Human Protein Function Prediction: application of machine learning for integration of heterogeneous data sources

Anna Lobley

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Biochemistry

University College London

July 2009

I, Anna Lobley, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Experimental characterisation of protein cellular function can be prohibitively expensive and take years to complete. To address this problem, this thesis focuses on the development of computational approaches to predict function from sequence. For sequences with well characterised close relatives, annotation is trivial, orphans or distant homologues present a greater challenge. The use of a feature based method employing ensemble support vector machines to predict individual Gene Ontology classes is investigated. It is found that different combinations of feature inputs are required to recognise different functions. Although the approach is applicable to any human protein sequence, it is restricted to broadly descriptive functions. The method is well suited to prioritisation of candidate functions for novel proteins rather than to make highly accurate class assignments.

Signatures of common function can be derived from different biological characteristics; interactions and binding events as well as expression behaviour. To investigate the hypothesis that common function can be derived from expression information, public domain human microarray datasets are assembled. The questions of how best to integrate these datasets and derive features that are useful in function prediction are addressed. Both co-expression and abundance information is represented between and within experiments and investigated for correlation with function. It is found that features derived from expression data serve as a weak but significant signal for recognising functions. This signal is stronger for biological processes than molecular function categories and independent of homology information.

The protein domain has historically been coined as a modular evolutionary unit of protein function. The occurrence of domains that can be linked by ancestral fusion events serves as a signal for domain-domain interactions. To exploit this information for function prediction, novel domain architecture and fused architecture scores are developed. Architecture scores rather than single domain scores correlate more strongly with function, and both architecture and fusion scores correlate more strongly with molecular functions than biological processes.

The final study details the development of a novel heterogeneous function prediction approach designed to target the annotation of both homologous and non-homologous proteins. Support vector regression is used to combine pair-wise sequence features with expression scores and domain architecture scores to rank protein pairs in terms of their functional similarities. The target of the regression models represents the continuum of protein function space empirically derived from the Gene Ontology molecular function and biological process graphs. The merit and performance of the approach is demonstrated using homologous and non-homologous test datasets and significantly improves upon classical nearest neighbour annotation transfer by sequence methods. The final model represents a method that achieves a compromise between high specificity and sensitivity for all human proteins regardless of their homology status. It is expected that this strategy will allow for more comprehensive and accurate annotations of the human proteome.

Acknowledgements

First and foremost I would like to acknowledge my supervisors Prof. D T Jones and Prof. C A Orengo for their encouragement and excellent guidance throughout my project.

I would also like to thank members of the Jones and Orengo groups especially Drs S. Lise and O. Redfern for helpful discussions. I acknowledge both UCL and CS department computing services for allowing me to gobble many CPU hours, abuse high memory machines and regularly destroy their exquisite computing infrastructure in order to complete my work.

Lastly and most importantly, I acknowledge close friends, family (Dad especially for proof reading) and Dr. Richard Myers for his support and patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The importance of protein function

Proteins play a central role in defining the behaviour within all biological systems. They are the fundamental work-horse components of living organisms. Ignoring fat, the average human being is composed of approximately 20% protein by dry weight. Proteins participate in almost all essential life processes; metabolism, growth and repair processes are all carried out by proteins. The disruption of normal protein functioning leads to a variety of disease states including Alzheimer's, Parkinson's, cancers and type II diabetes.

Proteins are synthesized in the Endoplasmic Reticulum of cells and are transported to different compartments, or tissues, to carry out their function. One of the largest classes of proteins are enzymes responsible for the catalysis of over four thousand documented reactions (Bairoch 2000). Enzymes catalyse the chemical reactions that are responsible for metabolism, the generation of energy from food sources, DNA repair and DNA synthesis. They are frequently cytoplasmic proteins and specifically catalyse only one or two reactions each. Other cytoplasmic proteins transport materials or transmit signals through the interior of the cell to the nucleus.

Extracellular proteins (existing outside of the cell) include hormones, acting as chemical messengers transmitting signals from the brain and other organs to destination cells and tissues. Hormones frequently act at receptors; proteins that reside in the cell membrane to carry out their functions. Signalling receptors transmit messages intracellularly by responding to chemical and biological stimulus (often ligand binding) at the cell surface. The messages are internalised, triggering secondary events that involve subsequent release or binding of another protein inside the cell. Other types of receptor (channels and pumps) transport ions into or out of the cell to maintain a balanced chemical environment. Antibodies, the effectors of the immune system are either extra-cellular proteins or can be tethered to the membranes of specialised blood cells. Here they mop up antigens (and foreign substances) targeting them for destruction. Another major class of proteins are structural proteins. The collagens and keratins that are found in skin, hair, teeth and bone. Many intracellular structures are composed of proteins, for example, ribosomes, the machinery responsible for protein synthesis are also composed of protein.

Each protein is a three-dimensional assembly of amino acids. Each amino acid is specified genetic material comprising triplet DNA codons. Interspersed regions of DNA that code for amino acids (exons) and are subsequently made into proteins comprise genes. Genes are first transcribed into an intermediate RNA molecule which is then synthesised into protein by the ribosome. The protein then adopts a stable three-dimensional structure and is transported to its site of action. It is both the quantity of protein in the cell dictated by the relative rates of transcription, translation and protein degradation at any one time, coupled with the subtle nature and diversity of interactions between proteins, DNA and small molecules that controls cellular behaviour and ultimately governs organism responses. Cataloguing the functions of proteins, the reactions they catalyse and the partners they interact with is therefore fundamental to furthering our understanding of physiological behaviour and offers valuable insights into the underlying mechanisms of disease.

1.2 The need for automated methods

The advent of high throughput DNA-sequencing technologies in the early 1980s enabled entire genome sequencing projects to be carried out rapidly and inexpensively. More recent third generation sequencing technologies are even higher throughput and less costly paving the way towards the 1,000 dollar human genome (Mardis 2006). Once raw sequence is obtained, automated techniques are required to identify the entire set of genes present in the organism; its genome. Subsequently the encoded proteins can be deciphered giving rise to the amino acid sequences that comprise the proteome.

Whole genome scale sequencing projects ensure that biological sequence databases continue to grow exponentially (Marshall 1995). In contrast, the numbers of completely functionally characterised sequences rises linearly (Baumgartner Jr. et al. 2007, Singh 2003). Clear and unambiguous functional characterisation of most sequences requires experimental validation. Crystal structure information can often provide clues to function by providing detailed 3 dimensional information about the fold of a protein. Bottlenecks in the process arise from the difficulty in obtaining sufficient quantities of pure and stable protein to form crystals. Other proteins do not express well *in vitro* particularly if they undergo post-translational modifications, or may exist in a disordered state requiring the presence of potentially unknown accessory proteins or ligands in order to fold or function. To observe the functions of these sequences, new experimental protocols must be developed, which can be labour intensive and prohibitively expensive.

The human genome sequence was completed in April 2003 some two years ahead of schedule (Pennisi 2003a). The number of genes present in the genome was estimated at around 30,000 although recent estimates are lower standing at around 25,000 (Pennisi 2003b). Ofran et al. (2005) estimated that of 2,000,000 known sequences, less than 25% were annotated to completion. Currently, there are approximately 2000 human protein sequences for which very little is known. For proteins with well characterised close relatives, it is trivial to infer function. Orphan proteins without discernible sequence relatives present a greater challenge. Here the task of experimental characterisation is blind and becomes unwieldy. It is highly unlikely that all known proteins will ever be completely experimentally characterised (Baumgartner Jr. et al. 2007). Thus there is a pressing need to develop fast and accurate computational approaches to fulfill this requirement.

1.3 Function annotation schemes

Fundamental to the task of annotation curation is a formalism of the concept of function. For a structural biologist it may constitute the fold of the protein, for a sequence analyst, its gene family and for a chemist the ligands or molecules that the protein can bind. Efforts to catalogue the functional repertoire of proteins include controlled vocabularies; Swissprot keywords (Bairoch and Apweiler 1996), the Gene Ontology (Ashburner et al. 2000), Multifun (Serres and Riley 2000), FunCat Functional Categories (Ruepp et al. 2004), and more recently the KEGG Brite Functional hierarchy, (Kanehisa et al. 2008). These classification schemes populate secondary biological knowledge-bases and permit high level analyses performed by grouping genes or proteins by functional class. The schemes emphasise different aspects of function, varying in specificity, coverage and simplicity of design. Each one attempts to provide a machine readable definition of function that can be exploited computationally in function prediction approaches.

The first attempts to formally describe the functions of proteins were Swissprot keywords. The keywords exist as free text labels assigned to one or more protein entities; ‘signal peptide’ or ‘kinase activity’ for example. The keyword system permits broad groupings of genes or proteins into biological pathways or by functional roles. Whilst offering flexibility, the relationships between keywords have not been defined. There is no means to interpret functional relatedness between annotations. The MultiFun hierarchy, MIPS (Munich Information Center for Protein Sequences) Functional Catalogue (FunCat), KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology and Gene Ontology Consortium represent more sophisticated approaches incorporating dependencies between protein functions in a controlled, machine readable format.

1.3.1 Multifun, FunCat and KEGG hierarchical schemes

The Multifun annotation system was the first function category hierarchy building on earlier work of Monica Riley characterising the E.Coli K12 genome (Riley 1993, Serres and Riley 2000). Function categories comprised 10 broad classes with finer sub-categories describing cellular roles for 66% of E. Coli genes. The MIPS Functional Catalogue (FunCat) employed a similar, but more detailed scheme based on the cellular roles of yeast proteins (Joshi et al. 2004). The initial hierarchy has now been extended and covers annotations for plants, higher eukaryotes and prokaryotes (Ruepp et al. 2004). Originally 28 broad categories described aspects of cellular metabolism and protein activity regulation. In the hierarchy specific annotations occupy sub-categories from the main branches (Table 1.1). The KEGG Brite system adopts a similar hierarchical structure for a series of descriptive schemes representing different aspects of function. The current system includes separate hierarchies for biological systems, pathway modules, human diseases and drug interactions. The KEGG Orthology (KO) groups proteins by their evolutionary history and provides a common unit used to navigate between hierarchies.

1.3.2 The Gene Ontology

The Gene Ontology (GO) (Ashburner et al. 2000) has become the de-facto standard in protein function annotation. The scheme represents the most complex and flexible annotation systems for describing protein functions. The annotation terms are modelled as a Directed Acyclic Graph (DAG) and provide a sophisticated model of functional complexity. Three independent contexts describe gene products; Molecular Function (MF), Biological Processes (BP) and Cellular Components (CC). The DAG structure is similar to a hierarchy. More specific child annotation terms inherit general annotations from their parents. Additional flexibility is incorporated by permit-

Table 1.1: Example of MIPS FunCat annotation scheme.

MIPS FunCat Functional Catalogue Entries
01 Metabolism
01.01 amino acid metabolism
01.02 nitrogen and sulphur metabolism
01.03 nucleotide metabolism
01.03.01 purine nucleotide metabolism
01.03.04 pyrimidine nucleotide metabolism
01.03.07 deoxyribonucleotide metabolism
01.03.10 metabolism of cyclic and unusual nucleotides
01.03.13 regulation of nucleotide metabolism
01.03.16 polynucleotide metabolism
01.03.16.01 RNA degradation
01.03.16.01 DNA degradation

MIPS FunCat scheme for Metabolism expanded for nucleotide and polynucleotide metabolism. Each level of the hierarchy is prescribed a two digit number and sub entry levels inherit the entry codes of their parents, separated by the dot notation ting annotation terms to share multiple parent-child relationships.

The MF graph describes the activity of the protein whereas the BP graph describes the pathway or cellular process in which proteins perform their role(s). The CC terms predominantly detail the localisation aspects of protein (see Figures 1.1, 1.2 and 1.3 for examples).

Molecular Function

The MF ontology formally describes the biochemical activities of a gene product and includes highly specific descriptions such as binding to ligands or structures (Ashburner et al. 2000). The annotation terms can be applied to a gene or its products alone, or may describe function(s) as part of a protein complex. The flexible DAG structure incorporates multiple annotations of a single protein (MCM4 in Figure 1.1) to different regions of the graph, emphasising different aspects of the proteins activity. The functional categories are defined for all organisms although some terms are lineage specific.

Biological Process

The BP ontology describes the biological objective to which the gene product contributes (Ashburner et al. 2000). Some of these terms correspond closely to MFs, or parts, or groups of MF categories. The protein MCM4 annotated with the activity of an ATP-dependent DNA helicase is further described as being involved in DNA dependent DNA replication, more specifically the DNA initiation, unwinding and pre-replicative parts of the process (Figure 1.2).

Cellular Components

CCs describe the part of a cell in which gene products perform their functional roles (Ashburner et al. 2000). Additionally, the component categories provide cellular protein complex information such as ‘interleukin-1 complex’ or ‘mRNA-editing complex’. Currently, some 679 terms describing complexes exist in the CC DAG. Proteins are dynamic cellular entities moving around within or between cells in order to function. The components graph captures this information appropriately for the MCM4 protein which forms part of the replication fork complex in the nucleoplasm and can also be found in the cytoplasm (Figure 1.3).

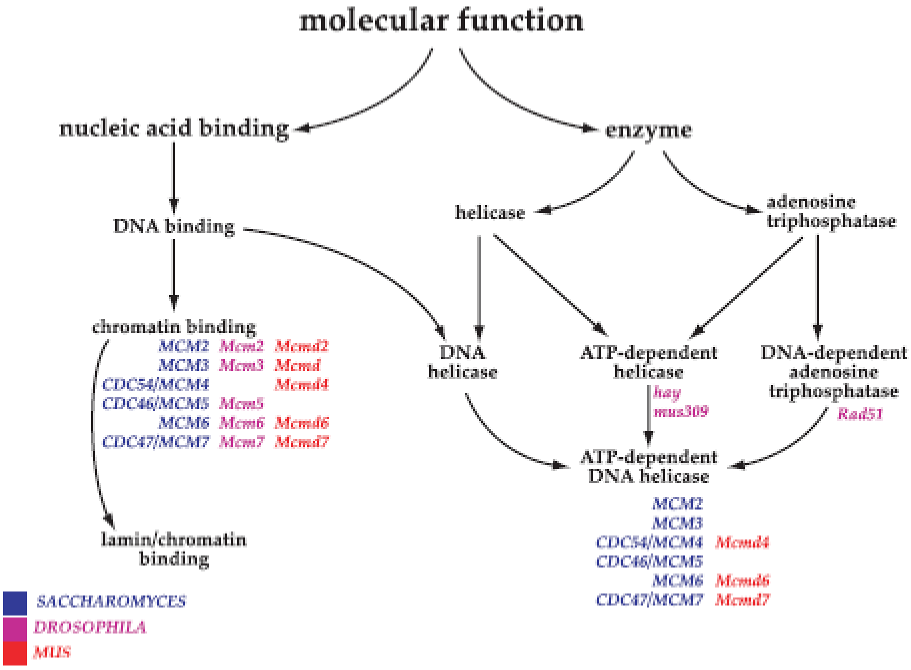


Figure 1.1: Example of Molecular Function graph taken from (Ashburner et al. 2000)

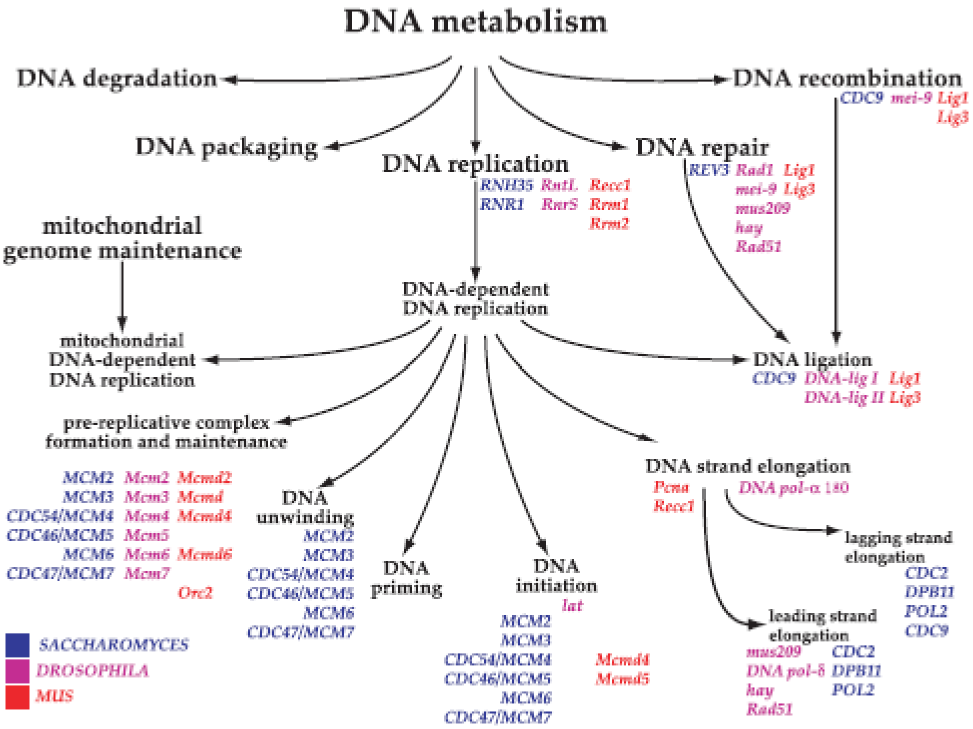


Figure 1.2: Example of Biological Process graph taken from (Ashburner et al. 2000)

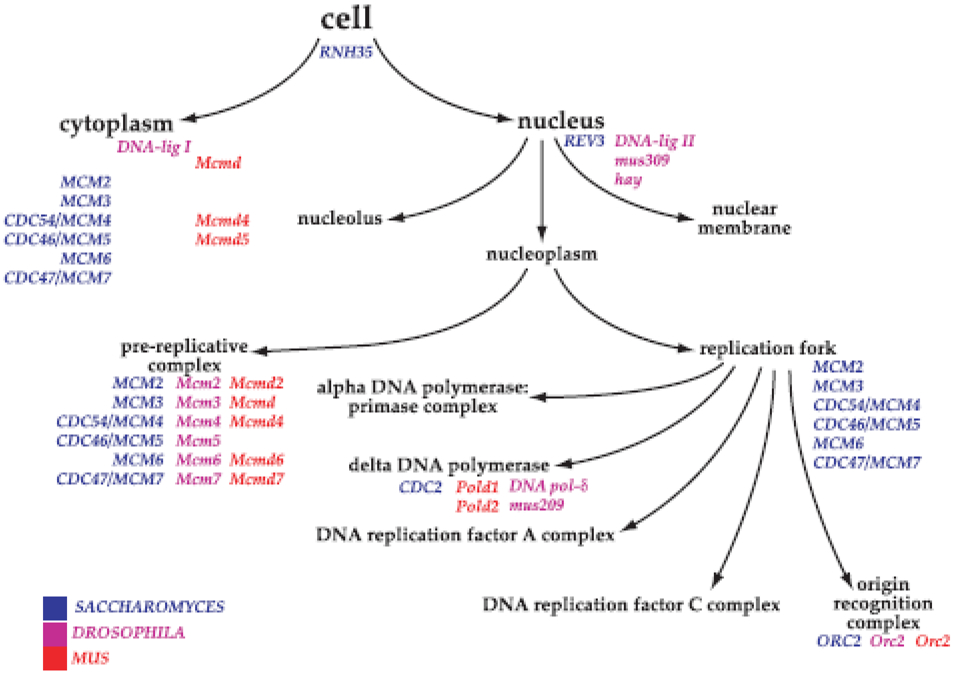


Figure 1.3: Example of Cellular Component graph taken from (Ashburner et al. 2000)

Gene Ontology Annotations

Genes and proteins are annotated with one or many GO terms by annotation consortia. Pieces of evidence from different sources are manually reviewed by the consortia before final annotation assignments are made. The annotation process is transparent; curators record evidence codes for each annotation (Table 1.2). Single annotations may be corroborated through multiple evidence sources, from a published journal article characterising the function of a gene product from an activity assay, or from high throughput microarray or yeast two hybrid experiments. The evidence codes are ranked by reliability (Table 1.2). Curator approved records from the literature with direct statements describing function(s) (TAS) and annotations from experimental assays (IDA) represent the most reliable sources. IEA and ND codes which have not been subject to human judgement are considered the least reliable pieces of evidence. Most of the current human annotations are sourced from the least reliable automated (IEA) codes. This is symptomatic of the manual efforts required to provide high quality function annotation.

GO term annotations can be considered at different levels of the DAG structure to perform meta-type analyses of biological data. Meta-analyses provide a higher level overview of the data by layering more generally descriptive information onto primary or raw data. Often this provides a means to interpret experimental outcomes where raw data is noisy or cannot be directly compared. In these cases trends may be observed at the meta-information level that cannot be determined at the primary raw data level. For example, an experimenter might wish to interpret a list of genes (primary data) that change between two conditions by comparing their function annotations (meta information). Alternatively, an experimenter might restrict an analysis to a particular function category of interest. Of vital importance when using GO terms to perform these analyses is the quality and completeness of dataset annotations. In the absence of such annotations, the power to detect experimental trends cannot be realised.

A more sophisticated interpretation of annotation categories that exploits predefined relationships between function annotations is the semantic information content or semantic similarity between annotation terms. These measures quantify the specificity of the annotation term by considering its frequency of occurrence. Rare terms are assigned a high specificity whilst general terms are assigned a low specificity value. Semantic content measures are useful for the Gene Ontology since the majority of annotations for genes and proteins are partial (only one of many functions of the protein has been annotated) and general (the annotation category is close

Table 1.2: GO evidence codes and their definitions. GOA human column represents the frequency of the evidence code in the human annotation files

Code	Definition	GOA human	Reliability
IC	Inferred by curator	109	-
IDA	Inferred from direct assay	953	1
IEA	Inferred from electronic annotation	27835	5
IEP	Inferred from expression pattern	164	3
IGI	Inferred from genetic interaction	14	1
IMP	Inferred from mutant phenotype	181	2
IPI	Inferred from physical interaction	1415	1
ISS	Inferred from sequence or structural similarity	705	3
NAS	Non-traceable author statement	3098	4
ND	No biological data available	1385	5
RCA	Inferred from reviewed computational annotation	0	-
TAS	Traceable author statement	6492	1
NR	Not recorded	1100	-

to the top of the hierarchy). Groupings of genes or proteins by annotation similarity provides a means to simultaneously evaluate functions of equivalent specificity as well as tolerating partial annotations.

1.3.3 Modelling function similarity

Protein similarity is well defined in terms of sequences using homology detection algorithms. Protein structural similarity is also well defined by comparing folds, architectures and or topologies using the CATH (Orengo et al. 1997) and SCOP (Murzin et al. 1995) nomenclatures. The continuum of protein function similarity can be defined based on the semantic similarities measured using one of the various annotation schemes. The benefits of defining functional similarity between proteins are manifold. If a protein has unknown function it is advantageous to be able to identify a functionally nearest neighbour. The relationships between sequence, structure and function can be more clearly defined and whole organisms can be viewed in terms of scale based networks representing protein function space.

Semantic similarity measures were first used in the WordNet project (Sigman and Cecchi 2002) to measure similarity between words existing as part of a network structure. Several semantic similarity measures (Table 1.3) have been formally applied to the Gene Ontology Graphs (Lord et al. 2003, Wang et al. 2007). The different methods share the common feature that they use linkages between terms to define a closest parent common ancestor term (pca) (Figure 1.4), which forms the basis of the similarity score. The functional similarity between proteins or genes has been evaluated using the different semantic similarity measures and compared with sequence similarity, microarray similarity and structural similarity (Lord et al. 2003, Pesquita et al. 2008, Wang et al. 2007, Zhang et al. 2006).

