing algorithm parameters and is used to create the classifier but not used in the final estimation of accuracy.

Bootstrap: Bootstrapping creates a training dataset through sampling with replacement of the whole dataset meaning that the training dataset can contain repeated instances (Efron & Tibshirani, 1993). The test set is composed of data not used in the training set. The benefit is that a good size training set can be created. Kohavi (1995) compared bootstrapping and cross validation and showed the best method to be ten-fold stratified cross validation in real-world datasets.

1.3 An overview of Network Analysis

Networks can be constructed from relationships that exist between a set of entities and can be used to represent many types of biological data at many levels, including gene expression, protein-protein interactions, signal transduction and metabolic pathways, phylogenetic, ecological and ecosystem data (Junker & Schreiber, 2008). Network analysis has only recently been applied to the world wide web, biological and social networks, and power grids. It is a rapidly growing field of research with the analysis of biological networks involving cross disciplinary research in biology, mathematics, physics and computer science. It is an important subject within the field of bioinformatics.

The rise of network approaches indicates a shift from a reductionist approach to a whole systems-level approach to understanding biology. This has only been possible in recent years due to the decrease in the cost of computation and the dramatic increase in biological data that has become available through projects such as the human genome sequencing project (Lander *et al.*, 2001; Venter *et al.*, 2001). The network based approach aims to assemble the ‘jigsaw’ of data produced through such initiatives.

Key work by Watts & Strogatz (1998) showed that many networks display common properties: they contain highly connected subgraphs and short path lengths. They termed these networks *small-world networks* due to the similar commonly known 'six degrees of separation' phenomonen seen between every person on earth. Barabasi & Albert (1999) created a model for these networks and called them *scale-free networks*. They found that they follow a power law distribution in terms of the number of edges incident to each node. These *scale-free networks* contain a small number of highly connected nodes and are very sturdy, being resilient to the random removal of nodes. Most studied biological networks follow these rules.

1.3.1 Graph theory

Networks are modeled as graphs in order to allow analysis. A graph is a mathematical object representing the networks as nodes and edges. Biological networks are represented as different types of graph models depending on the network. Networks modeled as graphs can be directed, undirected or mixed. An undirected graph contains edges where there is no edge direction. A protein-protein interaction (PPI) network is an example of a biological network that can be presented as an undirected graph $G = (V, E), v \in V, e \in E$ where the proteins are nodes (v) and the interactions are edges (e), with edge $e_{u,v}$ connecting nodes u and v (Junker & Schreiber, 2008). A directed graph represents an interaction where information passes from one node to the other, or one node has an effect on the other. Gene regulation networks are an example of a directed network. A mixed graph contains a combination of both types of interaction, directed and undirected. Multigraphs are those where multiple edges exist between a pair of nodes or vertices, which are in the same direction if the graph is directed.

When measuring properties of the graphs, the type of graph has to be considered. For example, a pair of vertices in a directed graph are strongly connected if a path exists between them when the direction of the edges is considered. The shortest path between a pair of vertices is the path containing the minimal number of edges. The path length is the number of edges. A connected component of a graph is the largest number of nodes where a path exists between each node pairing. As an example, Figure 1.4 displays the largest

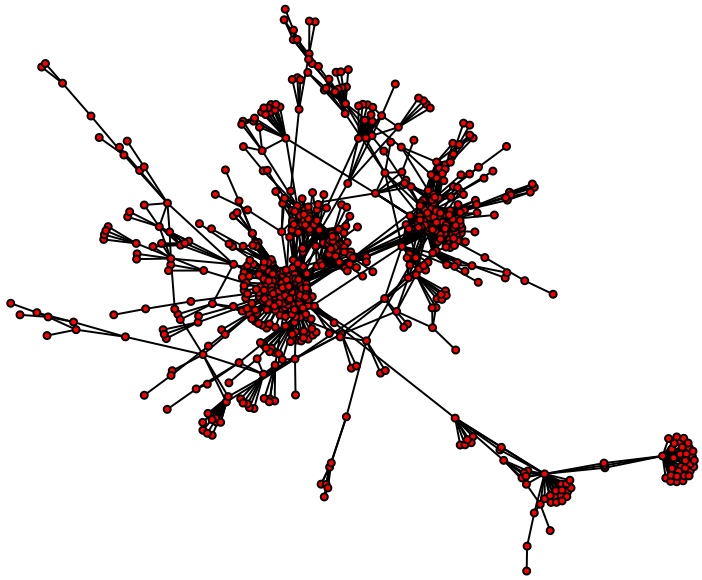


Figure 1.4: An example of a biological network, namely the largest connected component of the *Arabidopsis* protein-protein interaction network. A red node represents a protein and a connecting edge represents an interaction between a pair of proteins. The interactions were taken from the IntAct database (Kerrien *et al.*, 2007).

connected component of the *Arabidopsis thaliana* protein-protein interaction network.

Attributes are often associated with graph nodes and edges. Weights or distances can be applied to edges to quantify the relationship that exists between the nodes. In the construction of gene expression networks an edge can represent the level of coexpression between the nodes. Protein names and functional information can be added to protein nodes in a protein-protein interaction network.

Graphs are commonly stored as adjacency lists or matrices on a computer. In an adjacency matrix rows and columns represent nodes and a matrix element $G_{st} = 1$ if there is an edge between nodes s and t and $G_{st} = 0$ otherwise. Biological networks are often stored as adjacency matrices, however they are very memory intensive and adjacency lists are more appropriate when the number of edges is low. An adjacency list comprises a row for each node of a network. A row contains a list of all edges incident to node n.

Graph traversal algorithms are used to perform calculations on each node within a

network. Two search methods traditionally employed by graph algorithms are depth first searches (DFS) and breadth first searches (BFS) (Junker & Schreiber, 2008). In a depth first search the algorithm starts off with a particular node, then follows a path outwards as far as possible for each neighbour. A breadth first search visits each neighbour first before moving on to another vertex. Both methods can be encased within a loop to perform the search for each connected component.

There are a number of measurement types that can be used to describe the topology of graph models constructed to represent biological networks. These include: global network properties, centralities, motifs and clustering. Centralities are used to rank nodes in terms of their importance, motif analysis is the breakdown of sets of nodes into small units, and clustering analysis describes the organisation of the network on a number of levels. Clustering can be used to define functional modules and pathways in biological networks. Maybe one of the most commonly used graph measures in everyday life is the Google PageRank algorithm which is a variation of the common eigenvector centrality measure (Page *et al.*, 1998). The algorithm considers each web page to be a node with a link or edge between pages being a vote. Google looks at the number of votes a page receives. In addition it analyses the page casting the vote, votes from important pages (themselves having many votes) are upweighted.

The previous two sections described methods for extracting information and analysing datasets. The following sections describe areas where these technologies can be potentially applied to increase biological understanding.

1.4 An overview of Single Nucleotide Polymorphisms

A major challenge in the post-genomic era is to understand the relationship between genetic and phenotypic variation. A SNP is the most common type of variant in the human genome, they are frequently related to human diseases (Botstein & Risch, 2003). A SNP represents a single base pair difference (allele) between individuals of the same species at a particular position within the DNA sequence. In the world's population, there are thought to be about 10 million sites (one variant per 300 bases on average) where the minor allele frequency is greater than 1% (Ng *et al.*, 2008). These common SNPs consti-

tute 90% of the variation in the population and are commonly used to map phenotypes to genomic loci (Kruglyak & Nickerson, 2001; Reich *et al.*, 2003; WTCCC, 2007; Ng *et al.*, 2008). SNPs can be identified in an individuals genome by 'genotyping' a DNA sample. The associations between alleles in the population, is known as linkage disequilibrium (LD). There are often strong levels of LD between markers in close proximity to each other because the chance of a recombination event increases with distance from the SNP. A large amount of data now exists in public repositories such as dbSNP (Sherry *et al.*, 2001), HGVBASE (Fredman *et al.*, 2004) and SWISSPROT (Boeckmann *et al.*, 2003).

When SNPs or haplotypes associated with a particular phenotype are isolated it is necessary to identify the causative SNPs from the haplotype. This can be done using functional experiments, but theoretical knowledge in the first instance can helpful for both fine mapping and genotyping in the experimental design stage.

Single base changes in protein coding regions of DNA which lead to changes in an amino acid have the potential to effect protein structure and function. These are called non-synonymous single nucleotide polymorphisms (nsSNPs), and have been the subject of many recent studies (Ng *et al.*, 2008). Some nsSNPs are related to diseases but others are not associated with any change in the phenotype due to the change in the amino acid not being significantly disruptive and are thus regarded as neutral nsSNPs. Importantly, nsSNPs are the most frequent type of disease mutation (60%) (Botstein & Risch, 2003).

1.4.1 SNP Databases

A number of repositories exist in the public domain with SNP and related information. Four of the main SNP databases and their features are briefly described below:

- The dbSNP database is the most important public database of SNPs and currently includes approx 50 million SNPs from 44 organisms (`http://www.ncbi.nih.gov/SNP/`). The dbSNP database allows access to the data via a series of web pages as well as allowing bulk download in Extensible Markup Language (XML), FASTA format or MySQL dumps `http://www.mysql.org`. These SNPs have been detected either computationally in an automated manner by sequence comparison or have been determined experimentally and entered into the database via an

online submission process.

- HGVbase (`http://hgvbaseg2p.org/`) is a manually curated database of sequence variations aiming to provide links between genotypes and disease phenotypes (Fredman *et al.*, 2004). Submissions are accepted online via the website. HGVbase allows access to the data via a series of web pages and in basic tab delimited format for frequency and association data. There are plans to make table dumps of the relational database available for users.

- The Human Gene Mutation Database (`http://www.hgmd.cf.ac.uk/`) is a collection of locus-specific mutation and SNP databases (Stenson *et al.*, 2003, 2008). Individual entries can be accessed via the browser but there is no public access to a bulk download of the data. The public version of HGMD (free to academic and non profit organisations) contains 61,447 mutations (missense substitutions, insertions/deletions [indels], splicing variants etc) in 2,288 genes and provides 2,240 reference cDNA sequences as of December 2008.

- The SWISSPROT knowledgebase (`http://www.expasy.ch/sprot/`) is a high quality, manually curated protein-centric database that contains the SWISSPROT VARIANT pages. The SWISSPROT VARIANT pages contain detailed information related specifically to nsSNPs. The version used in this thesis contained 19,611 human nsSNPs annotated as either *disease* (57%), *polymorphism* (29%) or *unclassified* (10%). The term *disease* refers to SNPs that are causative in relation to disease as well as to disease-linked functional polymorphisms. The term *polymorphism* relates to mostly neutral polymorphisms. 3D structural information is provided using experimentally derived structures (>25% of the SNPs have corresponding 3D models).

1.4.2 Hapmap

The International HapMap Project (`http://www.hapmap.org`) aims to determine common gene variation and elucidate the haplotypes leading to the identification of tagging SNPs in the human genome by genotyping populations from Africa, Asia and Europe

(HapMap, 2003). When two markers are in high linkage disequilibrium (LD), which is a measure of how often alleles are inherited together, it is unnecessary to genotype both markers. Tagging SNPs are a subset of SNPs that can be selected, based on LD, to reduce the genotyping effort required to capture the majority of information within a region. The project has genotyped over 3.9 million SNPs (an average coverage of 1.3 SNPs/Kb), as of December 2008, in various populations from Africa, Americas, China, Europe and Japan. The allele frequencies, tagging SNPs and association between SNPs is also being annotated. The data is in the public domain and is available via the website and related tools such as 'Haploview' (Barrett *et al.*, 2005) and 'Tagger' (de Bakker *et al.*, 2005). These tools allow the identification of HapMap SNPs in a chosen region and selection of tagging SNPs to capture the majority of the variation with the minimum amount of redundancy between SNPs.

The Hapmap data is being used for association studies of candidate genes in the genome and for further analysis of regions suggested by family-based linkage analysis. More recently, whole genome association scans for variants that are causing common diseases have and are still being performed (WTCCC, 2007; Spencer, 2008).

1.4.3 SNP supervised classification

Several studies have attempted to predict the functional consequences of a nsSNP, namely whether it is disease related or neutral, based on attributes of the polymorphism. Some attributes depend only on the sequence information, for example the type of residue found at the SNP location. Structural attributes such as solvent accessibility can be chosen if the protein sequence containing the nsSNP has a known 3D structure or is highly similar to a protein sequence of known structure. As structural genomics projects gain momentum an increasingly large amount of protein 3D structural information is becoming available. Mapping nsSNPs onto the corresponding 3D structures or onto the structures of proteins which are highly similar at the sequence level immediately gives a structural context to the SNP and there are databases containing such models (Yip *et al.*, 2004).

Prior to the work completed in this thesis, a small number of studies had been done to try to identify rules by which a nsSNP could be predicted to be deleterious (affect protein

function) or neutral. These included the development of empirical rules (Wang & Moult, 2001; Ramensky *et al.*, 2002), the use of probabalistic methods (Chasman & Adams, 2001) and machine learning methods (Saunders & Baker, 2002; Krishnan & Westhead, 2003). Krishnan & Westhead (2003) compared the performance of two machine learning methods (support vector machines (SVMs) and decision trees) against the probabalistic methods employed by Chasman & Adams (2001) and found machine learning methods to be generally better performing. Machine learning methods were therefore considered to be a valuable tool in the classification of nsSNP status. The nsSNP datasets used, included data on known nsSNPs (Wang & Moult, 2001; Saunders & Baker, 2002; Ramensky *et al.*, 2002; Bao & Cui, 2005) and mutation data of bacteriophage T4 lysozyme and *E. coli* lac repressor (Chasman & Adams, 2001; Krishnan & Westhead, 2003). Databases of coding nsSNPs have also been developed by Karchin *et al.* (2005), Cavallo & Martin (2005). All SNPs contained within the SWISSPROT database have been manually annotated in terms of their functional status. Bao & Cui (2005) were able to perform the largest analysis to date using these annotated nsSNPs from SWISSPROT. They observed that structural information is useful when there is little information from homologous sequences. Some of the results that emerged from these approaches suggested that the majority of disease associated nsSNPs affect protein stability (Wang & Moult, 2001), they are located in surface pockets of protein structures (Stitziel *et al.*, 2004) and that conservation of the residue across species is an important predictive attribute (Saunders & Baker, 2002).

The availability of suitable datasets for analysis of annotated SNPs is constantly evolving in terms of the number of SNPs and the quality of SNP annotation. As these datasets grow, the performance of methods that aim to predict functionality of nsSNPs will continue to improve.

Once a disease associated nsSNP has been identified, the host gene and gene products becomes the focus of interest. If there is to be further understanding of the etiology of the disease the function of the protein isoforms must be ascertained. The next chapter focuses on methods for assigning function to a protein sequence.

1.5 An overview of Protein function

Proteins are macromolecular, organic compounds synthesised from genes which are fundamental units of heredity made up of sections of coding DNA. They form essential main structural components in every living cell and perform almost all cell functions.

The sequence of amino acids in a protein is defined by the DNA sequence of the gene. Protein synthesis is initiated through a transcription stage that involves genes being transcribed into messenger RNA (mRNA) by RNA polymerase. The mRNA is translated into amino acid sequence by ribosomes, transfer RNA (tRNA) recognizes the amino acids corresponding to each nucleotide triplet (codon) of the mRNA (Branden & Tooze, 1999). The amino acids are linked together forming a chain of peptides (polypeptide).

Protein sequences are composed of 'modular' domains whereby each domain has a specific function and is an independant folding unit. Some proteins are single domain and belong to one family, whereas others are multidomain proteins that can have more than one function and belong to more than one family (Orengo *et al.*, 1997). Domains belonging to a family often share function and are derived from a common ancestor. Similarities in amino acid sequences allow proteins to be grouped into families. Conserved amino acids within protein families are usually important for the function of a protein. The patterns of these conserved sequences can be used to assign proteins to functional families.

Protein structure is considered at 4 levels of organisation, the first being the primary structure and the remaining being 3D levels of folding. Understanding how proteins fold remains a major challenge within biology:

- The 'primary structure' is simply the amino acid sequence itself.
- The 'secondary structure' is the first level of folding and refers to the arrangement of the secondary structure components. The most common of these components are the alpha helix, beta sheets and coiled regions (Fletterick, 1992). Proteins can be composed of many sections of different secondary structure components.
- The 'tertiary structure' is the second level of folding and refers to the overall shape of a protein molecule produced by the combination of secondary components; the

spatial relationship of the secondary structures to one another. This controls the general function of the protein.

- The 'quaternary structure' is the resultant structure produced by a number of interacting proteins, forming a complex.

Sequence comparison by database searching is the most commonly used technique for assigning function to protein sequence, with the Gapped BLAST and PSI-BLAST programs (Altschul *et al.*, 1997) having a citation count of 26,793 at Google Scholar (`http://scholar.google.com`, January 2009). They work on the principal that homologous sequences will share a high level of sequence similarity and will relate to evolutionary distance between the sequences. If the search reveals a sequence which shares a large degree of similarity with the target sequence, an annotation can usually be transferred with some confidence.

1.5.1 Databases

1.5.1.1 Sequence databases

There are three main nucleotide databases:

- NCBI - GenBank database based at the National Institure of Health (NIH) `http://www.ncbi.nlm.nih.gov/Genbank/index.html`
- DNA DataBank of Japan (DDBJ) `http://www.ddbj.nig.ac.jp`
- European Molecular Biology Laboratory (EMBL) (Galperin, 2007). `http://www.ebi.ac.uk/embl/`

All three resources share their data and act as annotated collections of all publicly available DNA sequences. As of December 2008 there were over 85 billion base pairs in over 82 million sequences within Genbank. These databases can be searched by various BLAST tools.

The main protein databases include:

- Uniprot is a non-redundant database of amino acid sequences (Apweiler *et al.*, 2004). This database contains sequences from SWISSPROT, TrEMBL and PIR.

- SWISSPROT is a manually curated database of protein sequences whose source is the EMBL database. TrEMBL is an amino acid database from the same source but is automatically translated from EMBL and includes sequences not yet in SWISSPROT (Boeckmann *et al.*, 2003).

- PIR is a US based protein sequence database comprising comprehensively annotated non-redundant sets of sequences whereby entries are classified into family groups (Barker *et al.*, 1999).

1.5.1.2 Motif and Family databases

The main protein motif and family databases include the following resources:

- PROSITE is a collection of conserved motifs within protein families (Sigrist *et al.*, 2002). All motifs are extensively annotated with references to literature.

- The PFAM database contains multiple alignments and libraries of HMMs representing protein families (Bateman *et al.*, 2004). PFAM-A contains manually created protein familes whereas PFAM-B is automatically created and has greater coverage. PFAM can detect very rare instances of a motif.

- The ProDom database is automatically constructed from SWISSPROT and is a comprehensive collection of clustered domains. On the downside it lacks biological annotation and some of the cluster boundaries are unreliable (Servant *et al.*, 2002).

- The PRINTS and BLOCKS motif databases contain short multiple alignment fragments (Henikoff & Henikoff, 1996; Attwood *et al.*, 2003).

- Finally, Interpro is a database that aims to integrate data from most of the resources discussed above (Apweiler *et al.*, 2001; Mulder *et al.*, 2005). Each record contains links to the data sources in which it is present. It can be searched using InterProScan (Zdobnov & Apweiler, 2001).

1.5.2 Sequence comparison

It is preferrable when studying protein function to characterise the function in terms of the domain structure. Because of the modular nature of protein sequences *local alignment* methods are preferrable to *global alignment* methods. The latter is a term used to describe methods whereby sequences are compared over the entire length of the two sequences (Needleman & Wunsch, 1970). *Local alignment* methods were created later because of the need for an algorithm that could identify local regions of high similarity (Smith *et al.*, 1985). A number of tools exist for the assignment of protein function based on sequence comparison. As described, some databases contain annotated sequence whereas others contain extracted common motifs from the domains of each family.

1.5.2.1 Pairwise sequence alignment

The BLAST algorithm is a fast *local alignment* method for optimally aligning two sequences using dynamic programming (Altschul *et al.*, 1997). When performing a BLAST search of a query sequence against a database of sequences, a report is returned with a number of hits and associated statistical significance.

1.5.2.2 Multiple sequence alignment

Multiple sequence alignments can highlight patterns across families of sequences that are not obvious from pairwise alignments. A consensus alignment is created using observed residue frequencies at each position in the consensus. Sequences are usually weighted in order to remove over representation of similar sequences. The CLUSTAL algorithms are commonly used multiple alignment algorithms (Higgins *et al.*, 1992; Thompson *et al.*, 1994).

1.5.2.3 Sequence profiles

Sequence profiles perform better than comparing individual sequences for identifiying homologs. PSI-BLAST is an example of a profile searching method (Altschul *et al.*, 1997). A query sequence is initially searched against a database using BLAST. After the initial run a multiple sequence alignment is created and from this a *position specific scoring*

matrix (PSSM) is calculated whereby a score is held for each amino acid at each position within the sequence. The scores represent the observed frequencies of amino acids at each position of the multiple alignment. This PSSM is used to search the database again with hits being added to the PSSM. This continues for a limited number of rounds or until convergence of results. PSI-BLAST is a very sensitive searching technique but caution is required as unrelated sequences can be pulled in over multiple interations, distorting the PSSM and resulting in 'drift'.

1.5.2.4 Hidden Markov Models

Hidden Markov models can also be used to represent an alignment of proteins. Rather than creating a PSSM, the alignment is used to design a Markov chain and the transition probabilities are estimated (Durbin *et al.*, 1998). A probability can then be calculated as to whether a query sequence was emitted from a particular chain. HMMER and SAM are commonly used implementations of HMMs (Karplus *et al.*, 1997; Sonnhammer *et al.*, 1997).

1.5.2.5 Profile/profile comparisons

A recent extension to sequence/profile searching using PSI-BLAST or HMMs is profile/profile searching. This is now possible using PSSM/PSSM searches (Yona & Levitt, 2002; Sadreyev & Grishin, 2003; Soding, 2005).

1.5.3 Protein function supervised classification

There have been numerous genome-wide scans performed where the linked or associated region can be in excess of 30cM in size and thus contain hundreds of protein coding genes. In order to select genes and SNPs from these regions, it is important to have functional annotation for each gene and the protein it encodes in order to aid prioritisation of candidate genes for follow up studies. In the human genome, approximately 85% of protein coding genes are known genes (Consortium, 2004), of these 92 to 94% of human genes experience alternative splicing, with 86% having a minor isoform frequency of at least 15% (Wang *et al.*, 2008). Each protein isoform may have related, distinct or even

have opposing functions. Novel approaches to aid existing methods (described in section 1.5.2) for protein function annotation are still required as existing methods are not 100% effective.

Wilson *et al.* (2000) have estimated that broad biological function can be conserved down to about 25% sequence identity. However, there are a large number of sequences that cannot be annotated with current methods. This lack of annotation hinders the exploitation of some genome data, it also impacts on the understanding of biological systems as we do not have sufficient understanding of the constituent parts and how they might interact.

Machine learning methods have recently been used to explore the problem of protein function annotation. Rather than considering the sequences as strings to be compared at a character by character level, most of these methods seek to identify global features of the sequences that might be discriminative of function. Measures of function include the enzyme commission database (IUBMB, 1992), expert classifications from Riley for *Escherichia coli* (Riley, 1993), the Gene Ontology (Ashburner *et al.*, 2000) and categories from the Munich Information Centre for Protein Sequences (MIPS) (Mewes *et al.*, 2004).

Ding & Dubchak (2001) have explored the use of support vector machines (SVMs) for protein fold prediction using the SCOP protein structure database (Murzin *et al.*, 1995) as a benchmark. SCOP is a hierarchical categorization of protein structural domains where levels in the hierarchy correspond to class (reflecting the overall secondary structure composition of the protein, all α for example), fold (a general description of the spatial arrangement of the secondary structure elements), superfamily (related proteins) and family (closely related proteins where relationships are usually obvious from sequence similarity alone).

Support vector machines (SVM) have been used by Cai *et al.* (2003) to predict protein function for 54 functional families using attributes similar to those used by Ding & Dubchak (2001). The potential of the method for the prediction of distantly related proteins has also been explored by testing the method on 24 randomly selected distantly related proteins. This analysis achieved a prediction accuracy of 58.3%. Related studies on enzyme functional prediction found 72% of a set of 50 enzymes could be correctly as-

signed where there was no known sequence homolog available (Han *et al.*, 2004). SVMs have also been used to distinguish enzyme structures from non-enzyme structures (Dobson & Doig, 2003). The most useful features included secondary structure content and amino acid frequencies. Recently, Melvin *et al.* (2007) used SVMs for superfamily classification of distantly related proteins, but did not report the specific performance for each superfamily.

Clare & King (2003) and Clare *et al.* (2006) used decision trees with GO and MIPS functional categories for mining data on the *Saccharomyces cerevisiae* and *Arabidopsis thaliana* genomes. Predictions achieved 75% accuracy in the *Saccharomyces cerevisiae* study and 85% precision in the *Arabidopsis thaliana* study. Attributes used were those derived from PSI-BLAST, phenotypic properties, expression data, sequence and secondary structure.

Other sequence attributes that have been used for functional prediction relate to predicted properties of the sequences such as post translational modifications, subcellular localization and secondary structure (Jensen *et al.*, 2002) using the Riley functional classification (Riley, 1993) and Gene Ontology (Ashburner *et al.*, 2000).

After identifying disease causing nsSNPs and determining the function of the protein in which they reside, it becomes important to understand the environment and pathways through which the protein acts. This involves the study of proteins in the wider context of protein-protein interaction networks.

1.6 An overview of protein-protein interaction (PPI) networks

Protein-protein interaction (PPI) networks (the interactome) represent the relationships between protein molecules, the study of which is important as proteins acting as enzymes, channels and transporters perform almost all cell functions (Alberts *et al.*, 2002; Hwang *et al.*, 2008). The study of the interactome could help improve the understanding of complex diseases.

1.6.1 PPI database repositories

A number of PPI data repositories now exist in the public domain. A comprehensive list can be found at `http://tiny.cc/ppidatabases`. Some of the key resources are described below (details correct as of December 2008):

- HPRD - Human Protein Reference Database is a database of human protein information manually extracted from the literature by expert biologists who read, interpret and analyse the published data (Mishra *et al.*, 2006). The latest version contains 38,167 protein-protein interactions.

- IntAct - Interaction Database is a public repository of manually curated protein interaction data from the literature or through user submissions (Kerrien *et al.*, 2007). The site contains analysis tools, currently there are 174,078 interactions of which approximately 32,000 are human.

- DIP - Database of Interacting Proteins, combines experimentally derived interactions from a number of sources. The interactions are both manually and computationally curated. Currently there are 57,146 interactions, 2,070 of which are human interactions (Xenarios *et al.*, 2000).

- MINT - Molecular INTeraction Database contains protein interactions that have been verified experimentally (Chatr-aryamontri *et al.*, 2007). The interactions are extracted from the literature by expert curators. In total there are 111,847 interactions of which 21,357 are human.

- MIPS - Mammalian Protein-Protein Interaction Database contains literature derived, high-quality interaction data manually curated by experts (Pagel *et al.*, 2005).

- BioGRID - The Biological General Repository for Interaction Datasets database contains protein and genetic interactions from both high-throughput studies and conventional focused studies for key model organisms (Stark *et al.*, 2006). It currently contains over 198,000 interactions from six different species.

- BIND - Biomolecular Interaction Network Database is a database that stores details of interactions, molecular complexes and pathways (Bader *et al.*, 2001). BIND

accepts individual submissions as well as interaction data from the protein data bank (PDB) (Sussman *et al.*, 1998) and a number of large-scale high throughput interaction experiments.

- OPHID - Online Predicted Human Interaction Database is built by mapping high-throughput model organism data to human proteins and integrating data from yeast two-hybrid based, literature-based interaction and orthology-based interaction sources (Brown & Jurisica, 2005). The literature-derived human PPI are obtained from from BIND, HPRD and MINT. Predicted interactions are made from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*. The 23,889 predicted interactions currently listed in OPHID are evaluated using protein domains, gene co-expression and Gene Ontology terms. In total there are 48,222 interactions listed within OPHID.

- UniHI - The Unified Human Interactome is a unified repository based on 10 major interaction sources of computational and experimental derived interactions (Chaurasia *et al.*, 2007). It includes more than 150,000 distinct interactions for more than 17,000 human proteins. Scores for quality assessment are given based on co-annotation and co-expression of the interacting proteins.

- PIP - The Potential Interactions of Proteins web server contains interacting proteins constructed for the human genome using an orthology-based method (Jonsson & Bates, 2006a). The orthologous protein interactions were taken from DIP and MIPS. Each interaction was given a confidence score based on sequence similarity to proteins shown experimentally to interact and the amount of available experimental evidence for the interaction. There are 108,113 interactions in this database when a confidence score cut-off is applied that provides sensitivity of 85% and specificity of 82%.

1.6.2 Methods to identify protein-protein interactions

There are many approaches, experimental and theoretical, for detecting protein interactions, each varying in sensitivity and specificity. They include high-throughput meth-

ods such as yeast 2-hybrid experiments (Rual *et al.*, 2005; Stelzl *et al.*, 2005), manually curated and literature based interaction sources such as the Human Protein Reference Database (HPRD) (Mishra *et al.*, 2006) and Interaction Database (IntAct) (Kerrien *et al.*, 2007) as well as predicted interactions based on *in silico* methods such as Predictome (Mellor *et al.*, 2002), POINT (Huang *et al.*, 2004), Prolinks (Bowers *et al.*, 2004) and STRING (von Mering *et al.*, 2007).

1.6.3 PPI software

There are now many software applications available for the analysis of biological networks, a comprehensive survey, was recently described by Pavlopoulos *et al.* (2008). A selection of some of the popular tools are described below.

- APID - Agile Protein Interaction DataAnalyzer is a web based tool enabling the exploration and analysis of PPI data from BIND, BioGRID, DIP, HPRD, IntAct and MINT PPI resources (Prieto & De Las Rivas, 2006).

- Cytoscape is an open source bioinformatics Java software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data (Shannon *et al.*, 2003). Many user created plugins are available for specific analysis tasks.

- Osprey is a standalone application that runs on a range of platforms with a license for non commercial use (Breitkreutz *et al.*, 2003). Currently the source code is not available and it is not appropriate for large scale network analysis. Data can be loaded directly from BioGRID (Stark *et al.*, 2006) and there is support for a number of data formats. Osprey is a powerful tool for network manipulation and has the important ability to incorporate new interactions into an already existing network.

- VisANT is freely available (`http://visant.bu.edu`) and integrates, mines and displays hierarchical bio-network and pathway information (Hu *et al.*, 2008). It is supported by the Predictome database where much of the interaction data comes

from resources such as BioGRID, MIPS, BIND and HPRD. This tool is able to handle large-scale networks with millions of nodes and edges.

- Pajek is a standalone application (Batagelj & Mrvar, 1998). It is not an open source application and runs under Windows operating systems only, but it is free for non-commercial use. It is suitable for large scale networks, is highly interactive and incorporates many clustering methods. Pajek's main strength is the variety of layout algorithms.

- The Boost graph library (BGL) `http://www.boost.org/doc/libs/1_37_0/libs/graph/doc/index.html` is a C++ library for developers providing a generic interface for traversing graphs and accessing the graph's structure.

Other popular software includes Graphviz (`http://www.graphviz.org`), NetworkX (`http://networkx.lanl.gov`), cFinder (Adamcsek *et al.*, 2006), Guess (`http://graphexploration.cond.org`) and igraph (`http://cneurocvs.rmki.kfki.hu/igraph/`).

1.6.4 PPI networks and supervised classification of disease associated genes

Early work using decision tree based classifiers showed disease genes tend to be longer and more conserved than non-disease genes (Lopez-Bigas & Ouzounis, 2004). Subsequent work constructing supervised classifiers included additional sequence based attributes that included length, proximity to other genes, exon count, GC content, transmembrane and signal peptide domain content, CpG related properties and details of homologous and paralogous proteins (Adie *et al.*, 2005). Other annotation related attributes such as co-expression and similarity of Gene Ontology (GO) (Ashburner *et al.*, 2000) terms and text mining approaches have also been used for selection of disease gene candidates (Perez-Iratxeta *et al.*, 2005; Tiffin *et al.*, 2005; Adie *et al.*, 2006). More recently, attributes based on PPI have been used in supervised classification approaches (George *et al.*, 2006; Xu & Li, 2006).

PPI network based approaches for studying human diseases have shown that disease associated proteins often interact with other disease proteins or share interaction neighbours (Xu & Li, 2006). Specifically, there is a 10-fold increase in the likelihood of proteins interacting when they are associated with the same disease (Goh *et al.*, 2007). Goh *et al.* (2007) have also shown that 'essential' disease genes, in which mutations are lethal, form hubs (highly connected nodes) whereas 'non-essential' disease genes do not display this tendency. A k-nearest neighbours classifier using network features achieved a prediction accuracy of 0.76 using the OMIM dataset (Xu & Li, 2006). A disease is considered to result from the disruption of a specific cluster (functional module of interacting proteins) and is caused by mutations in one or more of the proteins resulting in a recognised phenotype (Loscalzo *et al.*, 2007). Different combinations of perturbed genes in a cluster can lead to the same phenotype. There is also data showing that some proteins are implicated in multiple phenotypes, that is there are disorders which can be termed connected in that they share associated proteins (Goh *et al.*, 2007; Loscalzo *et al.*, 2007; Sam *et al.*, 2007). Cancer is regarded as one of the most connected disorders (Goh *et al.*, 2007).

Analysis of protein-protein interaction networks has been used to explore several disease conditions including asthma (Hwang *et al.*, 2008), neurodegenerative diseases (Goni *et al.*, 2008), and with transcriptomics, human heart failure (Camargo & Azuaje, 2007, 2008). PPI network properties for Alzheimers related proteins from OMIM have been studied by Chen *et al.* (2006) who found these proteins form a highly connected subnetwork. They devised a metric that enabled the ranking of a protein for its biological relevance to Alzheimers pathways. Such analyses may be helpful in suggesting important single proteins or clusters, the disruption of which could lead to a variety of disease conditions. This can be particularly useful for adding weight to candidates identified through genome wide studies and could lead to a better understanding of the molecular basis of disease.

To date, many of the studies have been dependent on OMIM as a source of disease related or implicated genes. OMIM is a comprehensive catalogue of human genes and their associated genetic phenotypes. It provides 'full-text, referenced overviews on all known mendelian disorders and over 12,000 genes'. Although OMIM was initially cre-

ated to store details relating to mendelian traits its use has been extended to some extent to cover more complex traits. The resource is not available as a relational database but is available to download as formatted text. Studies such as van Driel *et al.* (2006) have created tools such as MimMiner in an effort to mine the natural language used in each record. MimMiner searches the data on a keyword basis using words found in the anatomy (A) and the disease (C) sections of the Medical Subject Headings vocabulary (MeSH) `http://www.nlm.nih.gov/mesh/`. However, OMIM is an incomplete resource that holds many speculative disease associations. There is a need for trait specific analyses to be performed on expertly curated datasets of disease implicated gene products.

With cardiovascular disease (cvd) set to become the number one cause of deaths worldwide, it is important to understand the etiologic mechanisms for cardiovascular related diseases such as hypertension, in order to identify new routes to improved treatment. There have only been a small number of cvd focused studies to date that have exploited the use of PPI networks. The approach of George *et al.* (2006) which employs PPI and pathway data together with sequence similarity, had no success in correctly identifying any of the putatively associated hypertension genes included in their dataset. These analyses were based on a small set of 5 hypertension related proteins extracted from OMIM.

Camargo & Azuaje (2007) undertook an analysis of genes implicated with human heart failure by studying PPI network connectivity in a human heart failure gene expression dataset. The network was constructed from interactions within the HPRD database. Relationships between co-expression and PPI connectivity were analysed showing that genes significantly differentially expressed were not always highly connected nodes. Though some traditional heart failure proteins were not differentially expressed, they sometimes interacted with differentially-expressed proteins. It was noted that network hubs can show weak co-expression with their directly interacting partners. The exploratory study aimed to identify patterns and trends, with the constructed network being available on request to the authors. However there was no metric or classifier described for prioritising candidate genes. In a recent study, Camargo & Azuaje (2008) focused on dilated cardiomyopathy, a leading cause of heart failure. Again, differentially expressed

genes were evaluated in terms of PPI networks. In this analysis classifier models were used to suggest novel dilated cardiomyopathy associated genes.

With cvd being such an important target, there is value in attempting to further develop such alternative approaches to predict potentially implicated genes. Such methods may be useful in identifying novel disease associated genes as well as complementing existing analysis strategies such as GWA studies.

1.7 Study Aims

Studies of the etiology and genetic contribution to complex diseases require methods to identify causative functional SNPs and the disease associated genes in which they reside. This study explores the utility of machine learning methods for predicting functional nsSNPs and the function for proteins where annotation using conventional homology based methods is absent. Such machine learning methods, combined with graph theoretic approaches are used to explore approaches to identify novel disease associated proteins and prioritise candidate gene lists through characterising the PPI network topology of implicated disease associated proteins.

1.7.1 Specific aims of thesis

1.7.1.1 nsSNP analysis

- To improve on previous methods for predicting disease associated nsSNPs by applying machine learning methods to look for patterns in the distribution of sequence and structural based attributes related to disease and neutral SNPs.

1.7.1.2 Protein function analysis

- To utilise machine learning methods for predicting protein superfamily membership using global sequence based attributes and a training set of protein domains from the SCOP classification scheme. Traditional homology based approaches work well where there is a high level of sequence similarity between the query sequence and a sequence of known function. This study focuses on sequences in the ‘twilight zone’

whereby sequence similarity is less than 30%.

1.7.1.3 Protein-protein interaction network analysis

- To characterise the topological properties of hypertension related proteins within the human interactome using protein-protein interaction data from OPHID and hypertension associated genes carefully selected from the OMIM database.

- To combine the identified hypertension protein network properties with simple sequence and functional based attributes to build a classifier for predicting novel hypertension related proteins.

- To analyse the topological properties of implicated cardiovascular (cvd) and cancer related proteins within the human interactome using protein-protein interaction data, and disease implicated proteins from publicly available sources.

- To quantify the influential nature of the cvd and cancer proteins, analyse community structures and show an approach for prioritising candidate gene products based on these network measures.

Chapter 2

nsSNP function analysis

This chapter focuses on using machine learning methods for predicting functional nsSNPs. All nsSNPs described in the SWISSPROT VARIANT web pages that mapped onto the Ensembl database (Hubbard *et al.*, 2002) were considered, allowing the application of Ensembl annotations to these variants. A number of sequence and structural attributes of nsSNPs were surveyed to see if previous trends of disease and neutrality are preserved in light of much larger datasets now available, the attribute of whether the nsSNP occurs in a protein binding site was also included (Bader *et al.*, 2003).

One of the problems with using the available collection of natural nsSNPs is the large difference in numbers of disease associated and neutral examples. To address this problem of class imbalance the effect of resampling and weighting on the prediction performance was assessed.

2.1 nsSNP analysis methods

2.1.1 SNP database creation

In order to create a resource to facilitate the prediction of functional nsSNPs, a SNP database was initially constructed by extracting SNP related data from the Ensembl database (Hubbard *et al.*, 2002) using a combination of structured query language (SQL) and the Ensembl perl application programming interface (API). Ensembl was used as it contains SNPs from the combined SNP resources described in section 1.4.1 and is a rich

source of annotation. This data was loaded into a MySQL database whereby the SNP 'rs identifiers' were used as keys. Pipelines were constructed in order to allow annotation from external sources to be added to the SNPs (figure 2.1). Data was included from:

- The manually curated protein knowledgebase SWISSPROT (Yip *et al.*, 2004).
- The interactions within the Biomolecular Interaction Network Database (BIND) and the Molecular Modeling Database (MMDBBIND) (Bader *et al.*, 2003).
- The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000).
- The Homology-derived Secondary Structure of Proteins database (HSSP) (Sander & Schneider, 1993).
- The Protein Data Bank (PDB) (Berman *et al.*, 2000).

Tools were built to parse, reformat, map and load data into the dababase from these sources. SWISSPROT was used to add information relating to the disease status of the SNP as well as information relating to functional sites within the protein sequence. MMDBBIND and BIND were used to provide information relating to protein interactions. BIND contains interactions/complexes and pathways but not at the atomic level. It provides residue ranges for the interacting regions. These entries are not dependent on structure as sequence identifiers are used. This database records *in-vivo* interactions being studied and references the experimental evidence that supports or disputes the occurrence of the interaction. MMDBBIND (www.bind.ca) contains atomic level details of interactions. These interactions are annotated automatically from MMDB entries (MMDB is a subset of PDB that excludes theoretical models). A contact is made when the van der waals radii of 2 atoms are within 0.5 Å. The KEGG database was used to provide information relating to pathways and both HSSP and PDB were used to add structural information relating to the SNPs.

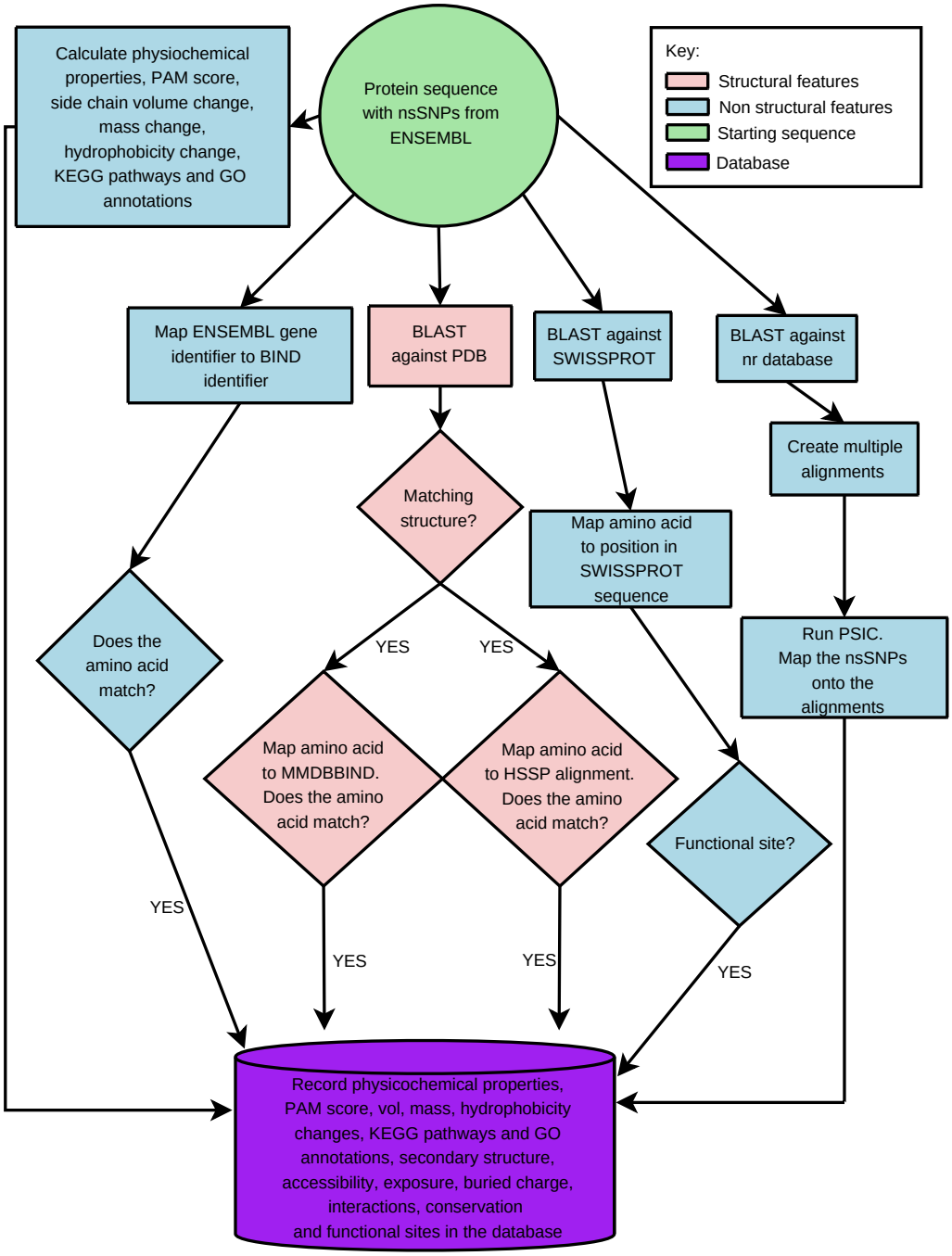


Figure 2.1: Overview of the nsSNP annotation pipeline for creating the SNP database.

2.1.2 nsSNP dataset

The SWISSPROT VARIANT web pages (Yip *et al.*, 2004) provide information on single amino acid polymorphisms associated with a given SWISSPROT entry. The variants are labelled as disease, unclassified or polymorphism. A subset of these SNPs were used in this study, namely those from *Homo sapiens* where the amino acid polymorphism was found to map onto the Ensembl human genome protein sequence. A SNP was considered

mapped where the amino acid was the same in both the SWISSPROT sequence and the Ensembl protein sequence and the aligned region using BLAST had an expectation (E) value $< 1e - 10$ over a region > 100 amino acids in length. Matches to known structure and to structural homologs were obtained in the following way:

- Each sequence containing a nsSNP was searched against all the sequences in the protein data bank using the PSI-BLAST program (Altschul *et al.*, 1997) with ten iterations.
- Only hits with an E value of less than 1e-10 where the amino acids at the position of the nsSNP were the same were stored.
- Each of these nsSNP containing SWISSPROT entries was aligned with the sequence in a relevant HSSP (Sander & Schneider, 1993) file. Where there were multiple PDB annotations in the SWISSPROT file, the PDB with the lowest E value was used.

2.1.3 nsSNP features

Structurally dependent features were considered separately from the set of features that were not dependent on structure because the subset of nsSNP containing proteins with associated 3D structures is considerably smaller than the set of all nsSNP containing proteins. A total of 17 features were used, 11 non structurally dependent and 6 structurally dependent.

2.1.3.1 Non structural features

The features chosen were largely based on those used by Ramensky *et al.* (2002) and Krishnan & Westhead (2003):

- The residue types of the original and mutated residues.
- The physiochemical properties of the original and mutated residues.
- Sequence conservation: is the nsSNP at a conserved position. The sequence was matched against a protein non redundant database using the BLAST program and

all hits with an E value less than 0.0005 were stored. A multiple alignment was constructed and sequence variation at the position of the nsSNP was described by calculating the position-specific independent counts (PSIC) score (Ramensky *et al.*, 2002).

- Point Accepted Mutation (PAM) score shift measured from the PAM120 matrix (Dayhoff *et al.*, 1978).
- Side chain volume change (Tsai J, 1999).
- Mass change. The molecular weights are those of the neutral, free amino acids.
- Hydrophobicity difference (Black SD, 1991).

In addition four further non structurally dependent attributes (described below) were used, these were taken from the SWISSPROT features table, pathway information, ontology classifications and interacting regions.

SWISSPROT features table

The SWISSPROT entry feature table may contain information about functional sites. A survey was carried out of functional site terms across all nsSNPs in the SWISSPROT VARIANT pages. Following Ramensky *et al.* (2002), nsSNPs located within the following labelled features were considered to be termed 'functional' sites for the benefit of the machine learning analysis:

- ACT_SITE - amino acid(s) involved in the activity of an enzyme.
- BINDING - binding site for any chemical group (co-enzyme, prosthetic group, etc.)
- MOD_RES - posttranslational modification of a residue.
- SITE - any interesting single amino-acid site on the sequence, that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids.
- LIPID - covalent bonding of a lipid moiety.

- METAL - binding site for a metal ion.
- DISULPHID - disulphide bond.
- CROSSLNK - posttranslationally formed amino acid bonds.
- TRANSMEM - extent of a transmembrane region.
- SIGNAL - extent of a signal sequence (prepeptide).
- PROPEP - extent of a propeptide.
- NP_BIND - extent of a nucleotide phosphate-binding region.
- MUTAGEN - Site which has been experimentally altered by mutagenesis.

KEGG pathways

In order to observe the distribution of disease and neutral nsSNPs within pathways we mapped the set of 16,352 nsSNPs to KEGG pathways (Kanehisa & Goto, 2000). For each pathway, i, we calculated the odds ratio P_i:

$$P_i = \frac{N^i_{dis}/N^i_{poly}}{N^{tot}_{dis}/N^{tot}_{poly}}$$

where N^i_{dis} is the number of disease nsSNPs in pathway i and N^{tot}_{dis} is the total number of disease nsSNPS in our dataset and similarly for polymorphic nsSNPs.

Gene Ontology

Each nsSNP containing protein sequence belongs to a number of Gene Ontology (GO) categories (Ashburner *et al.*, 2000). The odds ratio of neutral and disease nsSNPs were calculated for each of the GO categories.

Interactions

The BIND (Bader *et al.*, 2003) database was used to map nsSNPs to interacting regions. A potential interacting region was defined as a region from amino acid position n to amino acid position m. These interactions were generally regions observed experimentally and

were not considered structurally dependent annotations as the BIND database entries have sequence identifiers. The odds ratio P_i was calculated where N^i_{dis} is the number of sites containing disease nsSNPs in either an interacting region or non-interacting region i and N^{tot}_{dis} is the total number of sites containing disease nsSNPS in our dataset that map to BIND and similarly for polymorphic nsSNPs.

2.1.3.2 Structural features

Five structural attributes were extracted from the corresponding HSSP file (Sander & Schneider, 1993):

- Secondary structure conformation: residue is in an isolated beta-bridge (single pair beta-sheet hydrogen bond formation), 5 turn helix (pi helix), 3 turn helix (3/10 helix), 4 turn helix (alpha helix), bend, beta sheet in parallel and/or anti-parallel sheet conformation (extended strand), hydrogen bonded turn (3, 4 or 5 turn).
- Relative solvent accessibility.
- Normalised relative accessibility.
- Exposure (relative accessibility as 3 states).
- Buried charge.

Relative accessibility and normalised relative accessibility were calculated in the same manner as Chasman & Adams (2001). The maximum accessible surface area (Å^2) reference values are those calculated for residues in a Gly-Xaa-Gly tripeptide in extended conformation (Miller S, 1987). In order to group the relative accessibility, it was projected onto 3 states: buried (here defined as $<$9% relative accessibility), intermediate (9% $\leq$ rel. acc. $<$ 36%), exposed (rel. acc. $\geq$ 36%) (Rost & Sander, 1994). Buried charge is defined as K,R,D,E,H wild type amino acid and 'buried' exposure class (Krishnan & Westhead, 2003).

Interactions

The MMDBBIND database (Bader *et al.,* 2003) was used as a second source to map nsSNPs to interacting regions. MMDBBIND contains atomic level details of interactions.

These interactions are annotated automatically from MMDB (Chen *et al.*, 2003) which is a subset of experimentally determined PDB structures. This attribute is therefore dependent on structure as it requires a PDB identifier. MMDBBIND interactions are a much more precise interaction annotation than the BIND interactions as the BIND defined regions can sometimes be very large in amino acid length. Again, the odds ratio P_i was calculated where N_{dis}^{i} is the number of sites containing disease nsSNPs in either an interacting region or non-interacting region i and N_{dis}^{tot} is the total number of sites containing disease nsSNPS in our dataset that map to MMDBBIND and similarly for polymorphic nsSNPs.

2.1.4 Machine learning

All machine learning analysis was performed using the Weka package of machine learning algorithms (Witten & Frank, 1999).

2.1.4.1 Single attribute analysis

In order to identify the most effective classifier from all of the attributes, the 1R classifying algorithm was used (Holte, 1993). This classifier creates a single level decision trees for each attribute and measures the prediction error rate. It was used with a minimum bucket size of 14 and 10 fold cross validation on the fully balanced dataset containing all variables. The bucket size of 14 was chosen because bucket sizes below this value caused overfitting and/or an increase in the error rate. The attributes were then ranked in terms of their effectiveness as a predictor using the default ranker search method with this 1R attribute evaluator, they were also ranked in terms of the information gain (IG) they provide (Witten & Frank, 1999). Entropy is a measure of information and represents the amount of information that would still be needed to classify the nsSNP having used the attribute in question (Shannon CE, 1948). The information gain is the information required after using the attribute as a classifier subtracted from the information required before using the attribute as a classifier.

2.1.4.2 Attribute set analysis

It is of value to investigate the relative importance of attributes that require structure and those that can be obtained by sequence alone. The importance of sequence conservation has been previously noted (Saunders & Baker, 2002) so it was also important to observe whether the other non structurally dependent attributes could add to prediction quality achieved with conservation score alone. Hence, we compared predictions for the following sets of selected attributes:

- Set (1) - All variables (3821 nsSNPs).
- Set (2) - Structurally dependent variables (3821 nsSNPs).
- Set (3) - All non structurally dependent attributes (14,636 nsSNPs).
- Set (4) - Non structurally dependent variables excluding the conservation score (14,636 nsSNPs).
- Set (5) - The conservation score alone (14,636 nsSNPs).

Decision trees have been shown to perform well in a mixed cross validated training dataset (Krishnan & Westhead, 2003). They also provide a confidence score and intelligible rules to a prediction. Based on this knowledge we decided to use the J48 decision tree classifier to analyze the assembled sets of variables. J48 is the Weka implementation of C4.5 and was run with the default set of parameters and 10 fold cross validation. In performing 10 fold cross validation, the data was divided into 10 'folds' and each fold was treated as the test dataset in turn, with the remaining 9 being used as training data. The performance of the classifier on each fold was measured, and a final accuracy calculated based upon the average of all 10 folds.

2.1.4.2.1 Effect of imbalance There was a problem of imbalance (Al-Shahib *et al.*, 2005) within the dataset which would introduce skewing towards the avoidance of errors for the disease status as there are 2.5 times more disease nsSNPs than neutral. The imbalanced dataset applies a higher cost to getting a disease prediction wrong, meaning that the rules inferred by the imbalanced dataset are able to predict disease status but unable to

predict neutral nsSNPs accurately. The effect of imbalance depends on total set size, class heterogeneity, data complexity and the classification technique. To address the problem of imbalance in our dataset we applied cost-sensitive classification by either resampling or reweighting (Witten & Frank, 1999). Resampling can be used to either increase the number of the minority class (*oversample*) or reduce the number in the majority class (*undersample*) (Weiss & Provost, 2001). Reweighting can be used to apply a cost to an incorrectly classified minority class without altering the numbers in each class. The cost is directly proportional to the imbalance. This study compared results using both resampling and reweighting. We undersampled the disease class as oversampling would make exact copies of the neutral class, potentially resulting in overfitting of the data. Undersampling results in the loss of information so it was decided to randomly undersample at rates of 100%, 75%, 50%, 25% and 0%. This means that at each rate, '*n%* of the excess members of the majority class were randomly removed' (Al-Shahib *et al.*, 2005), resulting in a balanced dataset when undersampling at a rate of 100%.

2.1.4.2.2 Attribute redundancy Some attributes work well in combination leaving other attributes redundant and maybe even causing a reduction in prediction quality. The optimised subset of attributes for each attribute set at each level of imbalance was obtained using wrapper-based feature selection with J48 as the learning method with default option settings. The wrapper-based feature selection method in combination with the Genetic Search algorithm (Witten & Frank, 1999) produced the lowest error rates in tests. The genetic search algorithm was initialised with a population size of 20 and then 50 generations were evaluated.

2.1.4.2.3 Measure of prediction quality Matthews correlation coefficient (MCC) (Matthews, 1975) was used as the measure of prediction performance. Matthews correlation coefficient combines both sensitivity and specificity into one measure and lies in the range -1 to 1 with 1 meaning complete prediction accuracy, 0 meaning every prediction was randomly assigned. MCC is defined by

$$MCC = \frac{(TP.TN - FP.FN)}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}}$$

where TP is true positive, FP is false positive, TN is true negative and FN is false negative.

This is preferrable to using the error rate (E), defined below, because in a case where all samples are assigned to a majority class, E may still be low.

$$E = \frac{FP + FN}{TP + TN + FP + FN}$$

2.2 nsSNP analysis results

2.2.1 Distribution of attributes across the normal and disease associated nsSNPs

A set of 16,352 SWISSPROT nsSNPs (out of a potential 18,812) could be mapped onto the Ensembl database, of which 10,419 (64%) were disease associated, 4217 (26%) were labelled as being neutral and 1716 (10%) were unclassified. These disease and neutral nsSNPs were contained within 893 and 1256 proteins respectively. A total of 500 nsSNP-containing proteins had structural homologs, of which 299 proteins contained disease related nsSNPs and 295 contained polymorphic nsSNPs (a protein can contain both disease and polymorphic nsSNPs). The data is summarised in Table 2.1.

	Disease	**Polymorphism**	**Total**
Number of nsSNPS	10,419	4217	14,636
Number of nsSNPS within proteins with structural homologs	3212	609	3821
Number of Proteins with nsSNPs	893	1256	2149
Number of Proteins with nsSNPs having structural homologs	299	295	594

Table 2.1: Summary of SWISSPROT VARIANT training dataset

2.2.1.1 Analysis of non structural features

The distribution of sequence derived attributes suggests: tryptophan (W), tyrosine (Y) and cysteine (C) in the wild and mutated residues increases the chance of the nsSNP being disease related with odds ratios of 2.07, 2.03 and 2.03 respectively. This has previously been noted for tryptophan and cysteine by Vitkup *et al.* (2003). The likelihood of the nsSNP being deleterious increases as the volume, mass and hydrophobicity difference between the wild and mutated residue increases. The mean change in volume, mass and hydrophobicity between the wild and mutated residue was 1.29, 1.29 and 1.31 times greater for disease nsSNPs respectively. There appeared to be very little bias in the other physiochemical properties individually towards the status of the nsSNP. As previously observed, a nsSNP is much more likely to be deleterious with an increasing PSIC conservation score difference (Saunders & Baker, 2002). The mean PSIC conservation score was 2.2 times greater for disease related nsSNPs.

2.2.1.1.1 SWISSPROT features table Table 2.2 shows the most discriminatory terms from the SWISSPROT features table, namely those where over 90% of the corresponding nsSNPs are disease related. The annotation of a nsSNP in the SWISSPROT feature table is not a good discriminator between disease and polymorphic status. In this dataset, the feature table terms which are predominantly associated with disease related nsSNPs have very low counts, making it difficult to generalize about their utility in predicting whether a given nsSNP is disease related.

2.2.1.1.2 KEGG pathways Analysis of nsSNPs that map to KEGG pathways revealed that the odds ratio (P) is highest for the following 4 pathways: phenylalanine, tyrosine and tryptophan biosynthesis (15.6), methionine metabolism (15.16), carbon fixation (12.56), nucleotide sugars metabolism (12.33). Assignment to a KEGG map was not used as an attribute for machine learning prediction as this result may simply reflect that these are commonly studied pathways and the pathway was considered to be a property of the protein as opposed to the nsSNP.

Site	Disease	Polymorphism	Percentage (odds ratio) of nsSNPs within these sites that are disease
ACT_SITE	25	1	96.15 (10.12)
BINDING	13	0	100 (-)
DNA_BIND	352	20	94.62 (7.12)
METAL	38	0	100 (-)
MOD_RES	34	3	91.89 (4.59)
MUTAGEN	111	10	91.74 (4.49)
NP_BIND	108	8	93.1 (5.46)

Table 2.2: The number of disease and polymorphism nsSNPs within SWISSPROT feature table sites that contain $>$ 90% disease nsSNPs. ACT_SITE - amino acid(s) involved in the activity of an enzyme, BINDING - binding site for any chemical group (co-enzyme, prosthetic group, etc.), DNA_BIND - Extent of a DNA-binding region, METAL - binding site for a metal ion, MOD_RES - posttranslational modification of a residue, MUTAGEN - Site which has been experimentally altered by mutagenesis, NP_BIND - extent of a nucleotide phosphate-binding region.

2.2.1.1.3 Gene Ontology The ratio of deleterious nsSNPs was found to be the highest for the following GO biological processes: anti-inflammatory response (GO:0030236), peroxisome organization and biogenesis (GO:0007031), and peroxisomal membrane transport (GO:0015919). The GO cell location categories having the highest ratio of deleterious nsSNPs are the peroxisomal membrane (GO:0005778), integral to peroxisomal membrane (GO:0005779) and collagen type VII (GO:0005590) categories. The molecular function categories containing the highest ratio of disease to neutral nsSNPs are phenylalanine 4-monooxygenase activity(GO:0004505), alpha-galactosidase activity (GO:0004557) and pyruvate kinase activity (GO:0004743). GO categories were not used as machine learning attributes as they were considered to be properties of the protein as opposed to the nsSNP.

2.2.1.1.4 Interactions A total of 1,944 SWISSPROT nsSNPs mapped to proteins that have entries in BIND. A significant number of disease nsSNPs are within interacting regions (χ^2=32.85, p=0.001) within BIND. Table 2.3 shows 71.7% (odds ratio 1.29) of positions containing one or more nsSNPs that map to interacting regions are associated with disease (736 sites) as opposed to 28.3% (290 sites) which contain polymorphism nsSNPs.

2.2.1.2 Analysis of structural features

A total of 3,821 nsSNPs could be mapped to a homologous protein of known structure. Of the nsSNPs that could be mapped to structure, disease nsSNPs tended to be buried and neutral nsSNPs tend to be exposed. There was also a propensity towards nsSNPs causing disease occurring in beta sheets as previously noted (Sunyaev *et al.*, 2000) and a trend towards neutrality with increased accessibility.

2.2.1.2.1 Interactions A total of 3,028 SWISSPROT nsSNPs mapped to proteins that have structures or structural homologs in MMDBBIND (Bader *et al.*, 2003). Table 2.3 shows 86% (odds ratio 1.29) of positions containing one or more nsSNPs that map to interacting residues are associated with disease (294 sites) but also that 82% (odds ratio 0.97) of positions containing one or more nsSNPs that map to non-interacting residues are associated with disease. The difference between interacting sites containing disease nsSNPs and non-interacting sites containing disease nsSNPs was not significant (χ^2=3.17).

	Interacting sites (num)[odds ratio]	**Non-interacting sites (num)[odds ratio]**
Disease (BIND)	71.7%(736)[1.29]	58.6%(431)[0.72]
Polymorphism (BIND)	28.3%(290)	41.4%(304)
Disease (MMDBBIND)	86.0%(294)[1.29]	82.0%(1818)[0.97]
Polymorphism (MMDBBIND)	14.0%(48)	18.0%(398)

Table 2.3: Distribution of disease and neutral nsSNPs within locations (interacting or non-interacting) from BIND and MMDBBIND. Some sites may contain multiple nsSNPs

All attributes excluding the KEGG pathway and GO attributes were used for machine learning analysis.

2.2.2 Machine Learning

2.2.2.1 Single attribute analysis

The 1R algorithm identified the best single attribute in terms of predicting disease status. The attributes were ranked in terms of effectiveness as a predictor and were also ranked in terms of the information gain that they provided (Tables 2.4 and 2.5). The PSIC conservation score was identified as the best classifier in a balanced dataset achieving 72%

correctly classified instances with the rules that defined a nsSNP as being disease status with a score difference > 0.89 and neutral with a PSIC score difference $<= 0.89$. These classifiers compared favourably with the conservation score rules identified by Ramensky *et al.* (2002) in their study whereby a PSIC score difference $<= 0.5$ was classified as benign, 1.5 to 2.0 possibly damaging and $>= 2.0$ probably damaging.

1R Rank	Attribute
72.82	conservation score (PSIC)
67.49	normalised relative accessibility
63.46	MMDBBIND
62.64	mass change
62.56	relative accessibility
62.23	exposure
61.41	PAM score
60.67	mutation residue
60.34	volume change
59.19	wild type residue

Table 2.4: Top 10 attributes for predicting nsSNP function using 1R with 10 fold cross validation and bucket size 14.

Information gain (bits)	Attribute
0.2	conservation score (PSIC)
0.1	normalised relative accessibility
0.09	wild residue
0.07	relative accessibility
0.06	PAM score
0.06	mass change
0.05	mutation residue
0.05	exposure
0.04	volume change
0.04	hydrophobicity difference

Table 2.5: The information gain per attribute when predicing nsSNP function.

2.2.2.2 Attribute set analysis

The J48 decision tree algorithm was used to evaluate the predictive performance of the following subsets of attributes:

- Set (1) - All variables.
- Set (2) - Structural variables.

- Set (3) - Non structurally dependent variables.
- Set (4) - Non structurally dependent variables excluding the conservation score (PSIC).
- Set (5) - Conservation score alone.

2.2.2.2.1 Effect of Imbalance Attribute sets (1) and (2) contained 3,821 nsSNPs when imbalanced and 1,218 when balanced, both sets included structural variables. Datasets (3), (4) and (5) contained 14,636 nsSNPs when imbalanced and 8,434 when balanced. They contained more nsSNPs than sets (1) and (2) because they were not dependent on structure.

The MCC increased with increasing balance within each of the sets of attributes. There was a difference in the MCC score between 0% balanced and 100% balanced of 0.24 for dataset (1), 0.29 for (2), 0.08 for (3), 0.07 for (4) and 0.15 for (5). The performance of the weighted sets lay between the level of 25% and 50% balancing for each attribute set (Figure 2.2).

The 100% balanced dataset (1) achieved a MCC of 0.49. When weighted and imbalanced the MCC was 0.3 and 0.25 respectively for this same set. The balanced dataset (3) was equal second in the rankings with a 75% balanced set (1), performing better than dataset (2). The conservation score alone (set (5)) achieved a similar MCC score when considered separately (MCC 0.43) as it did when it was included in set (3) (MCC 0.44) when 100% balanced. When the conservation score is excluded there is a drop of 0.16 in the MCC of the 100% balanced dataset (3). When set (2) is balanced it performs better than (4) but when it is not 100% balanced it has a lower MCC. Dataset (3) actually performs better than the dataset (1) when the datasets are $<= 50\%$ balanced or weighted. The imbalanced dataset (2) achieved the lowest MCC score.

The rules learnt from the machine learning approach were then applied to make predictions on nsSNPs where the function was unknown. All nsSNPs within Ensembl (Build 27_1) were used as the unknown test dataset. The dataset was trained on the 100% balanced dataset of 609 neutral and 609 disease nsSNPs using all variables. This resulted in a predicted classification along with a confidence score for each of the 'unseen' nsSNPs

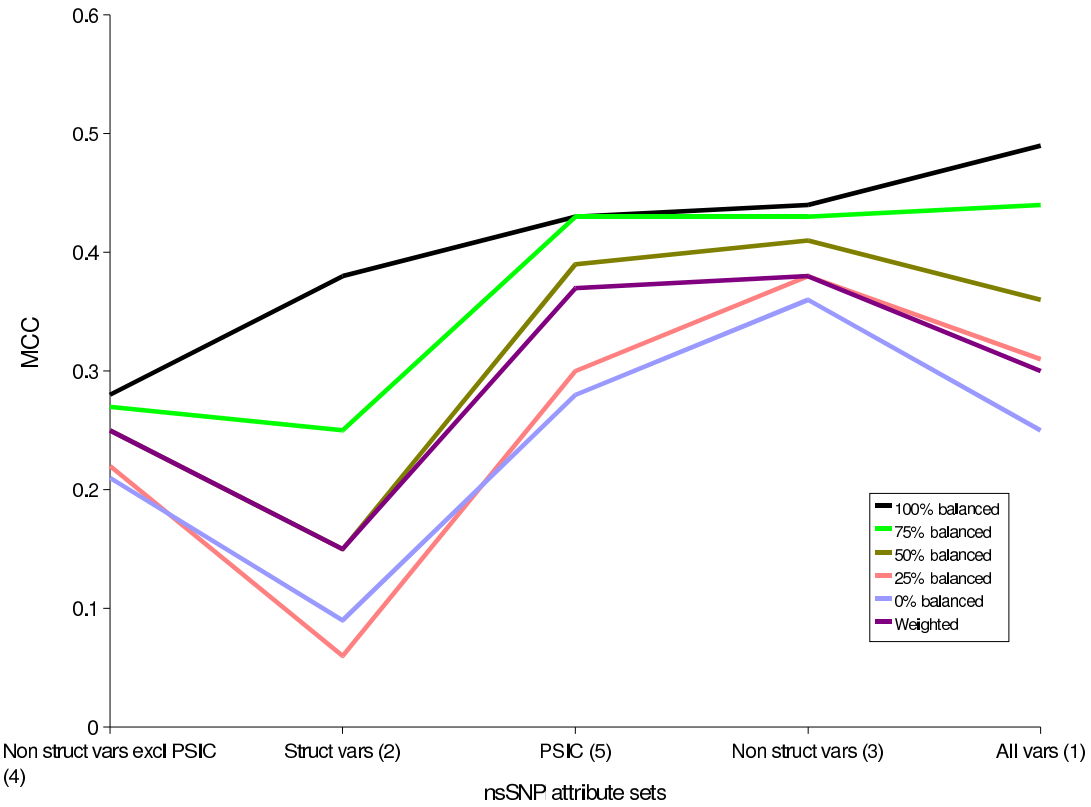


Figure 2.2: nsSNP function predictive performance of five attribute subsets measured using Matthews Correlation Coeffecient (MCC). Non struct vars excl PSIC - Non structurally dependent variables excluding the conservation score (PSIC); Struct vars - Structural variables; Non struct vars - Non structurally dependent variables; All vars - All variables.

within Ensembl. The predictions made for all of the Ensembl nsSNPs are available to be viewed within Ensembl as a Distributed Annotation System (DAS) source (Figure 2.3) (Dowell *et al.*, 2001).