To calculate a functional similarity score between two genes or proteins, they are first considered in terms of their annotations. Each protein can be annotated by multiple terms from each of the different Gene Ontologies; Molecular Function, Biological Process, or Cellular Component. Thus, a matrix of semantic similarities can be generated between all annotation pairs from each pair of proteins:

$$M_{GO} = PROT_A.GOterms \times PROT_B.GOterms$$

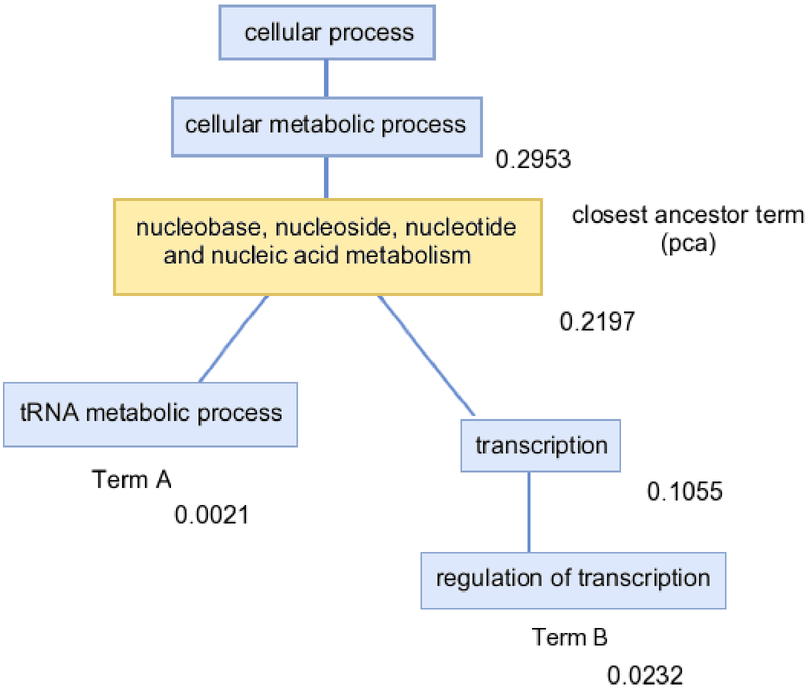

cellular process
cellular metabolic process
0.2953
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
closest ancestor term (pca)
0.2197
tRNA metabolic process
Term A
0.0021
transcription
0.1055
regulation of transcription
Term B
0.0232

Figure 1.4: Example of semantic similarity measure between GO terms ‘regulation of transcription’ and ‘RNA metabolism’. The closest common ancestor term is ‘nucleobase, nucleoside, nucleotide and nucleic acid metabolism’. The semantic information content of each term using its annotation frequency is shown alongside each node calculated as the probability of observing each term in the set of annotated human sequences.

Table 1.3: Semantic Similarity Scoring methods

Method	**Description**	**Equation**	**Range**
Resnik	scores GO term similarity as pca, pca defined by term frequency of occurrence	$-ln(pca)$	0 - inf
Lin	scores GO term similarity as normalised pca, pca defined by term frequency of occurrence	$-\frac{ln(pca)}{ln(A)+ln(B)}$	0 - 1
GFSST	scores GO term similarity as pms, pms defined by frequency of child terms	$-ln(pca)$	0 - inf
SimRel	hybrid of Resnik and Lin scores	$-ln(pca) \times \frac{ln(pca)}{ln(A)+ln(B)}$	0 - 1

Here pca represents the probability of the closest ancestor annotation in the GO term graph. In the Resnik, Lin and SimRel methods this probability is defined as the frequency of occurrence of the annotation in a population of sequences relative to the frequency of the root annotation; either Molecular Function or Biological Process. A and B refer to the probability of annotation A and B ocurring within the same sequence population .

$$= \begin{array}{c|cccc} & GOB_1 & GOB_2 & \dots & GOB_n \\ \hline GOA_1 & S(GOA_1, GOB_1), & S(GOA_1, GOB_2) & \dots & S(GOA_1, GOB_n) \\ GOA_2 & S(GOA_2, GOB_1), & S(GOA_2, GOB_2) & \dots & S(GOA_2, GOB_n) \\ GOA_3 & \vdots & \vdots & \ddots & \vdots \\ GOA_n & S(GOA_n, GOB_1), & S(GOA_n, GOB_2) & \dots & S(GOA_n, GOB_n) \end{array} \tag{1.1}$$

Functional similarities can be local (between the most similar annotation pairs), or global (between all annotation pairs) and symmetric where the resulting scores are the same for forwards comparisons (protein A vs protein B) and reverse comparisons (protein B vs protein A), or asymmetric where the resulting scores exhibit directional bias. The first application of function annotation similarity methods derived semantic similarities using average of pairwise similarity scores between all pairs of GO terms within the same Ontology. Resnik, Lin and Jiang methods were compared with sequence similarity for protein pairs. Similarities derived using the Resnik method were most highly correlated with sequence similarity (Lord et al. 2003). Subsequent studies reported superior correlations for the Resnik semantic similarity method compared with gene expression similarity (Sevilla et al. 2005) and domain architecture similarity (Bjorklund et al. 2005). A recent benchmark comparison suggested the use of the asymmetric maximum similarity score over the average similarity score or maximium similarity score (Pesquita et al. 2008).

1.4 Automated function prediction methods

Controlled vocabularies to describe protein function such as the GO and MIPS initiatives provide a framework for the development of function prediction algorithms. The GO graph structures are more complex than the straightforward four digit tree structures of either FunCat or Enzyme, but represent a compromise between the subtlety of relationships that can be described, the requirement for machine readability and standardised linguistics. The growth of automated function prediction servers and approaches has been considerable in recent years. In 2005 the first Automated Function Prediction Special Interest Group (AFP-SIG) meeting discussed benchmarking and quality measures for these methods (http://biofunctionprediction.org/AFP/). The result was a critical assessment for function prediction servers for given protein sequence and structure information to return a GO function prediction. Difficulties in providing a comprehensive gold standard experimentally validated dataset for prediction purposes has hampered progress in this area, since true and false negative predictions cannot be easily distinguished.

Methods of function prediction predominantly comprise annotation transfer methods utilising guilt-by-association approaches. Other methods are model-based single or multi-class function predictors. Some methods concentrate on single aspects such as the protein's sequence or structure, or combine different biological attributes to predict function. Homology based methods target the accurate annotation of similar protein sequences. Relatively few methods target the annotation of distantly related proteins, or proteins with no discernible sequence relatives - one of the greatest challenges in the function prediction field.

1.4.1 Homology based approaches

The largest group of methods rely on detection of evolutionary homologues by sequence similarity search. Annotation is transferred from well annotated sequences to uncharacterised queries by establishing the closest relative or closest group of consistently annotated sequences. Homologue detection is usually carried out by Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) or Position Specific Iterated BLAST PSI-BLAST (Altschul et al. 1997) algorithms. More precise orthologue detection is performed by aligning the unknown sequence with its closest homologues in a multiple sequence alignment and reconstructing a phylogenetic tree. Annotation is then transferred between orthologous sequences or clades. These approaches are referred to as phylogenomic approaches (Sjolander 2004).

BLAST or PSI-BLAST based approaches

BLAST is a heuristic sequence search algorithm used to identify similar sequences to a query of interest. The existence of a highly significant similarity score or multiple consistently annotated relatives is usually sufficient to transfer function annotation to the uncharacterised query sequence.

The premise of the method is that sequence relatives represent evolutionary homologues of the query sequence with common function. The greater the number of similar sequences returned with consistent annotation from the search, the greater confidence can be attributed to the annotation. Sequence similarities of $\geq 80\%$ identity are universally considered high confidence relationships for function assignment (Addou et al. 2009, Tian and Skolnick 2003). Many function prediction approaches have been developed that rely on BLAST or PSI-BLAST as the underlying method (Table 1.4).

The PSI-BLAST algorithm is an extended version of the BLAST algorithm designed to improve detection of distantly related sequences. The PSI-BLAST algorithm generates a position specific scoring matrix (PSSM) from the initial search results. The search is repeated using the PSSM model to query the same database. As more sequences are added to the profile, the iterations continue and power to detect remote relationships increases. The search terminates when no additional sequences are identified (convergence) or the number of iterations reaches a user defined threshold. The PSI-BLAST method is powerful where unambiguous sequence relationships cannot be detected by BLAST and the search capability must be extended.

The simplest methods Onto-BLAST (Zehetner 2003) and GOblet (Groth et al. 2004) search a database of well defined annotated proteins. The annotations of the matching sequences are collated and a statistical expectation value (E value) from the BLAST output used to score the occurrence of each GO annotation. The E value represents the number of times a match has occurred to a related sequence versus a match to a random sequence within a database of a certain size. The Onto-BLAST server summarises these annotations using the most significant similarity scores, the number of sequences carrying the annotation and number of species with an annotation from the matching sequences list. These servers do not attempt to indicate the correct annotation for an unannotated sequence but simply present the annotations returned via the similarity search ranked by score.

The GOFigure method (Khan et al. 2003) defines a minimum covering sub graph from the GO annotated sequence hits. Similarly, the GOtcha method (Martin et al. 2004) scores prospective GO terms in a sub-graph using the BLAST E value. The most probable sub-graphs are identified by the maximum score obtained by summing the E values from the leaves of the sub-graph to the root. The PFP server (Hawkins et al. 2006, 2009) additionally considers the most probable GO term annotations by including a probability term derived from a scoring matrix known as the FAM matrix. The FAM matrix represents conditional probabilities for co-occurrences of GO terms both within and between ontologies derived from the UniProt protein database. By considering these probabilities weak scores can be strengthened by the prediction of one or more

Table 1.4: Function prediction tools that rely on BLAST or PSI-BLAST algorithms

Method	**Authors**	**URL**
PFP	Hawkins et al. (2006)	http://dragon.bio.purdue.edu/pfp/pfp.html
OntoBlast	Zehetner (2003)	http://functionalgenomics.de/ontogate/
GOFigure	Khan et al. (2003)	http://udgenome.ags.udel.edu/gofigure/
GOPET	Vinayagam et al. (2006)	http://genius.embnet.dkfz-heidelberg.de/menu/cgibin/w2hopen/w2h.open/w2h.startthisSIMGOw̄2h.welcome
GOBLET	Groth et al. (2004)	http://goblet.molgen.mpg.de/
GOTCHA	Martin et al. (2004)	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php

of the co-occurring GO terms. The PFP server produces the top 10 most likely annotations for a protein ranked by the final score. The improved performance reported for this method is most likely due to the inclusion of the FAM matrix and thus it represents the most sophisticated of the automated sequence similarity transfer methods.

Phylogenomics approaches

The term phylogenomics refers to the application of phylogenetic information to the study of genomic data (Eisen 1998). Rather than assigning nearest sequence neighbours by similarity searching methods, phylogenomics approaches determine true evolutionary homologues by phylogenetic reconstruction (see Figure 1.5 for an example). The process involves identifying a set of homologous sequences, or family members, producing a sequence alignment between that can be used to construct a phylogenetic tree. Function labels are then overlaid onto the reconstructed tree of closely related sequence neighbours. Function can be inferred for an uncharacterised protein either by clade membership in the phylogenetic tree or by identification of the closest orthologous sequence in the tree.

In Figure 1.5 left hand side, sequences 2A and 2B represent cases where function has evolved in parallel. Unambiguous function assignments are determined by placement of the sequence alongside orthologues with common function. On the left hand side of Figure 1.5, the assignment of function is straightforward for uncharacterised sequences 2A and 2B as function has evolved in parallel and is preserved within clades following duplication. However, the case presented on the right hand side is more difficult. Here function has diverged within a clade either before or after the branch point of sequence 3 giving rise to orthologous sequences that do not share function. In this case the reconstruction cannot be used to make an unambiguous function assignment. Sequence 5 can be assigned function since it lies between two sequences (6 and 4) with shared function. These sequences are evolutionarily closer and more distant respectively, to the common ancestor with shared function.

The value of phylogenomic approaches are recognised in situations where convergent evolution gives rise to small changes in sequence that alter functional specificity, or where straightforward sequence similarities fail to correctly distinguish orthologous relationships from homologous relationships (Sjolander 2004). Automated methods employing phylogenomics approaches include SIFTER (Engelhardt et al. 2005), RIO (Zmasek and Eddy 2002) and ORTHOSTRAPPER

(Hollich et al. 2002, Storm and Sonnhammer 2002).

The SIFTER method uses a Bayesian phylogenomics approach to assign function to unannotated proteins (Engelhardt et al. 2005). It overcomes some of the time constraints involved in constructing accurate multiple sequence alignments by pre-computing family alignments to seed the searches. Uncharacterised sequences are added to the nearest family alignment and known functions overlaid onto the resulting phylogenetic tree. The conditional probability of unannotated protein having any of the functions is evaluated by considering the positions of the known functions in the tree. Additionally, reliability weights are applied to each annotation according to the annotation evidence codes. The authors demonstrated the applicability of their approach in deciphering annotations for the monophosphate/deaminase and lactate/malate dehydrogenase families quoting 96% accuracy, and considerably improved performances over the GOtcha and other sequence similarity based approaches. Both families present challenges in function prediction by representing multiple functions between closely related sequences. Hence these test datasets present cases where annotation transfer by sequence similarity methods are prone to errors.

RIO (Re-sampled Inference from Orthologues) (Zmasek and Eddy 2002) employs bootstrap and re-sampling procedures permuting sequence alignments and rebuilding the trees to estimate the reliability of function assignments. The ORTHOSTRAPPER method (Storm and Sonnhammer 2002) is very similar to the RIO method, however sequence similarity heuristic measures are used to construct a pairwise sequence distance matrix for tree building, as oppose to an evolutionary distance measure obtained from a phylogenetic tree.

All of the phylogenomic methods for inferring function rely upon close sequence family or domain assignments made through initial homology searches. They target highly accurate function assignments through the determination of evolutionary speciation and duplication events. The trade off with these methods is that they are time consuming. Bootstrap tree values and confidence estimates come at considerable computational cost and there is an implicit requirement that the orthologues of an unknown sequence are completely and correctly functionally annotated (Eisen and Fraser 2003). High quality multiple sequence alignments can be difficult to obtain without the use of expert knowledge or manual curation (Sonnhammer et al. 1997) and the quality of tree building methods relies explicitly on the breadth of annotated species present. As such phylogenomics techniques are restricted in terms of applicability and coverage.

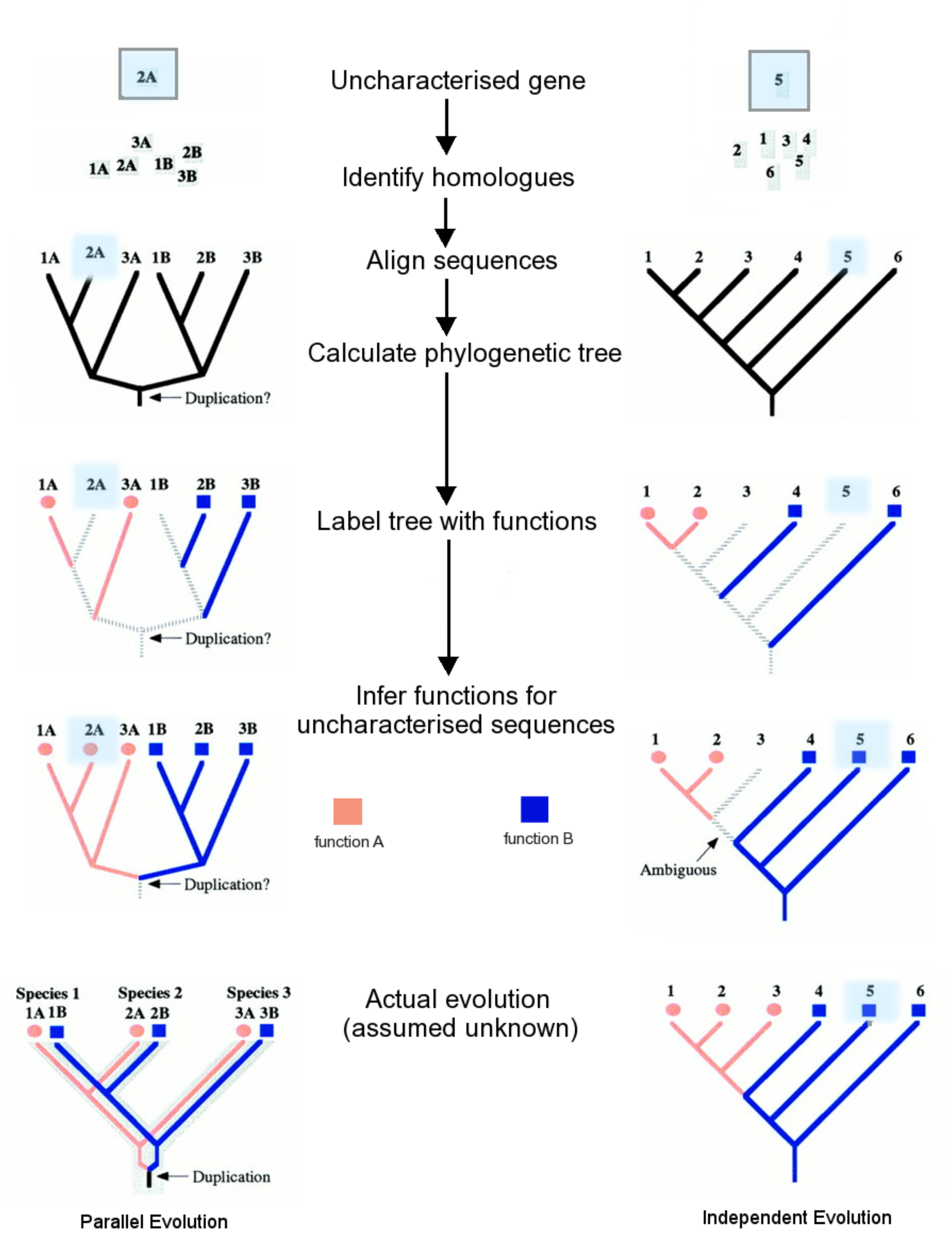


Figure 1.5: Schematic of phylogenomics based approaches. Sequences 2A and 1B can be unambiguously assigned function by clade membership, whilst sequences 5 and 3 on the right hand side represent less clear cut cases since orthologues have evolved independently within the different species.

Phylogenetic profiling

A related set of methods utilise phylogenetic profiles in order to transfer function annotations (Eisen and Wu 2002, Engelhardt et al. 2005, Ranea et al. 2007). Phylogenetic profiling techniques capture the evolutionary history of a gene or protein through the use of presence or absence species profiles constructed using sequence or domain relatives in other organisms (see Figure 1.6 for an example) (Eisen and Fraser 2003). The premise of the method is that pairs of sequences with similar phylogenetic profiles share common evolutionary history and are more likely to be functionally related (Loganantharaj and Atwi 2007).

There are two important steps in phylogenetic profiling: the construction of high quality profiles and the method used to compare between them. A variety of distance and similarity measures have been applied to the profiles in order to detect co-evolution, including mutual information, given by

$$MI(X:Y) = \sum \sum p(x,y) log \left[\frac{p(x,y)}{p_1(x)p_2(y)} \right] \tag{1.2}$$

, Pearson's correlation emphasising similarity of shape between two series and given by

$$r_{X:Y} = \frac{\sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \tag{1.3}$$

and Euclidean distance measuring the magnitude of difference between vectors

$$d_{X:Y} = \sqrt{\sum (x-y)^2} \tag{1.4}$$

(Wu et al. 2006). Whilst these methods are relatively quick to compute, their power is dictated by species diversity in the profile. Consequently, they perform better when applied to prokaryote genomes for which there exists approximately six times the number of fully sequenced species representatives (Cokus et al. 2007, Kyrpides 1999, Loganantharaj and Atwi 2007).

Domain or family based approaches

Historically, the domain has been cited as the primary unit of functional inheritance (Ponting and Russell 2002). Domains exist in proteins as independently folding units presumably selected for functional reasons. This information is captured directly in annotation schemes such as InterPro2GO, SCOP2GO and PFAM2GO where the explicit presence of an InterPro or PFAM domain

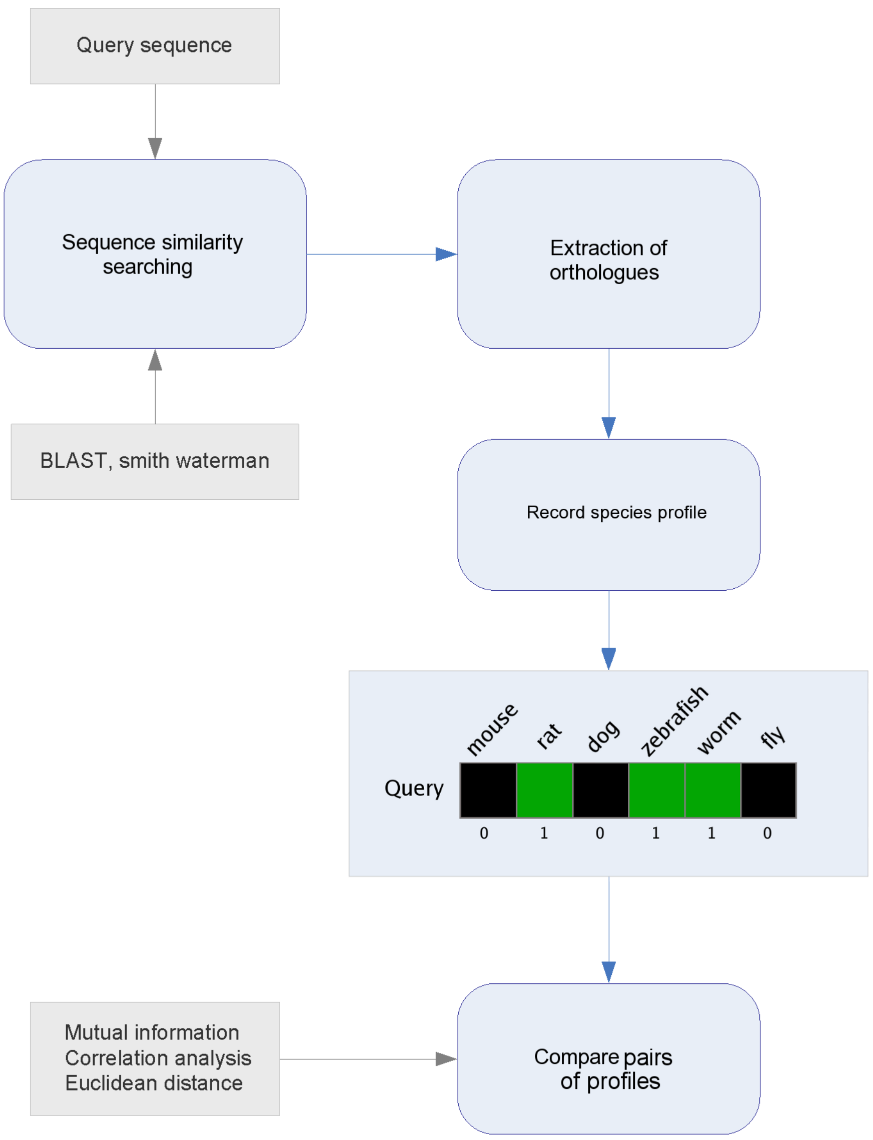


Figure 1.6: Workflow steps involved in phylogenetic profiling analysis. Sequence similarity searching is performed against a set of completely sequenced reference genomes to identify homologues or orthologues. Presence or absence of these relationships is then used to generate phylogenetic profiles. The profiles are then compared in order to identify cases of similar evolutionary history between sequences.

in a sequence is sufficient to indicate function. PFAM (Finn et al. 2003), SMART (Schultz et al. 1998), CDD (Marchler-Bauer et al. 2002) and PRODOM (Corpet et al. 1998) domains are represented as homologous sequence models delineated by the parts of each sequence representing the domain. In the CATH and SCOP databases these domains are defined by 3D structure representatives and classified into hierarchical schemes based on common topologies and architectures (Andreeva et al. 2008, Orengo et al. 1997). Function predictions can be carried out by screening uncharacterised sequences against libraries of domains and exploiting the different mapping schemes to transfer function annotations. Single domain annotation methods yield precise and accurate functional annotations for a limited number of functions and domain folds or families; just over 46% of PFAM families are mapped in the current version of PFAM2GO covering 2042 (22 - 25%) of function annotation classes. Many domains, the TIM Barrels for example, are functionally promiscuous (Basu et al. 2008). In these cases phylogenomic profiles can result in an expansion of function class candidates rather than narrowing the selection or providing specificity. Evolutionarily, domains evolve by fusion (joining of two domains) and fission (splitting of a single domain) events resulting in large numbers of combinatorially unique architectures from just a few individual domains (See Figure 1.7 for a schematic describing the process) (Snel et al. 2000, Vogel et al. 2004, Yanai et al. 2002). Multi-domain architectures are prevalent in eukaryotic proteomes constituting 65-80% of sequences (Bjorklund et al. 2005, Gerstein 1998). Sequences sharing common domain architectures therefore display more similar functionality. Pairs of architecturally distinct sequences can also be linked by the existence of a common ancestral fusion protein that contains domains from both sequences. These relationships are rare, but can be used to infer interaction partners or indicate shared biological pathways (Enright et al. 1999, Marcotte et al. 1999).

The majority of automated methods exploiting domain information for function prediction employ domain profiling techniques similar to the phylogenetic profiling techniques described in Section 1.4.3. The main difference is that sequence-based phylogenetic profiles are constructed from domain presence or absence rather than the presence or absence of sequence homologues (Marcotte et al. 1999, Ranea et al. 2007). The GO trees method (Hayete and Bienkowska 2005) models the entire functional domain content of proteins using PFAM domain definitions. Each sequence is represented as a profile containing a domain representation similar to the phylogenetic profile (Figure 1.8). Domain occurrence between pairs was scored either as a binary vector or by an integer vector encoding the frequency of occurrence of each domain. A decision tree classifier was then used to model function assignments from the domain encoded profile

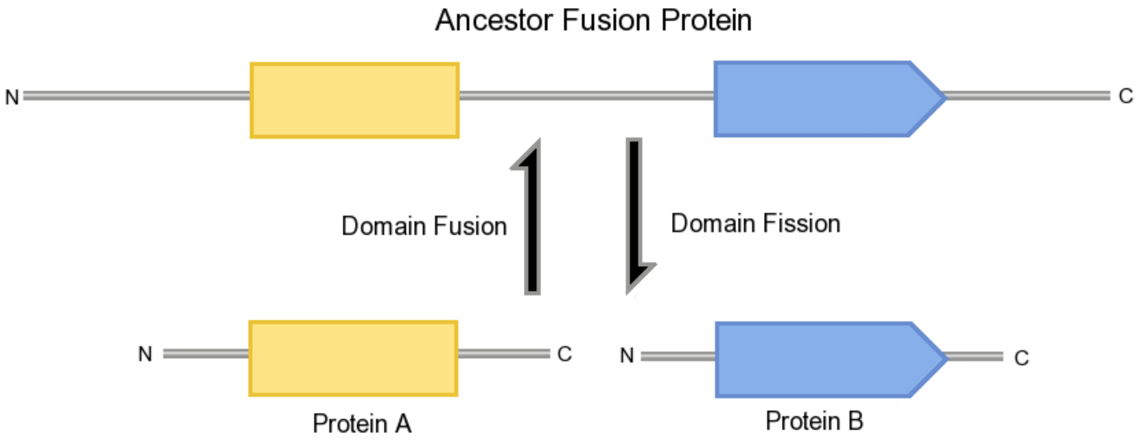


Figure 1.7: Diagrammatic representation of domain fusion and fissions

vectors. The approach improved sensitivity and specificity over and above that obtained using PFAM2GO mappings.

Forslund and Sonnhammer (2008) recently presented two approaches to infer function using domain architectures. The first method produced a strict mapping set between combinations of PFAM domains sufficient to infer function. The second Bayesian probabilistic approach evaluated the odds ratio for a function given a particular PFAM domain. The probability of annotation transfer given the full complement of domains in each sequence was then evaluated over all unique pairs of domains between two sequences. Performance assessment using Gene Ontology annotations showed that the probabilistic and direct mapping approaches were highly precise ($> 90\%$). However, much lower coverage was attained than annotations transferred using best or top BLAST hits (Forslund and Sonnhammer 2008).

Methods to predict protein function using domain fusion information have also been implemented (Enright and Ouzounis 2001). Ancestral fusion proteins are identified as pairs of domains in multi-domain proteins that are found to occur separately in another species (see Figure 1.7 for a diagram). The underlying rationale for the method is that in 3 dimensions two domains within close proximity to one another share at least one interacting surface (Chia and Kolatkar 2004). It then follows that proteins containing each of the domains may interact. This hypothesis has been validated using both fusions at the gene and domain level to suggest candidate protein interactions between unrelated sequences. Again profiling techniques can be used to capture these events at the sequence level and have proved useful specifying a weak signal for functional similarity (Serres and Riley 2005).

1.4.2 Non-homology based approaches

For a proportion of sequences (estimated at 33% of all currently known sequences (Ofran et al. 2005)) there are no annotated relatives. In some cases, detectable relatives are so distant that function assignments made via homology inference provide very general, low confidence annotations. A series of approaches to annotate these difficult cases utilise non-homology based features from sequence or structure. Other approaches incorporate information from experimental sources, expression data, or protein interaction data for example. One of richest sources of functional information that remains to be fully exploited lies in the scientific literature database MEDLINE accessible via the PubMed electronic gateway (Stewart et al. 2002). Automated ex-

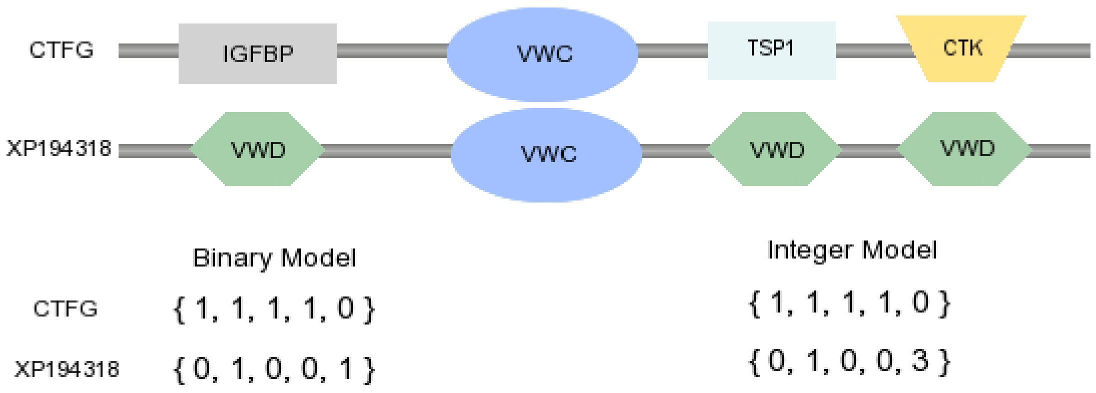


Figure 1.8: Domain representations for CTFG and XP_14318 sequences adapted from (Hayete and Bienkowska 2005). Domain abbreviations VWC and VWD represent Von Willebrand factor domains C and D respectively, IGFBP is the Insulin growth factor binding domain and CTK is a C terminal cysteine knot domain. TSP1 is the thrombospondin type I domain. The binary model represents domain presence and absence whereas the integer model records the frequency of each domain.

traction and language modelling of the relevant abstracts and articles from these vast resources constitutes one of the major challenges of the post-genomic era, text-mining.

Feature based methods

Feature based methods describe secondary characteristics of proteins predominantly obtained from sequence or structure. The sequence features do not directly exploit homology relationships by the use of sequence alignment methods, or attempt to define a neighbourhood of similar sequences. This makes these methods applicable to all proteins of known sequence regardless of their homology status.

The ProtFun method (Jensen et al. 2003) was one of the first methods specifically designed to target the annotation of orphan proteins. The approach employed neural network (NN) ensembles trained to recognise patterns of amino acid, localisation and secondary structure features to predict GO classes (see Figure 1.9 for a schematic of the process). 14 biological attributes were predicted from the amino acid sequence and encoded in feature vectors. Performance accuracies of $> 50\%$ coverage at error rates of less than 10% were obtained for 14 broad GO functional categories and 12 FunCat categories. Similar methods have been applied to enzyme function prediction using structural information (Dobson and Doig 2005). Here features such as surface accessibility, secondary structure and amino acid information were derived from crystal structure information and fed into Support Vector Machines (SVM's) to discriminate between different enzyme classes.

Whilst these approaches are applicable to all sequences or structures, limited information can be incorporated into the different features. Without alignments to identify conserved parts between sequences or structures, the features tend to comprise general characteristics describing the whole sequence or structure. This restricts the power of the method to provide discriminate functions between closely related sequences. Rather than relying on these methods to make accurate function assignments, they tend to be reserved for function candidate prioritisation for orphan sequences and sequences that cannot be aligned to well characterised proteins.

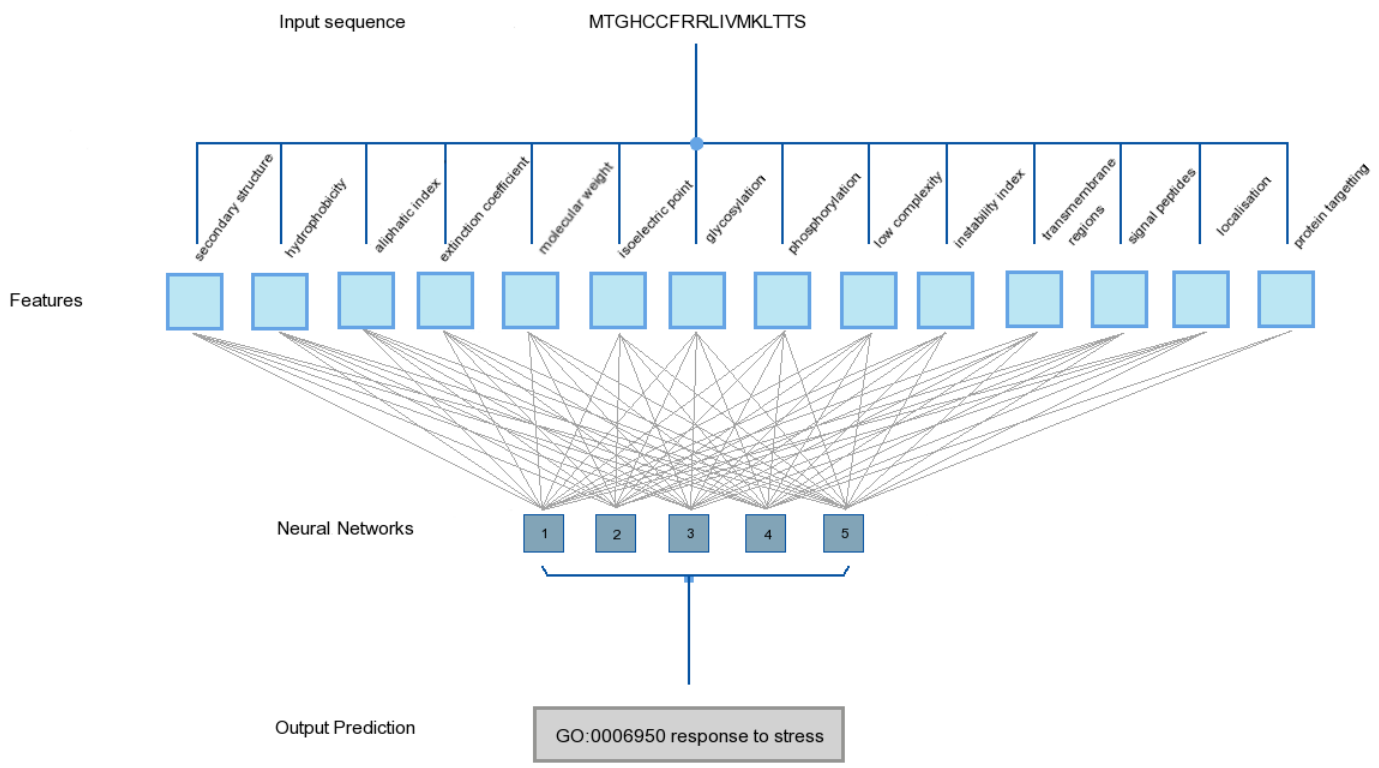


Figure 1.9: Schematic of the ProtFun method

Function prediction from expression information

The advent of DNA microarray technology has meant that thousands of genes can be simultaneously profiled for expression in a quantitative manner. Expression signatures in tissues and cell lines or responses to stimulus can be surveyed across whole genomes. Genes with similar expression profiles, tend to code for interacting proteins - a source of useful information regarding protein function (Ge et al. 2001, Jansen et al. 2002). Genes that react similarly to external events, ligand binding or stress conditions for example, tend to participate in similar pathways (Stuart et al. 2003).

Data stores have been set up to capture the results of expression experiments (Array Express (Parkinson et al. 2007), Gene Expression Omnibus (Barrett and Edgar 2006), RNA Abundance Database (Manduchi et al. 2004) and the Stanford Genome Database (Ball et al. 2005)) for viewing, querying and downloading these publicly available data. Single or multiple experiments can be re-analysed to gain insights into the behaviour of genes under different conditions and extract knowledge about their functions. Methods of annotating function from microarray data fall into two classes, unsupervised clustering approaches and supervised knowledge based approaches.

Unsupervised approaches

The most widely used methods for function classification from microarray data involve selecting groups of clustered, co-regulated or co-responsive genes from an experiment and examining their function annotations. Where an unannotated gene product is a member of a group with consistent or conserved annotations, its function can be inferred. Eisen et al. (1998) was one of the first to establish robust clustering methods for microarray data and publish analyses of co-regulated genes that were unrelated at the sequence level yet shared common functions. Subsequently, more sophisticated techniques have been applied to group genes with common expression behaviour such as eigen gene analysis, independent component analysis (Frigyesi et al. 2006, Lee and Batzoglou 2003, Liebermeister 2002) and bi-clustering algorithms. Unlike more traditional clustering approaches that identify similar expressions across all sets of tissues or samples in an experiment, bi-clustering techniques seek expression patterns that are conserved over subsets of conditions from a given experiment (Madeira and Oliveira 2004, Prelic et al. 2006). Once robust groupings of genes or transcripts are obtained from the data, their functional heterogeneity can be measured and broad level functions inferred using a guilt-by-association approach where deemed appropriate (see Figure 1.10 for a process overview).

One of the first steps in clustering transcript expressions is to establish a comparative measure between profiles. Two commonly used measures are Euclidean distance (Equation 1.4) which evaluates the magnitude of differences between expression intensities over each experimental condition, and Pearson's correlation (Equation 1.3) which measures the similarity of the shape of two expression profiles by considering the direction of the vectors between conditions (see Figure 1.11). Zhou et al. (2005) applied second order correlation coefficients to measure common behaviour between different experiments performed in yeast. These higher level correlations were determined from first order correlations measured within different experiments for pairs of transcripts. Applying thresholds to the first and second order correlations resulted in quadruplet groups of transcripts that were more likely to share common function than pairs of first order correlations. Using the resulting clusters, the authors were able to make function assignments to more than 60 uncharacterised genes. Several of these predictions could be supported by literature evidence. The method represents a straightforward way to integrate data from different experiments and microarray platforms together to increase the predictive power of the approach.

Supervised approaches

Supervised approaches involve building models of expression profiles or identifying gene signatures for particular functions. Uncharacterised transcript profiles can then be tested against the models to infer function with an associated confidence measure. Support vector machines (SVMs - see Appendix part I) have been used to classify function for unannotated yeast orfs (open reading frames) (Brown et al. 2000), and rule based approaches have been used to extract signature templates from time series microarray data for human GO term prediction (Lagreid et al. 2003). These approaches generated accurate function assignments for subsets of broad annotation classes determined by the type of experiments performed on the data.

For the yeast genome, 80 different hybridisation experiments at different time points from budding yeast were used. These included Diauxic shift, mitotic cell cycle division and sporulation experiments performed on custom built spotted arrays. The arrays were dual channel hybridisations performed using a fixed reference RNA sample for normalisation purposes. Expression measures were represented as log ratios calibrated using the corresponding reference RNA chan-

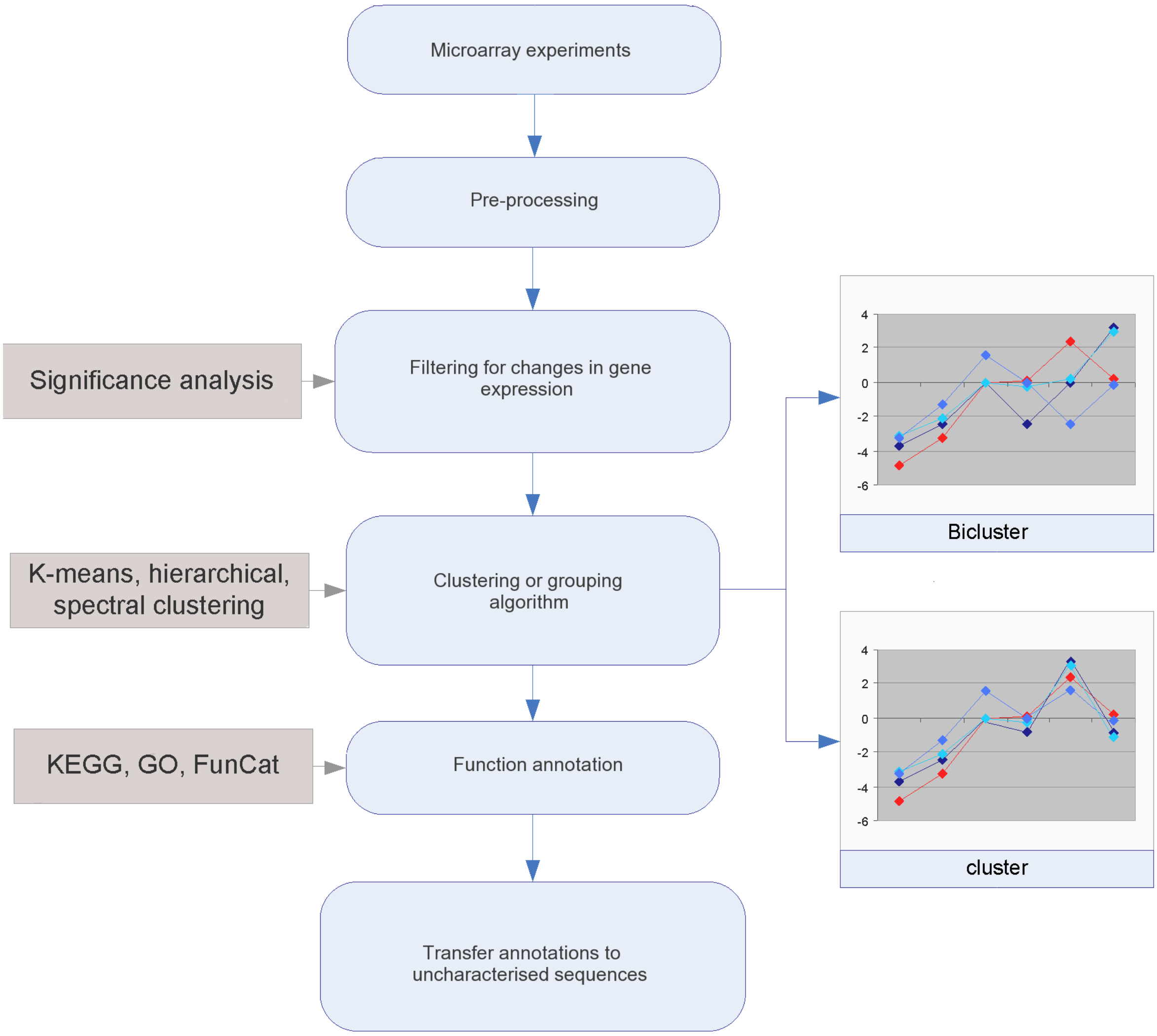

Microarray experiments
Pre-processing
Significance analysis
Filtering for changes in gene expression
K-means, hierarchical, spectral clustering
Clustering or grouping algorithm
KEGG, GO, FunCat
Function annotation
Transfer annotations to uncharacterised sequences
4
2
0
-2
-4
-6
Bicluster
4
2
0
-2
-4
-6
cluster

Figure 1.10: Workflow diagram of unsupervised methods for annotating functions from microarray data. Traditional clustering algorithms tend to group genes with common expression behaviour over all conditions whilst bi-clustering approaches group genes with common expression behaviour in just a few samples.

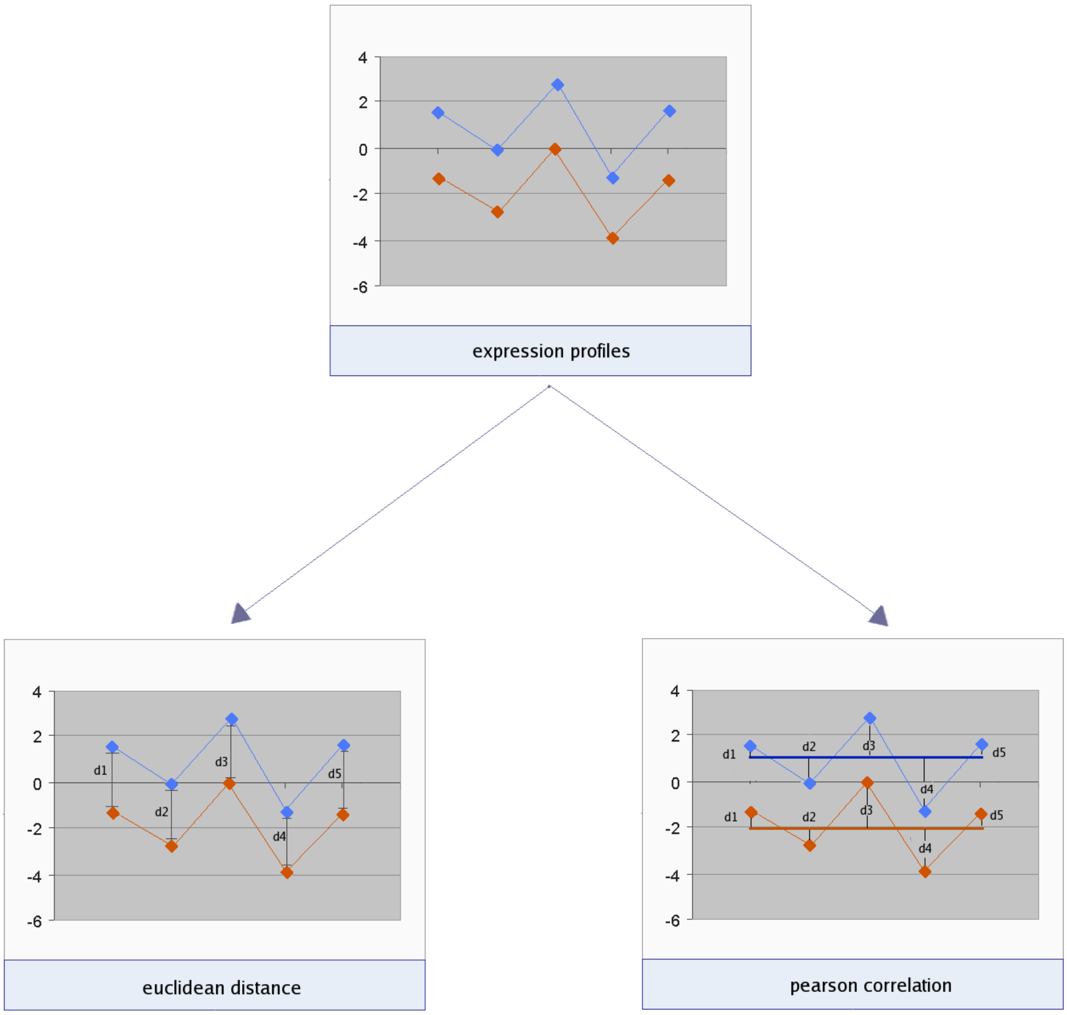


Figure 1.11: Different measures of similarity for expression behaviour across experimental samples. The Euclidean measure accounts for differences in magnitude between expression profiles (d1,d2,d3,d4 and d5) whilst Pearsons correlation coefficient looks for conserved shape by comparing the variation (d1,d2,d3 and d4) from the mean expression (dark blue and orange thick horizontal lines) measured within each expression series.

nel. A global normalisation

$$N_i = \frac{ln(E_i/R_i)}{\sqrt{\sum ln^2(E_i/R_i)}} \tag{1.5}$$

was then performed across all arrays and the resulting 80 dimensional vector of expression measures used as feature inputs to classifiers to recognise patterns of expression that were indicative of each functional class. Six function classes were predicted with high accuracy and low false positive rates. The best performances were observed for ‘ribosomal’ and ‘histone’ functions.