2.3 Discussion

The SNP database was created to observe how various sequence and structural based nsSNP attributes as well as the level of balance in the training dataset affect nsSNP functional prediction performance. Using the optimal set of attributes and level of balance in the training dataset was found to increase the Matthews correlation coefficient (MCC) and therefore increase the value of the predictions for use in the targeted studies of EH, and other diseases.

The use of a 100% balanced dataset dramatically increased the MCC and removed any bias towards building rules for prediction of the disease state. Complete undersampling is a better choice than reweighting in addressing an imbalanced dataset. When imbalanced,

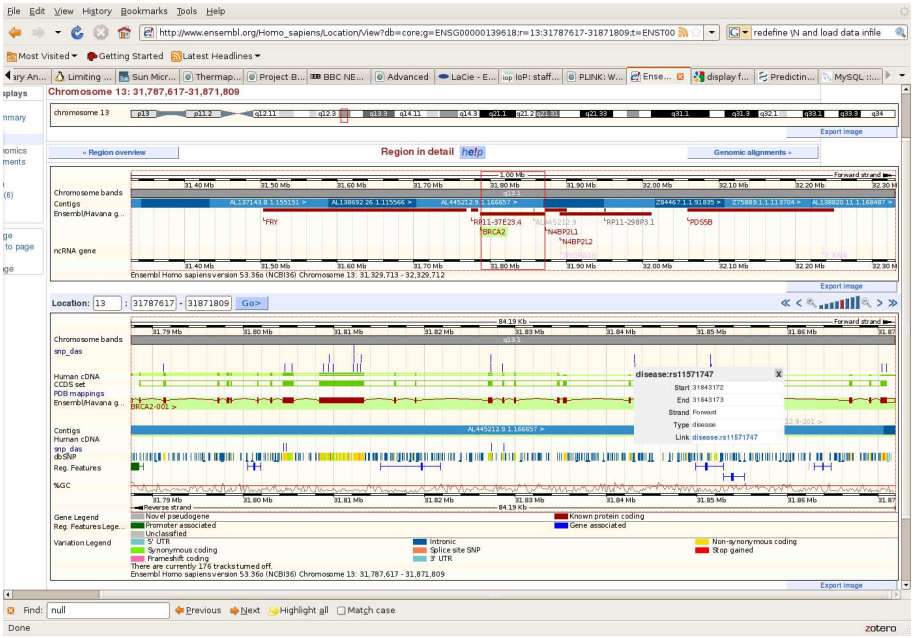


Figure 2.3: Screenshot of nsSNP function predictions (labelled as the snp_das track) integrated within the Ensembl browser as a DAS source (Dowell *et al.*, 2001)

performance using conservation alone (MCC 0.28) is close to that achieved by Bao & Cui (2005) (MCC 0.305) yet with a balanced dataset the MCC is greatly improved (MCC 0.43).

We saw a larger spread in the MCC when using the smaller datasets that included structural variables, because of the larger ratio of disease to neutral nsSNPs in these datasets. This explains why the MCC for the dataset of all variables performed best when $> 50\%$ balanced yet the performance drops below that of non structurally dependent variables when the level of balance falls below this figure. It also explains the similar pattern seen when comparing structurally dependent variables and non structurally dependent variables excluding conservation, except that the cut off lies at the 75% level of balance.

There are a number of caveats with the training dataset. The dataset may include nsSNPs predicted to be 'disease' where some of the nsSNPs may only be in linkage disequilibrium with the phenotype in question and may themselves not be causative. This 'pollutes' the training set and may lead to a higher error rate and lower MCC. Further filtering of the dataset would lead to a smaller but cleaner training set that could in turn lead to lower error rates and an increase in the MCC. Further complications could arise where molecular phenotypic changes that don't result in a physical phenotype and unstudied or

unobserved phenotypic changes may result in a nsSNP being classified as neutral that should be classified as disease. Improvements to the system could also be made if SNPs could be graded in terms of how damaging they are as opposed to the boolean states of disease and polymorphism that currently classifies them, in time databases may contain this information. Decision trees were used to build models for predicting functional nsSNPs due to their easily interpretable rules. Running Weka with all available classifiers and various configurations may identify a classifier that obtains improved accuracy.

Since completion of the nsSNP analysis, a number of further studies have been performed, thirteen of which cited the work in this thesis. Work has included a study focusing on 686 sequence based attributes (Hu & Yan, 2008) using a similar approach to that taken in this thesis. Performance was similar to results achieved in this study. Further work by Tian *et al.* (2007) created an SVM based application called Parepro which included sequence and evolutionary information surrounding a nsSNP and did not include structural attributes. A novel structure-based approach, Bongo (Bonds ON Graph), was introduced whereby protein structures were considered as residue-residue interaction networks (Cheng *et al.*, 2008). Graph theoretic approaches were applied to identify residues that are critical for maintaining structural stability within the network. The effect of a nsSNP change could then be evaluated. Performance was comparable to commonly used PolyPhen (Sunyaev *et al.*, 2001) and Panther (Thomas *et al.*, 2003) approaches. A study by Care *et al.* (2007) aimed to quantify effects of the different approaches used in the field. They concluded that the SWISSPROT training datatset used in this study was the preferred training dataset to date and some of the conclusions were based on findings from this thesis.

Reassuringly, previously observed trends can be seen in this study of a large number of nsSNPs. Disease nsSNPs tend to affect protein stability (Wang & Moult, 2001), are buried (Stitziel *et al.*, 2004) and often disrupt a conserved residue (Saunders & Baker, 2002). This work extends previous work by addressing the problems of imbalance and redundancy within the attributes for a large selection of natural nsSNPs and then goes on to make predictions on all Ensembl nsSNPs. Saunders & Baker (2002) and Bao & Cui (2005) showed that in the absence of a conservation score, structural

attributes are valuable predictors. Here it is affirmed, using machine learning methods that the sequence conservation measure is the most powerful single predictor and it has been shown that a high level of accuracy is achieved using the conservation score alone. It has also been shown that structural attributes in combination with the conservation score improves prediction accuracy but also that there are other non structurally dependent attributes that can reduce the error rate further and are valuable in the absence of a conservation score. The performance of all attribute subsets however, is very much dependent on how the datasets are configured. The maximum prediction accuracy can be achieved by combining all attributes of the nsSNP within a balanced dataset. The predictions based on all of these learnings are available for public use as a DAS source $http://www.brightstudy.ac.uk/das\_help.html$ (Figure 2.3) and as an annotation within the SNP function portal (Wang *et al.*, 2006).

Chapter 3

Protein function analysis

This chapter focuses on the performance of machine learning classifiers in predicting function for distantly related protein sequences. Typically, two approaches are used to address such a multi-class problem. The first involves adapting the algorithm to the multi-class problem directly. An example of an algorithm that can be easily generalized to cope with multi-class problems is the decision-tree classifier. The second approach involves creating several two-class problems and a class is assigned based on the predictions obtained from the two-class problems. This approach has the benefit of not requiring any changes to the underlying algorithm. Examples of this approach include error-correcting output codes (Dietterich & Bakiri, 1995) and pairwise classification (Fürnkranz, 2002). Here we experiment with a range of classifiers that implement varied approaches for addressing the multi-class problem.

Membership of a SCOP superfamily was used as a measure of functional relatedness (Murzin *et al.*, 1995). The SCOP database is a manually curated resource supported by a host of automated methods to provide comprehensive and accurate descriptions of the structural relationships between proteins where the structure is known. The relationship between sequence structure and function is indefinite and a number of studies have shown protein superfamilies within a single fold having diverse functions, an example being the aldo-keta reductases, a large hydrolase superfamily, and the thiol protein esterases which include the eye-lens and corneal crystallins (Hegyi & Gerstein, 1999). The TIM-barrel fold is an extreme example of divergent evolution with the fold functioning as a generic scaffold catalyzing 15 different enzymatic functions. Even at the superfamily level, there

can be difficulty in inferring function from structure. An example can be seen in the enolase superfamily where there are hundreds of sequences available. Known structures of this superfamily catalyze eight different overall reactions (Gerlt *et al.*, 2005). Despite this, proteins in the same SCOP superfamily are believed to be related from structural and other considerations and would therefore often be expected to have the same general functional role. However, they include proteins which are very diverse at the level of sequence similarity and for which relatedness would not be apparent from consideration of sequence alone.

3.1 Protein function analysis methods

This study was restricted to large and diverse SCOP superfamilies, namely those with more than 15 sequences that do not share more than 20% sequence identity. A range of popular machine learning methods as implemented in the Weka workbench (Witten & Frank, 1999) were employed and a web based clustered computing infrastructure was built to enable rapid identification of optimal classifiers and configurations (Figure 3.1). This tool parses and stores results in a MySQL database, whilst sending a summary to the user by email. A sequence enrichment step was introduced in order to increase the number of sequences available for training. The dataset provides a challenging benchmark but one which is very relevant to enhanced genome annotation strategies.

3.1.1 Protein domain dataset

Two datasets were created for analysis; the first comprised domains from single domain proteins exclusively and this was the main focus of this study. A second dataset included domains from multi domain proteins. The inclusion of SCOP domains from multi domain protein structures is useful for characterising domains but may present problems with the functional characterisation of a protein. Namely, the function of a multi domain protein (composed of 2 SCOP domains A and B for example) may not necessarily be the sum of the functions associated with the individual constituent domains, A and B. However including SCOP domains from multi domain protein structures does lead to many more

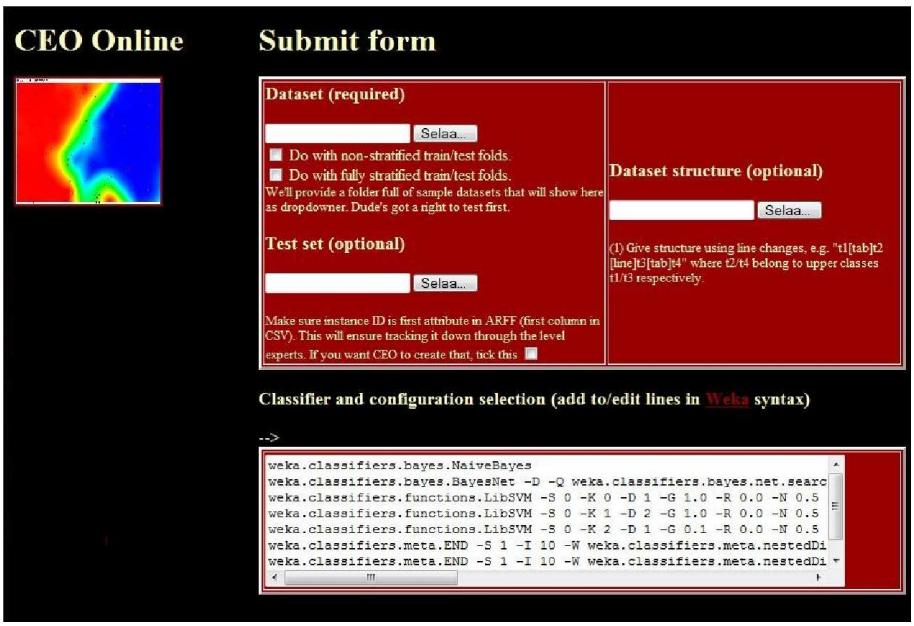


Figure 3.1: Screenshot of the web based version of Weka, which is integrated with a computing cluster.

examples.

Domain sequences were obtained from the Astral20 database which contains SCOP domain sequences sharing less than 20% sequence identity (Brenner *et al.*, 2000). Superfamilies containing fewer than 16 domains at this level of sequence redundancy were excluded. The datasets were split such that two thirds of instances from each superfamily were used for training and the remaining one third of instances for testing the models.

3.1.1.1 Superfamily enrichment

The SCOP database provides a gold standard structural resource with reliable comprehensive annotation, meaning that domains should be accurately classified at the level of superfamily despite being diverse at the sequence level. The aim was to extend this diverse set of domain sequences by including entries from sequence databases without known structure and therefore missing SCOP annotation. The reason for this was to boost the numbers of instances available for training the machine learning algorithms. It was necessary to be cautious, however, because if very remote relatives were included there was a danger they may not actually be part of the same superfamily. Sillitoe *et al.* (2005) have previously described an approach to recruit sequences into CATH domain superfamilies.

The following steps were performed to enrich the number of examples in each super-

family:

- A BLAST (Altschul *et al.*, 1997) search using each of the domain sequences from the diverse SCOP superfamilies was performed against the UniRef50 subset of the Uniprot database (Apweiler *et al.*, 2004).
- In order for a hit to be retained, the E value had to be <0.0005.
- Hits were excluded where $<80\%$ of the domain was aligned
- Hits were also excluded where the length of the aligned section of the UniRef50 hit was $<80\%$ of the length of the aligned section of the domain (to exclude hits that had long gaps within the alignment.)
- UniRef50 hits were further excluded that matched domains of more than one superfamily in order to reduce ambiguity in superfamily membership of the hit.
- BLASTClust (Dondoshansky, 2002) was then run against the resulting SCOP domains and UniRef50 hits for each superfamily to remove redundancy. For each cluster the SCOP domains were retained as the cluster representative when present.
- Results were compared where BLASTClust was used to remove redundant sequences at $>20\%$ and then $>30\%$ sequence identity. BLASTClust was set at these levels of sequence identity because below 25% similar function can not confidently be inferred by sequence alone (Wilson *et al.*, 2000). It was considered that 30% was a conservative cutoff where similar function could be more confidently inferred. At a cutoff of 20%, confidence in assumption of function was lower but it was considered to be of interest to compare to the 30% cutoff.

3.1.2 Protein domain features

Attributes selected for machine learning were based upon the properties explored by Dubchak *et al.* (1995) who analysed protein folds in the context of the SCOP classification. These attributes relate to the hydrophobicity, Van der Waals volume, polarity, polarizibility and predicted secondary structure of the amino acid sequence. The secondary structure (C=Coil, H=Helix, E=Strand) was predicted using PSIPRED (McGuffin

et al., 2000). Each amino acid was labelled as belonging to one of three groups for each of these descriptors (Table 3.1).

Property	**Group1**	**Group2**	**Group3**
Hydrophobicity	Polar R,K,E,D,Q,N	Neutral G,A,S,T,P,H,Y	Hydrophobic C,V,L,I,M,F,W
Normalized Van Der Waals	0-2.78 G,A,S,C,T,P,D	2.95-4.0 N,V,E,Q,I,L	4.43-8.08 M,H,K,F,R,Y,W
Polarity	4.9-6.2 L,I,F,W,C,M,V,Y	9.0-9.2 P,A,T,G,S	10.4-13.0 H,Q,R,K,N,E,D
Polarizibility	0.0.108 G,A,S,D,T	0.128-0.186 C,P,N,V,E,Q,I,L	0.219-0.409 K,M,H,F,R,Y,W
Secondary structure	C=Coil	H=Helix	E=Strand
Amino acid composition	n.a	n.a	n.a
Amino acid length	n.a	n.a	n.a

Table 3.1: Properties of each domain sequence that were used as attributes to predict superfamily membership using machine learning classifiers (n.a = not applicable). (The first 5 properties were taken from Dubchak *et al.* (1995))

All descriptors were analyzed in the context of their composition, distribution and transition along the amino acid sequence. Taking hydrophobicity as an example, the composition element comprised three attributes; the percentage composition of polar (P), neutral (N), hydrophobic (H) amino acids in the domain sequence. The transition was also composed of three hydrophobicity related attributes; the percentage frequency of P followed by N or N followed by P, the percentage frequency of P followed by H or H followed by P and the percentage frequency of N followed by H or H followed by a N. The distribution comprised 15 hydrophobicity related attributes describing the amino acid sequence in terms of the proportion of the length of the domain sequence that contained the first, 25%, 50%, 75%, 100% of each of the groups of amino acids (P, N or H). In addition to these previously studied properties, the amino acid sequence length (bins of length 20 amino acids) and amino acid composition were added as attributes. A total of 126 attributes were included in the machine learning analysis.

3.1.3 Machine learning

All machine learning analysis was performed using the Weka collection of tools and algorithms (Witten & Frank, 1999). The models were evaluated on an independent test dataset which comprised one third of the original non-enriched dataset.

3.1.3.1 Single attribute analysis

In order to identify the most effective attribute in the machine learning prediction, the 1R classifying algorithm (Holte, 1993) and the information gain (IG) attribute evaluator were used. The 1R classifying algorithm creates single level decision trees for each attribute and measures the prediction error rate. The IG evaluator measures the information required after using the attribute as a classifier subtracted from the information required before using the attribute as a classifier. In both algorithms the attributes were ranked in terms of their effectiveness as predictors using the default ranker search method (Witten & Frank, 1999).

3.1.3.2 Attribute set analysis

The performance of 32 machine learning classifiers in a total of 96 configurations were compared for the prediction of protein function based on assignment to one of 24 superfamilies using 126 amino acid based sequence attributes (Appendix C, Table 7). The clustered implementation of Weka (Witten & Frank (1999)) was used to rapidly identify the optimal classifiers and configurations.

The enrichment process was assessed by comparing the performance using the non-enriched resources and enriched resources using a BLASTClust cut-off of 20% and 30% sequence identity. In order to assess the performance of the length of the domain as an attribute, the prediction performances using all variables were compared with the performance of all variables excluding the length of the domain sequence.

3.1.3.3 Measure of performance

The performance of the machine learning methods was assessed using the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

The positive predictive value (precision) (P) of predictions can be described by

$$P = \frac{TP}{TP + FP}$$

and recall (sensitivity) (R) is considered to be

$$R = \frac{TP}{TP + FN}$$

The F-measure (F) combines the positive predictive value and recall measurements in the following manner

$$F = \frac{2(P * R)}{(P + R)}$$

3.1.4 Benchmarking

SCOP superfamilies chosen as classes in this study were specifically large diverse superfamilies whose domains shared no more than 20% sequence identity. It is typically difficult to classify members of such superfamilies using conventional sequence homology methods.

To evaluate the performance of the machine learning approach, the results of the PSI-BLAST program were compared with models built using the non-enriched datasets. PSI-BLAST was run using a similar method to Melvin *et al.* (2007). A database of UniRef90 sequences was initially used to create profiles for each of the studied SCOP domain sequences (Wu *et al.*, 2006). Each profile was then matched separately against a database composed solely of the Astral20 proteins from the studied superfamilies. Matches with an E-value $<$ 0.0005 over 5 iterations were identified.

A definitive comparison of PSI-BLAST with a model created by an SVM (no sequence enrichment) was difficult as various measures of performance could be used. For each PSI-BLAST query, we observed the number of matches with SCOP domains from the correct superfamily and incorrect superfamily using the defined threshold. We considered a query to be correctly assigned to a superfamily (TP) when the number of hits to domains from the true superfamily exceeded the number of hits from a false superfamily. We acknowledge that this approach is biased due to the size of the superfamily, but think that it is typical of the kind of approach taken when a user attempts to assign functional annotation.

3.2 Protein function analysis results

3.2.1 Protein domain datasets

Two datasets were created for analysis, a primary dataset comprising domains from single domain proteins only and a secondary dataset which included domains from multi domain proteins.

The exclusion of multi domain protein sequences in the single domain dataset reduced the number of domains included in the analysis from 4931 to 2867 (contained within 1136 superfamilies). Excluding superfamilies that contained fewer than 16 domains at this level of sequence redundancy further reduced the number of domains included to 573 contained within 24 superfamilies (columns 1 and 2 in Appendix B, Table 1). For the second dataset (which included domains from multi domain proteins), excluding superfamilies that contained fewer than 16 domains resulted in 1448 domains contained within 49 superfamilies (columns 1 and 2 in Appendix B, Table 2).

3.2.1.1 Superfamily enrichment

Increasing the number of diverse sequence examples in the training datasets involved taking entries from UniRef50 that showed distant homology yet similar function to a domain sequence within the superfamily being studied (Wu *et al.*, 2006). Appendix B Tables 1 and 2 show the number of instances per superfamily in the single and multi domain training datasets before and after the enrichment process using BLASTClust at 20% and 30% redundancy. The periplasmic binding protein-like II superfamily (id 53850) exhibited the biggest increase (10.38 fold single domain, 10.13 fold multi domain) in the number of instances after enrichment and the restriction endonuclease-like superfamily (52980) had the smallest increase (1.15 fold single domain, 1.13 fold multi domain).

3.2.2 Machine Learning

3.2.2.1 Single attribute analysis

The top ten attributes in the non-enriched datasets and at levels of 20% and 30% enrichment when using both the 1R and IG algorithms comprised attributes relating to the com-

position, transition and distribution of secondary structure elements (coil, helix, strand) and the length of the domain. The domain length was in the top 5 attributes in all but one training set and algorithm combination.

3.2.2.2 Attribute set analysis

3.2.2.3 Single domain dataset

Generally performance of the classifiers in both the single and multi domain datasets improved with the increasing level of enrichment in the training datasets. Best performing classifiers in the single domain dataset were the END classifier achieving 64.2% correctly classified instances on the non-enriched dataset, AdaBoostM1 obtaining 64.2% correctly classified instances on the dataset enriched at a level of 20% and LibSVM achieving 66.3% correctly classified instances with a dataset enriched at a level of 30%. END is a meta classifier for handling multi-class datasets with 2-class classifiers by building an ensemble of nested dichotomies (Dong *et al.*, 2005). AdaBoostM1 is a class for boosting a nominal class classifier using the Adaboost M1 method (Freund & Schapire, 1996). When excluding the domain length as an attribute, the best performing classifier was the END classifier obtaining 64.2% correctly classified instances using the 30% enriched training dataset.

The classifiers varied greatly in predicting each superfamily. Table 3.2 shows the performance per superfamily in the single domain dataset using the best performing LibSVM classifier and 30% enrichment. The model achieved the best performance in predicting membership to the ARM repeat superfamily (id 48371) (all alpha proteins class (id 46456)) with an F-measure of 0.91. The poorest performing superfamily was the nucleotide-diphospho-sugar transferases superfamily (id 53448) (alpha and beta proteins class a/b (id 51349)) with an F-measure of 0.

3.2.2.4 Multi domain dataset

There were 49 superfamilies (1448 domains) that had more than 15 domains within Astral20 when including multi domain proteins. Appendix B Table 3 shows the SCOP class and fold membership for the 49 superfamilies. The class alpha and beta proteins a+b (id

Superfamily	Precision	Recall	F-Measure
46458 a.1.1 sf Globin-like	0.8	0.8	0.8
46689 a.4.1 sf Homeodomain-like	0.67	0.67	0.67
46785 a.4.5 sf Winged helix DNA-binding domain	0.77	0.83	0.8
47266 a.26.1 sf 4-helical cytokines	0.86	0.75	0.8
48371 a.118.1 sf ARM repeat	1	0.83	0.91
49785 b.18.1 sf Galactose-binding domain-like	0.75	0.5	0.6
49899 b.29.1 sf Concanavalin A-like lectins/glucanases	0.67	0.57	0.62
50249 b.40.4 sf Nucleic acid-binding protein	0.57	0.89	0.7
50729 b.55.1 sf PH domain-like	0.75	0.5	0.6
51182 b.82.1 sf RmlC-like cupins	1	0.2	0.33
88633 b.121.4 sf Positive stranded ssRNA viruses	0.57	0.8	0.67
51445 c.1.8 sf (Trans)glycosidases	1	0.43	0.6
51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains	0.8	0.67	0.73
52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases	0.52	0.76	0.62
52833 c.47.1 sf Thioredoxin-like	0.86	0.67	0.75
52980 c.52.1 sf Restriction endonuclease-like	1	0.14	0.25
53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases	0.35	0.46	0.4
53383 c.67.1 sf PLP-dependent transferases	0.78	0.88	0.82
53448 c.68.1 sf Nucleotide-diphospho-sugar transferases	0	0	0
53474 c.69.1 sf alpha/beta-Hydrolases	0.75	0.82	0.78
53850 c.94.1 sf Periplasmic binding protein-like II	0.55	0.86	0.67
55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)	0.56	0.71	0.63
57059 g.3.6 sf omega toxin-like	0.8	0.57	0.67
57095 g.3.7 sf Scorpion toxin-like	0.63	0.83	0.71

Table 3.2: Performance in predicting the 24 SCOP superfamilies (excluding multi domain proteins) using Support Vector Machines (LibSVM) with enrichment at a redundancy cutoff of 30%.

53931) is better represented in this analysis.

The same configurations and evaluation method were applied to this dataset. Best performing classifiers for each dataset were AdaBoostM1 achieving 48.5% correctly classified instances using the non-enriched dataset, END obtaining 53.7% correctly classified instances on the training dataset enriched at 20%, and AdaBoostM1 achieving 55.6% on the training dataset enriched at 30%. The results of machine learning for each superfamily using AdaBoostM1 and enrichment at sequence identity of 30% can be seen in Table 3.3. When excluding the domain length as an attribute, the best performing classifier obtained 54.8% correctly classified instances using the 30% enriched training dataset and the END classifier.

Again, the success of the machine learning methods in predicting SCOP superfamily varied greatly depending on the superfamily with F-measure ranging from 0 to 0.92. The top performing superfamilies were the globin-like (id 46458) (all alpha protein class (id 46456)), and C2H2 and C2HC zinc finger (id 57667) (small proteins class (id 56992)) superfamilies, both with an F-measure of 0.92. The ARM repeat superfamily (id 48371) still performed well being ranked 4th in terms of F-measure (0.82). The poorest performing superfamilies were the restriction endonuclease-like (id 52980), the nucleotidylyl transferase (id 52374) (both belonging to the alpha and beta proteins class a/b (id 51349)) and the cysteine proteinases (id 54001) (Alpha and beta proteins a+b (id 53931)) superfamilies, all with F-measures of 0.

Table 3.3: Performance in predicting the 49 SCOP superfamilies (analysis including multi domain proteins) using AdaBoostM1 with enrichment at a redundancy cutoff of 30%.

Superfamily	**Precision**	**Recall**	**F-Measure**
46458 a.1.1 sf Globin-like	0.86	1	0.92
46626 a.3.1 sf Cytochrome c	0.67	0.75	0.71
46689 a.4.1 sf Homeodomain-like	0.57	0.67	0.62
46785 a.4.5 sf "Winged helix" DNA-binding domain	0.65	0.74	0.69
47266 a.26.1 sf 4-helical cytokines	1	0.38	0.55
47473 a.39.1 sf EF-hand	0.5	0.17	0.25
48371 a.118.1 sf ARM repeat	0.78	0.88	0.82
48726 b.1.1 sf Immunoglobulin	0.68	0.83	0.75
49265 b.1.2 sf Fibronectin type III	0.42	0.5	0.46
81296 b.1.18 sf E set domains	0.22	0.15	0.18
49503 b.6.1 sf Cupredoxins	0.67	0.5	0.57
49785 b.18.1 sf Galactose-binding domain-like	0.57	0.8	0.67
49899 b.29.1 sf Concanavalin A-like lectins/glucanases	0.57	0.73	0.64
50249 b.40.4 sf Nucleic acid-binding proteins	0.38	0.3	0.33
50729 b.55.1 sf PH domain-like	0.55	0.6	0.57
51011 b.71.1 sf Glycosyl hydrolase domain	0.4	0.4	0.4
51182 b.82.1 sf RmlC-like cupins	0.67	0.67	0.67
88633 b.121.4 sf Positive stranded ssRNA viruses	0.75	0.43	0.55

Continued on Next Page…

Table 3.3 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**
51445 c.1.8 sf (Trans)glycosidases	0.77	0.81	0.79
51569 c.1.10 sf Aldolase	0.67	0.33	0.44
51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains	0.5	0.7	0.58
51905 c.3.1 sf FAD/NAD(P)-binding domain	0.6	0.27	0.38
52317 c.23.16 sf Class I glutamine amidotransferase-like	0.5	0.4	0.44
52374 c.26.1 sf Nucleotidylyl transferase	0	0	0
52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases	0.36	0.49	0.42
52833 c.47.1 sf Thioredoxin-like	0.61	0.79	0.69
52980 c.52.1 sf Restriction endonuclease-like	0	0	0
53067 c.55.1 sf Actin-like ATPase domain	0.43	0.33	0.38
53098 c.55.3 sf Ribonuclease H-like	0.56	0.63	0.59
53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases	0.39	0.36	0.37
53383 c.67.1 sf PLP-dependent transferases	0.7	0.88	0.78
53448 c.68.1 sf Nucleotide-diphospho-sugar transferases	0.33	0.17	0.22
53474 c.69.1 sf alpha/beta-Hydrolases	0.42	0.62	0.5
53850 c.94.1 sf Periplasmic binding protein-like II	0.38	0.75	0.5
56784 c.108.1 sf HAD-like	0.5	0.4	0.44
54001 d.3.1 sf Cysteine proteinases	0	0	0

Continued on Next Page…

Table 3.3 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**
54211 d.14.1 sf Ribosomal protein S5 domain 2-like	0.6	0.33	0.43
54236 d.15.1 sf Ubiquitin-like	0.71	0.83	0.77
54373 d.16.1 sf FAD-linked reductases, C-terminal domain	1	0.6	0.75
54593 d.32.1 sf Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	0.8	0.67	0.73
55347 d.81.1 sf Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain	0.5	0.33	0.4
55486 d.92.1 sf Metalloproteases (“zincins”), catalytic domain	0.5	0.5	0.5
55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)	0.71	0.56	0.63
56672 e.8.1 sf DNA/RNA polymerases	0.8	0.8	0.8
57059 g.3.6 sf omega toxin-like	0.5	0.29	0.36
57095 g.3.7 sf Scorpion toxin-like	0.57	0.67	0.62
57196 g.3.11 sf EGF/Laminin	0.71	1	0.83
57667 g.37.1 sf C2H2 and C2HC zinc fingers	1	0.86	0.92
57716 g.39.1 sf Glucocorticoid receptor-like (DNA-binding domain)	0.71	0.71	0.71

3.2.3 Benchmarking

Performance of PSI-BLAST and SVMs (using the non-enriched datasets) was very variable, with the two methods often differing in performance for each superfamily (Appendix B Tables 4 and 5). We found that 8 out of 24 superfamilies achieved a better F-measure with SVMs in the single domain analysis and 10 out of 49 obtained a greater F-measure in the multi domain analysis. F-measures were comparable for many other superfamilies, especially in the single domain study. SVMs outperformed PSI-BLAST for all 5 of the studied superfamilies from the small protein class (id 56992) as well as performing better or comparably for superfamilies of the all alpha proteins class (id 46456). The mean performance measures per superfamily are shown in Table 3.4.

	SVM			**PSI-BLAST**		
Dataset	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Single domain	0.64	0.61	0.61	0.96	0.6	0.7
Multi domain	0.5	0.41	0.42	0.95	0.57	0.67

Table 3.4: The mean precision, recall and F-measure per superfamily produced by SVMs and PSI-BLAST using the unenriched datasets comprising 24 (domains from single domain proteins) and 49 superfamilies (including domains from multi domain proteins).

3.3 Discussion

The SCOP database provides a gold standard structural resource with reliable comprehensive annotation, meaning that domains should be accurately classified at the level of superfamily despite being diverse at the sequence level. It is desirable to be able to build machine learning models in order to be able to assign this functional annotation to domains where the structure is unknown and function is difficult to infer by traditional methods.

3.3.1 Superfamily enrichment

Machine learning methods benefit from having more training data. Our seed data sets, namely 24 and 49 large and sequence diverse (no two sequences sharing more than 20% sequence identity) superfamilies, provide a ‘ground truth’ since we know from SCOP

(which uses structural and other considerations) that the proteins are in fact related. However, the datasets were somewhat limited in size and the question of how to extend them was not trivial: adding very weakly related sequences detected by PSI-BLAST might contaminate the superfamily by introducing proteins which in fact did not belong to the superfamily; but being very restrictive with the cut off would only add more examples of close homologs. We observed that the performance of the machine learning algorithms improved when the SCOP superfamily datasets were enriched and that the percentage of sequence similarity used as a cutoff in the enrichment process effected prediction performance. The performance at the sequence identity cutoff of 30% was better than the lower cutoff of 20%. At the 20% level there was the possibility of contamination and alignment errors which would affect the predicted secondary structure attributes and may have led to lower performance. It is expected that this step could improve performance if applied to the many published fold prediction models (Ashburner *et al.*, 2000; Ding & Dubchak, 2001; Lin *et al.*, 2005; Shen & Chou, 2006; Melvin *et al.*, 2007; Shamim *et al.*, 2007; Damoulas & Girolami, 2008)

3.3.2 Single attribute analysis

Attributes vary in their contributions to the predictions of superfamily membership in the machine learning models. Jensen *et al.* (2002) previously concluded that secondary structure was the most important descriptor in their protein function prediction study. This study found predicted secondary structure was important for predicting function with the composition, transition and distribution of secondary structure elements being the most important attributes in the single attribute analysis. Jensen *et al.* (2002) concluded that protein length was not a valuable attribute in their studies. However, we found that the length of the sequence was a valuable attribute in the single attribute analysis for the 24 superfamilies of the single domain analysis. Figure 3.2 shows many superfamilies display a clustering with regards to length in the non-enriched single domain resource. All four domains over 550 residues belong to the ARM repeat superfamily (48371 a.118.1 sf), with the Importin beta domain (d1qgra_ a.118.1.1) being the longest domain (877 residues). However, when combining all attributes for use with the 32 applied classifiers

the models still performed well following the exclusion of domain length as an attribute. This suggests that the classifiers were not dependent on domain length as an attribute and other sequence properties were important in accurately classifying superfamilies.

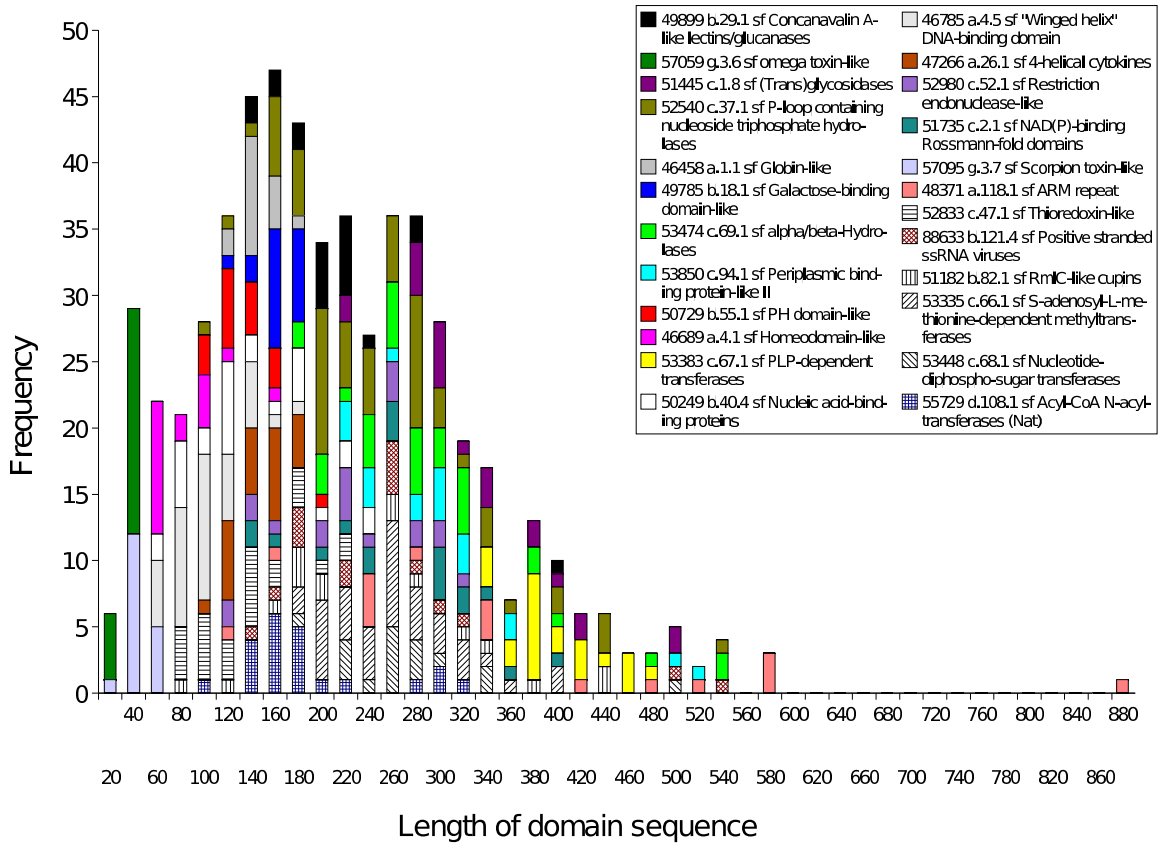


Figure 3.2: Domain sequence length for superfamilies from Astral20 that contain >15 domains (excluding multi domain proteins). The length is grouped into bins of 20 amino acids. This figure highlights the clustering of superfamilies by domain sequence length. The Importin beta domain (d1qgra_a.118.1.1) from the ARM repeat superfamily (48371 a.118.1 sf) is the longest domain (877 residues).

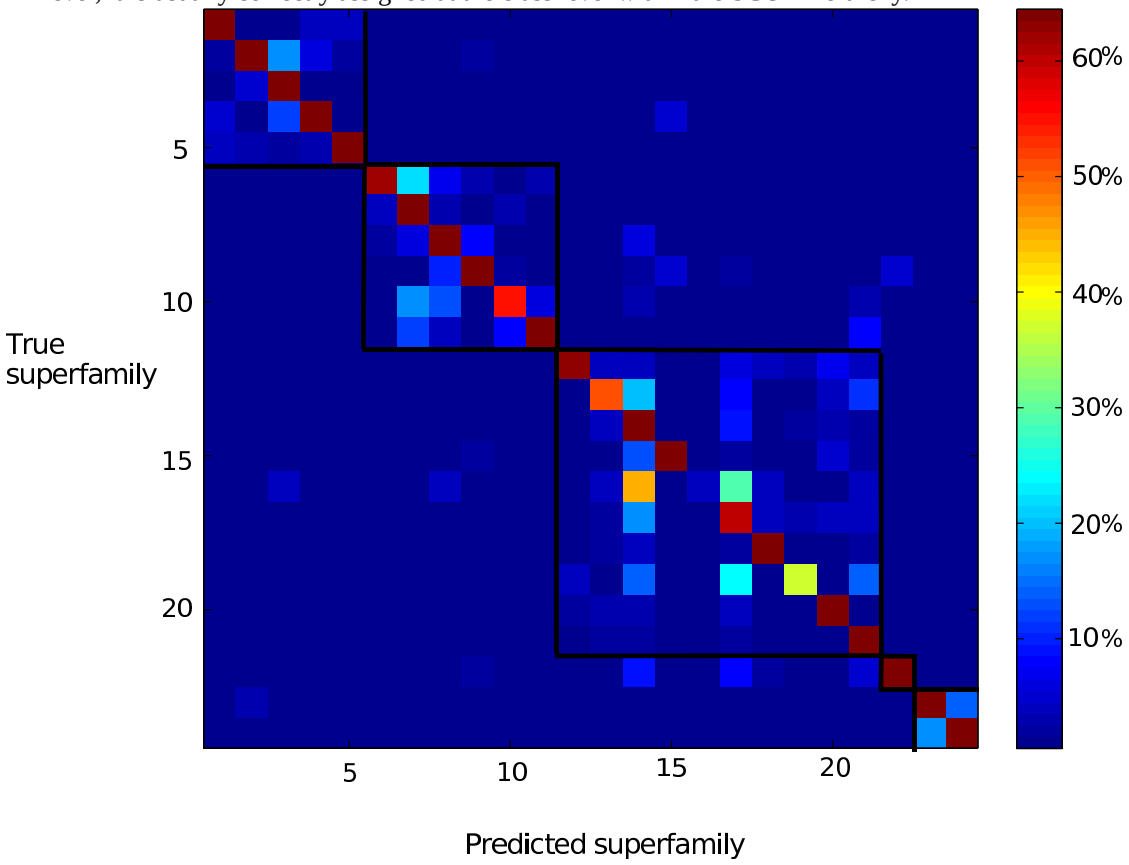
3.3.3 Attribute set analysis

3.3.3.1 Single domain dataset

In analysis of combined attributes in the single domain resource, the best performing classifier was LibSVM, obtaining 66.3% correctly classified instances in an independent test set using a training dataset that was enriched at a level of 30% sequence identity. The success of the machine learning methods in predicting SCOP superfamily varied greatly depending on the superfamily (Table 3.2). The P-loop containing nucleoside triphosphate hydrolases (id 52540) and S-adenosyl-L-methionine-dependent methyltransferases

(id 53335) had a large proportion of false positives. Thirty eight percent of instances belonging to the 53335 superfamily were classified as 52540 and 15% of 52540 instances were classified as 53335 suggesting that there is some similarity between these superfamilies or that the diversity of both groups means that classifying the two is difficult. Both superfamilies belong to the same alpha and beta proteins (a/b) (id 51349) class but are members of different folds within the SCOP classification. Figure 3.3 shows clearly (black bordered squares) that when the model misclassifies an instance, it usually classifies it correctly at the SCOP class level. This may reflect that the 'predicted secondary structure' attribute facilitated the correct class assignment (the SCOP class level represents the overall secondary structure composition of the protein). The alpha and beta proteins a/b (id 51349) class contains the largest number of superfamilies (10) in this study, resulting in some misclassifications among the superfamilies that it contains. The poorest performing superfamilies, nucleotide-diphospho-sugar transferases (id 53448) and restriction endonuclease-like (id 52980), both belong to this class. The 52980 superfamily also contains the smallest number of instances (15) in the training dataset enriched at 30%. The best performing superfamily, ARM repeat (id 48371), belongs to a class (all alpha proteins (id 46456)) containing only 5 superfamilies from this study and has 50 instances in the 30% enriched training dataset. It might be expected that there would be many misclassifications between the homeodomain-like superfamily (id 46689) and the "winged helix" DNA-binding domain superfamily (id 46785) as both of these superfamilies belong to the same SCOP fold (DNA/RNA-binding 3-helical bundle). Whilst 33% of 46689 instances were misclassified as 46785, only 16% of 46785 instances were misclassified as 46689. This may be explained by the large number instances belonging to the 46785 superfamily, 114 at 30% enrichment, compared to 35 for the 46689 superfamily (Appendix B, Table 1). The larger number of instances may have resulted in a better model being constructed. Therefore, it appears that the diversity of superfamilies at the class level as well as the number of instances available for training affect the performance of the classifiers.

Figure 3.3: Superfamily confusion matrix produced by an SVM model with a dataset enriched at 30% sequence identity (excluding multi domain proteins). Each small square represents the percentage of domains belonging to the superfamily on the y axis (true superfamily) that were predicted to belong to the superfamily on the x axis (predicted superfamily). The colour of each square relates to the predicted percentage of total instances according to the colour ramp on the right of the matrix. Black bordered squares represent the 5 classes that the 24 superfamilies are grouped into (Appendix B, Table 6). This figure highlights the fact that when a domain is misclassified at the superfamily level, it is usually correctly assigned at the class level within the SCOP hierarchy.



Axis labels:

(1) 46458 a.1.1 sf Globin-like, (2) 46689 a.4.1 sf Homeodomain-like, (3) 46785 a.4.5 sf "Winged helix" DNA-binding domain, (4) 47266 a.26.1 sf 4-helical cytokines, (5) 48371 a.118.1 sf ARM repeat, (6) 49785 b.18.1 sf Galactose-binding domain-like, (7) 49899 b.29.1 sf Concanavalin A-like lectins/glucanases, (8) 50249 b.40.4 sf Nucleic acid-binding proteins, (9) 50729 b.55.1 sf PH domain-like, (10) 51182 b.82.1 sf RmlC-like cupins, (11) 88633 b.121.4 sf Positive stranded ssRNA viruses, (12) 51445 c.1.8 sf (Trans)glycosidases, (13) 51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains, (14) 52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases, (15) 52833 c.47.1 sf Thioredoxin-like, (16) 52980 c.52.1 sf Restriction endonuclease-like, (17) 53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferase, (18) 53383 c.67.1 sf PLP-dependent transferases, (19) 53448 c.68.1 sf Nucleotide-diphospho-sugar transferases, (20) 53474 c.69.1 sf alpha/beta-Hydrolases, (21) 53850 c.94.1 sf Periplasmic binding protein-like II, (22) 55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat), (23) 57059 g.3.6 sf omega toxin-like, (24) 57095 g.3.7 sf Scorpion toxin-like

3.3.3.2 Multi domain dataset

The inclusion of multi domain proteins resulted in there being over twice the number of superfamilies available for study, with superfamilies from the alpha and beta proteins a+b class (id 53931) being better represented (Appendix B, Table 3). The AdaBoostM1 classifier obtained 55.6% accuracy with a training dataset enriched at 30%. The classifiers still performed well despite the increase in the number of superfamilies resulting from the inclusion of domains from multi domain proteins. Again, the success of the machine learning methods in predicting SCOP superfamily varied greatly depending on the superfamily (Table 3.3). The restriction endonuclease-like (id 52980), nucleotidylyl transferase (id 52374) and cysteine proteinases (id 54001) superfamilies all performed poorly with F-measures of 0. The top performing superfamilies were the globin-like (id 46458) and C2H2 and C2HC zinc finger (id 57667) superfamilies. The globin-like (id 46458) superfamily was ranked 3rd in the single domain analysis whereas the C2H2 and C2HC zinc finger (id 57667) superfamily was absent. The globin-like (id 46458) superfamily was ranked 29th in terms of the fold increase in the number of instances after the enrichment step at 30% and was ranked 35th in terms of the total number of instances after the enrichment. The C2H2 and C2HC zinc finger (id 57667) superfamily was ranked 19th in terms of the fold increase in the number of instances after the enrichment step at 30% and was ranked 29th in terms of the total number of instances after the enrichment. It therefore seems unlikely that performance was biased towards these superfamilies due to imbalance in the dataset. Again, the P-loop containing nucleoside triphosphate hydrolases (id 52540) had a large proportion of false positives (64%). Additionally 23% and 30% of domains from the superfamily E set domains (id 81296) were misclassified as superfamilies Immunoglobulin (id 48726) and fibronectin type III (id 49265) respectively. Both superfamilies belong to Immunoglobulin-like beta-sandwich fold (id 48725) which is part of the all beta proteins class (id 48724) and were excluded from the single domain dataset.

Generally, similar patterns were observed in the single and multi domain datasets with misclassifications at the superfamily level being correctly assigned at the fold or class level. Superfamilies that performed best in both the single and multi domain analysis

belonged to either the all alpha protein (id 46456) or small protein classes (id 56992). Poorest performers belonged to the alpha and beta classes (a/b or a+b) (ids 51349, 53931).

3.3.4 Benchmarking

For most superfamilies, PSI-BLAST did not detect unrelated domains with scores better than the threshold, although the program failed to detect all the possible correct matches (ie to related domains). For these superfamiles, the definition of a correct assignment, namely that the number of hits to domains from the true superfamily exceeded the number of hits from a false superfamily, meant that the precision was 1.0 leading to a high F-measure. A more exacting requirement for confident classification would be the identification of multiple (ideally all) related domains with scores better than the threshold. As an example, we describe the breakdown of PSI-BLAST results for 2 superfamilies. The globin-like superfamily (id 46458) performed well within the PSI-BLAST results in the single domain analysis (2nd top F-measure). Fourteen out of 15 domains in this superfamily were assigned to the true superfamily. However, of these 14, four were classified based on single matches, that is PSI-BLAST only detected a match to one other protein in the same superfamily. The "Winged helix" DNA-binding domain superfamily (46785) produced relatively poor results with PSI-BLAST (F-measure 0.49). Of the 37 domains within this superfamily, only 12 were assigned to the true superfamily. Matches for 5 of these 12 were based on single hits and the maximum number of correctly returned domains for any query was 8. So, almost half of the assignments were not confident classifications. The comparison with PSI-BLAST for the detection of these remotely related proteins shows that there are global sequence properties that can be used to successfully classify domains from superfamilies, with the performance in many cases depending on the class that the superfamily belongs to. As previously stated, a definitive comparison of PSI-BLAST with a model created by an SVM (no sequence enrichment) was difficult as various measures of performance could be used and performance of the models is further improved when we use BLAST as part of the enrichment process.

3.3.5 Summary

The protein universe does not contain only 24 or 49 superfamilies and we have not allowed for this possibility. The approach we describe does not allow for an extra category 'unknown superfamily'. One area of improvement would involve providing a method for identifying an instance that does not belong to any of the studied (24 or 49) superfamilies. This might evolve as a pre-process step. Additionally, the employed attributes are not expected to be optimal for detecting close sequence relationships for which good solutions already exist.

Whilst the methods described here do not provide a complete solution for superfamily prediction they show that machine learning methods that consider simple sets of global sequences based attributes may be useful for suggesting superfamily membership and hence narrow down the potential functional space, especially for superfamilies belonging to all alpha (id 46456) and small protein classes (id 56992). This study shows that machine learning approaches to predicting SCOP categories can be improved by performing a sequence enrichment step that exploits unannotated sequences within genomic sequence databases. As such these approaches may complement profile methods for detecting distant relationships.

Chapter 4

Combining protein-protein interaction network and sequence attributes for predicting hypertension related proteins

This chapter describes an exploratory study examining the properties of 65 proteins listed as being associated with hypertension in the Online Mendelian Inheritance in Man database (OMIM, Hamosh *et al.* (2002)). The performance of a classifier which includes protein-protein interaction (PPI) network, sequence and GO attributes for the detection of hypertension related candidate proteins is reported. Protein-protein interactions form networks which can be explored using graph theoretic approaches. The networks can be thought of as undirected cyclic graphs where the proteins are nodes and the interactions are edges. If proteins A and B directly interact then there exists an edge connecting nodes A and B.

4.1 Hypertension PPI and sequence analysis methods

4.1.1 Dataset

OMIM is a comprehensive catalogue of human genes and their associated genetic phenotypes. It provides a set of positive examples for machine learning approaches to build classifiers for predicting disease genes. Each record in the OMIM database is associ-

ated with a unique identifier which relates to a disease, the observed symptoms and the associated genes. The symptoms field of each OMIM entry was parsed for the term 'hypertension' and the results were manually filtered. The genes associated with OMIM entries displaying hypertension as a symptom were then mapped onto their SWISSPROT protein identifiers (Boeckmann *et al.*, 2003).

4.1.2 Protein-protein interaction network properties

Protein-protein interactions involving hypertension related SWISSPROT identifiers were extracted from the OPHID database (Brown & Jurisica, 2005). OPHID is an on-line database of human protein-protein interactions built by mapping high-throughput model organism data to human proteins. It also integrates data from yeast two-hybrid, literature-based interaction and orthology-based interaction sources. The hypertension related SWISSPROT proteins (nodes) present in OPHID are referred to as HTd (hypertension dataset). One thousand datasets, each containing the same number of proteins as HTd (65), were then generated by randomly selecting proteins (nodes) from OPHID, which included the HTd proteins. We refer to this group of datasets as Rd1..1000.

In order to investigate the PPI properties relating to hypertension, a two step approach was taken. Firstly, the 'general topology' of each HTd protein was investigated whereby PPI properties of each HTd protein were investigated in relation to all surrounding proteins. Secondly, network properties were investigated specifically in relation to other HTd proteins ('dataset topology'). Comparisons were made with the Rd1..1000 datasets. The aim of this analysis was to identify whether HTd proteins were better connected than random and whether any differences could be explained by their general background connectivity. For example, can short distances between HTd proteins be explained through HTd proteins being interaction hubs? A perl module (PPI.pm) was created that enables a graph to be created and written to disk with the benefit that the graph structure does not have to be created and read into memory each time a script is run. This saves a significant amount of time when performing repeated graph theoretical analysis on large graph structures.

4.1.2.1 General topology

Degree of nodes: The mean degree (total number of edges associated with protein (*p*)) was calculated for OPHID as a whole, for HTd and Rd1..1000. This measure was then extended to identify the number of proteins within a radius of 3 interaction steps from *p* (figure 4.1).

Clustering coefficient: The clustering coefficient (C) for protein *p* is the number of links between the proteins that directly interact with *p* divided by the number of links that could possibly exist between them (if the directly interacting proteins were a clique). This measure originates from Watts & Strogatz (1998) who used it to determine whether a network was 'small-world'. The clustering coefficient was calculated for each HTd and each Rd1..1000 protein.

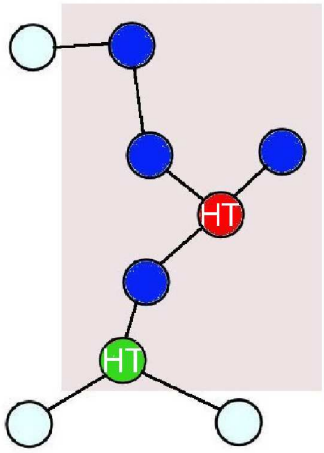


Figure 4.1: Illustration showing the number of proteins within a chosen radius of a selected hypertension related protein (red node). A radius of 2 is shown as an example (greyed area), the blue and green nodes are proteins falling within this radius. The blue node indicates that the protein is not hypertension related whereas the green node indicates a hypertension related protein.

4.1.2.2 Dataset topology

Degree of nodes: The mean degree was recalculated for each dataset (HTd, Rd1..1000) where only interactions with proteins from the same dataset were considered. This measure was then extended to identify the number of proteins from the same dataset within a radius of 3 interaction steps (green nodes in figure 4.1).

Geodesic distance: The length of the shortest connecting path between each pair of HTd proteins (HTd protein A to HTd protein B), and each pair of random proteins (Rd*x* protein

A to Rd*x* protein B) was calculated using Dijkstra's algorithm (Dijkstra, 1959).

Interaction subnetworks: We derived *expanded* subnetworks for each of the datasets, using the approach of Chen *et al.* (2006), whereby all the proteins and their directly interacting partners were selected. The proportion of all proteins from each of these *expanded* subnetwork datasets that were contained within the largest connected component were calculated. A connected component is a set of proteins whereby each protein can be reached from any other protein via a combination of interaction steps.

4.1.3 Hypertension pathways and protein function

To investigate pathway properties of hypertension related proteins, proteins from HTd were mapped to identifiers from the KEGG database (Kanehisa *et al.*, 2006). We excluded the following KEGG identifiers that related to types of interactions as opposed to pathways, although we are aware there is some subjectivity in this selection: ABC transporters, phosphotransferase system (PTS), two-component system, neuroactive ligand-receptor interaction, cytokine-cytokine receptor interaction, ECM-receptor interaction, cell adhesion molecules (CAMs), aminoacyl-tRNA biosynthesis, type II secretion system, type III secretion system, type IV secretion system, SNARE interactions in vesicular transport, ubiquitin mediated proteolysis, proteasome, cell cycle - yeast. The distribution of HTd proteins in the remaining pathways was investigated and compared to Rd1..1000.

The semantic similarity of Gene Ontology (GO) terms from each aspect (biological function, molecular process and cellular location) was obtained for the HTd proteins using the program G-Sesame (Wang *et al.*, 2007). A GO term's semantics (biological meanings) are encoded into a numeric value by aggregating the semantic contributions of all their ancestor terms in the GO graph. The similarity between GO terms is presented by aggregating the semantic contributions of their shared ancestor terms over the sum of the GO term semantic scores. An algorithm has been designed to measure the functional similarity of two genes based upon the semantic similarities among the GO terms that annotate these genes. The correlation between the semantic similarity of GO terms and geodesic distance apart in the PPI network was then measured for pairs of HTd proteins.

GO slims are cut-down versions of the GO categories containing a subset of the terms

in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms. The distribution of GO slim (Ashburner *et al.*, 2000) molecular functions and biological processes were studied in order to identify categories that were over-represented or under-represented in hypertension proteins compared to the Rd1..1000 datasets.

4.1.4 Classification

A machine learning approach was taken to predict hypertension related proteins using a combination of attributes from the PPI and GO analysis, combined with physicochemical properties of the protein sequences. The training dataset comprised the proteins contained within Rd1..30 (1950 instances) and the HTd dataset (65 instances).

The selected attributes relating PPI network properties of each protein were: the geodesic distance to the closest known HTd protein; the average and standard deviation of distances from each HTd protein; the number of direct interactions; the number of direct interactions with HTd proteins; the number of proteins up to 2 interactions away (up to one intermediary); the number of HTd proteins up to 2 interactions away; the number of proteins up to 3 interactions away (up to two intermediaries); the number of HTd proteins up to 3 interactions away; Attributes relating to molecular function and biological process were selected from GO slim categories that were found to be either over or underrepresented within the hypertension dataset, namely, 'response to stimulus' (GO:0050896), 'electron transport' (GO:0006118) and 'oxidoreductase activity' (GO:0016491). Physicochemical properties for each protein sequence were calculated using the Protparam program at Expasy (`www.expasy.org`). A bioperl (`www.bioperl.org`) module (Bio::Tools::Protparam) was created specifically for this purpose. Sequence properties used in the classifier were: amino acid length; number of negatively charged amino acids; number of positively charged amino acids; molecular weight; theoretical pI; number of carbon atoms; number of hydrogen atoms; number of nitrogen atoms; number of oxygen atoms; number of sulphur atoms; half life; instability Index; stability class; aliphatic index; GRAVY; amino acid composition; The GRAVY (Grand Average of Hydropathy) value for a peptide or protein was calculated as the sum of hydropathy values of all

the amino acids, divided by the number of residues in the sequence (Kyte & Doolittle, 1982). Various feature selection methods were tested using the Weka workbench (Witten & Frank, 1999) to remove redundancy and identify key attributes.

Because there was a large imbalance in the training dataset (many more random proteins than hypertension proteins), a CostSensitive classifier (Witten & Frank, 1999) was used as a wrapper around a Bagged PART classifier (Frank & Witten, 1998b; Breiman, 1996). A cost could then be applied for an incorrect HTd protein classification during ten fold cross validation in an attempt to address the imbalance. This weighted approach has been shown to be a succesful method for coping with class imbalance using a similar type of classifier and has an advantage over undersampling in that there is no loss of information (Chen *et al.*, 2004). Choosing a cost depends on priorities. For example, a researcher may be prepared to accept a high false positive rate (FPR) in order to obtain a high rate of recall for hypertension related proteins. The classifier was run 400 times with a range of cost matrices that applied varying penalties for incorrectly predicting a HTd protein using a key set of attributes. The bagged PART classifier is a decision list classifier that uses a separate-and-conquer approach. A partial C4.5 decision tree is built in each iteration and the 'best' leaf is made into a rule. It performs well in terms of speed because no postprocessing required. The runs were repeated using the simple majority-rule approach.

When benchmarking the classifier we wished to identify any sequence similar proteins as some of our attributes are sequence based. BLASTClust (at a level of 25% identity) (Dondoshansky, 2002) was used to identify sequence homologs within the HTd dataset.

4.2 Hypertension protein PPI and sequence analysis results

We isolated 96 hypertension related genes from OMIM, 90 of which could be mapped to SWISSPROT identifiers. Where an OMIM id had multiple associated proteins, we made the assumption that all were associated with hypertension and included them in the dataset as there was insufficient evidence to assume otherwise. Of the 90 ids, 65 were present within OPHID. These 65 proteins were associated with 47 diseases (distinct OMIM ids)

where hypertension was recorded as a symptom. The average number of proteins per OMIM id was 1.5. We refer to this dataset as HTd. The OPHID database used in this study contained 48,222 interactions.

4.2.1 Network properties

4.2.1.1 General topology

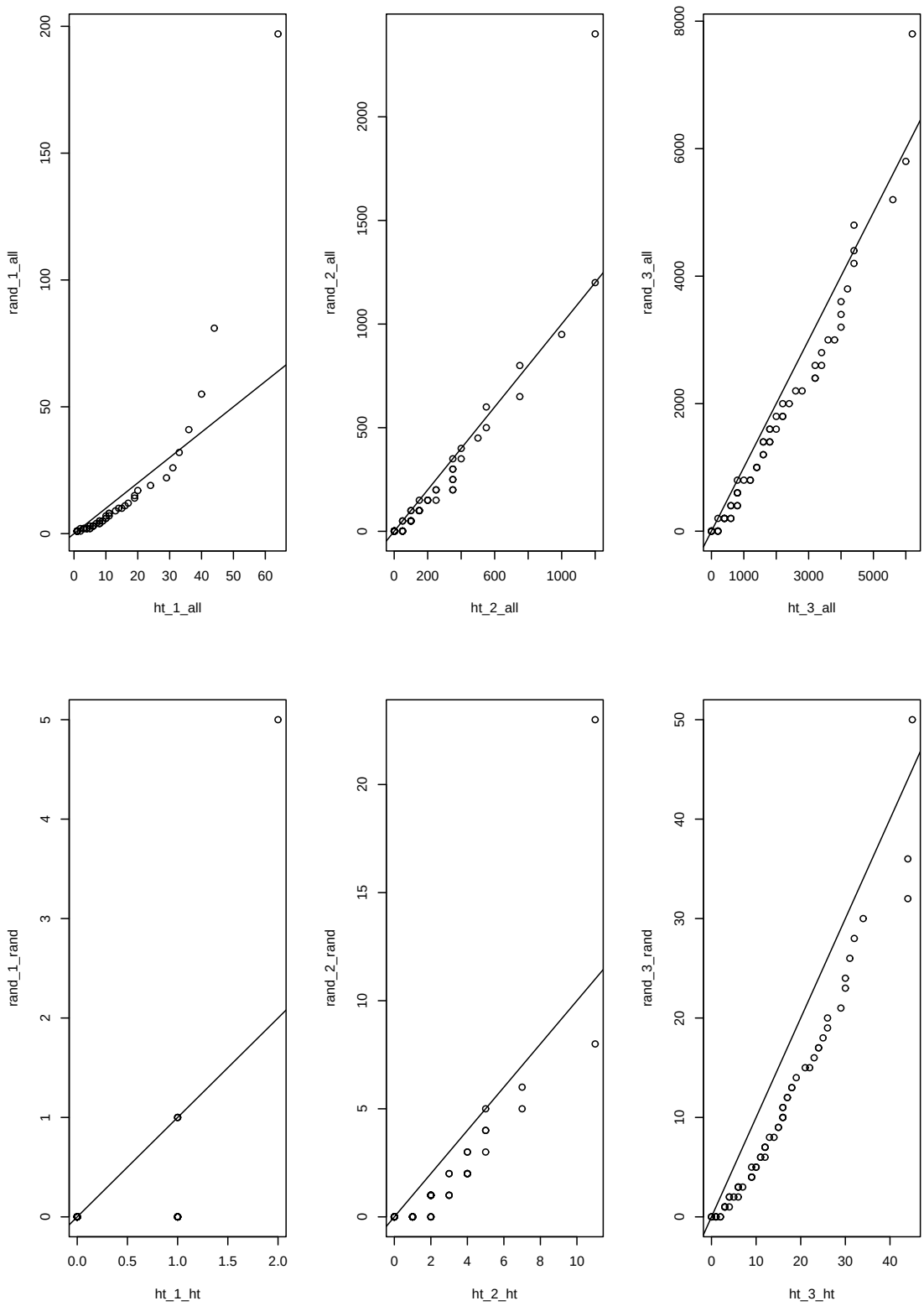
Degree of nodes: The average degree (number of direct interactions associated with a protein) for the whole of OPHID was 9.04. The HTd proteins had an average degree of 10.0615. The average degree for OMIM genes (that are present in OPHID) was 12.91. The number of proteins within radii of 1 (degree), 2 and 3 interactions from each protein is shown in the top row of quantile-quantile plots in Figure 4.2. The difference in distributions between HTd and Rd1..1000 was only marginally significant for direct interactions (degree) and was not significant for interactions within radii of 2 and 3 interactions when using the Wilcoxon rank sum test (p-values = 0.03, 0.09, 0.08 respectively), although there were outiers in the Rd1..1000 proteins acting as hubs.

Clustering coefficient: Figure 4.3 shows the quantile-quantile plot of clustering coefficients (C) for HTd and Rd1..1000. If they come from similar distributions, the distributions should align. Wilcoxon rank sum test with continuity correction shows that they come from the same distribution (p-value = 0.1085). However the Bartlett's K-squared test shows there is heterogeneity of variance (p-value = 0.001368) with the random genes having a wider variance of C. In terms of interacting partners that are involved in no further interactions (C=0), there was no significant difference between the two sets; 52.3% HTd proteins and 39% of Rd1..1000 proteins (Chi-squared = 1.857 p-value = 0.1730). There was no significant difference in the proportion of HTd and Rd1..1000 proteins that only have a single interacting partner with 18 HTd proteins and an average of 18.86 across the Rd1..1000 datasets.

4.2.1.2 Dataset topology

Degree of nodes: The second row of quantile-quantile plots in Figure 4.2 show the subset of interactions within radii of 1 (degree), 2, 3 interactions that belonged to the same

Figure 4.2: Quantile-quantile plots for the number of proteins up to a distance of 3 interactions away from HTd and Rd1..1000 proteins. The top row plots relate to all interactions and the second row plots limit to interactions with proteins belonging to the same dataset as the protein being studied.



Axis definitions: rand_1_all, rand_2_all, rand_3_all - number of proteins within radii of 1, 2, 3 interactions of each Rd*x* protein (*x*=1 to 1000); ht_1_all, ht_2_all, ht_3_all - number of proteins within radii of 1, 2, 3 interactions of each HTd protein; rand_1_rand, rand_2_rand, rand_3_rand - number of Rd*x* proteins within radii of 1, 2, 3 interactions of each Rd*x* protein; ht_1_ht, ht_2_ht, ht_3_ht - number of HTd proteins within radii of 1, 2, 3 interactions of each HTd protein.

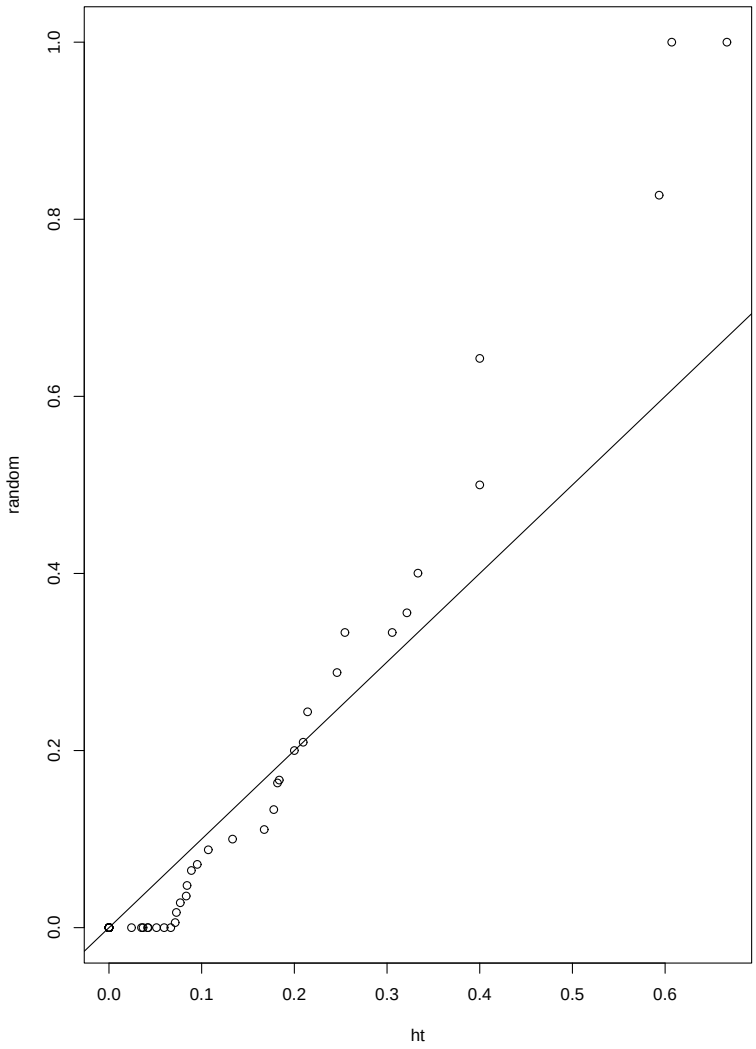


Figure 4.3: Quantile-quantile plot of clustering coefficients (C) for the HTd and Rd1..1000 proteins. Wilcoxon rank sum test with continuity correction shows that they come from the same distribution (p-value = 0.1085). ht=hypertension

dataset as the protein p under study. These plots can be compared with the first row plots displaying all interactions within similar radii. The difference in distributions between HTd and Rd1..1000 for these subsets of interactions up to a radius of 3 interactions is significant when using the Wilcoxon rank sum test (p-value = 2.49e-11, 3.842e-06, 0.0003 respectively). meaning there are larger numbers of HTd proteins surrounding any given HTd protein than there are Rd*x* proteins surrounding Rd*x* proteins (within the radii up to 3 interactions).