Another approach used a human fibroblast serum response time series dataset (Iyer et al. 1999) to predict GO terms (Lagreid et al. 2003). Initially, a set of functionally informative gene expressions were defined by considering the variance in expressions over different numbers of time intervals (minimum of 2 time points). Rules were generated from the function class templates and decision-based reasoning applied to predict class membership for test datasets. The rules were pruned using rough set theory to establish the minimum set required to classify each annotation category (Table 1.5). This method resulted in models for 16 BP categories from GO. The highest accuracies were reported for “chemotaxis”, “blood coagulation”, “cell embryogenesis” and “morphogenesis” categories.

The results from both these approaches suggest that gene function information can be reliably inferred from microarray data, given an appropriate model. In contrast, several studies suggest that co-expression signals determined by correlation analyses from microarray data serve only as a weak signal for function given that expression behaviour represents a cellular snapshot taken at the transcriptional level (van Noort et al. 2003, Yeung et al. 2004). Some transcripts may be rapidly degraded or be regulated by other cellular mechanisms such that they never reach their required destination to perform their functions. In this case transcriptional profiles do not provide an appropriate representation of the behaviour of the protein in the cell that can be used to make functional inferences. Other problems in interpreting expression patterns result from the technology itself. For example, some genes are expressed in very low copy numbers below the detection limits of the microarray whilst other genes can exhibit ‘hyper variable’ or erratic expression behaviour when characterised by microarray (Dozmorov et al. 2004).

The function categories that could be predicted from microarray data in the supervised approaches were restricted to broad categories with many representatives. In part this could be due to insufficient examples for robust model building, however unsupervised methods are not limited by annotation category size and have proved useful in predicting general functions such

Table 1.5: Expression rule set generated from time series microarray data

Biological Process: transport	
1	2Hr - 4Hr (Decreasing) AND 12Hr - 20Hr (Increasing)
2	2Hr - 6Hr (Decreasing)
3	12Hr - 20Hr (Increasing)

Rules generated for temporal expression patterns that correspond with the Biological Process "Transport" as ‘transcription factor’. Key considerations when dealing with microarray experimental data are the selection of experiments and types of expression profiles to be included. In the case of the yeast SVM study, these were selected by ‘experts’ (Brown et al. 2000).

Function prediction from protein-protein interactions

Protein-protein interaction data can be generated using experimental techniques such as high throughput yeast two hybrid (Y2H) screens, protein arrays, NMR, X-ray crystallography, pull down assays both *in vivo* and *in vitro*, co-immuno precipitation experiments and western blots. Most of these data are stored as binary relationships between two proteins and large data storage and access facilities have been set up for curation and deposition of these data. The INTACT (Hermjakob et al. 2004) database contains 66499 individual interactions and complexes from a variety of sources and organisms from 3464 distinct experiments. Other similar repositories include DIP (Xenarios et al. 2000) with 55000 interactions, BIND (Bader et al. 2003) with over 67000 imported interactions from high and low throughput experiments. The different types of interaction experiments result in varying data qualities. Some interaction experiments provide quantitative binding affinity data whilst others report qualitative relationships.

A variety of promising function prediction methods from protein-protein interaction data have been reported. The methods rely on exploiting binary relationship data from the repositories and constructing networks describing whole organism protein interactions. It has been observed that 70-80% of proteins share at least one function with their interacting partners (Titz et al. 2004). Additionally, proteins of a particular function are likely to interact with proteins of a restricted functional repertoire (Kelly and Stumpf 2008). These two concepts form the basis of protein function prediction from interpretation of protein interaction maps and networks. The simplest methods use a majority rule approach applied to the local interaction neighbourhood of an unannotated protein (Hishigaki et al. 2001) (local neighbourhood frequency approaches). More sophisticated probabilistic approaches have been developed in order to predict protein function by considering entire network architectures (Deng et al. 2002, Letovsky and Kasif 2003, Vazquez et al. 2003) (probabilistic whole network approaches).

Local neighbourhood frequency approaches

The local neighbourhood of an uncharacterised protein is defined by its position within a protein interaction map or network. Usually the local neighbourhood comprises only the immediate interacting proteins. A protein with unknown function can be annotated by its membership of a neighbourhood of common functions. The known and annotated protein functions in the neighbourhood are treated as the set of potential functions of the unknown protein. In the majority rule approach the function most commonly observed with the local neighbourhood is assigned to the unannotated protein. Other approaches have used over-representation statistics such as chi^2

$$S_i(J) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)} \tag{1.6}$$

to score prospective functions using the frequency of proteins assigned the function in the immediate interacting network compared with the frequency of occurrence of the function in the rest of the network (Hishigaki et al. 2001). Here, $n_i(j)$ is the number of interaction parters of protein Pi having function j. $e_i(j)$ is the expected number of partners having function j equal to $n_i(j)xp_i$ where p_i is the fraction of all proteins annotated with function j.

Local methods rely on the unannotated proteins residing in a part of the network that is heavily populated by well annotated proteins. Often this is not the case and when unannotated proteins do interact with proteins of known function, they are more likely to be part of the 20 - 30% of proteins that do not share common function with their neighbours (Titz et al. 2004). Additionally the statistical power of local analyses is reduced when the neighbourhood sizes are small.

Probabilistic whole network approaches

Whole network approaches consider the structure of the entire network to make function predictions. Typically the candidates are selected from the function annotations of immediate or local interaction partners but scored using a likelihood or probability measure considering the dispersion of annotation labels over the entire network.

Kirac et al. (2006) considered the local neighbourhood as all proteins sharing interaction paths leading to the uncharacterised protein of interest (see Figure 1.12). The likelihood of each annotation given each path is modelled using background frequencies of observing the path conditioned on a particular annotation category over samples of local neighbourhoods from the network. The final annotation score for an uncharacterised protein is then averaged over all local paths. Given the example in Figure 1.12, a simple majority rule or neighbourhood count would predict the yellow function as the most likely, however the green function or even orange functions might represent the more probable solution, should the paths be more frequently associated with these functions.

Other whole neighbourhood approaches (Deng et al. 2002, Letovsky and Kasif 2003) have used Markov random field theory to assign a probability to a protein having a particular function. Markov random field theory specifies mathematical functions describing the probability dependencies of annotation labels on positions of proteins in the network using a weighting scheme. The largest weights are assigned to nearest neighbours of the unannotated protein and decrease as path distance increases from the unannotated protein. The likelihood of a protein annotation is modelled against the prior probability of a function annotation (proportional to the frequency of annotations of all proteins).

Methods of predicting protein function from protein-protein interaction networks are robust where an unannotated protein resides in a densely populated area. Consequently the information content for the neighbourhood is high. For sparsely populated areas, even the more complex whole network approaches struggle. Another problem with these approaches lies in the quality of the interaction data. Most whole genome networks incorporate interaction data from multiple high and low throughput experimental techniques. Y2H experiments are notoriously noisy and frequently produce many false positive interactions (Deng et al. 2003). Networks constructed from low throughput high accuracy experiments, for example, tandem affinity purification (TAP), are sparse and incomplete. Careful integration of different data sources improves the quality of the network by reducing false positive associations (Hishigaki et al. 2001). However, it is impossible to distinguish proteins that genuinely do not interact from those whose interaction has not yet been observed because the required experimental conditions have not been met. Often this is because *in vitro* conditions cannot properly replicate *in vivo* conditions. Experimental data can produce false positive associations where two tested proteins are shown to interact *in vitro* yet under physiological conditions may never meet (Deng et al. 2003). These limitations can be addressed by overlaying information from other independent sources. Protein localisation information that reflects protein cellular compartmentalisation for example, has been shown to significantly improve accuracies obtained from function prediction methods utilising interaction networks (Nariai and Kasif 2007).

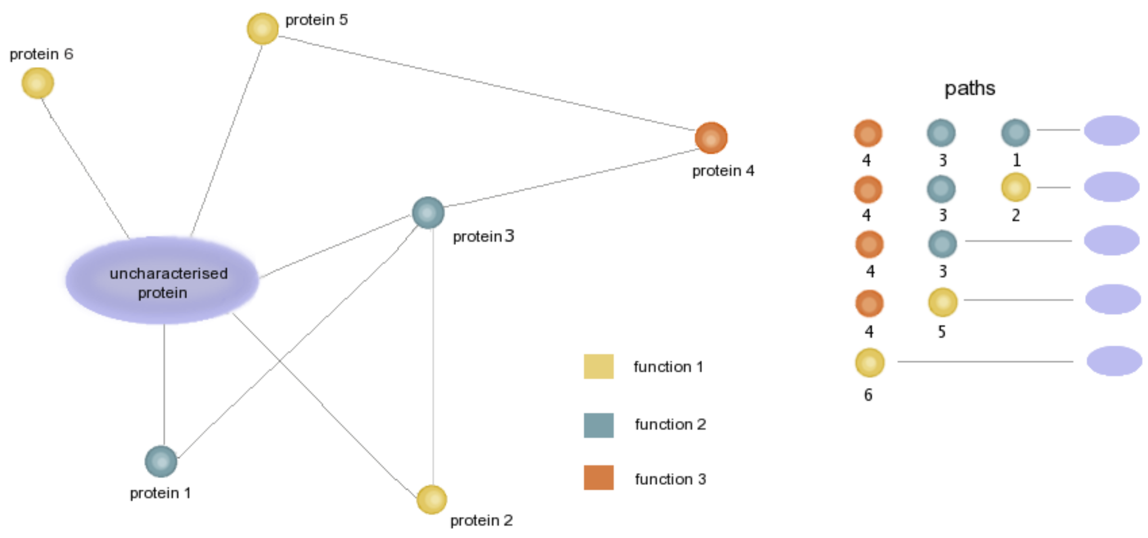


Figure 1.12: Hypothetical protein interaction showing all possible paths to the uncharacterised protein.

1.4.3 Integrated function prediction approaches

Individual methods of function prediction targeting homologous or non-homologous sequences typically use single data sources to make a prediction. Methods using homology information, structure, microarray, protein interaction and textual data as single sources to make a prediction are subject to annotation bias. They tend to perform best on particular sets of annotation categories. Few methods achieve both high accuracy and high coverage for all function categories; for example homology based methods only apply to proteins for which sequence similarities can be detected. These methods tend to yield high specificity but low overall coverage of protein sequences (Friedberg 2006). Methods using microarray information are biased towards the type of experiment performed and target annotation classes that are related to signalling and transcription control.

In theory, methods that combine different data sources together should provide more reliable predictions and achieve higher sequence and function class coverage. Several machine learning methods have been reported in the literature that target the integration of heterogeneous data sources. They can be classed as vector integration techniques, classifier integration techniques or kernel methods (Noble and Ben-Hur 2008). Bayesian networks (Troyanskaya et al. 2003), kernel methods (Lanckriet et al. 2004b) and unsupervised nearest neighbour algorithms (Yao and Ruzzo 2006) have all been applied to the task of function prediction by combining heterogeneous data sources. These methods show great promise for annotating whole genomes in a more reliable and consistent manner.

Kernel based methods

Kernel based methods are a recent advance in the field of machine learning (Taylor and Cristianini 2004). Kernel functions can be thought of as similarity functions (see Appendix I) operating between pairs of features characterising examples of interest (in this case proteins). Mathematically, a kernel function is any function that satisfies Mercer's theorem (defined in Appendix I) and produces positive semi-definite values. The simplest kernel is the dot or inner product between two numeric vectors describing a set of features. Practically, this might correspond to a set of measures describing characteristics of proteins, their size, shape or secondary structure. The kernel matrix then represents the similarity between pairs of proteins according to the feature characteristics.

The diffusion kernel is a popular choice for representing graphical or network topologies. For pairs of proteins this is similar to the probability of reaching one network node (or protein) by taking a set of random paths through the network from a starting node. One property of kernel functions is that they can be readily combined by applying simple transformations to produce a more complex similarity measure or kernel. This property makes kernel methods particularly suitable for integrating different data sources together.

To integrate information between different data sources, kernel matrices can be combined. The simplest and most commonly used integration technique for a set of kernel matrices is to combine them using a linear weighted sum.

$$Sim_{final} = c + w_i \times K_i + w_{i+1} \times K_{i+1} ... w_n \times K_n \tag{1.7}$$

The weights can be estimated from example data using a variety of optimisation strategies or machine learning approaches (Support Vector Machines or Artificial Neural Networks).

Lanckriet et al. (2004b) used this technique for predicting 13 broad functional classes in yeast. Individual kernel matrices were produced from five different sources (Table A-2) and combined using the linear weighting scheme. Semi-definite programming techniques were used to extract kernel weightings for each matrix. The final weights were 2.21:0.18:0.94:0.74:0.93 for PFAM, gen, phys, TAP and exp kernels respectively, suggesting that for most functions the PFAM kernel presented the most valuable information for classification.

Ben-Hur and Noble (2005) introduced the TPPK kernel (the tensor product pairwise kernel)

$$\begin{aligned} K((x_1, x_2), (x_3, x_4)) &= K(x_1, x_3) \times K(x_2, x_4) + \\ &\quad K(x_1, x_4) \times K(x_2, x_3) \end{aligned} \tag{1.8}$$

for handling protein gene or protein pair feature inputs. Kernel matrices measuring similarity between feature pairs should exploit information about the similarity between individual pairs of genes or proteins. Vert et al. (2007) further developed the idea by introducing the metric learning

Table 1.6: Heterogeneous data sources and types of kernel matrix integrated together by (Lanckriet, Deng, Cristianini, Jordan, and Noble 2004b)

Data source	**Kernel**	**Definition**
K_{PFAM}	inner product	binary feature matrix encoding presence/absence of PFAM domains
K_{phys}	diffusion kernel	network of physical interactions
K_{gen}	diffusion kernel	graph of genetic interactions
K_{TAP}	diffusion kernel	protein complexes from tandem affinity purification
K_{exp}	inner product	77 cell cycle control gene expression measures encoded as a pairwise binary matrix

pairwise kernel:

$$\begin{aligned} K((x_1, x_2), (x_3, x_4)) &= (K(x_1, x_3) + K(x_1, x_4) - \\ &\quad K(x_2, x_3) + K(x_2, x_4))^2 \end{aligned} \tag{1.9}$$

This kernel was used to reconstruct yeast metabolic networks from microarray expression data, localization information, PFAM and PSI-BLAST profile sequence similarity. The TPPK kernel permits comparisons between the first pair of genes and the second pair of genes whilst the metric kernel emphasizes feature differences between gene pairs. Pairs of pairs that are similar using the TPPK kernel can be classed as different using the MLPK kernel. The complementarity between these two kernels was exploited by combining them using a simple sum for each data source. The resulting kernel was then used to classify function using support vector machines. This resulted in superior prediction performance compared with either kernel alone, achieving a final accuracy of 95%.

Bayesian network approach

Bayesian networks are probabilistic models representing dependencies between data items as nodes of a directed acyclic graph. The nodes may represent continuous or binary variables and the probability dependencies between nodes are learnt from example datasets during a training phase. The posterior probabilities generated from the learning phase permits inferences to be made about the nodes. The likelihood of some behaviour based conditionally upon evidence from surrounding nodes can then be evaluated. Bayesian networks have been successfully applied to modelling gene regulatory networks, predicting protein structures as well as in data fusion.

The MAGIC algorithm (Troyanskaya et al. 2003) combined expression data, interaction datasets and promoter sequence information in the form of pairwise inputs to a Bayesian network to predict biological processes. The general network structure (Figure 1.13) incorporated both varied data sources and multiple representations of the same data source. For example, the topology for representing expression data permits both abundance and co-expression information to be captured using different clustering algorithms. To establish robust probabilities that a pair of genes shared the same biological process conditioned on each piece of evidence, training pairs of yeast proteins were passed through the network. Prior probabilities that a pair of genes or proteins shared the same function were established by expert consultation due to lack of available data. The results indicated superior performance of the network when applied to GO term prediction, however the improvements gained by incorporating microarray data were slight, suggesting either that the co-expression information was better represented by other data sources or that the microarray data itself was not very informative.

K-nearest neighbour approach

The K-nearest neighbour algorithm (K-NN) is one of the simplest machine learning approaches. K is a positive integer number describing the size of a neighbourhood determined by a distance or similarity measure. A K-value of 2 for example, would consider the closest two nearest neighbours of a data item; two most similar co-expressed genes, or top two BLAST hits for a sequence of interest. The magnitude of the measure used to assign the nearest neighbour is not considered, but is used to rank data items to establish their relative order. This assumption is appropriate for biological data considering that a large amount of missing information from experimental data sources can lead to low magnitude scores from prediction algorithms. For these cases it is advantageous to identify the current nearest neighbour to provide some information rather than negative prediction results.

Deng et al. (2003) used the K-NN method to assign pairwise independent similarity values to different data sources and estimated the likelihood of the pairs belonging to the same functional class. The algorithm was applied to *E. coli* microarray and sequence data using co-expression correlation, chromosomal proximity, shared paralogues and a block indicator for pairs of genes transcribed from the same operon. The naive K-NN method used Euclidean distance between features to produce a pairwise distance measure which was compared against an SVM to predict different functional classes. The methods yielded high numbers of false positives, approximately an order of magnitude larger than the number of true positives. At an accuracy threshold of 50%, the sensitivity of the SVM method was greater than that of the K-NN method. The dataset weights for each individual data source were 1.87 : 0.05 : 1.68 : 4.63 respectively for expression correlation, chromosomal distance, operon block indicator and paralogue indicator data sources demonstrating that the majority of information was obtained from the paralogue indicator.

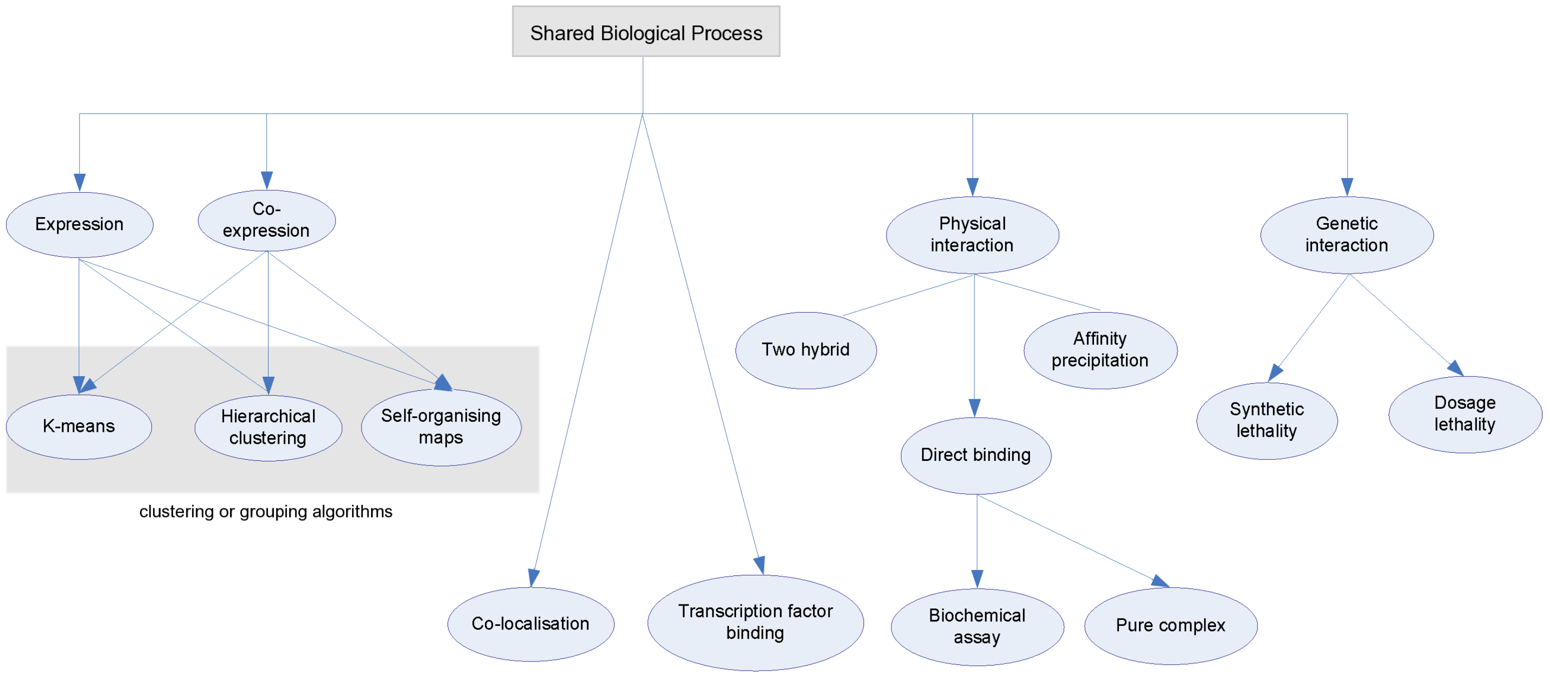


Figure 1.13: Bayesian network topology for integrating different biological data sources (adapted from (Troyanskaya et al. 2003).

Similar to the kernel methods, in the K-NN method, a single data source was responsible for the majority of the predictive power. One concern is that the dominance of these attributes masks information from other weaker features. Another challenge when combining heterogeneous data sources is missing data. Most datasets are incomplete or informative for just a few a genes or proteins. Restricting methods to cases where all data is present results in models with restricted applicability. Conversely data imputation techniques that replace missing values with estimates from a prior or randomised distribution can severely affect prediction quality (Deng et al. 2003). Poorly chosen or inappropriately estimated values can be detrimental to prediction performance by increasing false positives. Much work remains to be carried out in this area in order to achieve high quality high accuracy predictions of function.

1.5 Thesis aims

The main goal of the thesis comprised the development and implementation of methods to predict protein function using machine learning approaches. For development of methods and benchmarking procedures, human sequences were used. Human sequences were used in preference to other organisms, since human proteome annotations represent one of the largest, most challenging and well-studied eukaryotic annotation datasets. For definitions of function, the Gene Ontology system was selected due to its superior design in representing the biological behaviour of genes and proteins. The Gene Ontology provides evidence source information that can be exploited in benchmarking procedures to ensure unbiased testing. The second chapter characterises the current annotation status of the human proteome. The properties of the annotation system are described; annotation coverage and completeness estimates are provided. For subsequent benchmarking purposes, a baseline prediction accuracy for annotation transfer by sequence similarity is established.

To address the annotation of orphan and distantly homologous human proteins from sequence, the third chapter describes a feature based function prediction method based on the ProtFun approach. New feature attributes are derived and integrated into an ensemble sequence feature-based function prediction system. A performance benchmark for the system is established and applicability of the method to other eukaryotic genomes investigated. The latter part of the thesis involves the design of a separate function prediction methodology. In this approach sequence features are combined with information from microarray and predicted domain information. The design and integration work flow for incorporating pairwise feature inputs from microarray expressions is presented as a separate work. Similarly, the prediction of domain information and extraction of functionally informative features comprises another chapter. The final chapter evaluates different strategies for incorporating these feature attributes together to make probabilistic functional inferences.

Chapter 2

Characterising the system: the Gene Ontology and Human proteome annotations

2.1 Chapter aims

The Gene Ontology Annotation system currently comprises 2347 Cellular Components (CC), 16072 Biological Process (BP) and 9189 Molecular Function (MF) vocabulary terms to describe activities and component parts of all organisms. This chapter provides an in depth view of the GO Annotations currently available for human sequences. The number and types of annotations are quantified and coverage estimates for annotations and sequences are provided. Analysis of these features of the Gene Ontology and the process of annotation highlights important considerations for the design of a function prediction approach targeting recognition of GO classes from sequence.

The characteristics that are analysed include the shape and structure of each ontology, the specificity and completeness of human sequence annotations and the sequence annotation process. The shape and structure of each ontology relate to the type of modelling approaches that are useful for function prediction methods. Annotation specificities and completeness describe the current status of human sequence annotations. Sequences annotated with general functions, for example “Cellular processes”, or “Binding”, are of limited value in function prediction as little biological insight can be gained from making these assignments. By determining the volume of detailed and available annotation information for human sequences, realistic annotation coverage estimates for human sequences are defined and appropriate filters be put in place to generate higher quality data sets used to design and test a function prediction approach.

The process by which sequence annotation assignments are made provides information as to which biological characteristics might be useful to infer functions. Inflated performance estimates for some GO prediction methods has been attributed to bias in evidence sources from which annotations were made. One severe example of this is testing prediction methods that use homology on datasets that predominantly comprise homology based annotations (Rogers and Ben-Hur 2009). An analysis of the frequencies of annotation evidence sources for MF and BP category assignments to sequences reveals any bias in curated annotations so that it can be appropriately accounted for or removed from the data.

2.2 Shape and structure of the Ontology Graphs

The number of nodes in the MF and BP ontologies (more than 9000 and 16000 respectively) mean that visualisation of the entire structure to establish gross structural features of the graphs is impractical. However, several attributes can be analysed numerically using frequency information from the Directed Acyclic Graphs (DAGs). For example, surveying node (GO category) connectivity and position with respect to the root of each graph provides information describing the shape and organisation of each Ontology.

An overview of the shape of each graph using Biological Processes and Molecular Functions was obtained by recording the level of the GO annotation category from the root. This measure corresponds to the minimum path length between an annotation term and the graph root. Plotting the frequency of annotation terms at each level produced a histogram that approximates the layout of the graph (Figure 2.1 right panel). If the graphs contain increasing numbers of terms at each level the shape of the histograms would be triangular, whereas constant or low varying node frequencies at each level indicate a rectangular shape.

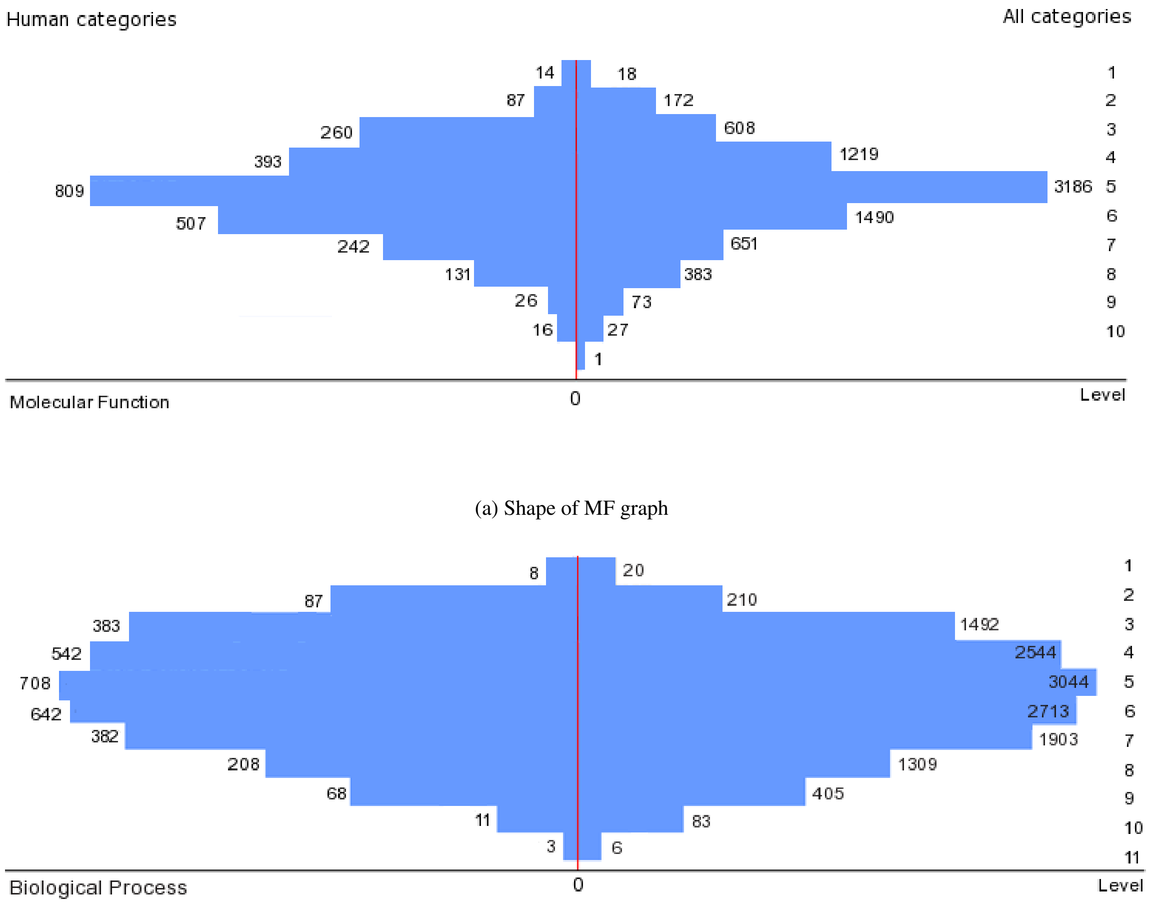
The MF Ontology tended towards a short wide pyramid shape (more tree like) whilst the BP Ontology was close to rectangular with more nodes populating lower levels from the root than in the MF graph. To determine whether the structure of each graph was influenced by organism specific annotation categories, the same node frequencies were calculated using only those categories that had been assigned to at least one human sequence and plotted as mirrored histograms (Figure 2.1 left side). Although the node frequencies for human annotation categories were smaller at each level there was clear symmetry between the left and right hand histograms confirming that the terms annotating human sequences represent an unbiased sample of all annotation terms, and that the shape of the graph constructed from nodes representing human only annotations was consistent with the shape of each entire graph. Additionally, this symmetry suggests that a large proportion of the human Gene Ontology assignments are from nodes close to the root of the each Ontology. If all annotations were low-level an asymmetry between left and right panels would be observed with no annotations at the upper levels.

To determine the degree of redundancy of annotation categories within each graph, connectivity was measured for each annotation term. This was determined by the number of parent annotation terms for each node (Figure 2.2). In this calculation the different types of relationship between GO terms; ”is_a” and ”part_of” were treated equivalently and considered evidence of parental linkages between annotations. This measure of connectivity within each graph provides topological information. A purely hierarchical annotation system, for example the FunCat or Enzyme schemes, support only single inheritance, that is each annotation category can only possess one parent.

Connectivity analysis revealed that BP annotation terms have on average more parents than MFs. This observation together with the triangular shape of the MF Ontology suggests that it is almost hierarchical. Only 1394 (15.36%) of annotation terms possess more than one parent linkage. Biological Process categories are more highly connected with over 6000 terms posessing more than one parent annotation term. This result may be a consequence of the different biological aspects described by the two Ontologies, and suggests that MFs describing activity and binding behaviour are rarely inter-linked whilst BPs are frequently inter-related. Combinations of specific BPs are responsible for a larger and more diverse set of more general BPs.

2.3 Human proteome annotations

Annotations for human protein sequences were obtained from the Gene Ontology Annotation release version 44. The human proteome sequences were sourced from the International Protein Index (IPI) database (Kersey et al. 2004). The IPI represents one of the most comprehensive protein resources amalgamating high quality and well characterised protein definitions from SwissProt (Bairoch and Apweiler 1996), RefSeq (Maglott et al. 2000) and Ensembl (Hubbard et al. 2002) with lower quality peptides from gene prediction algorithms (see Table 2.1). The dataset achieves a compromise between high coverage of annotated human sequences and quality of sequence information. Many of the sequences represent fragments of more complete products or are variants of highly similar sequences.



(a) Shape of MF graph

(b) Shape of BP graph

Figure 2.1: Mirrored histograms representing term distributions and position in the Molecular Function (MF) and Biological Process (BP) graphs. The x-axis represents annotation category frequency where the origin is marked in red. The scale for left and right hand panels are independent. The y-axis represents the minimum path length for an annotation term to the root of the Ontology graph. The histograms on the right hand side of each panel represent the total number of annotation terms at each level from the root term, whilst the left hand side records the frequency of categories annotating ≥ 1 human sequence.

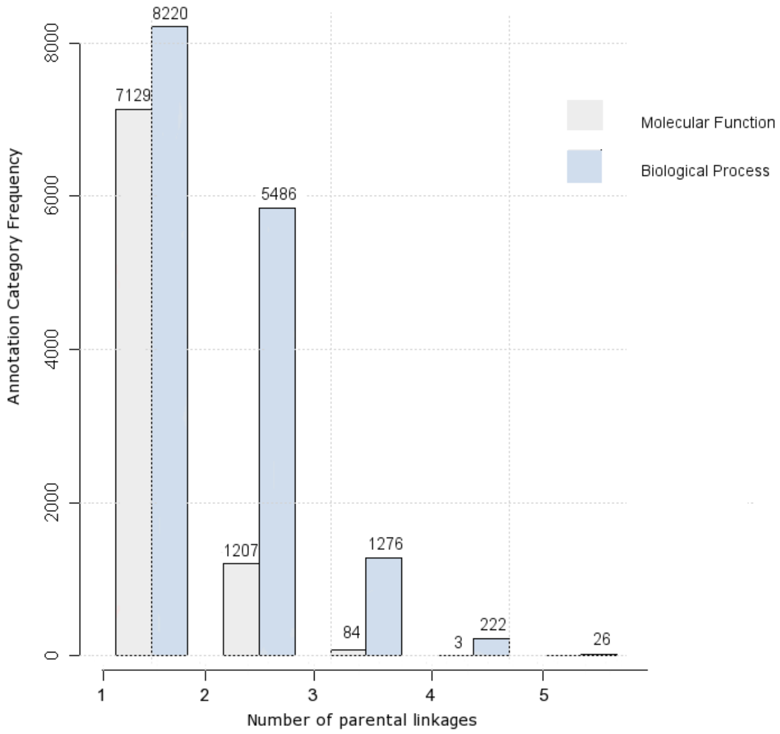


Figure 2.2: Connectivity graphs for annotation terms. The plot details the frequency of annotation terms (y-axis) occurring with different numbers of immediate parents (x-axis) within each Ontology.

Table 2.1: Composition of the IPI human dataset

Primary datasource	**Description**	**Number of proteins**
UniProtKB	High quality well characterised proteins	88451
Ensembl	Translations of mRNA verified gene predictions	46703
Vega	High quality manually annotated Vertebrate Genome Annotation database	51477
H-InvDB	Human Invitational database of full length cDNA clones translated into proteins	23204
RefSeq	Well annotated set of reference protein sequences	33268
Total	(unique proteins)	65,653

The counts in the third column represent redundant protein sequence counts.

2.4 Annotation specificity and completeness

The concept of annotation specificity applied to the Gene Ontology annotation terms relates to the generality of the annotation description. For example the annotation term “Receptor” is less specific than the annotation term “Serotonin receptor”. Completed annotations are those annotation assignments made to the most specific annotation categories; that is, those categories that are leaf-terms in the Ontology graphs.

Annotation specificity and completeness are two important concepts that affect the success of function prediction methods and benchmark performance assessments. Completeness affects the ability of specific functions to be predicted since the existence of few annotated sequences provides insufficient information for accurate model building. In prediction performance assessment, good results can be easily obtained for general annotation categories by chance alone, since they are often highly populated by sequence examples.

Mathematically, specificity can be expressed using a concept borrowed from information theory

$$GO_{spec} = ln(\sum x + 1) \tag{2.1}$$

where x may represent the popularity of each GO term in a set of protein annotation assignments or the relative position of the term with respect to the root term in the GO graph.

Both measures were calculated for the Gene Ontology category annotations for human sequences. Using the Gene Ontology Annotations for IPI sequences, there were 252262 protein-Molecular Function and 340374 protein-Biological Process term assignments respectively. A small proportion of these, 8.29% (20916) and 2.23% (7581) were complete leaf-term annotations.

The relationship between specificity of sequence annotations and the completeness was demonstrated for the two sets of annotation categories by plotting specificity measured using the frequency of child annotation terms against the frequency of sequence annotations for each annotation category (Figure 2.3). Completed annotations are assigned a specificity value of 0 and can occur with either high or low frequencies in a population of sequences. If all annotations were complete, then all datapoints would lie along the x=0 axis.

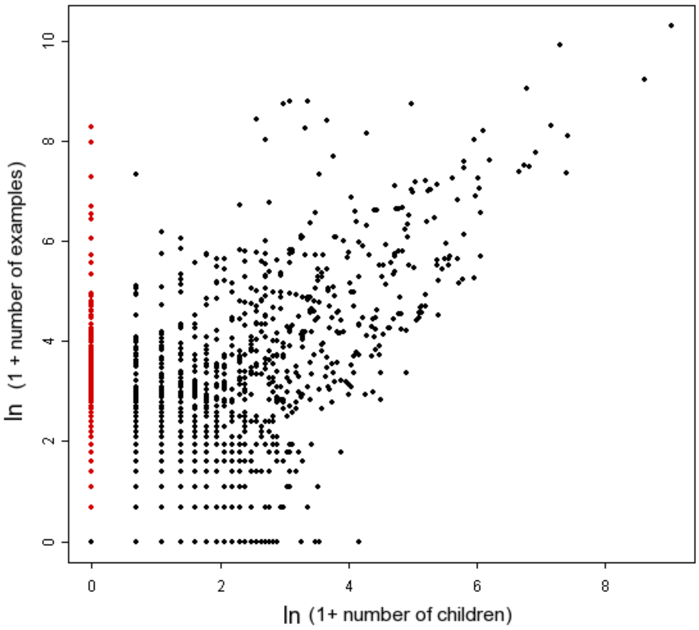
In the Molecular Function Ontology, there were more examples of completed annotations than for Biological Processes. This suggests an advantage in predicting Molecular Functions since more information is available for modelling each annotation category. The differences between MFs and BPs may also be linked to differences between the structure of the two ontologies determined by their descriptive nature. The lower connectivity for MFs suggests that even the most specific categories within this Ontology might be more general than completed leaf-term BPs (Figure 2.2).

2.5 Growth rates for human sequence annotations

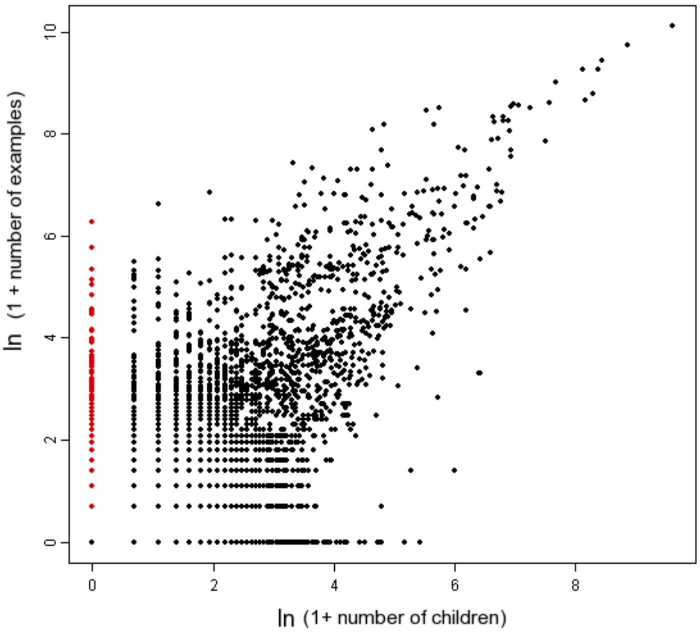
In higher eukaryotes, organism complexity arises from the functional plasticity of genome sequences (Lopez-Bigas et al. 2008). Consequently it is expected that human sequences carry out multiple functions and on average receive more than one distinct annotation. To verify whether this was the case, human sequence annotation frequencies were examined. The rate at which new annotations are made to sequences is of interest since prediction methods can become quickly out of date if the increase in annotations between database releases is high. To establish trends in the number of annotation assignments made to human sequences, annotation frequencies were recorded at 6 monthly intervals between December 2003 and December 2008 (Figure 2.4).

Currently there are 121,789 Molecular Function and 93,480 Biological Process assignments to 29879 and 25094 human IPI sequences respectively. On average each human sequence receives more than 3 distinct MF and BP annotations. Since 2003, the frequency of sequences with at least one annotation has plateaued (dark grey bars Figure 2.4) as the number of sequences in the human genome has stabilised (Pennisi 2003b). The increase in total annotations per release results primarily from the additional annotation of partially characterised sequences rather than from new annotations of uncharacterised sequences. The growth trend is linear for Biological Processes however has stabilised for Molecular Functions. This small increase may relate to the difficulty in de-orphanising novel sequences by experimental means. Some sequences are not experimentally tractable, either because they are difficult to clone or express or are unstable as monomers *in vitro*.

To determine the source of novel sequence annotations, similar frequency plots were made divided into the different evidence codes over the same time period (Figure 2.5). Electronic annotations (IEA), the major source of most annotations are unreviewed and predominantly result

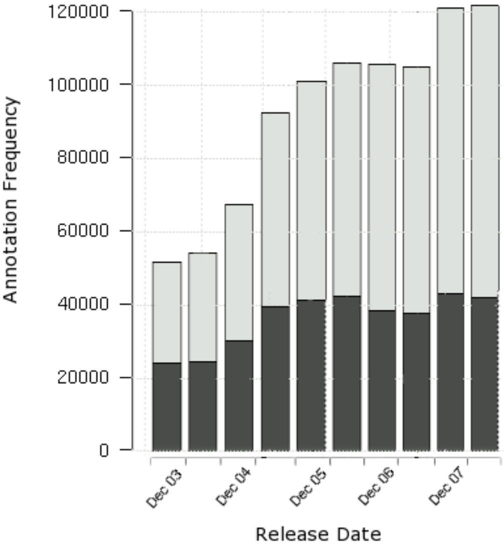


(a) Coverage of Molecular Function Graph

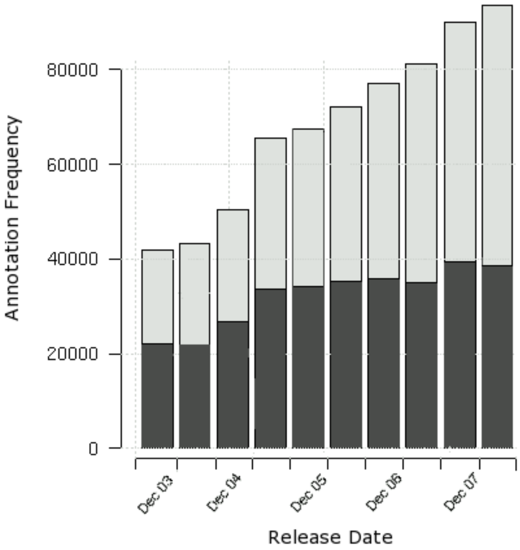


(b) Coverage of Biological Process Graph

Figure 2.3: Annotation completeness and specificity. Each datapoint represents an annotation category. The x-axis represents the term specificity approximated by the natural log of the number of child terms. The y-axis represents the natural log frequency of annotated sequences (number of examples) for each GO term. The red highlight (x-axis = 0) values denote completed annotations.



(a) Molecular Functions



(b) Biological Processes

Figure 2.4: Annotation growth rates. Growth in the number of human annotations at 6 monthly intervals since December 2003. The dark portion of the bars represents the number of distinct sequences that are annotated per release whereas the lighter shade represents the total number of sequence-annotation relationships.

from computational sequence similarity searches. Annotations from other sources are derived from experimental or literature studies and are reviewed prior to approval. However, the increase in novel annotations from this source has declined for MF categories, and is stabilising among BP and CC annotations. This may be due to quality control efforts that have discarded some of the IEA annotations. Traceable author statement annotations (TAS) are considered the most reliable source of annotations and comprise the second most frequent annotation class. The frequencies of other evidence sources were much lower; the proportion of Not Recorded (NRA), Non-traceable Author Statement (NAS), Not Determined (ND) and Protein Interaction (IPI) categories represented fewer than 2% of the total records.

In recent database releases, the frequency of TAS annotations has declined, perhaps as a consequence of the low rate at which new functional information arises in the literature. In contrast, annotations sourced from Direct Assays (IDA) have increased across all Ontologies. In part this reflects developments in experimental technologies that permit existing assay experiments to be run at scale, as well as suggesting the existence of new functional assays. Annotation assignments made from protein-protein interaction experiments have rapidly and recently increased in the Molecular Function Ontology. These are primarily attributed to 1185 terms representing molecular binding events.

This information implies that we are approaching the limits of computational annotation methods that use sequence similarity information. Other characteristics of sequences for example, protein interactions or expression information may in future play an increasingly important role in function assignments.

If annotation growth is set to continue in a consistent manner over subsequent annotation database releases, it is conceivable that a future status where all sequences are characterised by at least one annotation might never be reached. This justifies the need for accurate computational prediction methods that do not require sequence homology information.

2.6 Datasets for benchmarking

Considering features of the GO, the annotations described in the previous section, and their implications for function prediction methods, several working datasets were created. Function prediction methods tend to address two distinct challenges. The first is to correctly distinguish

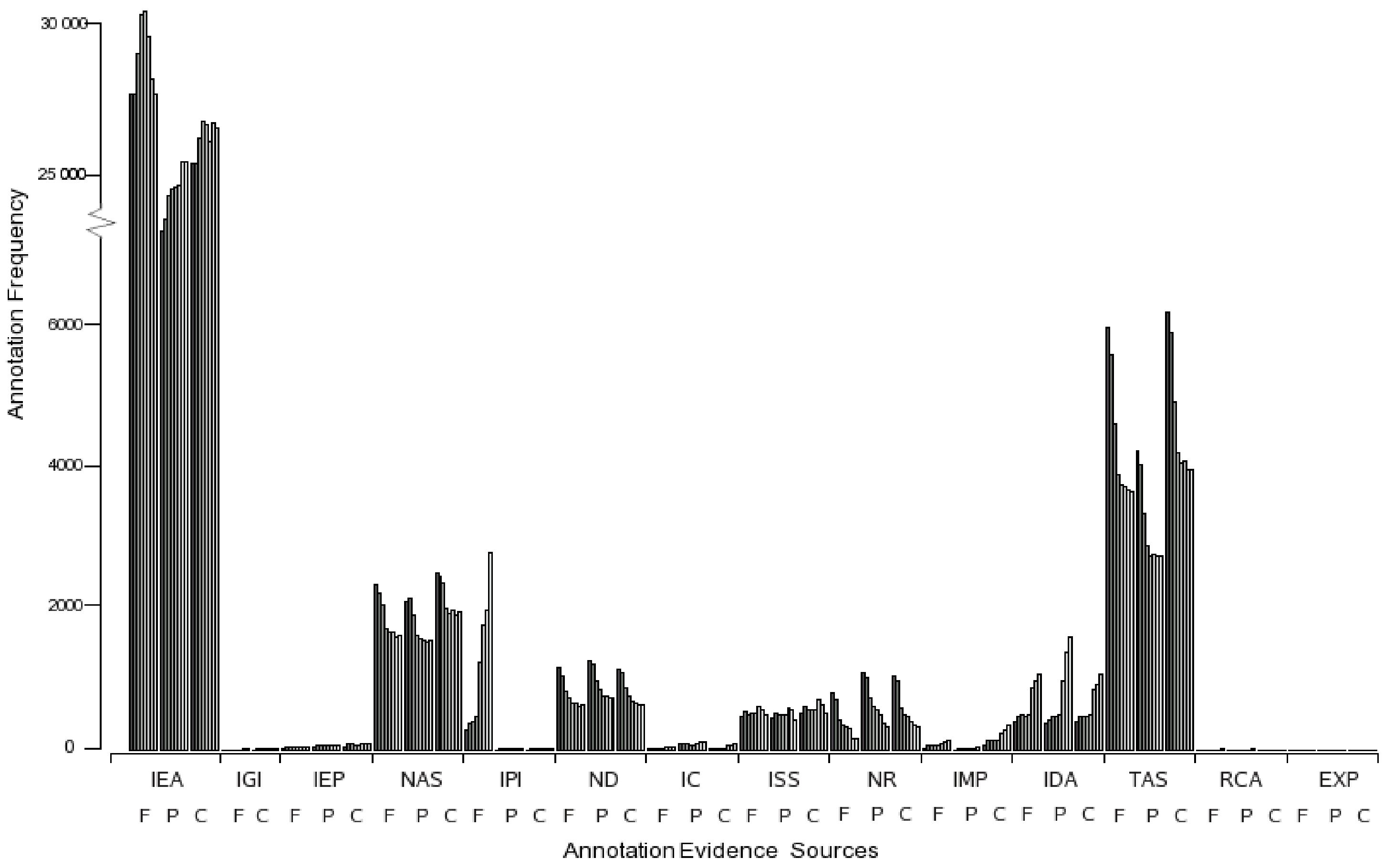


Figure 2.5: Annotation evidence sources for the Gene Ontology human annotation categories. For each of the different sources, annotation frequencies are shown for Molecular Function (F), Biological Process (P) and Cellular Components (C) categories for the database releases ordered by date. No Cellular Component annotations from IGI (Genetic Interaction) source were observed during this time period.

function between closely related sequences, and the second is to assign correct functions to sequences that are distantly related, or that are orphans. Function prediction methods using homology information should be assessed using test datasets reflecting their application area. Lower performance limits can also be established by testing these methods using datasets that do not contain homologues, or are at least filtered to reduce the occurrence of highly similar sequences. Those methods that address the annotation of non-homologous sequences must be tested on those sequences that cannot be annotated by homology based methods. For predictive modelling of annotation categories, a sufficient number of sequence examples are required per annotation category in order to determine patterns that correlate with function.