Geodesic distance: Figure 4.4 shows the geodesic distance between each pair of HTd proteins and each pair of proteins from Rd1..100. We limited to the first 100 random datasets due to the computationally expensive process involved in calculating the distance. The difference in the distribution of distances was significant (Wilcoxon rank sum p=0.004). Fifteen out of 65 (23%) HTd proteins are directly connected. In comparison, on average, only 3 out of every 65 (6%) Rd1..100 proteins are directly connected.

Interaction subnetworks: There were 623 proteins (646 interactions) in the dataset comprising the HTd proteins and their direct interaction partners. The average number of proteins and directly interacting partners for the Rd1..1000 datasets was 583 (std 109). The largest connected component in the *expanded* subnetwork involving the HTd proteins and their direct partners contained 550 of the 623 proteins (88%). The size of this subnet is in the upper 5% of the distribution over Rd1..1000 (Figure 4.5).

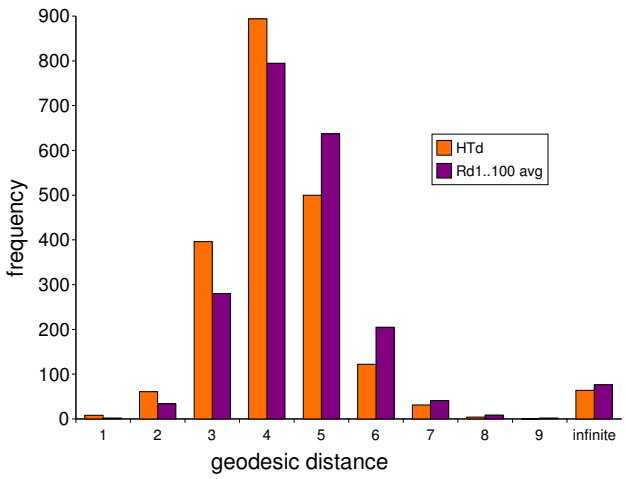


Figure 4.4: Illustration showing the geodesic distances between HTd protein pairs and Rd*x* protein pairs. Infinite relates to protein pairs that are unconnected, both directly and indirectly.

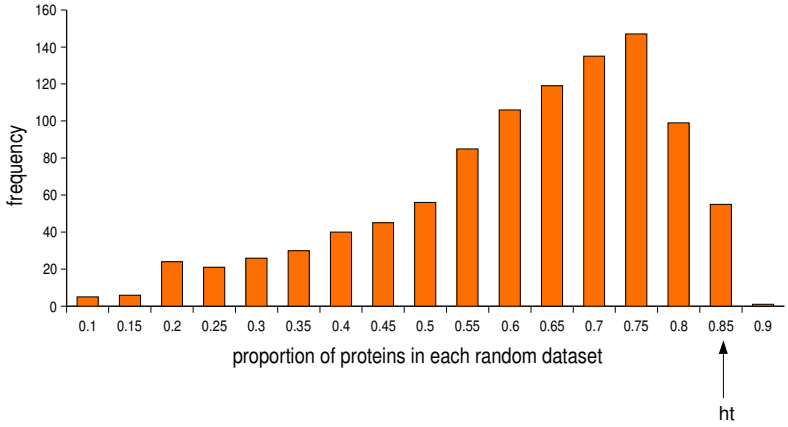


Figure 4.5: The proportion of proteins in the largest connected component for HTd and each Rd1..1000 *expanded* subnetworks. In the HTd *expanded* subnetwork, the largest connected component contains 88% of the proteins. ht=hypertension

4.2.2 Hypertension pathways and protein function

The HTd proteins are spread across 36 KEGG pathways. Three (8%) of these pathways contain 3 HTd proteins, 10 (28%) contain 2 HTd proteins and the remaining pathways (64%) contain single HTd proteins. Table 4.1 shows the pathways that contain multiple HTd proteins. By comparison, for the subset of 22 Rd1..1000 datasets that map to the same number of pathways (36), only 3% of the pathways contain 3 proteins, 15% contain 2 proteins and 82% contain 1 protein. The clustering of HTd proteins in KEGG pathways is significantly different to the pattern observed in the subset of Rd1..1000 datasets that map to 36 pathways (Wilcoxon rank sum test p=0.02).

It was important to investigate the origin of the observed high level of connectivity in 'dataset topological' properties of the HTd dataset. HTd proteins that clustered in pathways were investigated to see whether they originated from the same OMIM record. For those that did we noted the geodesic distance separating them. This might help identify any potential biases in the HTd dataset. Of the 3 pathways that each contain 3 HTd proteins, 2 pathways contain HTd proteins that map to the same hypertension related OMIM id. The first of these pathways is the human cell communication pathway

Pathway ID	Description	No. of HTd proteins
path:dhsa00500	Starch and sucrose metabolism	3
path:dhsa01430	Cell Communication	3
path:dhsa04610	Complement and coagulation cascades	3
path:dhsa00052	Galactose metabolism	2
path:dhsa00140	C21-Steroid hormone metabolism	2
path:dhsa00561	Glycerolipid metabolism	2
path:dhsa00600	Sphingolipid metabolism	2
path:dhsa03320	PPAR signaling pathway	2
path:dhsa04350	TGF-beta signaling pathway	2
path:dhsa04630	Jak-STAT signaling pathway	2
path:dhsa04640	Hematopoietic cell lineage	2
path:dhsa04742	Taste transduction	2
path:dhsa05216	Thyroid cancer	2

Table 4.1: The KEGG Homo sapiens pathways containing multiple HTd proteins

(path:dhsa01430). An OMIM id (215600 Cirrhosis, familial) is shared between 2 of the 3 HTd proteins in this pathway. The respective proteins are: K1C18_HUMAN [P05783] (Keratin, type I cytoskeletal 18 (Cytokeratin-18) and K2C8_HUMAN [P05787] (Keratin, type II cytoskeletal 8 (Cytokeratin-8). These proteins are separated by a geodesic distance of 4. The second pathway containing 3 HTd proteins is the complement and coagulation cascades pathway (path:dhsa04610). Again, one OMIM id (235400 hemolytic uremic syndrome) is shared between 2 of the 3 HTd proteins in this pathway. The proteins are: CFAH_HUMAN [P08603] (Complement factor H precursor (H factor 1)) and MCP_HUMAN [P15529] (Membrane cofactor protein precursor (Trophoblast leukocyte common antigen)). The geodesic distance between these proteins is 2. Only 1 of the 10 pathways that contain 2 HTd proteins have proteins that map to the same hypertension related OMIM id. This pathway is the taste transduction pathway (path:dhsa04742). The shared OMIM id is 177200 (Liddle syndrome). The 2 proteins in this pathway that share this OMIM id are: SCNNB_HUMAN [P51168] (Amiloride-sensitive sodium channel subunit beta (Epithelial Na(+) channel subunit beta)) and SCNNG_HUMAN [P51170] (Amiloride-sensitive sodium channel subunit gamma (Epithelial Na(+) channel subunit gamma)). These proteins directly interact in the PPI network.

There was not a strong correlation between GO semantic similarity and geodesic distance for HTd protein pairs. Correlations were calculated for each aspect of GO (molec-

ular function, biological process and cellular component).

Most of the HTd proteins fall into GO slim categories binding (GO:0005488), protein binding (GO:0005515) and catalytic activity (GO:0003824). The difference in the overall distribution of GO slim biological process categories between hypertension and Rd1..1000 proteins is significant (p-value = 0.01554) whereas the distribution of molecular function GO slim categories is not (p-value = 0.5369). In terms of biological processes, specific GO slim categories 'response to stimulus' (GO:0050896) and 'electron transport' (GO:0006118) are overrepresented within the hypertension dataset with p = 0.005277 and p = 0.0009852 repectively. In terms of molecular functions, 'oxidoreductase activity' (GO:0016491) is overrepresented within the hypertension dataset (p-value = 0.01219). These categories are still significantly overrepresented following the removal of 3 homologs in HTd.

4.2.3 Classification

The CfsSubsetEval evaluator used with the BestFirst search method identified seven key attributes: percentage amino acid composition of G; percentage amino acid composition of K; the geodesic distance to the closest HTd protein; the standard deviation of the geodesic distances to each HTd protein; whether the protein belonged to GO slim categories 'response to stimulus' (GO:0050896) and 'electron transport' (GO:0006118); the number of direct connections with HTd proteins. The BestFirst approach searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility (Witten & Frank, 1999). The CfsSubsetEval evaluator calculates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred (Hall, 1998). The Bagged PART classifier was run 400 times over a range of penalties (using a cost matrix) for incorrectly predicting a HTd protein using the 7 key attributes. The runs were repeated using the simple majority-rule approach but the TPR never exceeded the FPR. Figure 4.6 shows the true positive rate (TPR) plotted against the false postive rate (FPR) when predicting hypertension proteins.

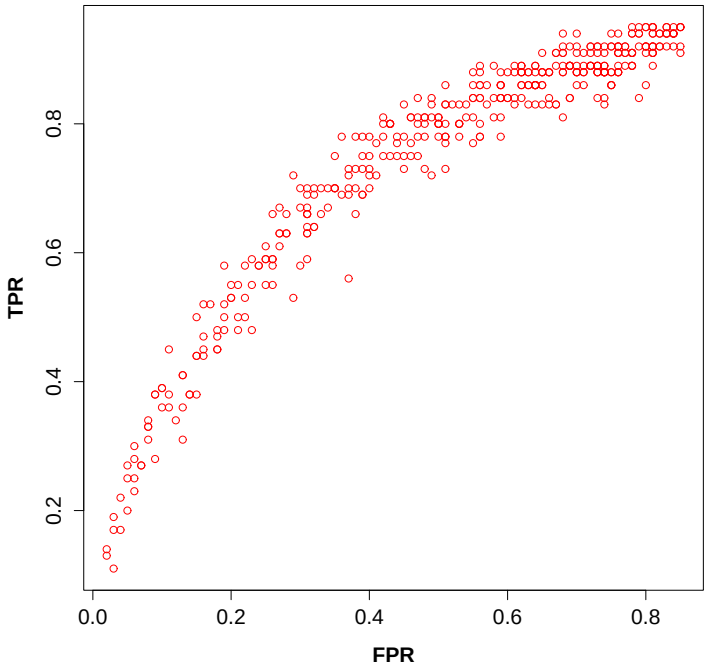


Figure 4.6: Illustration showing the true positive rate [TPR] against false positive rate [FPR] when predicting hypertension proteins using a weighted Bagged PART classifer. The penalty for an incorrect prediction was varied by using a CostSensitive classifier

BLASTClust (at a level of 25% identity) (Dondoshansky, 2002) showed that the HTd dataset was not heavily populated with sequence homologs. Only 2 pairs of proteins were found to share more than 25% identity. The first protein pair was: SCNNB_HUMAN [P51168] (Amiloride sensitive sodium channel subunit beta) and SCNNG_HUMAN [P51170] (Amiloride sensitive sodium channel subunit gamma). These proteins shared 34% sequence identity (E=3e-102). The second protein pair was: C11B1_HUMAN [P15538] (Cytochrome P450 11B1, mitochondrial precursor) and C11B2_HUMAN [P19099] (Cytochrome P450 11B2, mitochondrial precursor). These proteins shared 85% sequence identity (E=0.0). All proteins were included in the machine learning classification.

4.3 Discussion

This study found there to be little difference in the general background topological properties of HTd and Rd1..1000 proteins in protein-protein interaction networks. Hypertension related proteins do not form large hubs and they do not display high cluster coefficient

(C) scores. Previous studies including Rual *et al.* (2005); Stelzl *et al.* (2005); Jonsson & Bates (2006b); Xu & Li (2006) have suggested that disease genes were likely to form hubs. However, Goh *et al.* (2007) recently suggested that these studies included 'essential' genes in which any mutations are lethal. Once these genes had been excluded it was shown that the remaining 'non-essential' disease genes did not tend to form hubs. HTd are likely to be 'non-essential' genes and our findings are consistent with Goh *et al.* (2007). OMIM has an average degree of 13 which is higher than the hypertension proteins (10) and OPHID (9), possibly because OMIM includes these 'essential' disease genes.

Despite the insignificant differences in background network topology, we find that HTd proteins display greater connectivity in relation to each other than we might expect. HTd protein pairs exhibit shorter geodesic distances than random and the largest *expanded* subnet size lies within the top 5% of the distribution for the random datasets. This means that 88% of the proteins are connected (directly or indirectly) when a network is created using HTd proteins and their direct partners. It is similar to previously observed distributions in Alzheimers disease proteins where the largest subnet contained 83% of the proteins (Chen *et al.*, 2006). There is also a significant difference from random in the number of HTd proteins within a radius of 3 interactions from any other HTd protein.

The HTd proteins are spread over 36 KEGG pathways, reflecting the complex, locus rich nature of hypertension related proteins. We might have expected to see HTd proteins that cluster in the same pathway to have originated from the same OMIM id and be close in the PPI network. We found that this was not always the case. The proteins were usually associated with different diseases where hypertension was a symptom. Where proteins shared a pathway, originating from the same OMIM id, only 1 of the 3 HTd protein pairs were directly connected.

We expected to see a negative correlation between distance separating two HTd proteins in the PPI network and GO semantic similarity. However, we were unable to show this correlation in our dataset. The difference in the distribution of GO slim biological process categories between HTd proteins and the Rd1.1000 was significant. There were a number of notable molecular function and biological process categories that were over-represented in the hypertension dataset, namely 'response to stimulus' (GO:0050896),

'oxidoreductase activity' (GO:0016491), 'nucleic acid binding' (GO:0003676).

There are caveats with the OMIM hypertension dataset, the OMIM database is the most complete repository of diseases and their associated genes but of course it is not complete and is updated all the time. There was concern that the increased connectivity of the HTd proteins may be due to biases in the PPI resource. We might expect the hypertension related proteins to have been studied more than the randomly selected proteins and therefore to see a larger number of documented interactions. However, if this were the case, we would have expected more of them to be hubs. Potential interaction biases could be further investigated by considering interactions, such as those from high throughput experiments, separately. Because the sources of OPHID interactions vary in their reliability, we created a second weighted network, retaining the same proteins and interactions but assigning a weight (or distance) to each interaction in a similar manner to Chen *et al.* (2006). In this weighted network, proteins were separated by a distance relating to annotation confidence. Interactions with high quality annotation retained their default distance of 1, medium quality interactions were separated by a distance of 1.5 and low quality interactions a distance of 2. We then repeated the relevant analyses. Our results did not show significant trend differences to the unweighted analyses with respect to GO semantic similarity and geodesic distance correlations.

The methods described here could easily be applied to other disease datasets in OMIM. The hypertension dataset itself would be improved with the addition of validated hypertension related proteins. However, the model constructed shows that there are patterns within PPI networks, shared function and sequence based properties that can be used to aid prioritisation of candidate gene lists identified through experiments such as genome wide association studies. We anticipate that machine learning analyses that combine such attributes will be useful in helping to characterise disease related genes in future studies.

Chapter 5

Protein interaction networks associated with cardiovascular disease and cancer: shared network properties

The work in this chapter compares the protein-protein interaction network properties of two major diseases, cardiovascular disease (cvd) and cancer. For both diseases there are large curated datasets available. The study focuses on two sets of network descriptors, namely network centralities and network clusters. Centrality measures can be used to identify influential nodes in a graph and clustering analysis describes the organisation of the network on a number of levels and can be used to define functional modules and pathways in biological networks. Three measures of centrality were considered: degree centrality which simply counts the number of edges connected to a vertex; closeness centrality which considers communication to all other nodes by making use of the length of shortest paths to all nodes from a given node and betweenness centrality which 'measures' how much a node is involved in communication within the network, by identifying the number of shortest paths between pairs of nodes that pass through such a node. This measure identifies 'bottlenecks' within the network. The degree centrality only captures the local neighbourhood topology of a network and hence the influence of direct neighbouring proteins, whereas betweenness also captures the indirect influences of proteins distal to the subject protein. Betweenness is therefore a measure of importance within the wider context of the network. High betweenness and low degree has previously been

used to define the 'modularity' of various networks (Girvan & Newman, 2002; Guimerà & Amaral, 2004; Joy *et al.*, 2005). In addition to centrality, the following clustering properties were explored (i) whether proteins involved in cvd and cancer tended to be part of complex or simple processes, (ii) whether these processes are small or large, (iii) how the disease proteins are distributed across the processes and (iv) how many of the disease proteins are bridges between communities and therefore acting as interfaces between biological processes. A combination of centrality measures and clustering were used to describe the interactome topology of these diseases and demonstrate an approach that could be used to aid prioritisation of candidate genes.

5.1 PPI analysis methods

5.1.1 Dataset

Proteins thought to be implicated in cardiovascular disease were taken from the Vascular Disease 50k SNP Array Consortia chip (`http://bioinf.itmat.upenn.edu/cvdsnp/query.php`) (Keating *et al.*, 2008). Proteins on this chip were carefully selected as potential candidates for cardiovascular disease using information from quantitative trait loci studies, consideration of pathways important to vascular disease and the biomedical literature. The proteins were split into three categories: priority 1 proteins included significant known mediators of vascular disease and key findings from whole genome association studies (602 proteins). The two other categories included more speculative assignments (2015 priority 2 and 494 priority 3 proteins). We mapped these proteins onto a set of human PPI from the PIP webserver created by Jonsson & Bates (2006a) using an orthology based approach applied to the RefSeq dataset (Pruitt *et al.*, 2007). The cancer dataset came from a census conducted from the literature of genes that are mutated and causally implicated in cancer development ('cancer genes') (Futreal *et al.*, 2004). Futreal *et al.* (2004) had mapped these cancer proteins onto their PPI dataset.

5.1.2 Measures

The PPI network was considered as an undirected graph $G = (V, E), v \in V, e \in E$ where the proteins are nodes (v) and the interactions are edges (e), with edge $e_{m,n}$ connecting nodes m and n. We considered the following three measures of centrality using the Python package NetworkX (`https://networkx.lanl.gov`):

Degree centrality: the number of edges connected to a vertex were counted and normalised by dividing by the total number of possible interactions that could be made, that is the number of nodes minus one.

Closeness centrality: This could only be calculated for nodes belonging to the same connected component, the shortest path length for unconnected nodes is infinity. If $dist(m, n)$ is the length of the shortest path from m to n then the closeness centrality was defined as:

$$C_{close}(m) = \frac{1}{\frac{1}{(|V|-1)} \sum_{n \in V} dist(m, n)} \tag{5.1}$$

Shortest Path Betweenness centrality: In the following, let σ_{mn} denote the number of shortest paths between vertices m to n and $\sigma_{mn}(\nu)$ be the number of shortest paths from m to n that pass through ν (Junker & Schreiber, 2008). The rate of communication between m and n involving ν is given by $\delta_{mn}(\nu) = \sigma_{mn}(\nu)/\sigma_{mn}$. The shortest path betweenness centrality can then be defined as:

$$C_{betweenness}(\nu) = \sum_{m \in V \wedge m \neq \nu} \sum_{n \in V \wedge n \neq \nu} \delta_{mn}(\nu) \tag{5.2}$$

The average degree centrality was calculated for nodes associated with cancer and cvd to examine the hub like properties of disease associated proteins. The nodes of the human interactome were then ranked according to the three measures of centrality and identified putatively functionally important nodes in the interactome.

Following Jonsson & Bates (2006a), but also using the much larger cvd dataset, network clustering was investigated through the community structure with the *Cfinder* algorithm (Adamcsek *et al.*, 2006). This program uses the *k*-clique clustering method which defines communities in terms of overlapping cliques. A *k*-clique is a set of *k* nodes where there is an interaction between each pair of nodes. *Cfinder* identifies communities as the

union of *k*-cliques in which *k*-1 nodes are shared. Communities were identified at various *k*-values and the proportion of the member proteins that were cvd or cancer related. Generally at low values of *k* we would expect a large number of extensive communities of less tightly connected proteins with a large overlap. For higher *k* values fewer, more distinct communities appear. Analysis of the community structure identifies bridge nodes as nodes belonging to more than one community. These may participate in multiple processes and act as interfaces between processes.

5.2 Cardiovascular disease and cancer PPI analysis results

In total, there are 17,039 RefSeq protein IDs (108,113 interactions) in the PPI dataset (Jonsson & Bates, 2006a). We were able to map 2,249 cvd implicated protein IDs to this dataset, 19% being cvd priority 1 proteins, 63% cvd priority 2 and 18% cvd priority 3 (Keating *et al.*, 2008). Within the network, 439 protein IDs were annotated as being cancer from the cancer census (Futreal *et al.*, 2004). The number of cvd proteins mapped to the PPI network was therefore approximately 5 fold greater than the number of mapped cancer proteins. Of the cancer RefSeq proteins IDs mapped to the PPI dataset, 120 (27%) were also proteins implicated in cvd. Of these proteins, 28 were cvd priority 1 proteins.

5.2.1 Centrality

The betweenness, closeness and degree centrality measures were calculated for each protein within the PPI network. The average number of interactions (degree) for cvd priority 1 proteins was 19.6, cvd overall was 15.7 and cancer proteins was 22.6 (table 5.1). There is a notable difference between the average degree of cvd priority 1 proteins (19.6) and cvd priority 2/3 proteins (6.0). The distribution of each of the normalised centrality measures are shown in Figure 5.1. The distribution of betweenness and degree centrality values differ significantly when comparing each of the disease datasets against the non disease dataset using the Wilcoxon rank sum test. In general, the degree and betweenness values follow an exponential distribution and the closeness centrality conforms to a normal dis-

tribution. Using the probability density function we were able to calculate the probability for each centrality score. The proteins displaying 'high centrality' ($p<0.05$), particularly degree and betweenness centrality, are enriched approximately 2 fold with cancer and cvd proteins. Specifically, 24%, 26% and 18% of cvd proteins displayed high betweenness, degree and closeness centrality and 6%, 7% and 6% of cancer proteins displayed high betweenness, degree and closeness centrality. This compares to total presence in the interactome of 13% for cvd and 3% for cancer. Table 5.2 shows the 'top 10' proteins for each centrality measure, it shows that 29% of these are currently annotated as cvd proteins and 17% are already associated with OMIM morbidity accession numbers (Hamosh *et al.*, 2002).

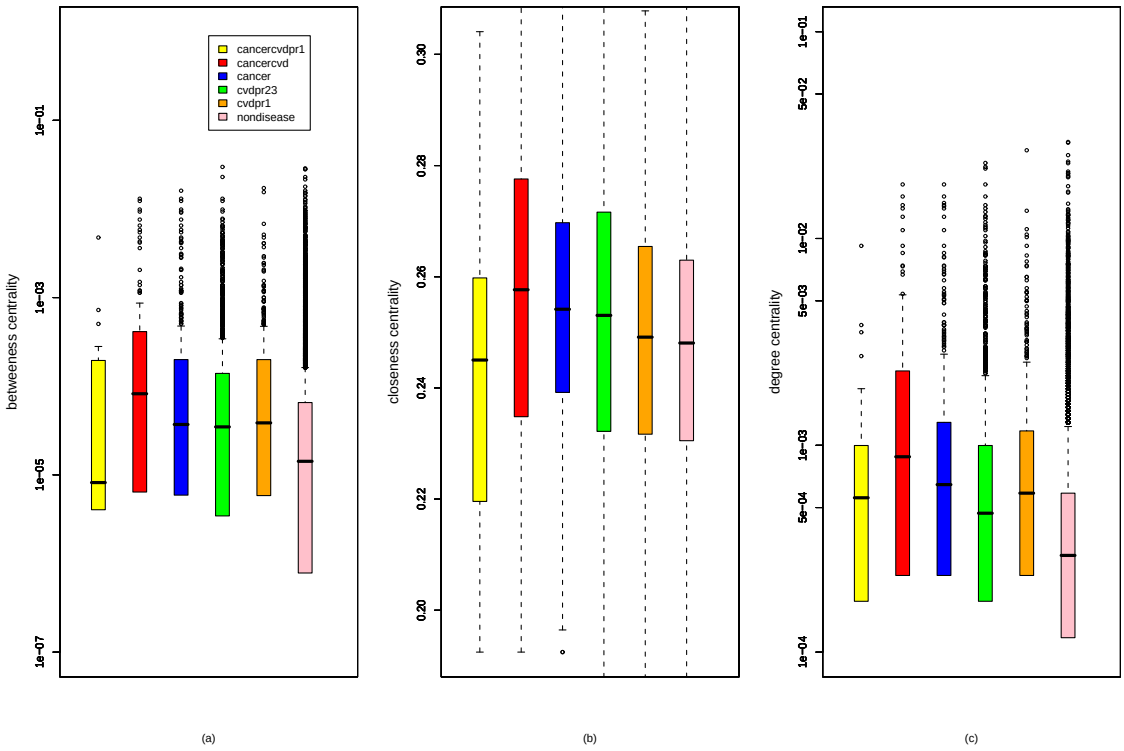


Figure 5.1: The (a) betweenness, (b) closeness and (c) degree centrality distributions for each studied subset of proteins. Cancercvdpr1 = proteins annotated as both cancer and cvd priority 1, cancercvd=proteins annotated as both cancer and cvd.

	cvd	non cvd	cvdpr1	non cvdpr1	cancer	non cancer
Average degree	15.7	6.4	19.6	6.0	22.6	9.9

Table 5.1: Connectivity of proteins: Average degree of cardiovascular (cvd), cvd priority 1 (cvdpr1) and cancer proteins.

The high degree observed by cvd and cancer proteins may partly be explained by the

Description (RefSeq Peptide ID)	cancer	cvd	cvd pr1	bc	dc	cc	OMIM Morbidity description
Filamin-A (Endothelial actin-binding protein) (NP_001447)		y		y	y		
Actin, cytoplasmic 1 (Beta-actin)(NP_001092)				y	y		Dystonia, juvenile-onset
Alpha-actinin-2 (NP_001094)				y	y		
60 kDa heat shock protein, mitochondrial precursor (Heat shock protein 60) (NP_002147)				y	y		Spastic paraplegia 13, autosomal dominant
Calmodulin (NP_001734)		y		y	y		
Cell division protein kinase 3 (NP_001249)				y	y		
Importin subunit alpha-7 (Karyopherin subunit alpha-6) (NP_036448)				y			
Transportin-1 (Importin beta-2) (Karyopherin beta-2) (NP_002261)				y			
Heat shock 70 kDa protein 1 (NP_005337)		y	y	y			
UDP-N-acetylglucosamine–peptide N-acetylglucosaminyltransferase 110 kDa subunit (NP_858058)				y			
Guanine nucleotide-binding protein G(k) subunit alpha (G(i) alpha-3) (NP_006487)					y		
Guanine nucleotide-binding protein subunit alpha-11 (G alpha-11) (NP_002058)					y		
Guanine nucleotide-binding protein G(i), alpha-2 subunit (NP_002061)		y	y		y		
Guanine nucleotide-binding protein G(q) subunit alpha (NP_002063)		y			y		
Glucose-fructose oxidoreductase domain-containing protein 2 precursor (NP_110446)						y	
Transcriptional enhancer factor TEF-3 (NP_958849)						y	
Glycoprotein hormones alpha chain precursor (NP_000726)						y	
Sphingomyelin phosphodiesterase 2 (NP_003071)		y				y	
Molybdenum cofactor biosynthesis protein 1 A(NP_620306)						y	
39S ribosomal protein L13, mitochondrial (NP_054797)						y	
Cytochrome c oxidase subunit 2 (NP_536846)						y	Mitochondrial complex IV deficiency
Bardet-Biedl syndrome 5 protein (NP_689597)		y				y	Bardet-Biedl syndrome
Interleukin-17 receptor A precursor (NP_055154)						y	
Uridine phosphorylase 1 (NP_003355)						y	

Table 5.2: The 'top ten' most influential interactome proteins for each of the centrality measures. The 'y' denotes possession of the quality defined by the column. Cancer=cancer protein, cvd=cardiovascular protein, cvdpr1=cardiovascular priority 1 protein. Of these proteins, 29% are currently annotated as cvd related and 17% are already associated with OMIM database morbidity identifiers (Hamosh *et al.*, 2002). The OMIM morbidity description relates to a disease which the protein is associated with. bc=betweeness centrality, dc=degree centrality, cc=closeness centrality.

promiscuous nature of their domains (Table 5.3). In order to calculate the general promiscuity of each domain we firstly extracted domain-domain interaction frequencies from PFAM (Bateman *et al.*, 2004). The promiscuity p values shown in the table were calculated based on this analysis of interaction frequencies of the PFAM domains, which conform to a probability density function as described by Jonsson & Bates (2006a). Sixteen of the top 20 most frequent occurring cvd domains are promiscuous and 7 are among the top 30 most frequent occurring cancer domains. Table 5.3 shows that domain promiscuity can generally be seen to increase with increasing cvd domain frequency. Two domains namely, zinc finger, C4 type PF00105 and ligand-binding domain of nuclear hormone receptor PF00104 are not promiscuous domains but are common to both disease conditions. Both families are steroid or nuclear hormone receptors implicated in DNA-dependent transcription regulation.

Most frequently occurring cvd domains (descending order)	PFAM id	Promiscuity (p)	In top 30 cancer domains	Promiscuous ($p<0.005$)
Protein kinase domain	PF00069	3.70E-013	♠	⋆
7 transmembrane receptor (rhodopsin family)	PF00001	4.67E-003		⋆
Immunoglobulin domain	PF00047	7.93E-011	♠	⋆
Protein tyrosine kinase	PF07714	4.67E-003		⋆
EGF-like domain	PF00008	1.33E-011	♠	⋆
Immunoglobulin V-set domain	PF07686	2.84E-009		⋆
Pleckstrin homology domain	PF00169	2.18E-005		⋆
Zinc finger, C4 type (two domains)	PF00105	1.67E-001	♠	
Ligand-binding domain of nuclear hormone receptor	PF00104	2.79E-002	♠	
Immunoglobulin I-set domain	PF07679	1.30E-004		⋆
Fibronectin type III domain	PF00041	3.64E-006	♠	⋆
SH2 domain	PF00017	3.64E-006		⋆
Leucine Rich Repeat	PF00560	1.33E-011		⋆
EGF-like domain	PF07974	7.80E-004		⋆
Trypsin	PF00089	0.00E+000		⋆
Collagen triple helix repeat (20 copies)	PF01391	1.67E-001		
ABC transporter	PF00005	7.80E-004		⋆
SH3-domain	PF00018	6.09E-007	♠	⋆
Calcium binding EGF domain	PF07645	3.64E-006		⋆
Variant SH3 domain	PF07653	1.00E+000		

Table 5.3: Promiscuity of the top 20 most frequently occurring cvd domains (in descending order). A ♠ is present for domains that are frequently present in cancer proteins. Domains marked with a ⋆ are thought to be promiscuous domains.

Joy *et al.* (2005) observed an abundance of proteins in the yeast interactome displaying high betweenness and low connectivity (degree) (HBLC). Such a feature is not found in randomly generated scale free networks. They suggested this was due to some modular organization of the network. We were able to show HBLC is a feature of the studied human PPI network (Figure 5.2) and note disease proteins seem to be evenly distributed in this figure. This measure of modularity through HBLC is observed where a link between modules is composed of 2 or more steps, the intermediate proteins will display low degree. Using a cutoff of $p<0.05$ for high betweenness and low degree there are no proteins that display 'extreme' HBLC. Identification of community bridges in the clustering analysis provides additional evidence of modularity within the network. Such community bridges whilst having high betweenness, may also display high degree due to the formation of interactions with proteins from multiple communities, therefore not displaying HBLC but still supporting the idea of modularity.

5.2.2 Clustering

The community structure of the network was then analysed with the *Cfinder* program. For low *k*-values there are a large number of highly overlapping communities. As *k* increases, the number of communities decreases as does the overlap between them (table 5.4). The disease proteins make up a larger proportion of the community proteins as *k*-value increases suggesting a greater presence in complex, distinct communities. There is no significant difference in the distribution of community based proteins with increasing *k*-values between the disease datasets using the Wilcoxon rank sum test.

For each *k*-value cancer proteins tend to be in the larger communities than cvd proteins which in turn are in larger communities than non-disease proteins (Figure 5.3). This suggests cvd and cancer proteins also take part in large processes. Are these proteins spread across a number of communities or do they cluster within specific communities? Figure 5.4 shows that cvd priority 1 and cancer proteins exhibit a greater tendancy to cluster within communities than cvd (all priorities) and non disease proteins.

Table 5.5 shows which proteins are present in multiple communities and therefore possibly act as interfaces between multiple processes further supporting the idea of mod-

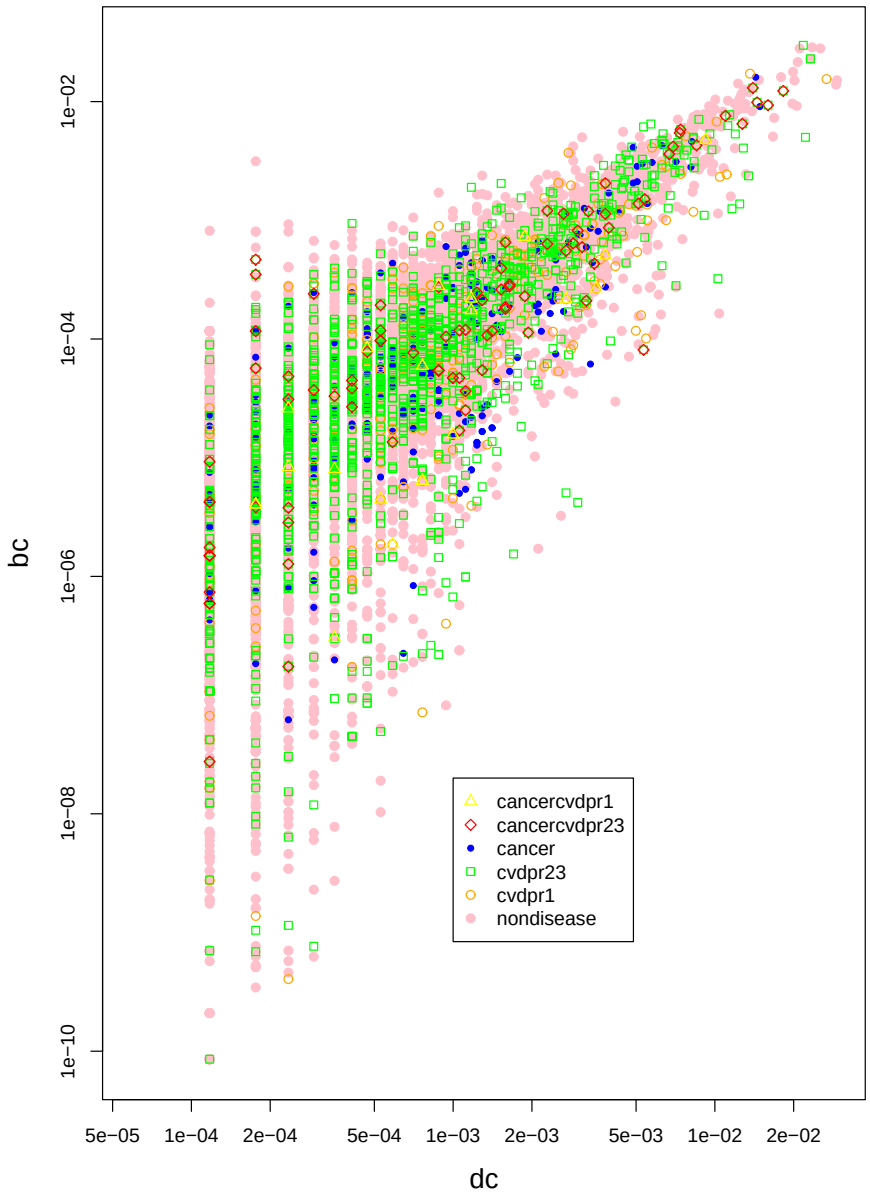


Figure 5.2: Degree (dc) versus betweenness centrality (bc) for the studied PPI network. Six subsets of proteins are shown; cancer and cvd priority 1 proteins, cancer and cvd priority 2 or 3 proteins, cancer proteins, cvd priority 2 or 3 proteins, cvd priority 1 proteins, proteins not implicated in either cvd or cancer

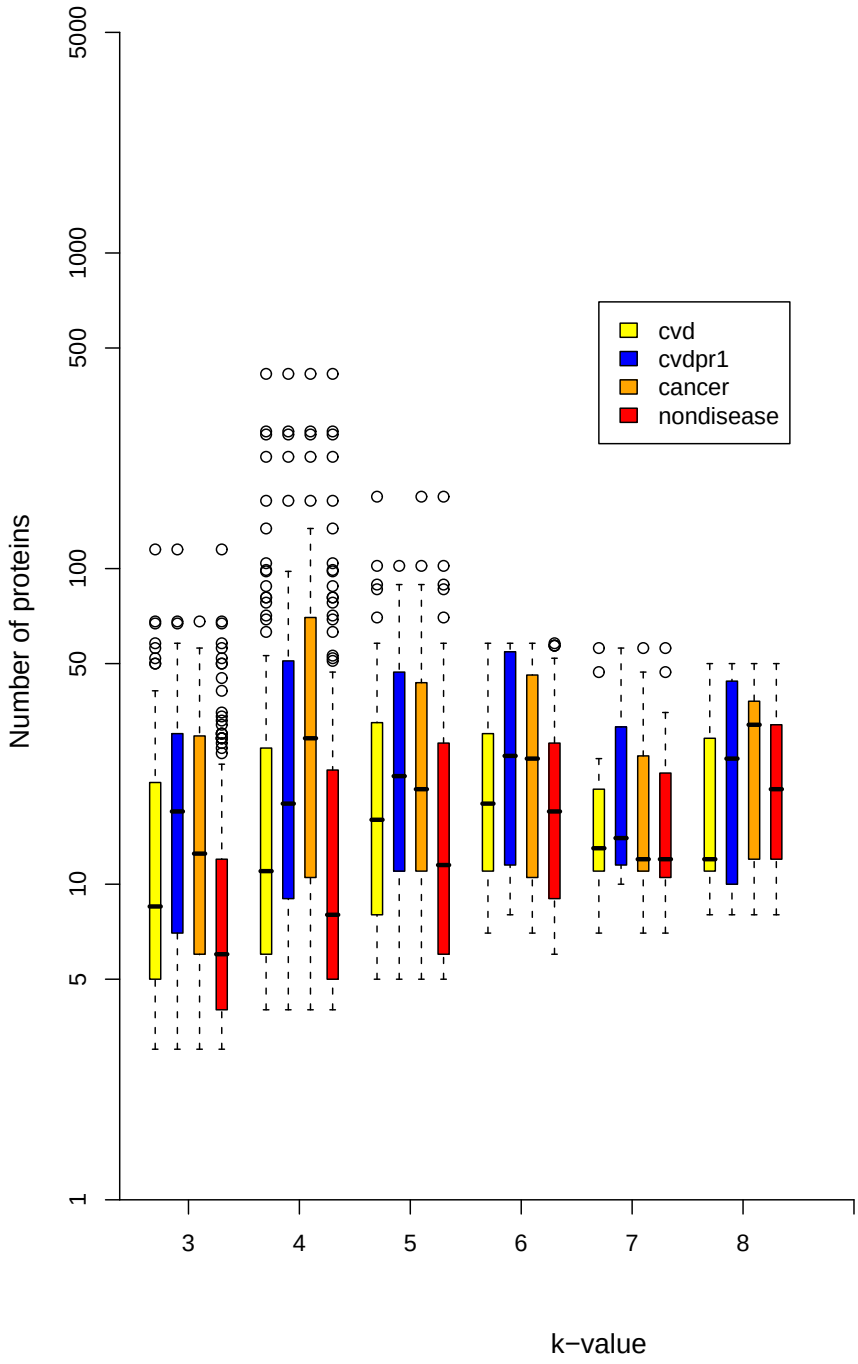


Figure 5.3: The distribution of community sizes (number of proteins) for the studied PPI network. Four types of communities are represented; those containing cvd proteins (all priorities), cvd proteins (priority 1 only), cancer proteins and communities that do not contain either cvd or cancer proteins (non disease).

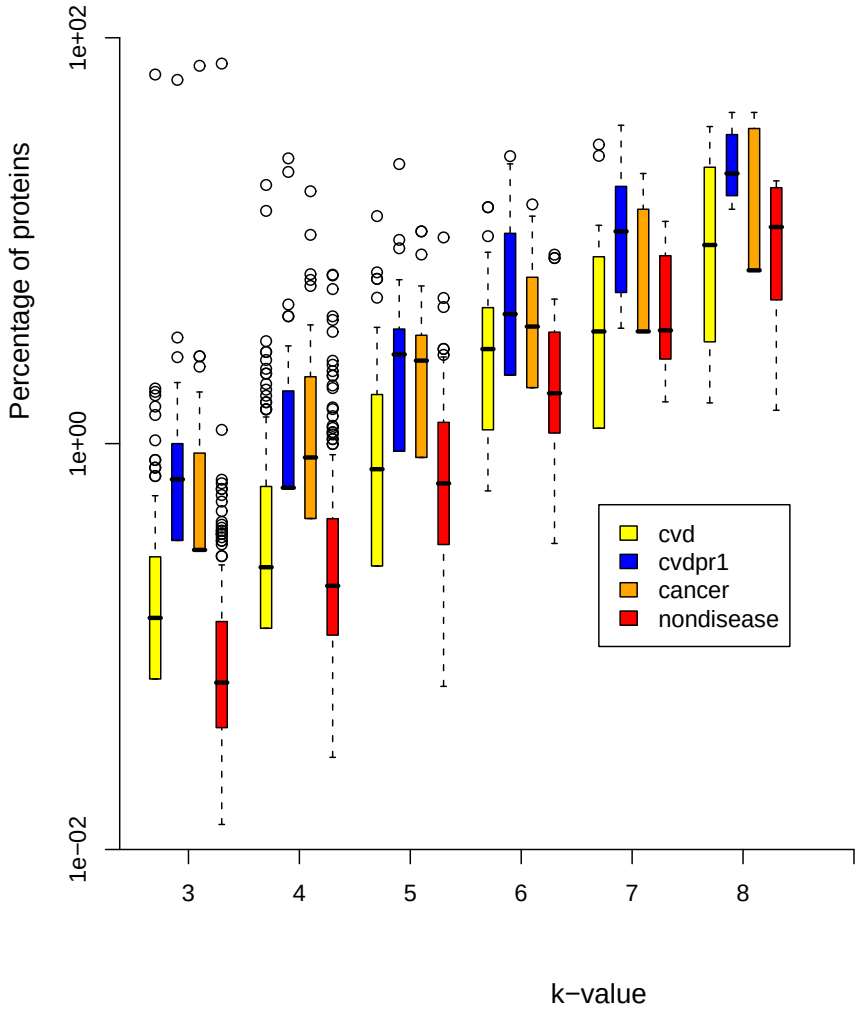


Figure 5.4: Illustration showing the percentage of proteins assigned to each community for 4 protein subsets. It shows any clustering within communities. The plots represent 4 datasets; communities containing cvd proteins (all priorities), cvd priority 1 proteins (cvdpr1), cancer proteins and communities that do not contain either cvd or cancer proteins (non disease)

k-value	Communities	% cvd proteins	% cvdpr1 proteins	% cancer proteins
3	222	15.2(0.50)	3.1(0.45)	3.5(0.45)
4	189	17.8(0.27)	3.8(0.25)	4.9(0.29)
5	98	18.4(0.13)	4.9(0.15)	5.4(0.15)
6	37	21.2(0.06)	5.5(0.07)	6.1(0.06)
7	19	22.4(0.03)	7.0(0.04)	7.0(0.04)
8	9	27.8(0.02)	8.8(0.03)	6.3(0.02)

Table 5.4: The percentage of proteins making up communities from cvd, cvd priority 1 and cancer protein datasets. The values in brackets represent the proportion of proteins for a chosen *k*-value as a fraction of all communities. Cardiovascular disease proteins make up 15.2% of proteins in *k*-value = 3 communities. This percentage accounts for 0.5 of all cvd proteins assigned to communities. There are more proteins in the larger communities (low *k*-value) but the disease proteins make up a larger proportion of the community proteins as *k*-value increases

ularity within the network. Both cvd and cancer proteins act as bridges more often than expected at each *k*-value. There is no significant difference in the distribution of proteins that are bridges between cvd, cvd priority 1 and cancer across the *k*-values using the Wilcoxon rank sum test.

k-value	cvd obs(%)	cvd exp(%)[fd]	cvdpr1 obs(%)	cvdpr1 exp(%)[fd]	cancer obs(%)	cancer exp(%)[fd]
3	12.29	7.95[1.55]	13.57	8.45[1.61]	12.67	8.46[1.50]
4	14.20	12.78[1.11]	9.66	13.16[0.73]	21.39	12.60[1.70]
5	14.97	9.33[1.61]	17.05	10.03[1.70]	12.37	10.26[1.21]
6	14.79	14.18[1.04]	21.62	13.88[1.56]	17.07	14.13[1.21]
7	13.51	6.64[2.03]	17.39	7.49[2.32]	17.39	7.49[2.32]
8	10.53	0.00[-]	16.67	1.60[10.43]	7.69	2.60[2.96]

Table 5.5: Community bridges - proteins that are present in multiple communities, acting as interfaces between processes. The percentage of proteins belonging to more than one community is shown for protein datasets cvd, cvd priority 1 and cancer (obs=observed). Expected (exp) values were based on non cvd priority 1 or non cancer proteins (fd=fold difference between observed and expected).

5.2.3 Combining centrality and clustering for novel candidate prioritisation

Combining measures relating to centrality and clustering may help identify or prioritise disease candidate genes. For example, here we describe, and show in Figure 5.5 how a set of proteins can be characterised in terms of their PPI topological environment and hence evaluated as potential cvd related proteins. We are not suggesting they are good candidates, merely showing how they can be evaluated. For this particular example we

show all proteins belonging to communities for *k*-value 8 which represent the most complex, tightly connected communities found in the interactome. The cvd proteins currently make up 27.8% of proteins in communities at this *k*-value, but this only represents 2% of all community based cvd proteins (Table 5.4). The shape of the node (protein) represents the disease status, the size is proportionate to the degree and the colour represents the betweenness centrality of the protein as shown in the colourbar. Of the 9 communities present, one is notable in that 6 of its 8 proteins (a single clique) are currently annotated as cvd proteins (Figure 5.6). The remaining 2 proteins could be of interest due to their presence in such a highly connected, cvd rich community. The first of these proteins is cell division cycle 37 protein CDC37 (NP_008996), which is thought to play a critical role in directing heat shock protein 90 (HSP90) to its target kinases. HSP90 is present in this clique, it is annotated as being both cancer and cvd related. The second protein is the TNF receptor-associated protein 1 TRAP1 (NP_057376) which is a mitochondrial HSP90 protein. A number of proteins from other *k*-value=8 communities are also of potential interest (Figure 5.5). Firstly, transportin 1 isoform 1 TNP01 (NP_002261) is currently not annotated as being cvd or cancer related, yet it interacts directly with a large number of cvd proteins and exhibits extremely high betweenness and degree centrality ($p<0.05$). It can also be seen to be bridging 2 communities and may therefore act as an interface between functional modules. This gene encodes the beta subunit of the karyopherin receptor complex which interacts with nuclear localization signals to target nuclear proteins to the nucleus. Secondly, leucine zipper transcription regulator 2 SEC16B (NP_149118). It currently has no known cvd or cancer association yet it displays very high degree and betweenness ($p<0.05$) and provides a link between 2 communities through its direct interaction with the nuclear factor of kappa light polypeptide gene NFKB2 (NP_002493), which is a cvd priority 1 and cancer associated protein. SEC16B is required for secretory cargo traffic from the endoplasmic reticulum to the golgi apparatus and for normal transitional endoplasmic reticulum (tER) organization. This protein is ubiquitous in terms of tissue specificity. Finally, protein phosphatase 2 catalytic subunit beta PPP2CB (NP_001009552) also stands out as an extremely influential interactome protein (both degree and betweenness centrality $p<0.05$), it also currently has no cvd or cancer anno-

tation. This protein is implicated in the negative control of cell growth and division, and the gene encodes the phosphatase 2A catalytic subunit. Protein phosphatase 2A is one of the four major serine/threonine phosphatases, and it is implicated in the negative control of cell growth and division. This protein maps to 8p21-p12, a region associated with a broad range of cancers (Imbert *et al.*, 1996).

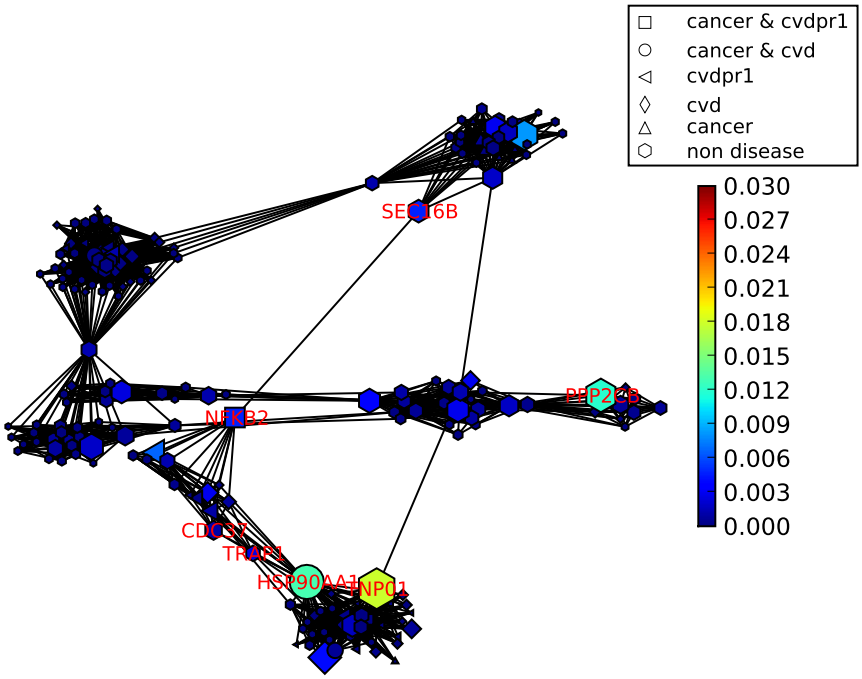


Figure 5.5: Network showing all proteins in communities for *k*-value = 8. The interactions between proteins of these communities are shown as edges linking the nodes. The node size = degree, colour = betweenness centrality and the node shape is defined by the disease status of the protein. A number of nodes are labelled with their protein identifiers.

5.3 Discussion

Network based approaches are providing important tools for systems biology. Simple graph theoretic measures such as degree and betweenness centralities are useful metrics for suggesting how influential particular proteins (nodes) in the network are, in relation to the network as a whole. We have found that both cvd and cancer proteins are over-represented within the set of proteins that are central to the interactome, particularly with respect to degree and betweenness, and that generally pleiotropic proteins tend to be most

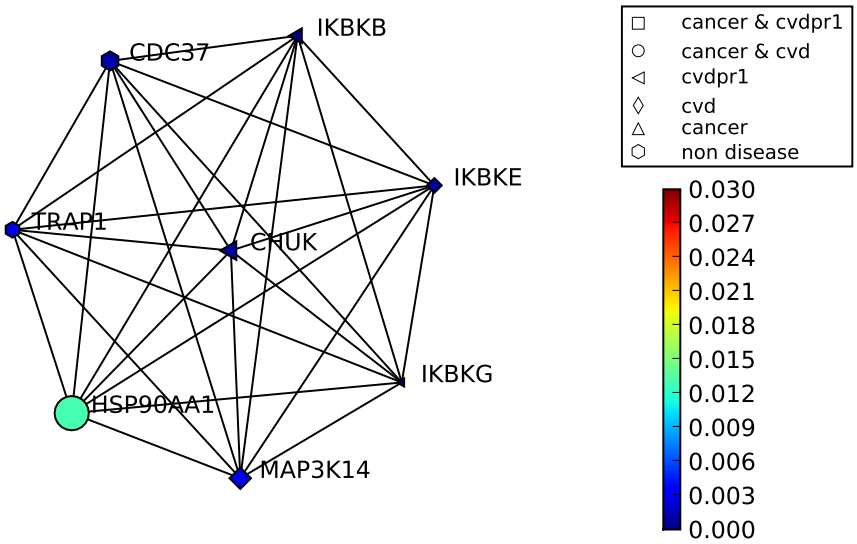


Figure 5.6: Network showing all proteins in a cvd rich *k*-value = 8 community comprising a single clique where 6 of the 8 component proteins are implicated in cvd. The node size = degree, colour = betweenness centrality and the node shape is defined by the disease status of the protein. Nodes are labelled with their protein identifiers. CHUK (NP_001269) is a conserved helix-loop-helix ubiquitous kinase, MAP3K14 (NP_003945) is mitogen-activated protein kinase, IKBKG (NP_003630) is an inhibitor of kappa light polypeptide gene, IKBKE (NP_054721) is a IKK-related kinase epsilon, IKBKB (NP_001547) is an inhibitor of kappa light polypeptide gene, CDC37 (NP_008996) is the cell division cycle 37 protein, TRAP1 (NP_057376) is the TNF receptor-associated protein 1, HSP90AA1 (NP_005339) is heat shock protein 90kDa alpha.

influential (Figure 5.1). A list of influential human interactome proteins, that is proteins with centrality scores with $p < 0.05$, is available from `http://compbio.mds.qmw.ac.uk/centralproteins.txt`. This list may be useful for the prioritisation of candidate gene lists.

Closeness centrality can only be calculated for connected proteins which leads to high closeness values for proteins belonging to small connected components. To overcome this caveat the closeness centrality could have been calculated for the single largest connected component with the unfortunate effect of reducing the number of proteins with closeness scores. This may partly explain why cvd is not as over-represented in the list of proteins exhibiting high closeness as this list contains a large number of proteins from small components.

Barabassi et al. (2007) have shown that 'essential' disease genes, in which mutations are lethal, often causing embryonic mortality, form hubs (highly connected nodes) whereas 'non-essential' disease genes do not display this tendency (Goh *et al.,* 2007). Analysis performed in chapter 4 supported this claim by showing hypertension related proteins, disruption of which would not be thought to be lethal, are generally not hub like proteins (Dobson *et al.,* 2008). In this study, we show that cvd and cancer proteins display a range of centrality scores but they are over-represented in the list of proteins displaying high degree and betweenness scores. Cardiovascular disease and cancer cover a wide range of disease phenotypes which may partly explain varying centrality scores. In addition, the cvd dataset contains many proteins which are thought to be good candidates for association with cvd but as yet many are unproven. Priority 2 and 3 proteins are more speculative suggestions than priority 1 proteins. Work is currently underway testing the cvd chip (Keating *et al.,* 2008) in a range of cardiovascular diseases including hypertension. This data will indicate whether any are actually causal in cvd.

Our results relating to network clustering add to the findings first shown in cancer by Jonsson & Bates (2006a). Importantly we are also able to show that similar properties are exhibited in a much larger dataset of cvd proteins. Cardiovascular related proteins, especially priority 1 proteins, tend to act through a small number of large, complex (tightly connected) processes and exist as interfaces between processes more often that would be

expected

We repeated the analysis on the cvd proteins in the I2D (previously OPHID) dataset (Brown & Jurisica, 2005) of PPIs in order to investigate whether the findings could be replicated. In an attempt to remove the effect of experimental bias, we only included interactions obtained through high throughout approaches by excluding those sourced from BIND, HPRD, MINT and MIPS. An obvious difference between the studied dataset and this I2D subset was the much greater range of clique sizes, extending to k=40, compared to k=12. It is reassuring to see the trends could be replicated in this dataset. We still find that there is an overrepresentation of bridges in the I2D dataset for cvd and that cvd proteins cluster together in larger communities for each k. Less obvious was the dramatic increase in the proportion of cvd proteins with increasing k, although the maximum proportion of cvd proteins can be seen at k=23 for cvdpr1 and k=31, 32, 33 for cvd (all priorities). In terms of centrality, observations were replicated with the average betweenness and degree centralities being 2.4 and 1.3 fold greater for cvdpr1 proteins compared to non cvd proteins.

The results presented here show that there are network properties common to both cancer and cvd which may also be reflected in other diseases. The shared properties relate to both network centrality and clustering. There are a number of proteins actually associated with both conditions and these may be general disease mediators.

The strategy of combining centrality measures with analysis of community structure is important when taking this wholeist perspective to understanding the etiological mechanisms of disease.

Chapter 6

General Discussion

Advances in methods for analysing genes and proteins related to disease have provided new opportunities for the application of biology to medical practice (Mathew *et al.*, 2007). The completeness, volume and interpretation of data produced by such methods, requires novel computational biology approaches. This thesis describes four different analyses, the results of which may aid in the interpretation and prioritisation of candidate disease genes in large scale molecular datasets. Machine learning and graph theoretic approaches were used. Some of the approaches combined heterogeneous data sources including such as PPI databases and curated databases such as OMIM.

Initial work focused on developing methods for identifying deleterious nsSNPs. The analysis found various sequence and structural properties were important. Sequence conservation was shown to be the most useful attribute in predicting functional nsSNPs in a large dataset from the SWISSPROT database. Structural attributes in combination with the conservation score improved the prediction accuracy, but other non structurally dependant attributes were found to reduce the error rate further and were valuable in the absence of a conservation score. The nsSNP function prediction analysis also showed the importance of balance within training datasets highlighting the importance of training dataset configuration. Currently, the SNP function prediction scores are being used for prioritising nsSNPs in current BRIGHT studies (`http://www.brightstudy.ac.uk`), and for the recent blood pressure meta analysis (Christopher Newton-Cheh *et al.*, 2009). The method could be extrapolated, by the creation of a tool to obtain on the fly predictions for novel SNPs identified through resequencing experiments. The user would be required to

submit a suspected SNP and its surrounding sequence. Since completion of the nsSNP analysis, a number of further studies have been performed, thirteen of which have cited the work in this thesis and are complementary. In light of these recent studies (described in Section 2.3), further work could be performed, for example: some of the attributes found to be important in studies such as those of Tian *et al.* (2007) and Hu & Yan (2008) could be incorporated into the method; a meta server that collects predictions from all prediction servers would also be useful. A classifier could then be trained which uses predictions from each of the component servers as attributes. It is currently difficult to evaluate the worth of such studies without performing functional work to prove that the SNP was truly a functional SNP. These predictions should be considered a guide to help prioritise gene candidates. With time more prediction servers may become available to complement servers such as PolyPhen (Sunyaev *et al.*, 2001) and SIFT (Ng & Henikoff, 2003). The focus of most SNP studies so far has been on predicting nsSNPs whereas disease associated SNPs often fall in regulatory regions. Such SNPs have only been considered in a small number of studies such as Mottagui-Tabar *et al.* (2005) and Torkamani & Schork (2008), exploring methods to predict these would also be of interest.

Machine learning approaches were also used to develop a method for the functional annotation of proteins belonging to large diverse superfamilies. Our analysis found that global sequence properties of protein domains are useful in determining the protein superfamily. These properties have previously been used to predict protein folds (Ding & Dubchak, 2001). Such an approach can be used to complement traditional homology alignment based approaches. In performing the analysis an enrichment approach step was explored, this resulted in significant improvements to the classifier performance. The enrichment step involved carefully choosing and adding sequences to the training dataset that are currently absent from SCOP. It is expected that this addition could improve performance if applied to the many published fold prediction models (Ashburner *et al.*, 2000; Ding & Dubchak, 2001; Lin *et al.*, 2005; Shen & Chou, 2006; Melvin *et al.*, 2007; Shamim *et al.*, 2007; Damoulas & Girolami, 2008). To extend this work a selective ensembling algorithm is currently under construction for multi-classifier, multi-subspace classification tasks such as the superfamily and fold prediction problem. This should pro-

vide improved prediction performance over single classifiers. Essentially, the approach deploys a large number of different 'base' classifiers (eg. neural networks and decision trees) that are trained with various feature subspaces (amino acid bigrams or composition features for example) and selects the best classifier/subspace pair for each target class, in this case the protein fold or superfamily. This is achieved rapidly through the use of the cluster implementation of Weka. These 'class winners' are ordered by the number of predictions made by the winner on the class it represents. This results in a list of rules that are applied in sequence. This approach has been found to improve previously reported state of the art approaches in terms of classification accuracy, reported by Lin *et al.* (2005) on the SCOP PDB-40D benchmark fold dataset (Ding & Dubchak, 2001).

The PPI topology of hypertension implicated proteins was also investigated and models were produced for predicting novel hypertension proteins. The models showed there are patterns within PPI networks, as well as shared function and sequence based properties that can be used to aid prioritisation of candidate gene lists. Predicted hypertension related proteins are closer and better connected in the interactome than would be expected by chance, despite not being hubs or having a highly connected local environment. We thought that geodesic distance between hypertension protein pairs might correlate with GO semantic similarity but were unable to find a significant correlation. In addition to the attributes used in this study, data from other sources, such as expression data, could also be integrated into the model. A number of recent studies have combined expression data and literature derived data with information relating to PPI networks to perform integrated analyses for the study of human heart failure (Camargo & Azuaje, 2007). In this study, a PPI network was assembled representing heart failure-relevant interactions. The relationships between protein connectivity and expression were analysed and co-expression and connectivity. High connectivity did not always correlate with high differential expression and genes may exhibit weak expression correlation with their interacting partners. The study was very much an exploratory, hypothesis-free, data driven study. A next stage for this study, would be to develop a web based application, to enable users to prioritise novel candidates based on these properties. This could involve either the construction of a pipeline to calculate the attributes required by the classifier on the fly or precalculating

the attributes for each RefSeq protein (Pruitt *et al.*, 2007) and storing the attributes in a database.

Finally we performed an interactome analysis of a large dataset of manually curated cvd and cancer implicated proteins, the data showed that these disease implicated proteins tend to act through a small number of large, complex (tightly connected) processes and exist as interfaces between processes more often that would be expected. The proteins also had a tendancy to be influential proteins within the interactome and there were network properties in common which may also be reflected in other diseases. A recent genome wide meta analysis identified 8 loci associated with blood pressure (Christopher Newton-Cheh *et al.*, 2009). All the gene products from these regions are being assessed for priority based on their centrality scores, community structure and proximity to other cvd implicated proteins. A few of the proteins from each of the 8 loci have been found to belong to the same community, possibly highlighting a common pathway or functional module. These proteins are priority targets for further investigation. A useful aid to prioritising candidates based upon the studied topological properties would be a metric that combined values relating to centrality and clustering into a single score. In addition, a tool could be developed that allows a user to enter a RefSeq protein identifier and return a visual representation of the protein in its topological environment with cvd and cancer implicated proteins highlighted along with the centrality and community structure displayed, in a similar manner to figure 5.5. Such an application could be created as a plugin to the open source package, Cytoscape (Shannon *et al.*, 2003).

Data integration is a rapidly growing field that combines data from wholist heterogeneous biological sources providing many opportunities to develop more complete models of systems. Methods that combine and integrate data with clinical data and pathway data enable the investigation of how perturbations lead to disease and should provide a clearer understanding of the pathways through which they act. As such, these methods could aid accurate diagnosis and prognosis as well as enabling better prevention and therapy in the future. Most current cancer treatments, for example, have low specificity leading to aggressive side effects (Mathew *et al.*, 2007). Identifying cancer specific molecular changes could lead to the identification of disease sub types, and molecular markers, this could

help in the development of more targeted therapies with fewer side effects. Multiple drug combinations are often used to treat hypertension, and there are varied patient responses (Chobanian *et al.*, 2003). A tailored drug combination for a particular patient determined at diagnosis would improve the patient experience and reduce care costs.

Comprehensive integrated pathway information is vital for studying biological processes and how they are affected in disease. However, pathway data exists in a range of diverse pathway databases (`http://pathguide.org`). The datasets are often incomplete and sparsely populated (Cary *et al.*, 2005). PPIs networks created through a combination of manual curation and high throughput screening can contribute towards knowledge of pathway structure.

The integration of genotype and microarray data with PPI and pathway information is an important challenge showing promise for predicting the effect of a mutation on disease and identifying therapeutic targets through vulnerable points in particular pathways. In a study performed by Chuang *et al.* (2007), expression data was combined with PPI networks in order to identify differentially expressed combinations of trancripts (modules) based on interactome proximity and mutual information. The analysis showed improved ability over single transcript markers to predict metastasis within breast cancer patients. We are currently developing an approach for identifying disease associated modules within PPI networks based on SNP association scores from GWA studies. This could help identify multiple SNPs that by themselves do not display significant association but have a significant combined effect (figure 6.1). This assumes that the biological pathways have multiple vulnerable points that can lead to the same disease phenotype (Mathew *et al.*, 2007). These are examples of integrated approaches that could enable significant advances in the study and understanding of the etiology of complex diseases.

This thesis utilised static, qualitative presentation of integrated genome scale data, through the identification and study of relationships that exist between component parts. A quantitative analysis that aims to understand the relationships or network dynamics (understanding the nature of the links) within a system is a bigger challenge and it is receiving growing interest (Luscombe *et al.*, 2004). Progress in this area is being made by the measuring of gene expression through microarrays and pathway simulations have

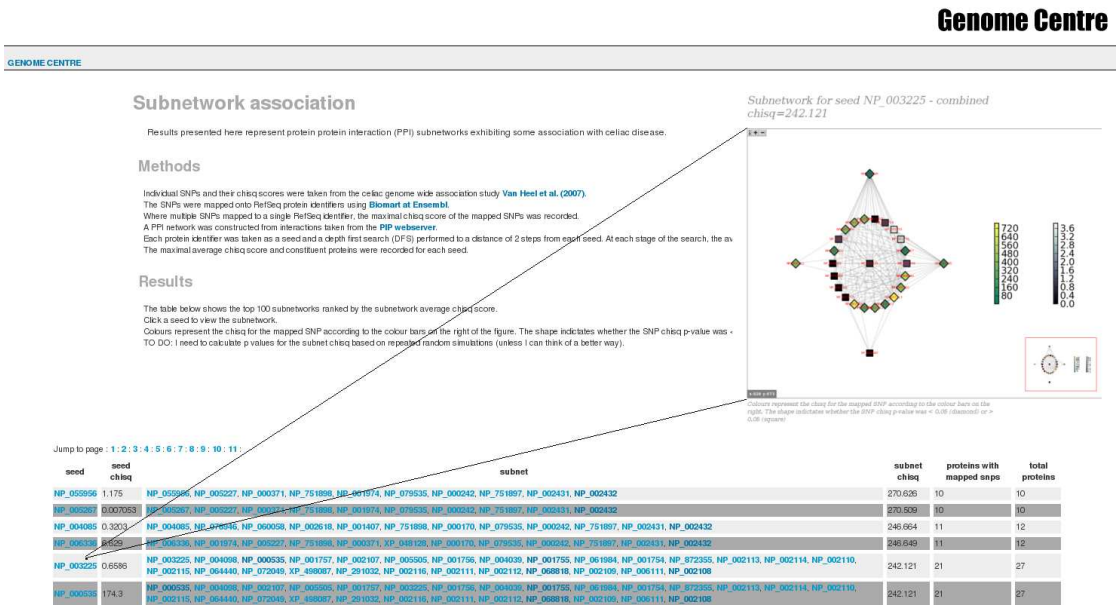


Figure 6.1: An approach that combines SNP association Chi-squared scores to identify disease associated PPI subnetworks. In this example, SNPs were taken from the coeliac genome wide association study by van Heel *et al.* (2007). Proteins (nodes) are coloured according to their Chi-squared value with diamonds representing proteins with Chi-squared scores having p-values < 0.05

been used in model organisms such as *Escherichia coli* and budding yeast to find pathway regulators (Chen *et al.*, 2000). The speed of advances within these key areas of computational biology suggests that such integrated, wholist approaches will extend long into the future.