Focussing on the considerations highlighted in the previous sections, several working datasets were designed for use in modelling and testing the performance of different function prediction methods (Table 2.2). The nr80 and nr35 datasets are designed for performance testing of prediction methods on easy and difficult cases. The nr80 sequences are similar but not identical, whilst the nr35 sequences contain more distant homologues. An additional annotation category specificity filter was applied to the both datasets so that annotations comprised leaf terms or terms more than four levels from the root. These datasets were also balanced in terms of annotation evidence sources by sampling the IEA annotations such that they contributed no more than 50% of the total sequence annotation assignments.

The nr60 dataset was not filtered for annotation specificity since it is designed for modelling annotation categories. Instead the dataset required maximal reduction in similar sequences for a minimal loss in annotation coverage. This threshold was determined by characterising the relationship between sequence identity and representation of annotation categories (see Figure 2.6). At more stringent similarity thresholds, the representation of function categories was compromised. All identity filters were made using the BLASTCLUST algorithm (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt 2007) using a coverage threshold of 30%.

2.7 Chapter Summary

By characterising the properties of the GO and the process by which annotation assignments are made to sequences, several constraints in designing and evaluating function prediction methodologies were exposed. MF and BP Ontologies differ substantially in their topologies. MF annotation terms are arranged in a broad, flat tree like hierarchical structure implying some similarity

Table 2.2: Dataset definitions and sizes

Dataset	Description	MF Sequences	BP Sequences
all	Full set for characterising features of all annotations	29879	25094
nr80	Contains close homologues yet distinct protein sequences that are annotated with terms at least 4 levels from the root, or leaf terms. Evidence source balanced.	5469	7322
nr60	Optimal dataset minimising size whilst retaining representation of the majority of function categories.	14086	14098
nr35	Homology reduced set containing sequences annotated with terms at least 4 levels from the root, or leaf terms. Evidence source balanced.	4096	5813

Each count represents the total sequences within each dataset annotated with Molecular Functions and Biological Processes.

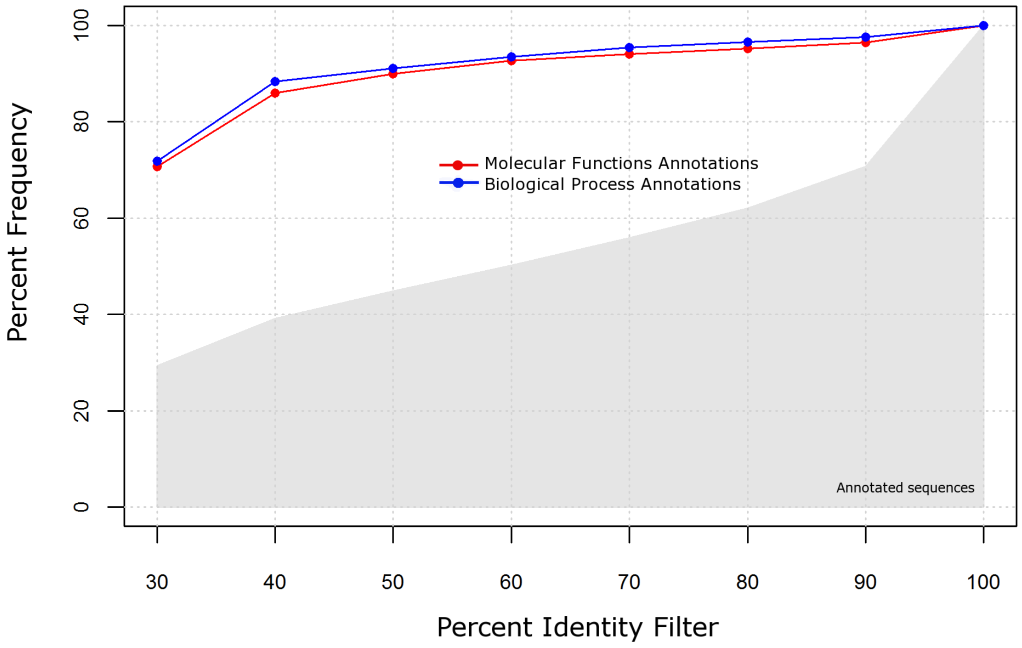


Figure 2.6: Effect of dataset reduction on representation of functions. The y-axis details the percentage of total sequences retained (gray shaded area) and the percentage of annotations retained for Molecular Function and Biological Process annotations (red and blue lines respectively) using different sequence identity filters (x-axis).

with other well defined annotation schemes like the Enzyme or FunCat hierarchies (see Chapter 1 Section 3). For these MF categories, more sequence examples exist for specific annotation categories. Consequently more information is present for robust modelling procedures. Methods that have been successful in predicting Enzyme or FunCat annotations are likely to be applicable to the task of predicting MFs.

The BP Ontology is more complex in structure than the MF Ontology, possessing more specific (low level) annotation categories with fewer sequence examples. Since less information is available for specific BPs, and the frequency of inter-node connections are higher, it is anticipated that prediction of these function categories might present a more challenging task.

For benchmarking purposes, bias of ascertainment in annotation sources can strongly affect prediction performance. Frequently, approaches are assessed using datasets originally annotated by methods similar to those being tested. This can produce mis-leading performance statistics that do not truly reflect the additional value of an approach in reliably predicting novel annotations. A recent study reported a performance difference of over 10% when electronically sourced annotations were removed from test datasets (Rogers and Ben-Hur 2009). Evidence source bias was controlled by randomly sampling sequences that were annotated electronically as a post filter to homology based reduction.

One striking observation about the GO was the current state of annotations for human sequences. Considering that the human genome represents one of the most highly studied amongst higher eukaryotic genomes, the occurrence of just 2 and 8% complete MFs and BPs annotations was much smaller than expected. This demonstrates that much functional information remains to be uncovered to increase our understanding of human biology. Consequently, there is a great need to develop fast and accurate computational function prediction approaches to meet this challenge.

Chapter 3

Quantifying homology based annotation transfer

3.1 Chapter introduction and aims

The simplest, most widely used automated methods for function annotation involve annotation transfer between similar sequences. Identifying the nearest neighbour of an unannotated sequence is carried out by sequence homology searches followed by subsequent transfer of annotations from well characterised relatives. The approach implies that the annotated neighbour is an orthologue, or paralogue of the query sequence that has retained common function throughout evolution (Copley et al. 2002).

Generally, the functions of similar sequences are conserved. However sequences with shared ancestry can acquire new functions propagating erroneous annotations throughout biological sequence databases (Devos and Valencia 2000, 2001, Joshi and Xu 2007, Wilson et al. 2000). One difficulty lies in the use of sequence similarity to infer orthology or paralogy. This is not straightforward since most sequence alignment algorithms struggle to discriminate familial relationships in the twilight zone (between 25% to 35% sequence identity) (Jaroszewski et al. 2000, Joshi and Xu 2007, Rost 1999).

Quantification of the relationship between sequence similarity and function has been the subject of numerous recent publications (Gerlt and Babbitt 2000, Joshi and Xu 2007, Punta and Ofran 2008, Sangar et al. 2007, Wilson et al. 2000). The degree of sequence conservation within Enzyme families (defined in the Enzyme Classification scheme) varies in the literature (Devos and Valencia 2000). One study claimed that 3rd level Enzymes Classifications were conserved at 25% sequence identity (Wilson et al. 2000), whilst others reported 50% sequence identity. For conservation of 4th level classifications, threshold identities of 40%, 50% and 70% have been reported (Gerlt and Babbitt 2000, Hegyi and Gerstein 2001). These ambiguities suggest that sequences evolve at different rates within enzyme families. The nature of the annotation (its depth of description) also relates to the degree of sequence similarity that is useful for inferring function. Wall et al. (2005) suggested the different evolutionary rates of sequences are related to the dispensability of their function(s). This implies that sequences participating in essential cellular processes evolve more slowly and demonstrate a high degree of conservation that other sequences. Sequences whose functions can be carried out by other molecules or those whose absence subtly influences phenotype are dispensable and subject to more rapid evolution.

Further ambiguity in quantifying the relationship between sequence and function arises from subtle differences in scoring methods applied to annotation transfer. This is likely to be the cause of the conflicting performance statistics and thresholds that have been reported in the literature. Frequently, the transfer of only a single closest annotation between two sequences is considered. This results in enthusiastic accuracies since false positives are not accounted for. However, the definition of a false positive annotation assignment itself is ambiguous since an unverified annotation might represent a correct result.

In Chapter 2, it was observed that on average human sequences are annotated to multiple function classes. This further complicates the relationship between sequence and function since sequence similarity is typically defined as the single best alignment score between a pair of sequences. Because function assignments are made to sequences without positional information, it is difficult to relate a region of sequence similarity with a single function using a single quantity.

Another consideration is the origin of sequence information from which annotations are transferred. Chervitz et al. (1998) and Mika and Rost (2006) have suggested that protein-protein and protein-DNA interactions, which are important determinants of function, are more conserved within than between species. If this property is a feature of sequences with common GO classes, it is important to determine to what extent this might affect the accuracy of annotation transfers. Most homology based function prediction approaches transfer annotations between orthologous sequences, whereas most integrated function prediction approaches are carried out using the sequences of single genomes (Friedberg 2006, Lanckriet et al. 2004b, Troyanskaya et al. 2003). It is therefore important to quantify information loss that results from searches restricted to intra-species comparisons.

Recent studies characterising the relationship between sequence similarity and GO classes have focused on theoretical aspects of the relationships rather than the practical issues surrounding annotation transfer practices (Joshi and Xu 2007, Sangar et al. 2007). For example, sequence identities derived from local sequence alignments have been used to measure sequence similarity, however sequence relatives are more commonly ranked by statistical Expectation scores (E-values). Realistic error rates for the relationships are rarely given, and many studies have used subsets of function annotations or a handful of well defined functions to make an assessment. For example, performance obtained for enzymes cannot be generalised to other function categories since enzymes represent an unusual case of function. Their specific nature means that the majority are responsible for catalysis of a single reaction and rarely receive multiple function annotations. In contrast signalling molecules are functionally diverse and can interact with many different partners, therefore participate in multiple functions and processes (Lopez-Bigas et al. 2008).

The purpose of this Chapter is to investigate the limitations of homology-based annotation transfer using standard procedures. Two algorithms are frequently used to identify similar sequences, BLAST (Basic Local Alignment Search Tool) and PSI-BLAST (Position Specific Iterated BLAST) (Altschul et al. 1990, 1997). The PSI-BLAST algorithm extends the ability of the BLAST algorithm to detect remote relationships by performing iterated database searches to produce sequence profiles. By comparing performance between the two algorithms, the accuracy of the sequence profile alignment scores is compared with pairwise sequence alignment scores to detect function.

The effect of function heterogeneity on annotation transfer performance is also investigated by applying both locally optimal scoring thresholds and a single global score threshold to sequence relationships. These measures are also considered for human-human sequence relationships and multi-species relationships. The results of these comparisons are used to determine best practices in homology based annotation transfer for detecting MFs and BPs. In doing so a baseline performance for the method is established that can be used for comparing performance of other prediction methods.

3.2 Conducting homology searches

3.2.1 Datasets

The human proteome and UniRef Gene Ontology annotations from the GOA database were used as the basis of this study. The corresponding fasta sequences were obtained from human IPI (Kersey et al. 2004) and UniRef (Leinonen et al. 2004) databases comprising 28,966 and 4,002006 annotated sequences respectively. For intra-species comparisons, comparisons were performed using each human protein as a query sequence against the database of human sequences. Inter-species comparisons represent the results for searches computed between human and UniRef sequences after human sequences (those with taxonomy code 9606) were removed. For the human database search, the database size parameter (-Y) was fixed at a value equivalent to size of the UniRef database in order to ensure that the resulting E-value statistics were comparable between the two searches. The E-value threshold for BLAST and PSI-BLAST, and inclusion threshold (h) for PSI-BLAST were set to 0.001.

3.2.2 Scoring sequence similarity

For all searches, the top scoring local alignment (the highest scoring pairwise match) between two sequences was used to derive three sequence similarity measures, bit score, E-value and sequence identity. The bit score given as

$$S\prime = \frac{\lambda S - lnK}{ln2} \tag{3.1}$$

and represents a normalised alignment score S derived from the sum of pairwise amino acid substitution scores measured between pairs of aligned amino acids. The normalisation factors Kappa (K) and Lambda (λ) are statistical parameters estimated from the scoring system used and the background amino acid frequencies of the sequences being compared. The E-value is given by

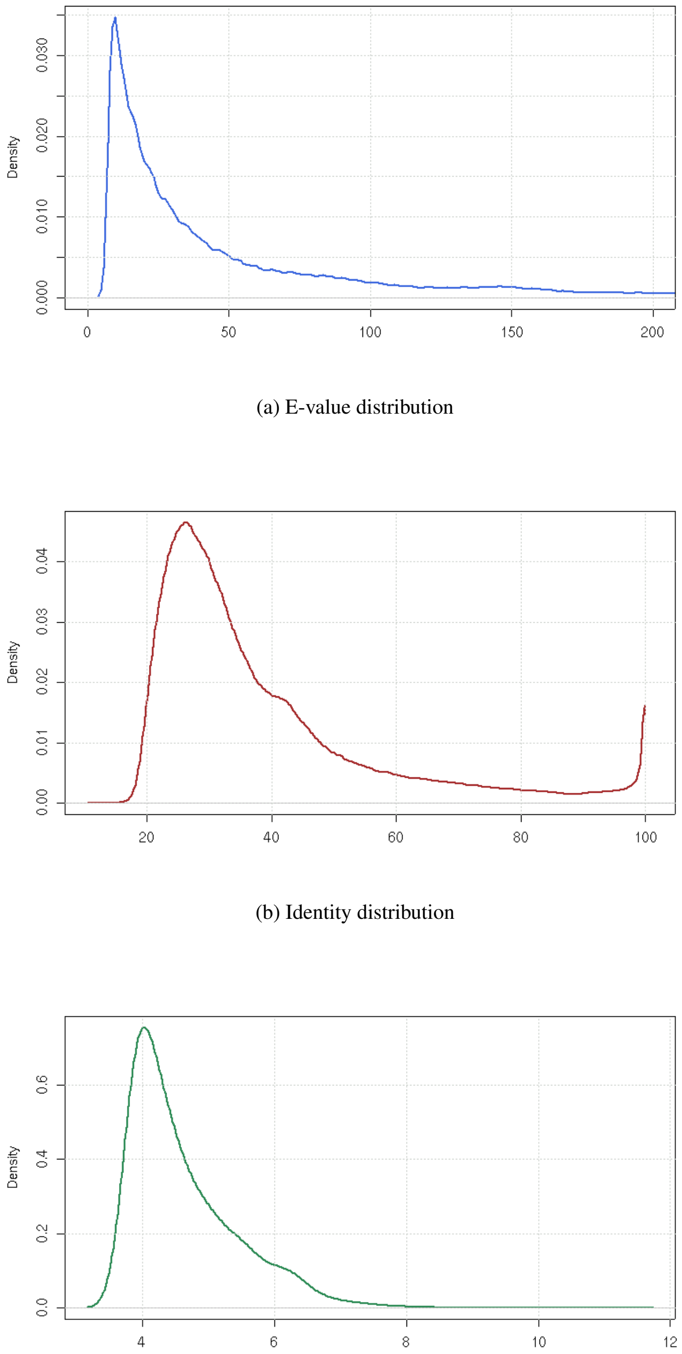
$$E = mn2^{-S\prime} \tag{3.2}$$

and represents the significance of the bit score $S\prime$. The E-value is the frequency of the observed bit score obtained during a database search of a given size when two sequences are related, compared against the occurrence of random sequence matches during the search. Sequence identity is given by

$$Identity = \sum_{i}^{i=1} \frac{1}{len} \tag{3.3}$$

and corresponds to the number of identical pairs of amino acids aligned between two matching sequences.

To determine the most appropriate measure for pairwise sequence relationships, the distributions of bit scores, E-values and identities were compared across all BLAST searches performed between human sequences and the UniRef database (Figure 3.1). Within results from a single query sequence, pairwise alignments ranked by sequence identity are similar to those obtained using E-values and bit scores. However, in this approach to evaluate the relationship between sequence similarity and function, each candidate measure of sequence similarity was compared from alignments resulting from all query sequence searches performed against the same target database. Sequence identity did not discriminate values at the lower end of the spectrum, whilst the E-value, designed to evaluate significance and not a measure of similarity between two sequences, was insensitive at the higher end of the distribution. For bit scores greater than 635, the precise E-value was so small that it approximated 0 which meant that high scoring alignments could not be effectively discriminated by E-value (Figure 3.1a). Additionally, the E-value statistic depends on the length of each pair of sequences so that highly similar alignments computed between different length sequence pairs obtained a different statistical significance value. Bit scores represented an intermediary scale possessing greater resolution than the E-value and the desirable property that the same alignment between two different sequence pairs obtained a consistent value (Figure 3.1c). Unlike sequence identity, the bit score increases with alignment length. This is a practically useful distinction when measuring sequence similarity since longer alignments are more likely to represent genuine evolutionary relationships (Healy 2007). Finally, bit scores are stable and can be compared between searches carried out against different databases using the same alignment algorithm.



(a) E-value distribution

(b) Identity distribution

(c) Ln bit score distribution

Figure 3.1: Distribution of sequence similarity measures, Expectation value (E-value), identity and natural log of the bit score computed between human sequences and the UniRef database.

3.2.3 Scoring annotation transfers

The scoring of successful annotation transfers would at first appear a simple task, however practically requires careful interpretation since annotations are related to one another. Subtle differences in scoring approaches can greatly influence the results of an analysis especially in cases where pairs of matched annotation terms are related to one another, but not identical. This forces consideration of inheritance within the GO term graphs. To control for the effects of partially and redundantly annotated sequences, only the most specific leaf terms or $\geq$ level 4 annotation matches were considered.

In the scoring procedure adopted by this benchmark (Figure 3.2) identical pairs of annotations between sequences were recorded as correct matches. The study was designed to assess the transfer of annotations from well characterised sequences to poorly characterised sequences, consequently the directionality between matches is important. If an annotation term from the matched sequence (Sequence B, Figure 3.2) was a parent term for a query sequence annotation, it was considered a correct match. If a term from the matched sequence (Sequence B Figure 3.2) was a child term of any of the query sequence annotations it was considered ambiguous and the result omitted from the assessment. Terms from the matched sequence that were not related to any of the query sequence annotations were considered incorrect and penalised. These matches represent a mixture of false positive annotations and potentially correct, but not verified transfers. Since it is impossible to discriminate between these cases, the occurrence of false positive annotations reported in the results is less important than the recognition of true positives as they conceivably represent correct novel annotations.

Annotation transfer performance was judged by comparing actual true positive and false positives between two test datasets, all human sequences and human sequences filtered at 35% identity. In addition, two scoring methods were considered, scoring all annotation matches, and annotations from the highest scoring sequence match only. Scoring transferable annotations for all sequence relationships examined the ability of the bit score to correctly identify functionally equivalent sequences. Considering annotations transferred between close relatives represents common practice in whole genome annotation. In this approach the rank of a sequence relationship is important regardless of the magnitude of the similarity measure. Performance measured using the nr35 dataset determined the behaviour of the approach when detecting more distant sequence relationships. Transferring annotations between all sequences enabled performance

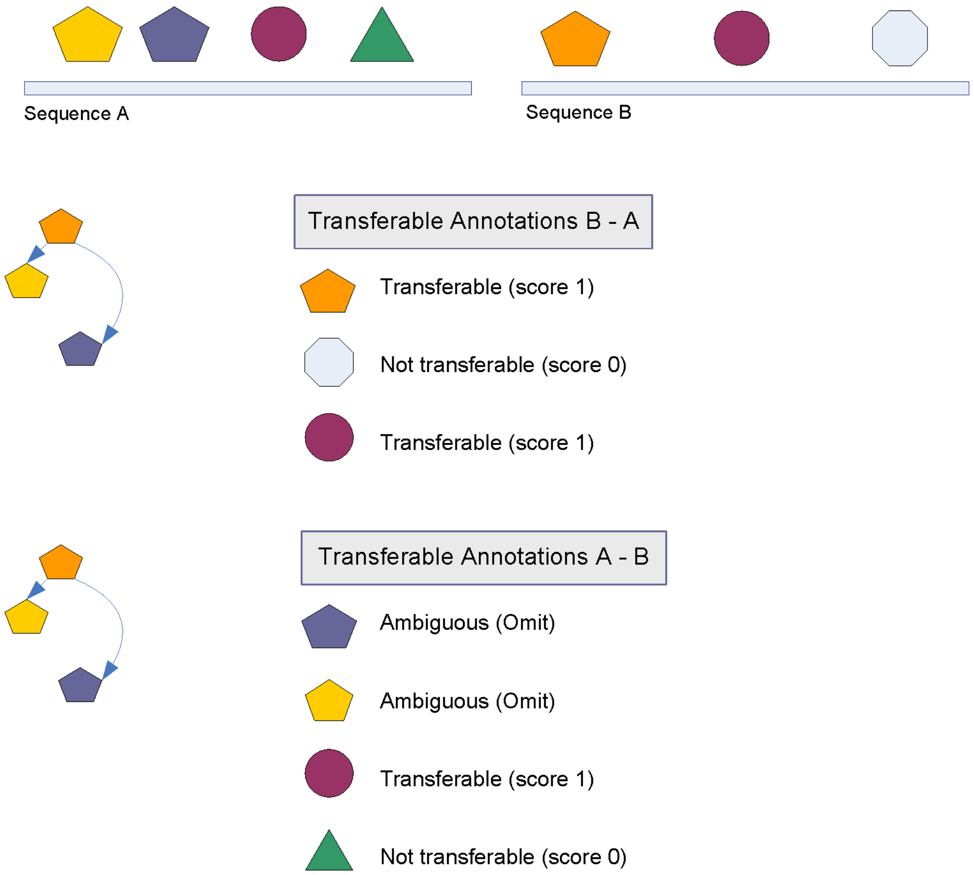


Figure 3.2: Directional scoring method for annotation transfers. Similar shapes are related to one another within the GO hierarchy. Ambiguous transfers are not scored whilst multiple similar transfers are only scored once.

between highly similar sequences to be determined. The use of inter and intra-species relationships to make function assignments permitted the degree of conservation between functionally equivalent sequences between and within species to be assessed.