6.1 Publications

Directly arising from this thesis:

Papers

- **Richard JB Dobson**, Patricia B Munroe, Mark J Caulfield, Mansoor AS Saqi. (2009). Protein interaction networks associated with cardiovascular disease and cancer: shared network properties. *Journal of Theoretical Biology*. In revision.

- **Richard JB Dobson**, Patricia B Munroe, Mark J Caulfield, Mansoor AS Saqi. (2009). Global sequence properties for superfamily prediction: a machine learning approach *Journal of Integrative Bioinformatics*. 6(1):109, 2009

- **Richard JB Dobson**, Patricia B Munroe, Charles A Mein, Mark J Caulfield, Mansoor AS Saqi. (2008). Combining PPI network and sequence attributes for predicting hypertension related proteins. In the proceedings of the second *Bioinformatics Research and Development International Conference*, BIRD 2008 Vienna, Austria, July 7-9, 2008. Communications in Computer and Information Science, 377-391. Springer.

- **RJ Dobson**, PB Munroe, MJ Caulfield, MAS Saqi. (2006). Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics*, 2006 Apr 21;7(1):217.

Posters and Meeting abstracts

- **Richard JB Dobson**, Patricia B Munroe, Mark J Caulfield, Mansoor AS Saqi. (2009). Global sequence properties for superfamily prediction: a machine learning approach *ISMB/ECCB*. Accepted 2009.

- **Richard JB Dobson**, Patricia B Munroe, Mark J Caulfield, Mansoor AS Saqi. (2008). Protein interaction networks associated with cardiovascular disease and cancer: shared network properties. *William Harvey Day*, Barts and The London, Queen Mary University of London, October 2008.

Produced collaboratively during this thesis:

Papers

- Mansoor AS Saqi, **Richard JB Dobson**, Preben Kraben, David Hodgson, David Wild. (2009) An approach to pathway reconstruction using whole genome metabolic models and sensitive sequence searching, *Journal of Integrative Bioinformatics*. In Press.

- Org E, Eyheramendy S, Juhanson P, Gieger C, Lichtner P, Klopp N, Veldre G, Doring A, Viigimaa M, Sober S, Tomberg K, Eckstein G; KORA, Kelgo P, Rebane T, Shaw-Hawkins S, Howard P, Onipinla A, **Dobson RJ**, Newhouse SJ, Brown M, Dominiczak A, Connell J, Samani N, Farrall M; BRIGHT, Caulfield MJ, Munroe PB,

Illig T, Wichmann HE, Meitinger T, Laan M. (2009) Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Hum Mol Genet* 2009 Mar 20. [Epub ahead of print].

- Stephen Newhouse, Martin Farrall, Chris Wallace, Mimoza Hoti, Beverley Burke, Philip Howard, Abiodun Onipinla, Kate Lee, Sue Shaw-Hawkins, **Richard Dobson**, Morris Brown, Nilesh J. Samani, Anna F. Dominiczak, John M. Connell, G. Mark Lathrop, Jaspal Kooner, John Chambers, Paul Elliott, Robert Clarke, Rory Collins, Maris Laan, Elin Org, Peeter Juhanson, Gudrun Veldre, Margus Viigimaa1, Susana Eyheramendy, Francesco P. Cappuccio, Chen Ji, Roberto Iacone, Pasquale Strazzullo, Meena Kumari, Michael Marmot, Eric Brunner, Mark Caulfield and Patricia B. Munroe. (2009) Polymorphisms in the WNK1 gene are associated with blood pressure variation and urinary potassium excretion. *PLoS ONE*. In Press.

- C. Sinclair, E.A. O'Toole, D. Paige, H. El Bashir, J. Robinson, **R. Dobson**, N. Lench, H.P. Stevens, G.A. Hitman, R. Booy, C.A. Mein, D.P. Kelsell Filaggrin mutations are associated with ichthyosis vulgaris in the Bangladeshi population *British Journal of Dermatology*. Early View, Date: February 2009

- Mark J. Caulfield, Patricia B. Munroe, Deb O'Neill, Kate Witkowska, Fadi J. Charchar, Manuel Doblado, Sarah Evans, Susana Eyheramendy, Abiodun Onipinla, Philip Howard, Sue Shaw-Hawkins, **Richard J. Dobson**, Chris Wallace, Stephen J. Newhouse, Morris Brown, John M. Connell, Anna Dominiczak, Martin Farrall, G. Mark Lathrop, Nilesh J. Samani, Meena Kumari, Michael Marmot, Eric Brunner, John Chambers, Paul Elliott, Jaspal Kooner, Maris Laan, Elin Org, Gudrun Veldre, Margus Viigimaa, Francesco P. Cappuccio, Chen Ji, Roberto Iacone, Pasquale Strazzullo, Kelle H. Moley, Chris Cheeseman. (2008). SLC2A9 Is a High-Capacity Urate Transporter in Humans. *PLoS Med* 2008 Oct 7;5(10):e197.

- Chris Wallace; Stephen J Newhouse; Peter Braund; Feng Zhang; Martin Tobin; Mario Falchi; Kourosh Ahmadi; **Richard J Dobson**; Ana Carolina B Marcano; Cother Hajat; Paul Burton; Panagiotis Deloukas; Morris Brown; John M Connell; Anna Dominiczak; G Mark Lathrop; John Webster; The Wellcome Trust Case Con-

trol Consortium; Martin Farrall; Tim Spector; Nilesh J Samani; Mark J Caulfield; Patricia B. Munroe. (2008). Genome-wide association study identifies novel genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet*. 2008 Jan 10; 82(1) pp. 139 - 149.

- Wallace C, **Dobson RJ**, Munroe PB, Caulfield MJ. (2007). Information capture using SNPs from HapMap and whole-genome chips differs in a sample of inflammatory and cardiovascular gene-centric regions from genome-wide estimates. *Genome Research*. 2007 Sep 25.

- Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007 Jun 7;447(7145):661-78.

- Chris Wallace, Ming-Zhan Xue, Stephen J. Newhouse, Ana Carolina B. Marcano, Abiodun K. Onipinla, Beverley Burke, Johannie Gungadoo, **Richard J. Dobson**, Morris Brown, John M. Connell, Anna Dominiczak, G. Mark Lathrop, John Webster, Martin Farrall, Charles Mein, Nilesh J. Samani, Mark J. Caulfield, David G. Clayton, and Patricia B. Munroe. (2006). Linkage Analysis Using Co-Phenotypes in the BRIGHT Study Reveals Novel Potential Susceptibility Loci for Hypertension. *Am J Hum Genet*. 2006 Aug;79(2):323-31. Epub 2006 Jun 19.

- Munroe PB, Wallace C, Xue MZ, B Marcano AC, **Dobson RJ**, Onipinla AK, Burke B, Gungadoo J, Newhouse SJ, Pembroke J, Brown M, Dominiczak AF, Samani NJ, Lathrop M, Connell J, Webster J, Clayton D, Farrall M, Mein CA, Caulfield M. (2006). Increased Support for Linkage of a Novel Locus on Chromosome 5q13 for Essential Hypertension in the British Genetics of Hypertension Study. *Hypertension*. 2006 Jul;48(1):105-11.

- Bell JT, Wallace C, **Dobson R**, Wiltshire S, Mein C, Pembroke J, Brown M, Clayton D, Samani N, Dominiczak A, Webster J, Lathrop GM, Connell J, Munroe P, Caulfield M, Farrall M. (2006). Two-dimensional genome-scan identifies novel epistatic loci for essential hypertension. *Hum Mol Genet*. 2006 Apr 15;15(8):1365-74.

- Sandosh Padmanabhan Ph.D, Chris Wallace Ph.D, Patricia B. Munroe Ph.D, **Richard Dobson BSc**, Morris Brown F.R.C.P, Nilesh Samani F.R.C.P, David Clayton B.A, Martin Farrall FRCPath, John Webster F.R.C.P, Mark Lathrop Ph.D, Mark Caulfield F.R.C.P, Anna F. Dominiczak F.R.C.P, John M. Connell F.R.C.P. (2006) Chromosome 2p shows significant linkage to anti-hypertensive response in the British Genetics of Hypertension (BRIGHT) study. *Hypertension*. 2006 Mar;47(3):603-8.

- S.J. Newhouse, C. Wallace, **R. Dobson**, C. Mein, J. Pembroke, M. Farrall, D. Clayton, M. Brown, N. Samani, A. Dominiczak, J.M. Connell, J. Webster, G.M. Lathrop, M. Caulfield, P. B. Munroe. (2005). Haplotypes of the WNK1 gene associate with blood pressure variation in a severely hypertensive population from the British Genetics of Hypertension (BRIGHT) study. *Human Molecular Genetics*. Jul 1;14(13):1805-14.

- Mark Caulfield, Patricia Munroe, Janine Pembroke, Nilesh Samani, Anna Dominiczak, Morris Brown, Nigel Benjamin, John Webster, Peter Ratcliffe, Suzanne O'Shea, Jeanette Papp, Elizabeth Taylor, **Richard Dobson**, Joanne Knight, Stephen Newhouse, Joel Hooper, Wai Lee, Nick Brain, David Clayton, G Mark Lathrop, Martin Farrall, John Connell, for The MRC British Genetics of Hypertension Study. (2003). Genome-wide mapping of human loci for essential hypertension. *Lancet* 2003;361:2118-23.

Posters and Meeting abtracts

- **Richard JB Dobson**. (2004) MRC BRIGHT study: Demographics of the sibling pair resource. *British Journal of Pharmacology*, 2004.

- Newhouse S, **Dobson R**, Wallace C, Pembroke J, Garcia E, Mein C, Clayton D, Samani N, Dominiczak A, Brown M, Webster J, Lathrop G, Farrall M, Connell J, Caulfield M, Munroe P. (2004). No association of the WNK1 gene with essential hypertension in the MRC BRIGHT study. *European Society of Hypertension* (Paris) June 2004 and *Human Genetics Mapping* meeting Berlin, April 2004.

- **Richard JB Dobson**, Edwin Garcia, Steve Newhouse, Mark Caulfield, Patricia Munroe. (2003) Identification of tagging single nucleotide polymorphisms in haplotype datasets. *Royal Statistical Society Statistical Genetics and Bioinformatics*, Limburgs Universitair Centrum, Hasselt, Belgium. 14-17 July 2003.

- Newhouse S, Garcia E, **R.Dobson**, M.Caulfield, P.Munroe. (2003). Haplotype structure of the WNK1 gene and association studies in hypertensive populations. *London Hypertension Society* 2003.

- Newhouse S, Garcia E, **R.Dobson**, M.Caulfield, P.Munroe. (2003). Haplotype structure of the WNK1 gene and association studies in hypertensive populations. *London Hypertension Society*, 2003.

Reviews

- Mein CA, Caulfield MJ, **Dobson RJB**, Munroe PB. (2004). Genetics of essential hypertension. *Hum Mol Genet*. 2004 Apr 1;13 Spec No 1:R169-75

Produced as part of the Wellcome Trust Case Control Consortium:

- Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, K IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H, WTCCC. (2007). Genomewide association analysis of coronary artery disease. *N Engl J Med*. 2007 Aug 2;357(5):443-53.

- Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, Knig IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H,

Wellcome Trust Case Control Consortium (WTCCC). (2007). Genomewide association analysis of coronary artery disease. *N Engl J Med*. 2007 Aug 2;357(5):443-53.

Appendices

Appendix A. Environment, Parameters and Specification

Hardware

The majority of the data processing performed in this thesis was done using Linux and Unix computers available at the Genome Centre, St Barts and The London, Queen Mary University of London. Both the Genome Centre and Queen Mary University of London High Throughput Computing (HTC) clusters were used to perform analyses. Large amounts of filespace on the Genome Centre servers were used to store data and software. All Genome Centre servers and cluster were administered and maintained by the thesis author.

Programming Languages and Databases

The main programming languages used for data harvesting, parsing, manipulation, results collection and evaluation were Perl 5 (`http://www.perl.org`) and Python (`http://www.python.org`). MySQL version 5.0.5 (`http://www.mysql.com`) was used to create databases for storing data and results of the analyses. R version 2.6.2 was used for statistical tests including the Wilcoxon sign rank test, Pearsons and Spearmans and to produce most of the plots within the thesis (R Development Core Team, 2008).

Software and Operating Systems

- Linux and Unix operating systems including Ubunbtu, CentOS, Debian, Scientific Linux and Sun Solaris 8 distributions were used throughout this thesis. All programs were written and tested using these operating systems.
- The Weka machine learning workbench was used for machine learning classification (Witten & Frank, 1999). It consists of Java implementations of many machine algorithms, that can be applied directly to a dataset. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualisation. LibSVM was integrated into the Weka Environment using WLSVM (EL-Manzalawy & Honavar, 2005).

- The Apache 2 web server (`http://www.apache.org/`) and Perl CGI modules (`http://perldoc.perl.org/CGI.html`) were used to create the web based, cluster implementation of Weka.

- The PSIC (position-specific independent counts) program was used to calculate conservation at the position of a nsSNP (Ramensky *et al.*, 2002). Profiles are extracted from sequence alignments with position-specific counts of independent observations.

- The LDAS (Lightweight Distributed Annotation Server) framework was used for implementing the DAS server to server up nsSNP function predictions (Dowell *et al.*, 2001)

- PSI-BLAST and BLAST were used for various analyses that required sequences to be aligned. This included the enrichment step of the protein function analysis (Altschul *et al.*, 1997).

- BLASTClust was used to cluster sequences based on their sequence similarity (Dondoshansky, 2002).

- Secondary structure prediction of a protein was performed with PSIPRED (McGuffin *et al.*, 2000).

- The Python API implemented by Casbon *et al.* (2006) as part of the Biopython project was used to manipulate and parse ASTRAL and SCOP files to construct datasets for the protein superfamily analysis.

- G-Sesame was used to calculate the semantic similarity of GO terms associated with sets proteins (Wang *et al.*, 2007).

- The NetworkX Python routines were used to calculate centrality scores and produce network figures (`https://networkx.lanl.gov`)

- The *Cfinder* program was used to identify communities for the analysis of cvd and cancer implicated proteins (Adamcsek *et al.*, 2006). This program uses the *k*-clique clustering method.

Appendix B. Supplementary tables relating to protein superfamily prediction

Superfamily	**No. of domains (D)**	**No. of domains (20E)**	**No. of domains (30E)**	**30E/D**
46458 a.1.1 sf Globin-like	11	22	31	2.82
46689 a.4.1 sf Homeodomain-like	12	35	35	2.92
46785 a.4.5 sf "Winged helix" DNA-binding domain	25	81	114	4.56
47266 a.26.1 sf 4-helical cytokines	15	19	27	1.8
48371 a.118.1 sf ARM repeat	11	35	50	4.55
49785 b.18.1 sf Galactose-binding domain-like	13	17	21	1.62
49899 b.29.1 sf Concanavalin A-like lectins/glucanases	14	19	28	2
50249 b.40.4 sf Nucleic acid-binding proteins	19	37	58	3.05
50729 b.55.1 sf PH domain-like	11	27	27	2.45
51182 b.82.1 sf RmlC-like cupins	11	16	20	1.82
88633 b.121.4 sf Positive stranded ss-RNA viruses	11	23	23	2.09
51445 c.1.8 sf (Trans)glycosidases	15	30	45	3
51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains	13	24	56	4.31
52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases	43	89	138	3.21
52833 c.47.1 sf Thioredoxin-like	17	36	41	2.41
52980 c.52.1 sf Restriction endonuclease-like	13	14	15	1.15
53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases	25	71	92	3.68
53383 c.67.1 sf PLP-dependent transferases	15	36	72	4.8
53448 c.68.1 sf Nucleotide-diphospho-sugar transferases	11	20	45	4.09
53474 c.69.1 sf alpha/beta-Hydrolases	23	41	130	5.65
53850 c.94.1 sf Periplasmic binding protein-like II	13	31	135	10.38
55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)	15	37	49	3.27
57059 g.3.6 sf omega toxin-like	15	19	19	1.27
57095 g.3.7 sf Scorpion toxin-like	12	22	22	1.83

Table 1: Number of domains per superfamily (in the analysis that excluded multi domain proteins) from Astral20 before enrichment (D) and after enrichment at 20% (20E) and 30% (30E) sequence identity cutoffs

Table 2: Number of domains per superfamily (in the analysis that included multi domain proteins) from Astral20 before enrichment (D) and after enrichment at 20% (20E) and 30% (30E) sequence identity cutoffs

Superfamily	**D**	**20E**	**30E**	**30E/D**
46458 a.1.1 sf Globin-like	11	19	24	2.18
46626 a.3.1 sf Cytochrome c	15	20	20	1.33
46689 a.4.1 sf Homeodomain-like	23	114	115	5
46785 a.4.5 sf "Winged helix" DNA-binding domain	55	144	175	3.18
47266 a.26.1 sf 4-helical cytokines	15	20	24	1.6
47473 a.39.1 sf EF-hand	13	25	34	2.62
48371 a.118.1 sf ARM repeat	17	43	53	3.12
48726 b.1.1 sf Immunoglobulin	36	103	105	2.92
49265 b.1.2 sf Fibronectin type III	21	60	61	2.9
81296 b.1.18 sf E set domains	26	52	53	2.04
49503 b.6.1 sf Cupredoxins	15	33	40	2.67
49785 b.18.1 sf Galactose-binding domain-like	21	37	43	2.05
49899 b.29.1 sf Concanavalin A-like lectins/glucanases	22	41	56	2.55
50249 b.40.4 sf Nucleic acid-binding proteins	39	68	91	2.33
50729 b.55.1 sf PH domain-like	19	47	48	2.53
51011 b.71.1 sf Glycosyl hydrolase domain	19	24	24	1.26
51182 b.82.1 sf RmlC-like cupins	12	19	22	1.83
88633 b.121.4 sf Positive stranded ssRNA viruses	15	23	23	1.53

Continued on Next Page…

Table 2 – Continued

Superfamily	**D**	**20E**	**30E**	**30E/D**
51445 c.1.8 sf (Trans)glycosidases	33	64	99	3
51569 c.1.10 sf Aldolase	12	16	36	3
51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains	47	87	137	2.91
51905 c.3.1 sf FAD/NAD(P)-binding domain	21	37	42	2
52317 c.23.16 sf Class I glutamine amidotransferase-like	11	19	23	2.09
52374 c.26.1 sf Nucleotidylyl transferase	13	24	31	2.38
52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases	70	150	227	3.24
52833 c.47.1 sf Thioredoxin-like	28	56	71	2.54
52980 c.52.1 sf Restriction endonuclease-like	15	16	17	1.13
53067 c.55.1 sf Actin-like ATPase domain	17	27	36	2.12
53098 c.55.3 sf Ribonuclease H-like	15	28	37	2.47
53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases	29	69	110	3.79
53383 c.67.1 sf PLP-dependent transferases	16	32	66	4.13
53448 c.68.1 sf Nucleotide-diphospho-sugar transferases	12	19	45	3.75
53474 c.69.1 sf alpha/beta-Hydrolases	27	54	145	5.37

Continued on Next Page…

Table 2 – Continued

Superfamily	**D**	**20E**	**30E**	**30E/D**
53850 c.94.1 sf Periplasmic binding protein-like II	15	33	152	10.13
56784 c.108.1 sf HAD-like	11	23	47	4.27
54001 d.3.1 sf Cysteine proteinases	18	28	33	1.83
54211 d.14.1 sf Ribosomal protein S5 domain 2-like	11	21	23	2.09
54236 d.15.1 sf Ubiquitin-like	11	21	24	2.18
54373 d.16.1 sf FAD-linked reductases, C-terminal domain	11	19	19	1.73
54593 d.32.1 sf Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	12	19	19	1.58
55347 d.81.1 sf Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain	11	23	32	2.91
55486 d.92.1 sf Metalloproteases (“zincins”), catalytic domain	18	44	54	3
55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)	11	25	41	3.73
56672 e.8.1 sf DNA/RNA polymerases	11	23	56	5.09
57059 g.3.6 sf omega toxin-like	15	19	19	1.27
57095 g.3.7 sf Scorpion toxin-like	12	20	20	1.67
57196 g.3.11 sf EGF/Laminin	11	55	55	5
57667 g.37.1 sf C2H2 and C2HC zinc fingers	15	40	40	2.67
57716 g.39.1 sf Glucocorticoid receptor-like (DNA-binding domain)	13	22	22	1.69

Table 3: The 49 superfamilies in the multi domain analysis with their respective folds and classes within the SCOP hierarchy. False positive superfamily predictions are often correctly assigned at the level of protein class.

Class	Fold	Superfamily
46456 a cl All alpha proteins	46457 a.1 cf Globin-like	46458 a.1.1 sf Globin-like
	46625 a.3 cf Cytochrome c	46626 a.3.1 sf Cytochrome c
	46688 a.4 cf DNA/RNA-binding 3-helical bundle	46689 a.4.1 sf Homeodomain-like
		46785 a.4.5 sf "Winged helix" DNA-binding domain
	47265 a.26 cf 4-helical cytokines	47266 a.26.1 sf 4-helical cytokines
	47472 a.39 cf EF Hand-like	47473 a.39.1 sf EF-hand
	48370 a.118 cf alpha-alpha superhelix	48371 a.118.1 sf ARM repeat
48724 b cl All beta proteins	48725 b.1 cf Immunoglobulin-like beta-sandwich	48726 b.1.1 sf Immunoglobulin
		49265 b.1.2 sf Fibronectin type III
		81296 b.1.18 sf E set domains
	49379 b.6 cf Cupredoxin-like	49503 b.6.1 sf Cupredoxins
	49784 b.18 cf Galactose-binding domain-like	49785 b.18.1 sf Galactose-binding domain-like
	49898 b.29 cf Concanavalin A-like lectins/glucanases	49899 b.29.1 sf Concanavalin A-like lectins/glucanases
	50198 b.40 cf OB-fold	50249 b.40.4 sf Nucleic acid-binding proteins

Continued on Next Page…

Table 3 – Continued

Class	Fold	Superfamily
	50728 b.55 cf PH domain-like	50729 b.55.1 sf PH domain-like
	51010 b.71 cf Glycosyl hydrolase domain	51011 b.71.1 sf Glycosyl hydrolase domain
	51181 b.82 cf Double-stranded beta-helix	51182 b.82.1 sf RmlC-like cupins
	88632 b.121 cf Nucleoplasmin-like/VP (viral coat and capsid proteins)	88633 b.121.4 sf Positive stranded ssRNA viruses
51349 c cl Alpha and beta proteins (a/b)	51350 c.1 cf TIM beta/alpha-barrel	51445 c.1.8 sf (Trans)glycosidases
		51569 c.1.10 sf Aldolase
	51734 c.2 cf NAD(P)-binding Rossmann-fold domains	51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains
	51904 c.3 cf FAD/NAD(P)-binding domain	51905 c.3.1 sf FAD/NAD(P)-binding domain
	52171 c.23 cf Flavodoxin-like	52317 c.23.16 sf Class I glutamine amidotransferase-like
	52373 c.26 cf Adenine nucleotide alpha hydrolase-like	52374 c.26.1 sf Nucleotidylyl transferase
	52539 c.37 cf P-loop containing nucleoside triphosphate hydrolases	52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases
	52832 c.47 cf Thioredoxin fold	52833 c.47.1 sf Thioredoxin-like
	52979 c.52 cf Restriction endonuclease-like	52980 c.52.1 sf Restriction endonuclease-like

Continued on Next Page. . .

Table 3 – Continued

Class	Fold	Superfamily
	53066 c.55 cf Ribonuclease H-like motif	53067 c.55.1 sf Actin-like ATPase domain
		53098 c.55.3 sf Ribonuclease H-like
	53334 c.66 cf S-adenosyl-L-methionine-dependent methyltransferases	53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases
	53447 c.68 cf Nucleotide-diphospho-sugar transferases	53448 c.68.1 sf Nucleotide-diphospho-sugar transferases
	53473 c.69 cf alpha/beta-Hydrolases	53474 c.69.1 sf alpha/beta-Hydrolases
	53849 c.94 cf Periplasmic binding protein-like II	53850 c.94.1 sf Periplasmic binding protein-like II
	56783 c.108 cf HAD-like	56784 c.108.1 sf HAD-like
53931 d cl Alpha and beta proteins (a+b)	54000 d.3 cf Cysteine proteinases	54001 d.3.1 sf Cysteine proteinases
	54210 d.14 cf Ribosomal protein S5 domain 2-like	54211 d.14.1 sf Ribosomal protein S5 domain 2-like
	54235 d.15 cf beta-Grasp (ubiquitin-like)	54236 d.15.1 sf Ubiquitin-like
	54372 d.16 cf FAD-linked reductases, C-terminal domain	54373 d.16.1 sf FAD-linked reductases, C-terminal domain
	54592 d.32 cf Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	54593 d.32.1 sf Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase

Continued on Next Page. . .

Table 3 – Continued

Class	Fold	Superfamily
	55346 d.81 cf Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain	55347 d.81.1 sf Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain
	55485 d.92 cf Zincin-like	55486 d.92.1 sf Metalloproteases (“zincins”), catalytic domain
	55728 d.108 cf Acyl-CoA N-acyltransferases (Nat)	55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)
56572 e cl Multi domain proteins (alpha and beta)	56671 e.8 cf DNA/RNA polymerases	56672 e.8.1 sf DNA/RNA polymerases
56992 g cl Small proteins	57015 g.3 cf Knottins (small inhibitors, toxins, lectins)	57059 g.3.6 sf omega toxin-like
		57095 g.3.7 sf Scorpion toxin-like
		57196 g.3.11 sf EGF/Laminin
	57666 g.37 cf C2H2 and C2HC zinc fingers	57667 g.37.1 sf C2H2 and C2HC zinc fingers
	57715 g.39 cf Glucocorticoid receptor-like (DNA-binding domain)	57716 g.39.1 sf Glucocorticoid receptor-like (DNA-binding domain)

Table 4: The precision, recall and F-measure produced by PSI-BLAST and SVMs on the unenriched dataset containing 24 superfamilies (domains from multi domain proteins excluded).

	SVM			**PSI-BLAST**		
Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
46458 a.1.1 sf Globin-like	0.71	1	0.83	1	0.94	0.97
46689 a.4.1 sf Homeodomain-like	0.67	0.67	0.67	1	0.33	0.5
46785 a.4.5 sf "Winged helix" DNA-binding domain	0.71	0.83	0.77	1	0.32	0.49
47266 a.26.1 sf 4-helical cytokines	1	0.63	0.77	1	0.3	0.47
48371 a.118.1 sf ARM repeat	0.8	0.67	0.73	1	0.65	0.79
49785 b.18.1 sf Galactose-binding domain-like	0.4	0.67	0.5	1	0.32	0.48
49899 b.29.1 sf Concanavalin A-like lectins/glucanases	0.5	0.57	0.53	1	0.67	0.8
50249 b.40.4 sf Nucleic acid-binding proteins	0.75	0.33	0.46	1	0.32	0.49
50729 b.55.1 sf PH domain-like	0.5	0.5	0.5	1	0.59	0.74
51182 b.82.1 sf RmlC-like cupins	0.75	0.6	0.67	1	0.81	0.9
88633 b.121.4 sf Positive stranded ssRNA viruses	0.8	0.8	0.8	1	0.63	0.77

Continued on Next Page. . .

Table 4 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
51445 c.1.8 sf (Trans)glycosidases	0.8	0.57	0.67	1	0.59	0.74
51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains	0.8	0.67	0.73	1	0.95	0.97
52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases	0.5	0.71	0.59	1	0.75	0.86
52833 c.47.1 sf Thioredoxin-like	0.86	0.67	0.75	1	0.85	0.92
52980 c.52.1 sf Restriction endonuclease-like	0.5	0.57	0.53	1	0.1	0.18
53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferase	0.2	0.23	0.21	1	0.84	0.91
53383 c.67.1 sf PLP-dependent transferases	0.56	0.63	0.59	1	1	1
53448 c.68.1 sf Nucleotide-diphospho-sugar transferases	0	0	0	1	0.59	0.74
53474 c.69.1 sf alpha/beta-Hydrolases	0.55	0.55	0.55	1	0.91	0.95
53850 c.94.1 sf Periplasmic binding protein-like II	0.71	0.71	0.71	1	0.8	0.89
55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)	0.67	0.29	0.4	1	0.95	0.98

Continued on Next Page. . .

Table 4 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
57059 g.3.6 sf omega toxin-like	0.86	0.86	0.86	0	0	0
57095 g.3.7 sf Scorpion toxin-like	0.83	0.83	0.83	1	0.11	0.2

Table 5: The precision, recall and F-measure produced by PSI-BLAST and SVMs on the unenriched dataset containing 49 superfamilies (domains from multi domain proteins included).

	SVM			PSI-BLAST		
Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
46458 a.1.1 sf Globin-like	0.78	0.82	0.8	1	0.94	0.97
46626 a.3.1 sf Cytochrome c	0.76	0.7	0.73	1	0.83	0.9
46689 a.4.1 sf Homeodomain-like	0.51	0.63	0.56	1	0.43	0.6
46785 a.4.5 sf "Winged helix" DNA-binding domain	0.72	0.77	0.74	1	0.35	0.52
47266 a.26.1 sf 4-helical cytokines,	0.74	0.61	0.67	1	0.3	0.47
47473 a.39.1 sf EF-hand,	0.63	0.26	0.37	1	0.58	0.73
48371 a.118.1 sf ARM repeat,	0.64	0.64	0.64	1	0.44	0.61
48726 b.1.1 sf Immunoglobulin	0.54	0.7	0.61	1	0.78	0.88
49265 b.1.2 sf Fibronectin type III	0.58	0.48	0.53	1	0.71	0.83
81296 b.1.18 sf E set domains	0.19	0.18	0.18	1	0.28	0.44
49503 b.6.1 sf Cupredoxins	0.07	0.04	0.05	1	0.78	0.88
49785 b.18.1 sf Galactose-binding domain-like	0.43	0.48	0.46	1	0.42	0.59
49899 b.29.1 sf Concanavalin A-like lectins/glucanases	0.37	0.49	0.42	1	0.64	0.78

Continued on Next Page. . .

Table 5 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
50249 b.40.4 sf Nucleic acid-binding proteins	0.33	0.49	0.39	1	0.39	0.56
50729 b.55.1 sf PH domain-like	0.63	0.59	0.61	1	0.59	0.74
51011 b.71.1 sf Glycosyl hydrolase domain	0.59	0.55	0.57	0.89	0.28	0.42
51182 b.82.1 sf RmlC-like cupins	0.14	0.06	0.08	1	0.78	0.88
88633 b.121.4 sf Positive stranded ssRNA virus	0.64	0.41	0.5	1	0.5	0.67
51445 c.1.8 sf (Trans)glycosidases	0.53	0.82	0.65	1	0.71	0.83
51569 c.1.10 sf Aldolase	0	0	0	1	0.72	0.84
51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains	0.44	0.67	0.53	0.92	0.87	0.9
51905 c.3.1 sf FAD/NAD(P)-binding domain	0.46	0.47	0.46	0.97	0.91	0.94
52317 c.23.16 sf Class I glutamine amidotransferase-like	0.6	0.19	0.29	1	0.88	0.93
52374 c.26.1 sf Nucleotidyl transferase	0	0	0	1	0.89	0.94
52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases	0.34	0.75	0.46	1	0.82	0.9

Continued on Next Page. . .

Table 5 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
52833 c.47.1 sf Thioredoxin-like	0.77	0.57	0.66	1	0.83	0.91
52980 c.52.1 sf Restriction endonuclease-like	0.33	0.09	0.14	1	0.09	0.16
53067 c.55.1 sf Actin-like ATPase domain	0.57	0.46	0.51	1	0.27	0.42
53098 c.55.3 sf Ribonuclease H-like	0.52	0.48	0.5	1	0.57	0.72
53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases	0.23	0.33	0.27	1	0.86	0.93
53383 c.67.1 sf PLP-dependent transferases	0.67	0.42	0.51	1	1	1
53448 c.68.1 sf Nucleotide-diphospho-sugar transferases	0	0	0	1	0.61	0.76
53474 c.69.1 sf alpha/beta-Hydrolases	0.59	0.58	0.58	1	0.93	0.96
53850 c.94.1 sf Periplasmic binding protein-like II	1	0.35	0.52	1	0.74	0.85
56784 c.108.1 sf HAD-like	1	0.06	0.12	1	0.19	0.32
54001 d.3.1 sf Cysteine proteinases	0	0	0	1	0.59	0.74
54211 d.14.1 sf Ribosomal protein S5 domain 2-like	0.31	0.19	0.23	1	0.53	0.69
54236 d.15.1 sf Ubiquitin-like	0.33	0.12	0.17	1	0.31	0.48

Continued on Next Page. . .

Table 5 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
54373 d.16.1 sf FAD-linked reductases, C-terminal domain	0.57	0.25	0.35	1	0.94	0.97
54593 d.32.1 sf Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	1	0.59	0.74	1	0.33	0.5
55347 d.81.1 sf Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain	0.31	0.22	0.26	1	0.41	0.58
55486 d.92.1 sf Metalloproteases (“zincins”), catalytic domain	0.8	0.24	0.36	0.96	0.85	0.9
55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)	0.5	0.37	0.43	0.92	0.75	0.83
56672 e.8.1 sf DNA/RNA polymerases	0	0	0	1	0.81	0.9
57059 g.3.6 sf omega toxin-like	0.5	0.55	0.52	0	0	0
57095 g.3.7 sf Scorpion toxin-like	0.63	0.56	0.59	1	0.11	0.2
57196 g.3.11 sf EGF/Laminin	0.71	0.63	0.67	1	0.06	0.12
57667 g.37.1 sf C2H2 and C2HC zinc fingers	0.9	0.82	0.86	1	0.14	0.24

Continued on Next Page…

Table 5 – Continued

Superfamily	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
57716 g.39.1 sf Glucocorticoid receptor-like (DNA-binding domain)	0.5	0.35	0.41	0	0	0

Table 6: The 24 superfamilies in this study with their respective folds and classes within the SCOP hierarchy. False positive superfamily predictions are often correctly assigned at the level of protein class.