The results from the benchmark were judged using standard performance metrics, sensitivity, specificity and Matthew's Correlation Co-efficient (MCC). These metrics are derived from confusion matrices which assess the numbers of True positive (TP), False positive (FP), True negative (TN) and False negative (FN) results obtained at a particular score threshold. Sensitivity ($\frac{TP}{TP+FN}$) is defined as the proportion of true positive values obtained at a score threshold against the total number of positive test cases. Specificity represents the proportion of false positives that are correctly recognised at a particular score threshold and is given as $\frac{FP}{TN+FP}$. Precision is defined as $\frac{TP}{TP+FP}$ and represents the proportion of correct test cases observed at a given score threshold out of all test cases identified at that threshold. Finally MCC defined as

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)} \tag{3.4}$$

is the class based equivalent of Pearson's correlation coefficient used to quantify performance on a scale between -1 and 1. A value of 0 implies random performance, and 1 indicates perfect classification.

3.3 Global scoring for annotation transfer

A global score threshold for annotation transfer represents the application of a single bit score threshold to all sequence relationships above which common function is assumed. This implies that all relationships between sequence similarity and function are homogeneous. The threshold was selected by varying the bit score and determining the value at which the Matthew's Correlation Coefficient (MCC) was maximised. Performance obtained using different datasets and methods was compared at each threshold using sensitivity (coverage, or proportion of true positives), specificity and precision measures (Table 3.1). MCC coefficients were not comparable between datasets when scoring all annotation matches since the measure is sensitive to the different numbers of sequence relationships detected. Specificity values (proportion of true negatives recovered) were also uninformative for performance comparisons between methods. Specificity

Table 3.1: Performance statistics for annotation transfer using a global threshold.

Dataset	**MCC**	**Score**	**Sensitivity**	**Specificity**	**Precision**	**Total Pos**	**Total Neg**
Molecular Function (top hit)							
Human	0.22	142	92.5	24.0	81.2	26761	6196
Human psi	0.36	58.9	97.9	24.8	78.1	28419	7950
Other	0.22	125	96.4	15.8	87.2	33186	4879
Other psi	0.23	95.9	97.7	16.0	88.5	32954	4533
Human nr35	0.32	166	75.2	57.5	77.6	4577	1322
Human psi nr35	0.31	151	62.0	43.8	57.6	4297	3163
Other nr35	0.44	154	89.3	43.9	86.7	6888	1059
Other psi nr35	0.35	145	89.8	43.8	86.9	7066	1062
Biological Process (top hit)							
Human	0.20	151	90.8	24.2	60.0	23660	15758
Human psi	0.31	151	98.4	34.3	58.4	23974	17043
Other	0.23	149	96.4	15.8	72.8	26266	10248
Other psi	0.17	155	93.8	13.2	71.1	26425	10728
Human nr35	0.31	151	76.4	43.8	57.6	4297	3163
Human psi nr35	0.41	110	78.5	63.7	58.8	4514	3591
Other nr35	0.34	155	92.4	65.0	73.4	7622	2765
Other psi nr35	0.26	123	94.7	25.6	74.5	8083	2769
Molecular Function (all hits)							
Human	0.28	81.6	72.2	42.8	80.2	1061470	261379
Human psi	0.57	41.6	68.9	88.5	67.0	2188890	1077954
Other	0.48	41.6	95.4	43.1	84.3	5451861	1016800
Other psi	0.17	139	68.6	53.0	86.9	5802124	869653
Human nr35	0.24	70.5	75.6	81.5	72.7	73430	49903
Human psi nr35	0.37	34.5	66.3	15.2	40.6	268397	396640
Other nr35	0.47	41.6	89.1	54.4	72.2	893597	326724
Other psi nr35	0.17	189	41.8	76.5	81.1	605520	140745
Biological Process (all hits)							
Human	0.31	209	37.0	88.5	83.7	514560	160042
Human psi	0.49	57.7	66.7	87.7	45.2	1713753	1516805
Other	0.35	43.5	94.5	13.2	67.6	3237157	1550746
Other psi	0.11	137	38.0	25.1	69.9	3212318	1385444
Human nr35	0.20	92.4	40.2	79.8	42.0	42613	58963
Human psi nr35	0.30	67.0	32.8	95.4	61.8	113881	184151
Other nr35	0.40	46.2	53.9	49.0	59.7	540348	364973
Other psi nr35	0.18	223	21.7	69.9	77.3	314721	135706

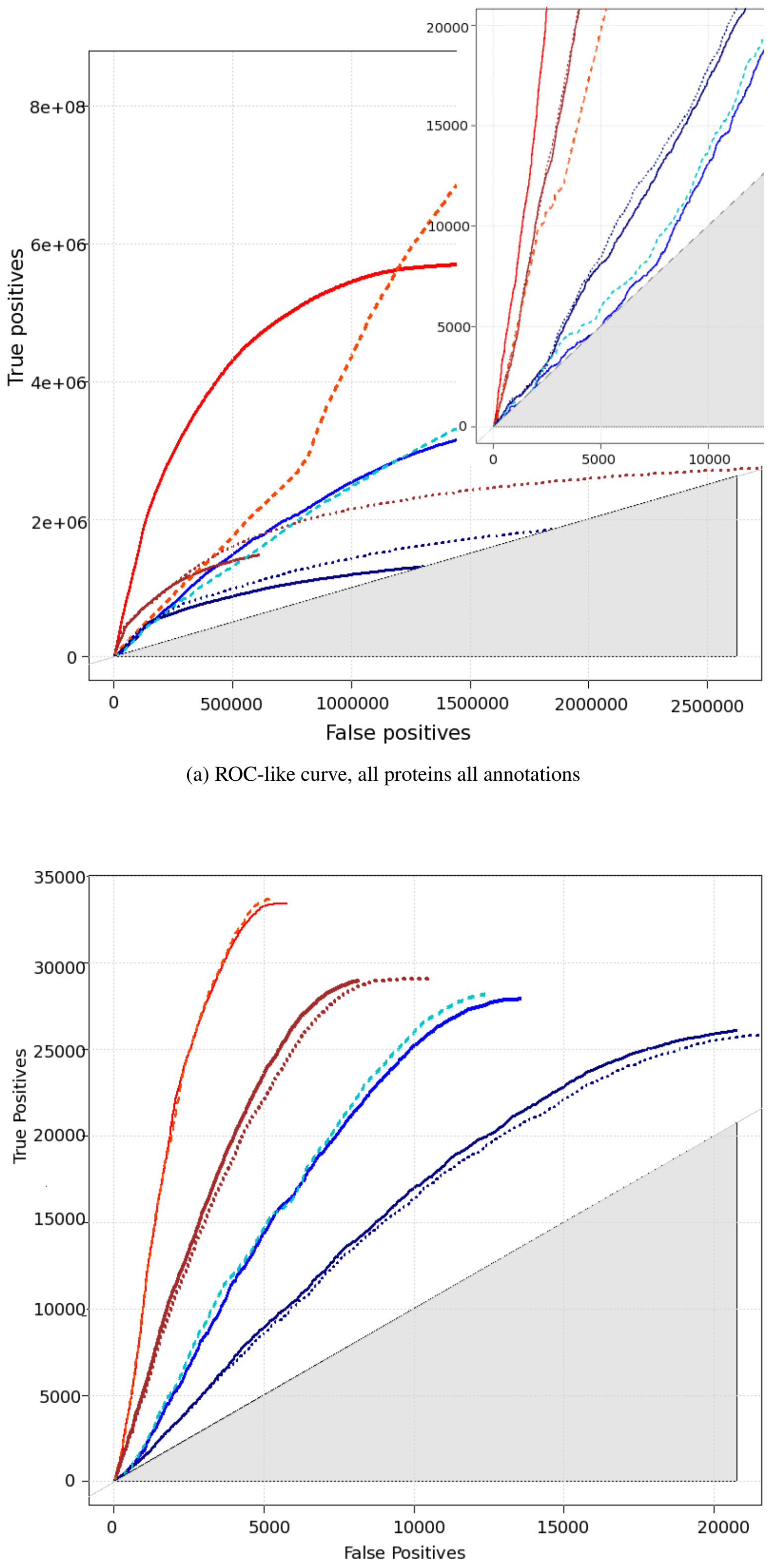
Results obtained using the PSI-BLAST algorithm rather than BLAST are labelled as “psi”. Human only annotation transfers are termed “Human” and reported separately from other species transfers (Other). The statistics reported are Matthew’s Correlation Coefficient (MCC), sensitivity, specificity and precision. The results are separated into two parts. The first represents annotation transfer between all sequence relationships (all hits) and the second represents closest relatives (top hit) only. Within these results datasets are filtered at 35% identity (nr35) or unfiltered (all sequences).

is influenced by the frequency of detected relationships that could be classed as negatives. Both low and high specificity values are tolerable providing the ratio of true positives to false positives (overall precision) is high. The actual number of true positive and false positive values provided comparable estimates of the performance of each method. Between 62 and 98.4% of annotations were correctly recovered at optimal bit scores using top hit annotations for both MF and BP categories. Overall precision ranged between 43 and and 87.1%. The poorest performance statistics were observed using PSI-BLAST relationships to detect MFs and BPs for sequences in the nr35 dataset. The most accurate results used the BLAST algorithm to identify inter-species relationships for annotation transfer. Overall the total number of correct results reported at this optimal threshold was high. However, the occurrence of false positives was also high, particularly for Biological Process categories where the number of negatives obtained when transferring all annotations exceeded 10,000. Despite consideration of the fact that some of these false positives represented novel and correct predictions, these numbers are unacceptably high for practical use on a whole genome scale.

Annotation transfer using all sequence relationships (lower half of Table 3.1) emphasised the differences between dataset sizes for intra and inter-species sequence similarity searches. More than five times the number of relationships could be detected using the optimal bit score threshold applied to relationships between human and other species sequences for MFs using BLAST. A six fold increase was observed for relationships detected between human and other species for Biological Process categories. This difference emphasised that the statistical likelihood of detecting a functionally equivalent sequence relationship between species exceeded that observed for searches performed between human sequences. This results from greater numbers of sequences present with common function. Despite this fact, when the top annotated hit only was considered, the improvement in recognition of correct annotations was smaller than expected; 2606 additional Biological Process and 6425 additional Molecular Function annotations were identified using inter-species than intra-species BLAST relationships.

3.3.1 Comparing Molecular Function and Biological Process annotation transfer

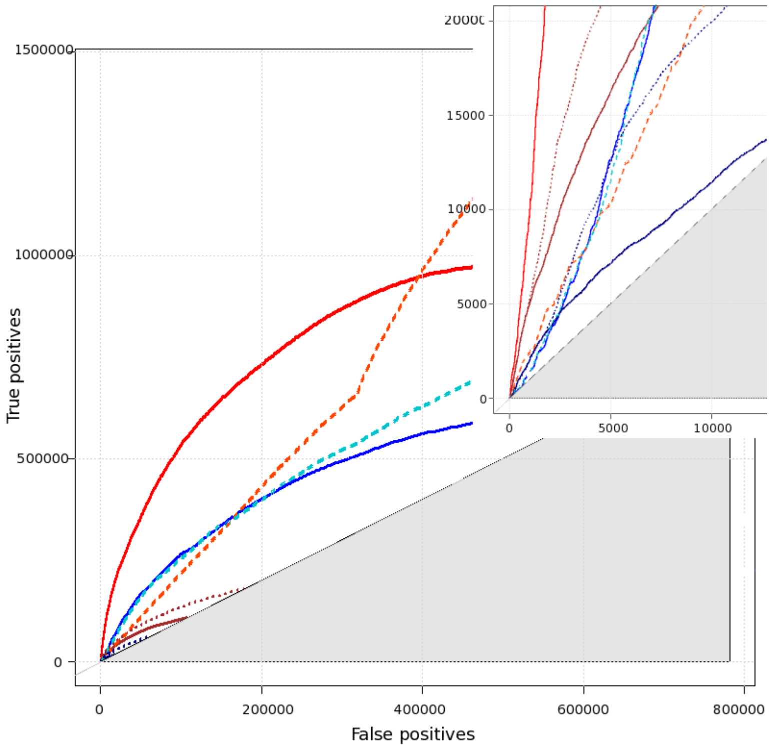
Sequence similarity proved a much stronger indicator of MF than BP categories across all datasets (shown by lower numbers of correct annotations and greater numbers of incorrect annotations in Table 3.1 for equivalent methods and performance curves in Figure 3.3). Consideration of all relationships above the bitscore threshold for MFs resulted in far greater numbers



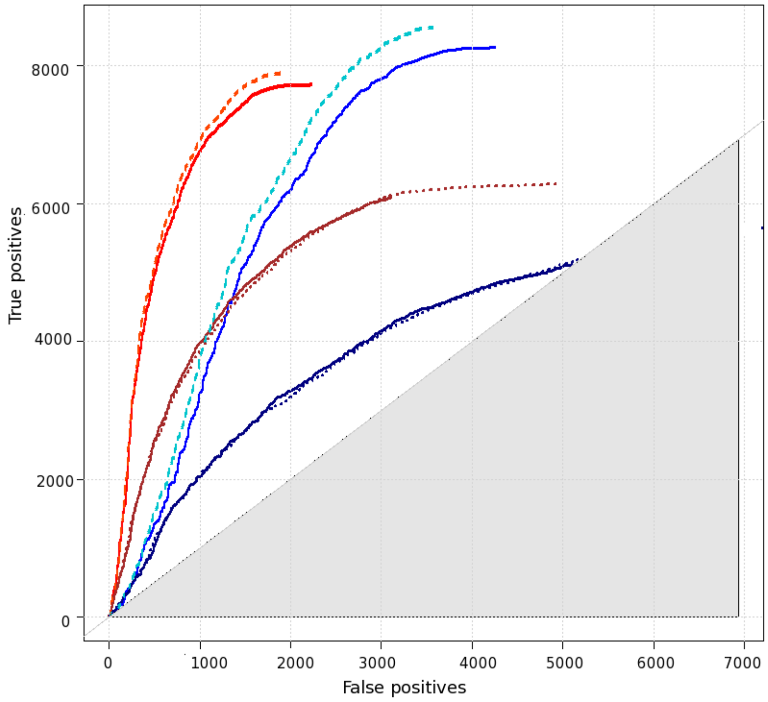
(a) ROC-like curve, all proteins all annotations

(b) ROC-like curve for all proteins, top hit

Figure 3.3: Performance plots for annotation transfer using between species relationships (red and blue) and within species relationships (navy and brown) identified by BLAST (solid) and PSI-BLAST (dashed) algorithms. The red and brown series represent MF annotation transfers whilst the blue and navy series represent BPs.



(a) ROC-like curve for nr35 proteins all annotations



(b) ROC-like curve for nr35 proteins, top hit

Figure 3.4: Performance plots for annotation transfer using between species relationships (red and blue) and within species relationships (navy and brown) identified by BLAST (solid) and PSI-BLAST (dashed) algorithms. The performance results were obtained using a dataset containing sequences filtered at 35% identity. The red and brown series represent MF annotation transfers whilst the blue and navy series represent BPs.

of true positives than the equivalent number for BPs at reduced numbers of false positives (Figure 3.3a). This trend was also observed using the nr35 dataset where the number of correctly detected MF annotations was roughly double the equivalent number for BPs.

Top scoring annotations were used to determine closest ancestral sequence relationships for annotations transfer. The frequency of true positive annotations was similar between MFs and BPs. However, the occurrence of false positives was much greater for BPs. Using the nr35 dataset the performance difference between Ontologies was reduced, but was still superior for MFs. Better performance was obtained for recognition of BPs from distant relationships than for all sequences. Precision values increased from 71.85% to 73.95% when averaged for BLAST and PSI-BLAST results. This was only observed using inter-species annotation transfer, which suggests that more general BPs are better preserved between species at greater evolutionary distances.

3.3.2 Performance of BLAST and PSI-BLAST algorithms

ROC-like curves were used to compare the performance of the different methods by applying a threshold to the bitscore and plotting the number of true positive and potential false positive annotation pair results. This method resulted in directly comparable curves that emphasized the different total numbers of relationships recovered during the sequence similarity searches.

To determine whether annotation transfer was improved by increasing the detection of remote sequence relationships, performance was compared using PSI-BLAST and BLAST algorithms (see Figures 3.3 a) and b) and 3.4 a) and b)). PSI-BLAST improved the recognition of MFs at very low bit scores and BPs to a lesser extent (Figure 3.3a). Within BPs, the additional information provided by the more powerful search algorithm only slightly improved the detection of correct functional relationships (Figure 3.3a). Using the nr35 dataset (Figure 3.4a) PSI-BLAST algorithm performance was only slightly different to BLAST for detecting BPs, and achieved poorer performance than BLAST for detecting MFs using inter-species sequence relationships.

The purpose of the PSI-BLAST algorithm is to increase sensitivity in detecting remote sequence relationships. After each iteration, a profile is constructed derived from an alignment between all sequences related to the query. Alignments between profile and sequence are then used to detect new relationships. This procedure can result in a different ranking of close homologues by alignment score, since the score can be up-weighted by the presence of related sequences in the profile. This results in a loss in sensitivity in measuring the degree of similarity between a pair of sequences, but improves sensitivity at low bit scores.

This property explains the observed trends for annotation category recognition where at low bit scores accuracy significantly improved. When the highest scoring sequence relationships only were considered, PSI-BLAST improved the performance of inter-species transfers only (Figures 3.3b and 3.4b). Considering human sequence relationships only, the extra search iterations served only to introduce false positives that could be correctly distinguished using BLAST. This likely results from the presence of fewer numbers of homologues in the species specific dataset, and suggests that in general, function is preserved over shorter evolutionary time scales within species.

3.3.3 Within and between species transfer

Comparing between all datasets and methods, transfer of MFs between species was more successful than within species (Figure 3.3). Detection of equivalent BPs within species was more accurate at high bit scores (Figures 3.3a and 3.4a) when all sequence relationships were considered. This was surprising since the number of annotated sequences and possible relationships at any threshold score is much greater in the multi-species sequence dataset. The results suggest that highly similar sequences more frequently share processes within species, perhaps due to organism specific biology. This same trend was not observed when only the highest scoring sequence relationships were considered. Annotation transfers between top hits were more accurate using inter-species relationships regardless of the algorithm or level of sequence redundancy within each dataset.

3.4 Annotation specificities

In the previous sections, each annotation was considered with equal importance regardless of its specificity. Consequently, correctly annotated specific terms, “5 to 3 prime oxidoreductase”, for example, could not be differentiated from correct general annotations such as “Receptor”. General annotation categories occur with higher frequencies in a population of sequences, therefore the probability of identifying a sequence annotated as “Receptor” in a similarity search is much greater.

To determine the extent specificity influenced the interpretation of results, GO term specificity distributions were obtained for each method using the natural log of the frequency of each annotation term in a population of sequences ($Spec = ln(x)$) to measure specificity. Low scores close to 0 imply that an annotation class is specific since it occurs rarely in a population of annotated sequences. High scores denote general annotation terms that can be recognised by their greater popularity in a pool of sequence annotations. Each distribution was obtained using correct annotations only to assess whether algorithm or species datasets were correlated with a bias in annotation specificity.

Non-parametric statistics were used to make comparisons between methods since the distribution of annotation specificities were multi-modal. Results from the Wilcoxon rank test were used to assess the significance of difference between pairs of distributions (Table 3.2). PSI-BLAST annotations were on average less specific than those reported using BLAST. This was expected since the PSI-BLAST algorithm improved the ability to detect distant sequence relationships. These distantly related sequences are more likely to represent correct results for larger more general categories. Annotations from intra-species transfers were on average more general than those obtained from inter-species relationships. This may be determined by the frequency of specific annotations represented by sequences from other species compared with human annotations.

Specificity distributions were also compared between species, however in this case the natural log frequency of child terms in the GO graph was used to calculate specificity. This was necessary to avoid species bias in the specificity scores introduced by the differing numbers of annotated sequences and belonging to each model organism. Again, similar Wilcoxon tests were performed between specificity distributions from each model organism to determine whether human annotations were on average more or less specific than other species (Table 3.3).

Significant differences were obtained between some distributions using the Wilcoxon pair test despite identical median values. This was caused by the fact that the distributions were more or less categorical. This had the effect of enabling the median values to appear equal whilst the rank changes might exhibit a preference in either direction resulting in a significant p-value. In the case of the worm, the MF distribution was skewed towards annotations that were either highly specific or very general causing the median to be more specific whilst the majority of rank

Table 3.2: Average specificity for annotation transfers

Ontology	Dataset	Method1	Method2	Median1	Median2	P value
Inter Species Comparison						
Function	all sequences	BLAST	PSI-BLAST	6.06	6.11	$<$ 1e-200
Process	all sequences	BLAST	PSI-BLAST	5.93	5.96	$<$ 1e-200
Function	nr35	BLAST	PSI-BLAST	5.38	5.56	7.96e-92
Process	nr35	BLAST	PSI-BLAST	5.22	5.35	2.02e-67
Intra Species Comparison						
Function	all sequences	BLAST	PSI-BLAST	6.74	7.99	$<$ 1e-200
Process	all sequences	BLAST	PSI-BLAST	7.01	7.01	$<$ 1e-200
Function	nr35	BLAST	PSI-BLAST	7.28	7.99	$<$ 1e-200
Process	nr35	BLAST	PSI-BLAST	6.86	7.01	$<$ 1e-200

Specificity comparisons for different results sets. The median specificity is reported alongside P-values representing the results of a 2 tailed Wilcoxon rank test.

changes were positive indicating worm terms were more general.

Human Molecular Function annotations were on average more specific than annotations from mouse, worm and fly. Annotations for rat and yeast were on average more complete. Biological Process annotations were more general for human than for all other tested species. These differences affect interpretation of performance measures. Lower specificity values resulted from correct inter-species than intra-species annotation transfers. For Biological Process categories, this is likely to result from the observation that human annotations are less complete than in other species. However, for Molecular Functions, the greater specificity of correct annotation transfers between species is more likely to reflect evolutionary process since annotations for human sequences display similar or higher specificities than those from other eukaryotes.

Different degrees of completion and organism specific biology means that annotation categories have different information value between species. This makes performance comparisons between different approaches applied to sequences from multiple species difficult to interpret. Here, it has been shown that the frequency of correct results is determined by the number of homologous sequence relationships detected, the quality of annotations for those sequences and the heterogeneity of a population of function categories. The application of a global score threshold to determine sequence similarity relationships therefore provides a best general case that is sub-optimal for highly accurate annotation transfer. The procedure likely introduces errors related to the lack of homogeneity of function conservation.

3.5 Sources of errors

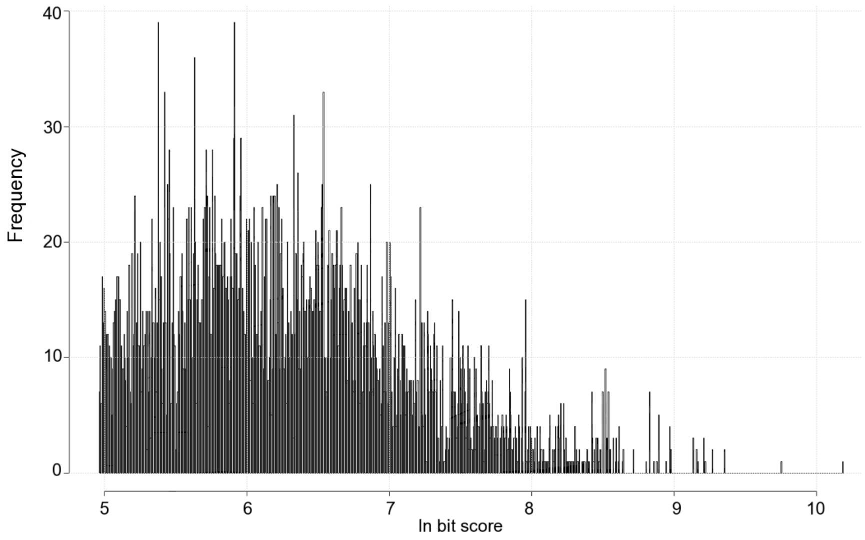
Despite the fact that the definition of negatives for this study was difficult, an attempt was made to determine whether the incorrect annotation transfers resulted from inconsistencies arising between sequence annotations or reflected genuine function differences. The occurrence of ‘errors’ was reported at different bit scores for the top scoring sequence relationships identified using BLAST (Figures 3.4 and 3.5).