Class	**Fold**	**Superfamily**
46456 a cl All alpha proteins	46457 a.1 cf Globin-like	46458 a.1.1 sf Globin-like
	46688 a.4 cf DNA/RNA-binding 3-helical bundle	46689 a.4.1 sf Homeodomain-like
		46785 a.4.5 sf "Winged helix" DNA-binding domain
	47265 a.26 cf 4-helical cytokines	47266 a.26.1 sf 4-helical cytokines
	48370 a.118 cf alpha-alpha superhelix	48371 a.118.1 sf ARM repeat
48724 b cl All beta proteins	49784 b.18 cf Galactose-binding domain-like	49785 b.18.1 sf Galactose-binding domain-like
	49898 b.29 cf Concanavalin A-like lectins/glucanases	49899 b.29.1 sf Concanavalin A-like lectins/glucanases
	50198 b.40 cf OB-fold	50249 b.40.4 sf Nucleic acid-binding proteins
	50728 b.55 cf PH domain-like	50729 b.55.1 sf PH domain-like
	51181 b.82 cf Double-stranded beta-helix	51182 b.82.1 sf RmlC-like cupins
	88632 b.121 cf Nucleoplasmin-like/VP (viral coat and capsid proteins)	88633 b.121.4 sf Positive stranded ssRNA viruses

Continued on Next Page…

Table 6 – Continued

Class	**Fold**	**Superfamily**
51349 c cl Alpha and beta proteins (a/b)	51350 c.1 cf TIM beta/alpha-barrel	51445 c.1.8 sf (Trans)glycosidases
	1734 c.2 cf NAD(P)-binding Rossmann-fold	51735 c.2.1 sf NAD(P)-binding Rossmann-fold domains
	52539 c.37 cf P-loop containing nuceoside triphosphate hydrolases	52540 c.37.1 sf P-loop containing nucleoside triphosphate hydrolases
	52832 c.47 cf Thioredoxin fold	52833 c.47.1 sf Thioredoxin-like
	52979 c.52 cf Restriction endonuclease-like	52980 c.52.1 sf Restriction endonuclease-like
	53334 c.66 cf S-adenosyl-L-methionine-dependent methyltransferases	53335 c.66.1 sf S-adenosyl-L-methionine-dependent methyltransferases
	53382 c.67 cf PLP-dependent transferases	53383 c.67.1 sf PLP-dependent transferases
	53447 c.68 cf Nucleotide-diphospho-sugar transferases	53448 c.68.1 sf Nucleotide-diphospho-sugar transferases
	53473 c.69 cf alpha/beta-Hydrolases	53474 c.69.1 sf alpha/beta-Hydrolases
	53849 c.94 cf Periplasmic binding protein-like II	53850 c.94.1 sf Periplasmic binding protein-like II
53931 d cl Alpha and beta proteins (a+b)	55728 d.108 cf Acyl-CoA N-acyltransferases Nat)	55729 d.108.1 sf Acyl-CoA N-acyltransferases (Nat)

Continued on Next Page. . .

Table 6 – Continued

Class	**Fold**	**Superfamily**
56992 g cl Small proteins	57015 g.3 cf Knottins (small inhibitors, toxins, ectins)	57059 g.3.6 sf omega toxin-like
		57095 g.3.7 sf Scorpion toxin-like

Appendix C. Weka classifier lineup

Table 7: The lineup of classifiers and configurations chosen to run as a batch job on the clustered implementation of Weka. This lineup was used to identify the best classifier configuration for predicting SCOP superfamily.

Classifier
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0'
weka.classifiers.bayes.BayesNet – -D -Q weka.classifiers.bayes.net.search.local.K2 – -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator – -A 0.5
weka.classifiers.bayes.BayesNet – -D -Q weka.classifiers.bayes.net.search.local.HillClimber – -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator – -A 0.5
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 2.0'
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 4.0'
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01'
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.1'
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.05'
weka.classifiers.functions.SMO – -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K 'weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.001'
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 30 -W weka.classifiers.rules.PART

Continued on Next Page...

Table 7 – Continued

Classifier
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 50 -W weka.classifiers.rules.PART
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 30 -W weka.classifiers.trees.J48
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 50 -W weka.classifiers.trees.J48
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.REPTree
weka.classifiers.bayes.BayesNet – -D -Q weka.classifiers.bayes.net.search.local.RepeatedHillClimber – -U 10 -A 1 -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator – -A 0.5
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 30 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 50 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 107 -S 1
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 87 -S 1
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 67 -S 1
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART

Continued on Next Page. . .

Table 7 – Continued

Classifier
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 30 -W weka.classifiers.rules.PART
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 50 -W weka.classifiers.rules.PART
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 30 -W weka.classifiers.trees.J48
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 50 -W weka.classifiers.trees.J48
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 10 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 30 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.DataNearBalancedND – -S 1 -W weka.classifiers.meta.Bagging – -P 100 -S 1 -I 50 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 0 -S 1
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48 – -C 0.25 -M 2
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART – -M 2 -C 0.25 -Q 1
weka.classifiers.bayes.BayesNet – -D -Q weka.classifiers.bayes.net.search.local.TabuSearch – -L 5 -U 10 -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator – -A 0.5

Continued on Next Page. . .

Table 7 – Continued

Classifier
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.REPTree – -M 2 -V 0.0010 -N 3 -S 1 -L -1
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.rules.PART – -M 2 -C 0.25 -Q 1
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 30 -W weka.classifiers.rules.PART – -M 2 -C 0.25 -Q 1
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 50 -W weka.classifiers.rules.PART – -M 2 -C 0.25 -Q 1
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.J48
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 30 -W weka.classifiers.trees.J48
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 50 -W weka.classifiers.trees.J48
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 30 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 50 -W weka.classifiers.trees.REPTree
weka.classifiers.meta.END – -S 1 -I 10 -W weka.classifiers.meta.nestedDichotomies.ClassBalancedND – -S 1 -W weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.SimpleCart – -S 1 -M 2.0 -N 5 -C 1.0
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 107 -S 1
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 87 -S 1
weka.classifiers.meta.AdaBoostM1 – -P 100 -S 1 -I 10 -W weka.classifiers.trees.RandomForest – -I 10 -K 67 -S 1
weka.classifiers.functions.LibSVM – -S 0 -K 0 -D 1 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B

Continued on Next Page. . .

Table 7 – Continued

Classifier
weka.classifiers.functions.LibSVM – -S 0 -K 0 -D 1 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 0 -D 1 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 0.1 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 0 -D 1 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 10.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 1 -D 2 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 1 -D 2 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 0.1 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 1 -D 2 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 10.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 1 -D 4 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 2 -D 1 -G 0.005 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 2 -D 1 -G 0.001 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 2 -D 1 -G 0.01 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.LibSVM – -S 0 -K 2 -D 1 -G 0.1 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B
weka.classifiers.functions.MultilayerPerceptron – -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
weka.classifiers.functions.MultilayerPerceptron – -L 0.6 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
weka.classifiers.functions.MultilayerPerceptron – -L 0.8 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
weka.classifiers.functions.RBFNetwork – -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1
weka.classifiers.functions.RBFNetwork – -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.3

Continued on Next Page. . .

Table 7 – Continued

Classifier
weka.classifiers.functions.RBFNetwork – -B 4 -S 1 -R 1.0E-8 -M -1 -W 0.1
weka.classifiers.functions.RBFNetwork – -B 4 -S 1 -R 1.0E-8 -M -1 -W 0.3
weka.classifiers.functions.SimpleLogistic – -I 0 -M 500 -H 50 -W 0.0
weka.classifiers.functions.SimpleLogistic – -I 0 -M 500 -H 50 -W 0.0 -A
weka.classifiers.trees.LMT – -I -1 -M 5 -W 0.0
weka.classifiers.trees.LMT – -I -1 -M 5 -W 0.0 -A
weka.classifiers.bayes.BayesNet – -D -Q weka.classifiers.bayes.net.search.local.TAN – -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator – -A 0.5
weka.classifiers.bayes.NaiveBayes
weka.classifiers.lazy.IB1
weka.classifiers.lazy.IBk
weka.classifiers.lazy.KStar
weka.classifiers.lazy.LWL
weka.classifiers.misc.HyperPipes
weka.classifiers.rules.ConjunctiveRule
weka.classifiers.rules.DecisionTable
weka.classifiers.rules.JRip
weka.classifiers.rules.NNge
weka.classifiers.rules.OneR
weka.classifiers.rules.PART
weka.classifiers.trees.SimpleCart – -S 1 -M 2.0 -N 5 -C 1.0
weka.classifiers.rules.ZeroR
weka.classifiers.trees.DecisionStump
weka.classifiers.trees.J48
weka.classifiers.trees.REPTree
weka.classifiers.trees.RandomForest
weka.classifiers.trees.RandomTree

Continued on Next Page. . .

Table 7 – Continued

Classifier
weka.classifiers.bayes.ComplementNaiveBayes
weka.classifiers.bayes.NaiveBayesMultinomial
weka.classifiers.misc.FLR
weka.classifiers.trees.NBTree

Bibliography

-omes and -omics glossary taxonomy (2009). `http://www.genomicglossaries.com/content/omes.asp`.

Adamcsek, B., Palla, G., Farkas, I.J., Derenyi, I. & Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**, 1021–3.

Adie, E., Adams, R., Evans, K., Porteous, D. & Pickard, B. (2006). SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–4.

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.

Al-Shahib, A. (2005). *Addressing the Core Challenges in Predicting Protein Function from Sequence Using Machine Learning*. Ph.D. thesis, Department of Computing Science, University of Glasgow.

Al-Shahib, A., Breitling, R. & Gilbert, D. (2005). Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl Bioinformatics*, **4**, 195–203.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). *Molecular Biology of the Cell, Fourth Edition*. Garland.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J. & Zdobnov, E.M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, **29**, 37–40.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. & Yeh, L.S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, **32**, D115–9.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–9.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*, **31**, 400–2.

Bader, G., Betel, D. & Hogue, C. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, **31**, 248–50.

Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T. & Hogue, C.W. (2001). BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res*, **29**, 242–5.

Bao, L. & Cui, Y. (2005). Prediction of the phenotypic effects of nonsynonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*.

Barabasi, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–12.

Barandela, R., Snchez, J.S., Garca, V. & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, **36**, 849 – 851.

Barker, W.C., Garavelli, J.S., McGarvey, P.B., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S., Ledley, R.S., Mewes, H.W., Pfeiffer, F., Tsugita, A. & Wu, C. (1999). The PIR-International Protein Sequence Database. *Nucleic Acids Res*, **27**, 39–43.

Barrett, J., Fry, B., Maller, J. & Daly, M. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–5.

Batagelj, V. & Mrvar, A. (1998). Pajek – program for large network analysis.

Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., Studholme, D., Yeats, C. & Eddy, S. (2004). The Pfam protein families database. *Nucleic Acids Res*, **32**, D138–41.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**, 235–42.

Black SD, M.D. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal Biochem*, **193**, 72–82.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**, 365–70.

Botstein, D. & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, **33 Suppl**, 228–37.

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O. & Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol*, **5**, R35.

Branden, C.I. & Tooze, J. (1999). *Introduction to Protein Structure*. Garland Publishing.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32.

Breitkreutz, B.J., Stark, C. & Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol*, **4**, R22.

Brenner, S.E., Koehl, P. & Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, **28**, 254–6.

Brown, K. & Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, **21**, 2076–82.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, **97**, 262–7.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. & Chen, Y.Z. (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res*, **31**, 3692–7.

Camargo, A. & Azuaje, F. (2007). Linking gene expression and functional network data in human heart failure. *PLoS ONE*, **2**, e1347.

Camargo, A. & Azuaje, F. (2008). Identification of dilated cardiomyopathy signature genes through gene expression and network data integration. *Genomics*.

Care, M.A., Needham, C.J., Bulpitt, A.J. & Westhead, D.R. (2007). Deleterious SNP prediction: be mindful of your training data! *Bioinformatics*, **23**, 664–72.

Cary, M.P., Bader, G.D. & Sander, C. (2005). Pathway information for systems biology. *FEBS Lett*, **579**, 1815–20.

Casbon, J., Crooks, G. & Saqi, M. (2006). A high level interface to scop and astral implemented in python. *BMC Bioinformatics*, **7**, 10+.

Cavallo, A. & Martin, A. (2005). Mapping SNPs to protein sequence and structure data. *Bioinformatics*, **21**, 1443–50.

Chang, C.C. & Lin, C.J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chasman, D. & Adams, R.M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*, **307**, 683–706.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. & Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res*, **35**, D572–4.

Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E.E. & Futschik, M.E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, **35**, D590–4.

Chen, C., Liaw, A. & Breiman, L. (2004). Using random forest to learn imbalanced data. Tech. Rep. 666, Department of Statistics, University of California, Berkeley, http://www.stat.berkeley.edu/tech-reports/666.pdf.

Chen, J., Anderson, J., DeWeese-Scott, C., Fedorova, N., Geer, L., He, S., Hurwitz, D., Jackson, J., Jacobs, A., Lanczycki, C., Liebert, C., Liu, C., Madej, T., Marchler-Bauer, A., Marchler, G., Mazumder, R., Nikolskaya, A., Rao, B., Panchenko, A., Shoemaker, B., Simonyan, V., Song, J., Thiessen, P., Vasudevan, S., Wang, Y., Yamashita, R., Yin, J. & Bryant, S. (2003). MMDB: Entrez's 3D-structure database. *Nucleic Acids Res*, **31**, 474–7.

Chen, J., Shen, C. & Sivachenko, A. (2006). Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput*, 367–78.

Chen, K.C., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B. & Tyson, J.J. (2000). Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell*, **11**, 369–91.

Cheng, T.M.K., Lu, Y.E., Vendruscolo, M., Lio', P. & Blundell, T.L. (2008). Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol*, **4**, e1000135.

Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jr, Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T., Jr & Roccella, E.J. (2003). Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*, **42**, 1206–52.

Christopher Newton-Cheh, Toby Johnson, Vesela Gateva, Martin D Tobin, Murielle Bochud, Lachlan Coin, Samer S Najjar, Jing Hua Zhao, Simon C Heath, Susana Eyheramendy, Konstantinos Papadakis, Benjamin F Voight, Laura J Scott, Feng Zhang, Martin Farrall, Toshiko Tanaka, Chris Wallace, John C Chambers, Kay-Tee Khaw, Peter Nilsson, Pim van der Harst, Silvia Polidoro, Diederick E Grobbee, N Charlotte Onland-Moret, Michiel L Bots, Louise V Wain, Katherine S Elliott, Alexander Teumer, Jian'an Luan, Gavin Lucas, Johanna Kuusisto, Paul R Burton, David Hadley, Wendy L McArdle, Wellcome Trust Case Control Consortium, Morris Brown, Anna Dominiczak, Stephen J Newhouse, Nilesh J Samani, John Webster, Eleftheria Zeggini, Jacques S Beckmann, Sven Bergmann, Noha Lim, Kijoung Song, Peter Vollenweider, Gerard Waeber, Dawn M Waterworth, Xin Yuan, Leif Groop, Marju Orho-Melander, Alessandra Allione, Alessandra Di Gregorio, Simonetta Guarrera, Salvatore Panico, Fulvio Ricceri, Valeria Romanazzi, Carlotta Sacerdote, Paolo Vineis, Inês Barroso, Manjinder S Sandhu, Robert N Luben, Gabriel J. Crawford, Pekka Jousilahti, Markus Perola, Michael Boehnke, Lori L Bonnycastle, Francis S Collins, Anne U Jackson, Karen L Mohlke, Heather M Stringham, Timo T Valle, Cristen J Willer, Richard N Bergman, Mario A Morken, Angela Döring, Christian Gieger, Thomas Illig, Thomas Meitinger, Elin Org, Arne Pfeufer, H Erich Wichmann, Sekar Kathiresan, Jaume Marrugat, Christopher J O'Donnell, Stephen M Schwartz, David S Siscovick, Isaac Subirana, Nelson B Freimer, Anna-Liisa Hartikainen, Mark I McCarthy, Paul F O'Reilly, Leena Peltonen, Anneli Pouta, Paul E de Jong, Harold Snieder, Wiek H van Gilst, Robert Clarke, Anuj Goel, Anders Hamsten, John F Peden, Udo Seedorf, Ann-Christine Syvänen, Giovanni Tognoni, Edward G Lakatta, Serena Sanna, Paul

Scheet, David Schlessinger, Angelo Scuteri, Marcus Dörr, Florian Ernst, Stephan B Felix, Georg Homuth, Roberto Lorbeer, Thorsten Reffelmann, Rainer Rettig, Uwe Völker, Pilar Galan, Ivo G Gut, Serge Hercberg, G Mark Lathrop, Diana Zeleneka, Panos Deloukas, Nicole Soranzo, Frances M Williams, Guangju Zhai, Veikko Salomaa, Markku Laakso, Roberto Elosua, Nita G Forouhi, Henry Völzke, Cuno S Uiterwaal, Yvonne T van der Schouw, Mattijs E Numans, Giuseppe Matullo, Gerjan Navis, Göran Berglund, Sheila A Bingham, Jaspal S Kooner, Andrew D Paterson, John M Connell, Stefania Bandinelli, Luigi Ferrucci, Hugh Watkins, Tim D Spector, Jaakko Tuomilehto, David Altshuler, David P Strachan, Maris Laan, Pierre Meneton, Nicholas J Wareham, Manuela Uda, Marjo-Riitta Jarvelin, Vincent Mooser, Olle Melander, Ruth JF Loos, Paul Elliott, Gonçalo R Abecasis, Mark Caulfield & Patricia B Munroe (2009). Eight blood pressure loci identified by genome-wide association study of 34,433 people of European ancestry. *Nature Genetics*, in Press.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, **3**, 140.

Clare, A. & King, R.D. (2003). Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics*, **19 Suppl 2**, II42–II49.

Clare, A., Karwath, A., Ougham, H. & King, R.D. (2006). Functional bioinformatics for Arabidopsis thaliana. *Bioinformatics*, **22**, 1130–6.

Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–45.

Corte, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, **20**, 273–297.

Damoulas, T. & Girolami, M.A. (2008). Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, **24**, 1264–70.

Dayhoff, M., Schwartz, R. & Orcutt, B. (1978). Atlas of Protein Sequence and Structure. *National Biomedical Research Foundation, Washington, DC*, **5**, 345–348.

de Bakker, P., Yelensky, R., Pe'er, I., Gabriel, S., Daly, M. & Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet*, **37**, 1217–23.

Dietterich, T.G. & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, **2**, 263–286.

Dijkstra, E.W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.

Ding, C. & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–58.

Dobson, P.D. & Doig, A.J. (2003). Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol*, **330**, 771–83.

Dobson, R., Munroe, P., Mein, C., Caulfield, M. & Saqi, M. (2008). Combining protein-protein interaction (PPI) network and sequence attributes for predicting hypertension related proteins. In *Proceedings Bioinformatics Research and Development, BIRD 2008. In Press*.

Dondoshansky, I. (2002). Blastclust (NCBI Software Development Toolkit), 6.1 edition. *NCBI, Bethesda, MD.*.

Dong, L., Frank, E. & Kramer, S. (2005). Ensembles of balanced nested dichotomies for multi-class problems. 84–95.

Dowell, R., Jokerst, R., Day, A., Eddy, S. & Stein, L. (2001). The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Dubchak, I., Muchnik, I., Holbrook, S.R. & Kim, S.H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A*, **92**, 8700–4.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.

Efron, B. & Tibshirani, R. (1993).

EL-Manzalawy, Y. & Honavar, V. (2005). *WLSVM: Integrating LibSVM into Weka Environment*. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

Fawcett, T. & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, **1**, 291–316.

Fletterick, R.J. (1992). Introduction to protein structure, by carl branden and john tooze. new york: Garland publishing company, 302 pages, $27.95 (paper), 1991. *Proteins: Structure, Function, and Genetics*, **12**, 200+.

Frank, E. & Witten, I.H. (1998a). Generating accurate rule sets without global optimization. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, 144–151, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Frank, E. & Witten, I.H. (1998b). Generating accurate rule sets without global optimization. In *Proc. 15th International Conf. on Machine Learning*, 144–151, Morgan Kaufmann, San Francisco, CA.

Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. (2004). Data mining in bioinformatics using weka. *Bioinformatics*, **20**, 2479–2481.

Fredman, D., Munns, G., Rios, D., Sjoholm, F., Siegfried, M., Lenhard, B., Lehvaslaiho, H. & Brookes, A. (2004). HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res*, **32**, D516–9.

Freund, Y. & Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 148–156.

Fürnkranz, J. (2002). Round robin classification. *J. Mach. Learn. Res.*, **2**, 721–747.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M.R. (2004). A census of human cancer genes. *Nat Rev Cancer*, **4**, 177–83.

Galperin, M.Y. (2007). The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res*, **35**, D3–4.

George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D. & Wouters, M.A. (2006). Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*, **34**, e130.

Gerlt, J.A., Babbitt, P.C. & Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Archives of Biochemistry and Biophysics*, **433**, 59 – 70, highlight issue on Enzyme Mechanisms.

Girvan, M. & Newman, M.E. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, **99**, 7821–6.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. & Barabasi, A.L. (2007). The human disease network. *Proc Natl Acad Sci U S A*, **104**, 8685–90.

Goni, J., Esteban, F.J., de Mendizabal, N.V., Sepulcre, J., Ardanza-Trevijano, S., Agirrezabal, I. & Villoslada, P. (2008). A computational analysis of protein-protein interaction networks in neurodegenerative diseases. *BMC Syst Biol*, **2**, 52.

Guimerà, R. & Amaral, L.A.N. (2004). Modeling the world-wide airport network. *European Physical Journal B*, **38**, 381–385.

Hall, M. (1998). Correlation-based feature selection for machine learning.

Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D. & McKusick, V.A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, **30**, 52–5.

Han, L., Cai, C., Ji, Z., Cao, Z., Cui, J. & Chen, Y. (2004). Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res*, **32**, 6437–44.

Hand, D., Manilla, H. & Smythe, P. (2001). *Principles of Data Mining*. MIT press.

HapMap (2003). The International HapMap Project. *Nature*, **426**, 789–96.

Hastie, T. & Tibshirani, R. (1998). Classification by pairwise coupling. *Proceedings of the 1997 conference on Advances in neural information processing systems 10*, 507 – 513.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*, **288**, 147–64.

Henikoff, J.G. & Henikoff, S. (1996). Blocks database and its applications. *Methods Enzymol*, **266**, 88–105.

Higgins, D.G., Bleasby, A.J. & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci*, **8**, 189–91.

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. *Mach. Learn.*, **11**, 63–90.

Hu, J. & Yan, C. (2008). Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information. *BMC Bioinformatics*, **9**, 297.

Hu, Z., Snitkin, E.S. & DeLisi, C. (2008). VisANT: an integrative framework for networks in systems biology. *Brief Bioinform*, **9**, 317–25.

Huang, T.W., Tien, A.C., Huang, W.S., Lee, Y.C., Peng, C.L., Tseng, H.H., Kao, C.Y. & Huang, C.Y. (2004). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273–6.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Res*, **30**, 38–41.

Hwang, S., Son, S.W., Kim, S.C., Kim, Y.J., Jeong, H. & Lee, D. (2008). A protein interaction network associated with asthma. *J Theor Biol*, **252**, 722–31.

Imbert, A., Chaffanet, M., Essioux, L., Noguchi, T., Adelaide, J., Kerangueven, F., Le Paslier, D., Bonaiti-Pellie, C., Sobol, H., Birnbaum, D. & Pebusque, M.J. (1996). Integrated map of the chromosome 8p12-p21 region, a region involved in human cancers and Werner syndrome. *Genomics*, **32**, 29–38.

IUBMB (1992). *Enzyme nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Academic Press, New York.

Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., Valencia, A. & Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol*, **319**, 1257–65.

Jonsson, P.F. & Bates, P.A. (2006a). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**, 2291–7.

Jonsson, P.F. & Bates, P.A. (2006b). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**, 2291–2297.

Joy, M.P., Brock, A., Ingber, D.E. & Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol*, **2005**, 96–103.

Junker, B.H. & Schreiber, F. (2008). *Analysis of Biological Networks (Wiley Series in Bioinformatics)*. Wiley-Interscience.

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**, D354–7.

Karchin, R., Diekhans, M., Kelly, L., Thomas, D., Pieper, U., Eswar, N., Haussler, D. & Sali, A. (2005). LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L. & Sander, C. (1997). Predicting protein structure using hidden Markov models. *Proteins*, **Suppl 1**, 134–9.

Katagiri, F. (2003). Attacking complex problems with the power of systems biology. *Plant Physiol*, **132**, 417–9.

Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., Price, T.S., Glessner, J.T., Galver, L., Barrett, J.C., Grant, S.F., Farlow, D.N., Chandrupatla, H.R., Hansen, M., Ajmal, S., Papanicolaou, G.J., Guo, Y., Li, M., Derohannessian, S., de Bakker, P.I., Bailey, S.D., Montpetit, A., Edmondson, A.C., Taylor, K., Gai, X., Wang, S.S., Fornage, M., Shaikh, T., Groop, L., Boehnke, M., Hall, A.S., Hattersley, A.T., Frackelton, E., Patterson, N., Chiang, C.W., Kim, C.E., Fabsitz, R.R., Ouwehand, W., Price, A.L., Munroe, P., Caulfield, M., Drake, T., Boerwinkle, E., Reich, D., Whitehead, A.S., Cappola, T.P., Samani, N.J., Lusis, A.J., Schadt, E., Wilson, J.G., Koenig, W., McCarthy, M.I., Kathiresan, S., Gabriel, S.B., Hakonarson, H., Anand, S.S., Reilly, M., Engert, J.C., Nickerson, D.A., Rader, D.J., Hirschhorn, J.N. & Fitzgerald, G.A. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS ONE*, **3**, e3583.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. & Hermjakob, H. (2007). IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*, **35**, D561–5.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. 1137–1143, Morgan Kaufmann.

Krishnan, V.G. & 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, evaluation Studies.

Kruglyak, L. & Nickerson, D. (2001). Variation is the spice of life. *Nat Genet*, **27**, 234–6.

Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105–32.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum & et al (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lin, C.Y., Lin, K.L., Huang, C.D., Chang, H.M., Yang, C.Y., Lin, C.T., Tang, C.Y. & Hsu, D.F. (2005). Feature selection and combination criteria for improving predictive accuracy in protein structure classification. In *BIBE '05: Proceedings of the Fifth IEEE Symposium on Bioinformatics and Bioengineering*, 311–315, IEEE Computer Society, Washington, DC, USA.

Lopez-Bigas, N. & Ouzounis, C. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, **32**, 3108–14.

Loscalzo, J., Kohane, I. & Barabasi, A.L. (2007). Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol*, **3**, 124.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–12.

Mathew, J.P., Taylor, B.S., Bader, G.D., Pyarajan, S., Antoniotti, M., Chinnaiyan, A.M., Sander, C., Burakoff, S.J. & Mishra, B. (2007). From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Comput Biol*, **3**, e12.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, **405**, 442–51.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P. & Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, **9**, 356–69.

McGuffin, L.J., Bryson, K. & Jones, D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–5.

Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J. & DeLisi, C. (2002). Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, **30**, 306–9.

Melvin, I., Ie, E., Kuang, R., Weston, J., Stafford, W.N. & Leslie, C. (2007). SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics*, **8 Suppl 4**, S2.

Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. & Ruepp, A. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, **32**, D41–4.

Michalski, R.S., Carbonell, J.G. & Mitchell, T.M. (1983). *Machine Learning: An Artificial Intelligence Approach*. Tioga Publishing Company.

Miller S, L.A.C.C., Janin J (1987). Interior and surface of monomeric proteins. *J Mol Biol*, **196**, 641–56.

Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K.S., Sharma, S., Chandrika, K.N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H.G., Nagini, M., Kumar, G.S., Jose, R., Deepthi, P., Mohan, S.S., Gandhi, T.K., Harsha, H.C., Deshpande, K.S., Sarker, M., Prasad, T.S. & Pandey, A. (2006). Human protein reference database–2006 update. *Nucleic Acids Res*, **34**, D411–4.

Mottagui-Tabar, S., Faghihi, M.A., Mizuno, Y., Engstrom, P.G., Lenhard, B., Wasserman, W.W. & Wahlestedt, C. (2005). Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC Genomics*, **6**, 18.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J., Silventoinen, V., Studholme, D.J., Vaughan,

R. & Wu, C.H. (2005). InterPro, progress and status in 2005. *Nucleic Acids Res*, **33**, D201–5.

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536–40.

Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443–53.

Ng, P. & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, **31**, 3812–4.

Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L. & Venter, J.C. (2008). Genetic variation in an individual human exome. *PLoS Genet*, **4**, e1000160.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. & Thornton, J. (1997). Cath - a hierarchic classification of protein domain structures. *Structure*, **5**, 1093 – 1108.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W., Ruepp, A. & Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–4.

Pavlopoulos, G.A., Gap, Wegener, A.L., Aw & Schneider, R., Rs (2008). A survey of visualization tools for biological network analysis. *BioData Min*, **1**, 12.

Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. (2005). G2D: a tool for mining genes associated with disease. *BMC Genet*, **6**, 45.

Platt, J. (1998). Machines using sequential minimal optimization. In B. Schoelkopf, C. Burges & A. Smola, eds., *Advances in Kernel Methods - Support Vector Learning*, MIT Press.

Prieto, C. & De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, **34**, W298–302.

Pruitt, K.D., Tatusova, T. & Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61–5.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Quinlan JR (1993). *C4. 5: Programs for Machine Learning*. Morgan Kaufmann.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*, **30**, 3894–3900.

Reich, D., Gabriel, S. & Altshuler, D. (2003). Quality and completeness of SNP databases. *Nat Genet*, **33**, 457–8.

Riley, M. (1993). Functions of the gene products of Escherichia coli. *Microbiol Rev*, **57**, 862–952.

Roberts, M. & King, T. (1987). *Biology: A Functional Approach*. Nelson Thornes,.

Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–26.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–8.

Sadreyev, R. & Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, **326**, 317–36.

Sam, L., Liu, Y., Li, J., Friedman, C. & Lussier, Y.A. (2007). Discovery of protein interaction networks shared by diseases. *Pac Symp Biocomput*, 76–87.

Sander, C. & Schneider, R. (1993). The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res*, **21**, 3105–9.

Saunders, C. & Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol*, **322**, 891–901.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. & Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Brief Bioinform*, **3**, 246–51.

Shamim, M.T., Anwaruddin, M. & Nagarajaram, H.A. (2007). Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics*, **23**, 3320–7.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498–504.

Shannon CE (1948). Mathematical theory of communication. *Bell System Tech*.

Shen, H.B. & Chou, K.C. (2006). Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **22**, 1717–22.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–11.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, **3**, 265–74.

Sillitoe, I., Dibley, M., Bray, J., Addou, S. & Orengo, C. (2005). Assessing strategies for improved superfamily recognition. *Protein Sci*, **14**, 1800–10.

Smith, T.F., Waterman, M.S. & Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucleic Acids Res*, **13**, 645–56.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–60.

Sonnhammer, E.L., Eddy, S.R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–20.

Spencer, G. (2008). 1000 Genomes Project.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, **34**, D535–9.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. & Wanker, E.E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–68.

Stenson, P., Ball, E., Mort, M., Phillips, A., Shiel, J., Thomas, N., Abeysinghe, S., Krawczak, M. & Cooper, D. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*, **21**, 577–81.

Stenson, P.D., Ball, E., Howells, K., Phillips, A., Mort, M. & Cooper, D.N. (2008). Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet*, **45**, 124–6.

Stitziel, N., Binkowski, T., Tseng, Y., Kasif, S. & Liang, J. (2004). topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res*, **32**, D520–2.

Sunyaev, S., Ramensky, V. & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, **16**, 198–200.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A.S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum Mol Genet*, **10**, 591–7.

Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O. & Abola, E.E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr*, **54**, 1078–84.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*, **13**, 2129–41.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80.

Tian, J., Wu, N., Guo, X., Guo, J., Zhang, J. & Fan, Y. (2007). Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics*, **8**, 450.

Tiffin, N., Kelso, J., Powell, A., Pan, H., Bajic, V. & Hide, W. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*, **33**, 1544–52.

Torkamani, A. & Schork, N.J. (2008). Predicting functional regulatory polymorphisms. *Bioinformatics*, **24**, 1787–92.

Tsai J, C.C.G.M., Taylor R (1999). The packing density in proteins: standard radii and volumes. *J Mol Biol*, **290**, 253–66.

van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. & Leunissen, J.A. (2006). A text-mining analysis of the human phenome. *Eur J Hum Genet*, **14**, 535–42.

van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K., Feighery, C., Jewell, D., Kelleher, D., Kumar, P., Travis, S., Walters, J.R., Sanders, D.S., Howdle, P., Swift, J., Playford, R.J., McLaren, W.M., Mearin, M.L., Mulder, C.J., McManus, R., McGinnis, R., Cardon, L.R., Deloukas, P. & Wijmenga, C. (2007). A genome-wide association study

for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet*, **39**, 827–9.

Venter, J.C., Adams, M.D. & Myers, E.W.e.a. (2001). The sequence of the human genome. *Science*, **291**, 1304–51.

Vitkup, D., Sander, C. & Church, G. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biol*, **4**, R72.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B. & Bork, P. (2007). STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, **35**, D358–62.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. & Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–6.

Wang, J., Du, Z., Payattakool, R., Yu, P. & Chen, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–81.

Wang, P., Dai, M., Xuan, W., McEachin, R.C., Jackson, A.U., Scott, L.J., Athey, B., Watson, S.J. & Meng, F. (2006). SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, **22**, e523–9.

Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat*, **17**, 263–270.

Watts, D. & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–2.

Weiss, G. & Provost, F. (2001). The Effect of Class Distribution on Classifier Learning: An Empirical Study. *Technical Report ML-TR-44*, **Deptartment of Computer Science, University of Glasgow**.

Wilson, C., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, **297**, 233–49.

Witten, I. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–78.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N. & Suzek, B. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, **34**, D187–91.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res*, **28**, 289–91.

Xu, J. & Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, **22**, 2800–5.

Yip, Y.L., Scheib, H., Diemand, A.V., Gattiker, A., Famiglietti, L.M., Gasteiger, E. & Bairoch, A. (2004). The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum Mutat*, **23**, 464–470.

Yona, G. & Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, **315**, 1257–75.

Zdobnov, E.M. & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–8.

Index