More errors were obtained at all bit scores for transfer of Biological Processes (Figures 3.5b and 3.6b) than Molecular Functions (Figure 3.5a) and (Figure 3.6a). Here the most heavily populated regions of the distributions were between 5 and 6 ln bit score units rather than 6 to 7 for Biological Processes. Greater numbers of false positives were observed at low bit scores for

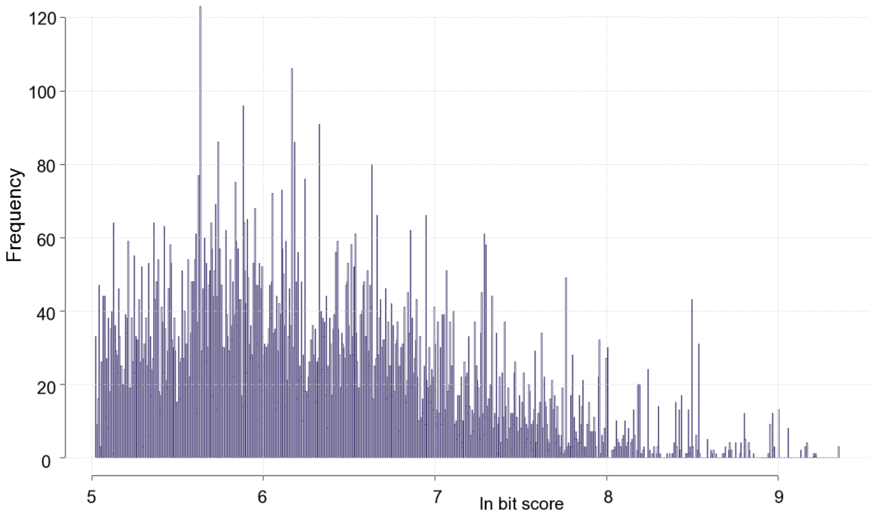
Table 3.3: Comparison of species annotation specificity.

Ontology	**Human**	**Mouse**	**Rat**	**Fly**	**Worm**	**Yeast**
Molecular Function						
Median	3.664	3.664	3.466	3.664	2.197	3.664
P value		5.42e-08	1	2.38e-18	4.51e-70	0.018
Biological Process						
Median	3.638	2.485	2.890	1.946	1.099	2.303
P value		1	1	1	1	1

The first row for each ontology represents the median specificity for the species and the second row details the associated p-values from the normal approximation to the Wilcoxon one-tailed rank test.

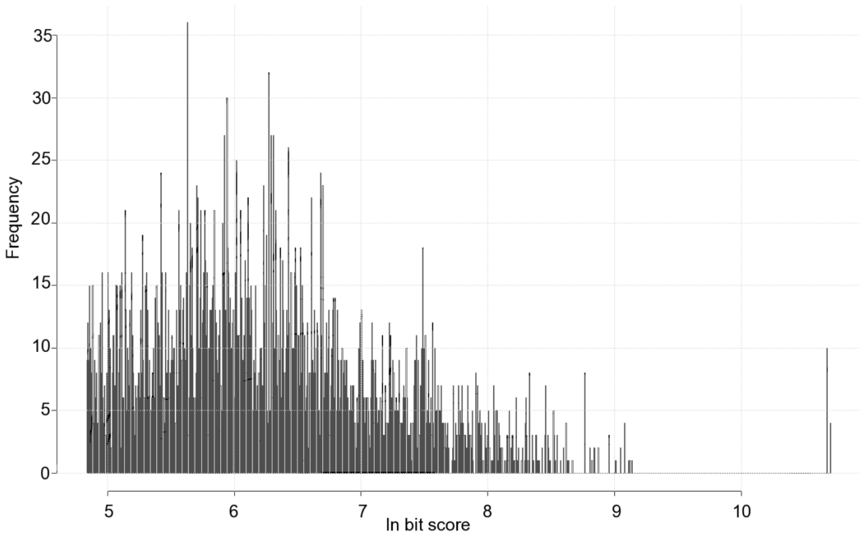


(a) Human Molecular Functions

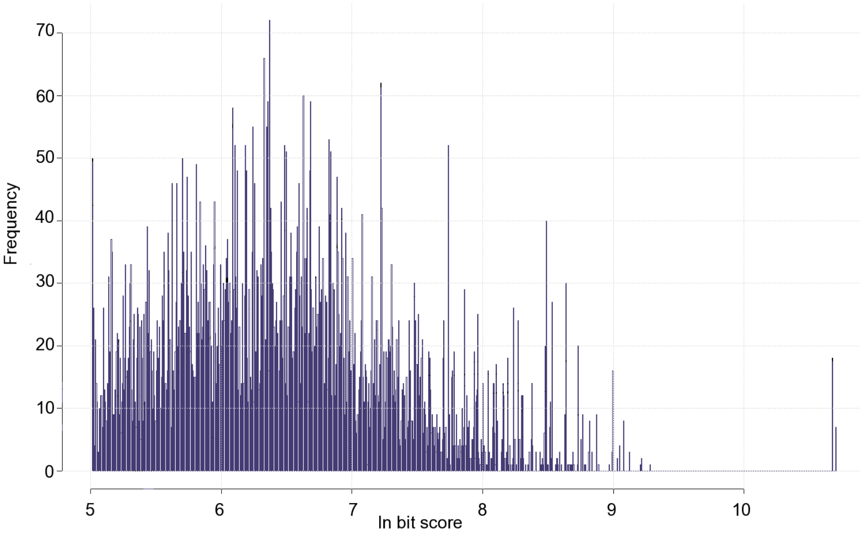


(b) Human Biological Processes

Figure 3.5: Characterising false positives. Each plot represents the frequency histogram of false positives obtained at different bit scores above the optimal threshold for each method. The x-axis in each case represents the natural log of the bit score value.



(a) Multi-species Molecular Functions



(b) Multi-species Biological Processes

Figure 3.6: Characterising false positives. Each plot represents the frequency histogram of false positives obtained at different bit scores above the optimal threshold for each method. The x-axis in each case represents the natural log of the bit score value.

Molecular Function transfer than for Biological Processes, and greater numbers of false positives were observed at high bit score values (≥ 9) using intra-species transfers (Figure 3.4).

False positive annotations can occur at high sequence similarities for several reasons. First they might result from incomplete or even incorrect annotations. They may also reflect true functional differences between sequences, for example highly similar sequences that are sequence variants frequently have different functionality because the part of the sequence responsible for the function, a catalytic residue or domain is disrupted. These cases are impossible to discriminate by sequence similarity measures. The fact that false positives occurred across the bit score range may indicate that they are a genuine reflection of biology rather than predominantly annotation errors. However, a previous study carried out on microbial genome annotations estimated that these cases can affect between 5% and 40% of highly specific genome annotation assignments (Devos and Valencia 2001). A more recent study of human annotations reported a figure close to 6% .

The false positive transfer observed at a very high bit score (Figure 3.5a) comprised a relationship between two Mucin sequences. IPI000646572 and IPI00103552 represent sequence variants of Mucin 16. The first sequence possesses an annotation of ATP-binding whereas the second example does not. The database entry for IPI000646572 reports an ATPase domain that is absent from the IPI00103552 variant. This suggests that the assignment of a false positive annotation in this case is appropriate for the data.

A similar false positive transfer resulting from an inter-species relationship represents a case of incomplete annotation. IPI00759754 (human) and A2ASS6 (mouse) represent isoform1 of the titin gene. The mouse sequence has been annotated with the term GO:005509 calcium ion binding that is absent in the IPI human GOA mapping files (Submission Date: 18th Sept 2008). However, the correct annotation is present in the equivalent Swiss-Prot database annotation entry for the human sequence. This highlights important inconsistencies between different annotation curation efforts that affects the results of analyses performed using the GO Annotations. This problem is more likely to affect inter-species annotation transfers since different Consortia and curator groups tend to produce annotations varying in quality and consistency for different genomes.

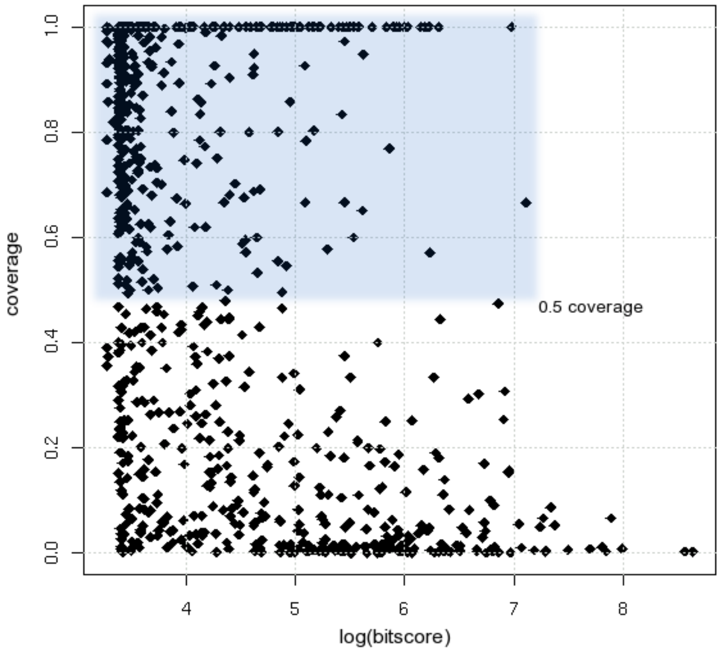
3.6 Local scoring and functional heterogeneity

One way to address functional heterogeneity is to consider each annotation category independently. The relationship between sequence conservation and function is then quantified conditionally on each annotation. If sequence conservation is highly correlated with function annotation specificity, and the rate at which sequences have evolved is determined by function, local score thresholds applied to sequence relationships should provide superior results for function annotation.

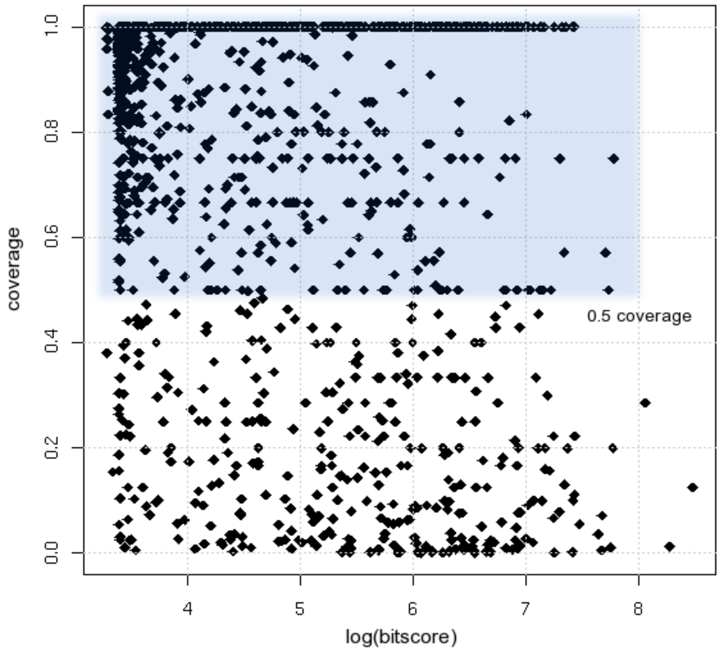
To test this hypothesis, a single optimal bit score threshold was selected by maximising the Matthew's correlation co-efficient for annotation transfers on an individual per category basis. The degree of variability between these thresholds relates to the degree of sequence variability within each annotation category. To visualise these thresholds, the optimal bit score cut-off was plotted against the resulting annotation coverage calculated as the proportion of correctly recovered annotations for Molecular Functions and Biological Processes (Figure 3.7).

Annotation categories with high coverage and high bit score thresholds represent cases where highly similar sequences populate an annotation category. These are predominantly specific categories with few sequence representatives. For example, "GO:0004352 glutamate dehydrogenase" for which two human sequences could be identified with an alignment score of 1077 (6.98 on the natural log scale). Categories with low coverage at low bit score thresholds represented cases where sequence divergence between homologues exceeded the detection limits of the BLAST algorithm or where sequence similarity was not an important determinant of function. For example "GO:0031494 chloride ion binding" for which the optimal threshold at bit score 29.2 was sufficient to recover just 17% of example sequences.

Overall locally optimal thresholds varied considerably between different annotation classes within both Ontologies. The bit score range occupied by categories achieving greater than 50% coverage was broader for Biological Processes than Molecular Functions, suggesting that sequence relationships are more variable between Biological Process Categories than Molecular Functions. Optimal thresholds for Molecular Function categories achieved either low or high coverage with few thresholds obtaining mid-range bit-scores (Figure 3.7a). The high frequency of low bit score thresholds suggested improvements might be gained for these categories by the use of the PSI-BLAST algorithm and indicated that more Molecular Function than Biological



(a) Molecular Function



(b) Biological Process

Figure 3.7: Evidence for functional heterogeneity; each data point represents an annotation category with associated optimal bit score thresholds determined from BLAST intra-species relationships. The natural log the bit score threshold is represented on the x-axis against the resulting coverage values (y-axis). The shaded part represents the area comprising annotation categories that achieve a fractional coverage of $\geq$ 0.5.

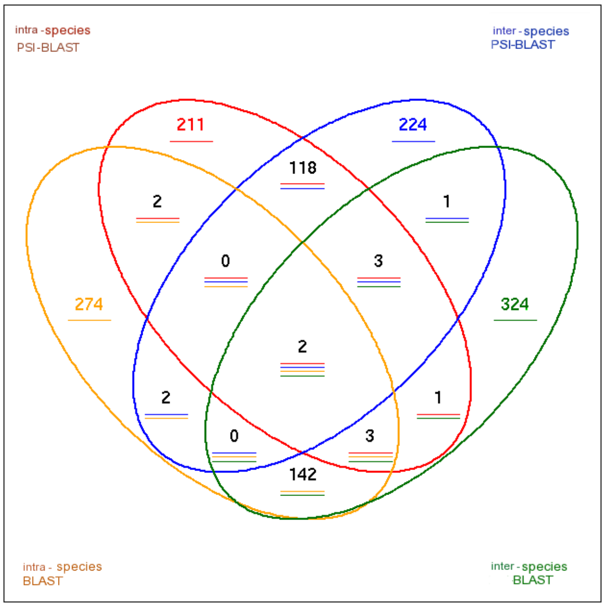
Process categories are conserved to a greater degree.

A greater number of Biological Process categories were present with very high coverage than Molecular Functions, and fewer categories were observed with low optimal bit score thresholds (Figure 3.7b). This suggests that for a larger proportion of Biological Processes, the annotated sequences are very similar and could easily be recovered by BLAST search. For the remaining Biological Process categories, it is unlikely that sequence relationships are sufficient to determine function since coverage was low at high bit score thresholds. This analysis supports the appropriate use of local sequence similarity score thresholds for different annotation categories, given the observed heterogeneity of function with respect to sequence relationships.

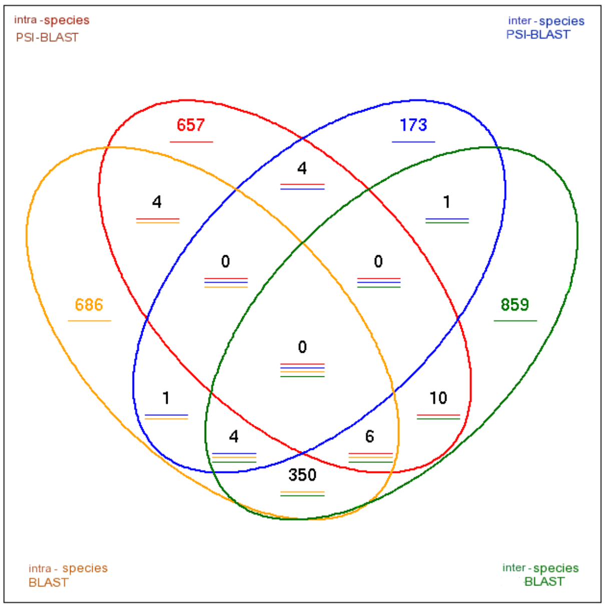
To determine the benefit of applying a local sequence similarity score threshold to each annotation category, performance between alignment algorithms (PSI-BLAST and BLAST) and inter and intra-species datasets were compared. For consistency, the Matthew's Correlation Coefficient (MCC) was calculated to measure performance of each method for a particular annotation category. The best method for each category corresponded to the method with the highest MCC. In total, 1308 of 2623 Molecular Function categories obtained an MCC value of > 0. This threshold represents performance obtained above random and is a deliberately permissive threshold used for the purpose of comparing numbers of classifiers between the datasets rather than to indicate high quality classification performance. The corresponding figure for Biological Process annotations was 2756 of 4676. The results show that different datasets and algorithms produce better performance for different annotation categories (Figure 3.8).

In both Molecular Function and Biological Process Ontologies, the method obtaining the greatest annotation performance was PSI-BLAST detecting inter-species relationships, although this majority was slight for Molecular Functions. Few methods attained equivalent performance using the different datasets. 485 Molecular Functions, and 1343 Biological Processes were better determined by inter-species relationships using sequence similarity scores from either BLAST or PSI-BLAST searches. Equivalent numbers for intra-species comparisons were 548 and 1032 respectively. This result suggests that intra sequence relationships are more useful than inter species sequence relationships in recognising at least 50% of Molecular Functions and Biological Processes.

Statistical tests were used to determine which annotation categories obtained significantly better performance using the different methods. This ensured that any interpretation of the results



(a) Molecular Function



(b) Biological Process

Figure 3.8: Venn diagrams showing best performance using local bit score thresholds. Each frequency represents the number of GO annotation categories for which the highest MCC value amongst all other methods was observed. Intra-species comparison results are on the left whilst inter-species results are on the right hand side. BLAST algorithm performance is represented by the outer rings and PSI-BLAST by the innermost sets.

related more to biology than random chance. Annotation categories were filtered according to the significance values of a t-test applied to Fisher's Z score transform of the correlation value. The Fisher transform is defined as

$$Z = 0.5 \times \frac{ln(1+r)}{ln(1-r)} \tag{3.5}$$

and operates on correlation values (Pearson's or Matthew's) termed r. Subsequently, a t-test can be performed on the Z scores given by

$$\begin{aligned} t &= \frac{z_1 - z_2}{\sigma_{1,2}} && (3.6) \\ \sigma_{1,2} &= \sqrt{\left(\frac{1}{N_1 - 3}\right) + \left(\frac{1}{N_2 - 3}\right)} && (3.7) \end{aligned}$$

.

The t-test results represent the significance of correlation difference between two MCC values.

41 Molecular Function Categories and 26 Biological Process Categories were significantly better predicted by intra-species sequence relationships, whilst 36 Function and 58 Process categories were identified that were better predicted by inter-species relationships. This confirms that during annotation transfer, the origin of the related sequence is important. Methods restricted to either dataset are likely to produce sub-optimal performance where sequence relationships alone are used.

Categories that were better annotated by inter-species transfer included Kinase and GPCR (G-Protein Coupled Receptor) sub types, antigen presentation, chemical and sensory stimulus and amino acid biosynthesis pathways (see Appendix II). These functions represent a mixture of cases of organism specific biology resulting in functions that are not represented in other species, and cases where sequences have diverged more rapidly within other species giving rise to homologous sequences with related but not identical function.

Annotation categories that were better recognised using relationships determined between species were generally describing functions common to all species, for example, "Transcription" and related categories, "Cell Death" and "Regulation of metabolism". There was little convincing evidence that sequences encoding interacting proteins (best described by Molecular Function binding categories and some Biological Process categories) were more conserved within than between species, however evidence for this hypothesis might be buried if this trend was present for a subset of sequence relationships that do not correspond to annotation category definitions.

Local and global thresholds applied to sequence relationships were compared using annotation category coverage and sequence coverage statistics. The global score threshold permitted more annotation classes to be covered for a similar coverage of sequences (Table 3.4). Fewer annotation categories could be modelled using local score thresholds. However, the method achieved greater depths of annotation coverage. This is reflected by comparing the coverage ratio between number of annotations and number of categories. On average 19.2 and 20.6 correct Molecular Function annotations per category could be recovered using local thresholds for annotation transfer compared to 15.9 and 13.2 correct annotations per category using the global threshold. This effect was reduced for Biological Process categories. The equivalent ratios were 7.82 and 7.61 for transfers made using a global score cut-off, and 8.23 and 7.69 for transfers made using local score cut-offs. In total, annotation category coverage was at most 62% and 63% respectively for Molecular Functions and Biological Process Ontologies. These figures show that a large portion of annotations remain inaccessible to methods using sequence similarity relationships to infer function.

3.7 Discussion

The aim of the chapter was to investigate the use of sequence similarity relationships to annotate function. To detect sequence relationships, the BLAST algorithm was used despite the fact that other algorithms (Smith-Waterman (Smith and Waterman 1981) for example) produce more rigorous alignments between sequences. The BLAST algorithm represents an explicit trade-off between speed and accuracy and is both the biologist's and bioinformatician's method of choice for performing large scale sequence similarity searches. Throughout this benchmark common practices were followed wherever possible so that the results and interpretation related to the most widely used approach in genome annotation.

The study results demonstrated that overall, sequence similarity is a strong indicator of function,

Table 3.4: Annotation coverage

Method	Category coverage	Sequence coverage	Correct annotations
Molecular Function			
Global intra	1150	18313	26068
Global inter	1538	20248	26959
Local intra	805	15426	20925
Local inter	927	19133	25125
Total	**2485**	**22306**	**38137**
Biological Process			
Global intra	2208	17131	20318
Global inter	2181	16743	23938
Local intra	1485	12221	16786
Local inter	2032	15426	20657
Total	**3492**	**21176**	**47084**

Performance statistics represent values obtained at optimal (maximal MCC) bit score thresholds. Coverage obtained using global thresholds represents the number of GO categories for which there was at least one true positive annotation. Sequence coverage is the number of unique sequences for which at least one correct annotation was recovered, and correct annotations are th number of true positive assignments however common functionality cannot be observed for all similar sequences, or even those that are closest ancestors. One problem lies in the ability of BLAST and PSI-BLAST alignment scores to detect homologues. Discriminating genuine sequence relationships from random becomes difficult where few distant homologues exist (Koski and Golding 2001). Additionally, the closest BLAST relationship may not represent the closest phylogenetic ancestor. This can be problematic where many close homologues are identified and represents cases where function has diverged after speciation events giving rise to similar sequences with different functions that cannot be correctly distinguished using alignment scores (Gerlt and Babbitt 2000). One study reported that 27% of closest BLAST relationships for *E. coli* sequences did not represent the nearest phylogenetic ancestor sequence (Koski and Golding 2001). This phenomena is even apparent between close species, for example, estimated rates of gene loss and divergence between *S. cerevisiae* and *S. pombe* sequences are both 7% (approximately 300 genes) (Aravind et al. 2000). In higher eukaryotes, these figures may are expected to be amplified due to increased functional complexity arising from signalling and developmental processes, and the presence of large gene families that have evolved by successive duplication events (Lespinet et al. 2002).

Since the definition of a sequence that does not have a particular function is ambiguous, it was difficult to determine to what degree these factors affected the results. However, it is conceivable that the majority of false positive annotations reported were genuine as it is unlikely that large proportions of sequences are annotated with high specificity function categories. More than one third of annotation transfers for Molecular Function and close to 40% of Biological Process categories were considered false positives. These numbers likely result from incorrect identification of the closest ancestral sequence, or from cases where the closest ancestor sequence has functionally diverged. This assumption can be supported by the fact that in other studies the observed difference in performance obtained from expert sequence family based analysis ranges between 8 and 10% compared to a basic approach using BLAST (Brenner 1999, Devos and Valencia 2001).

The higher false positive rate observed for annotation transfer of Biological Process categories compared to Molecular Functions is in agreement with a recent study demonstrating that regulatory processes (described by Biological Process annotations) display a high degree of plasticity whilst the core components of metabolism, transport, and protein synthesis (described in the Molecular Function Ontology) are conserved (Caron et al. 2001).

The lack of formal definition of functionally equivalent categories meant that the results of annotation transfers for both general and specific categories were equally weighted by score. This affects performance measures obtained across a set of function categories. Statistics can be erroneously high and biased towards transfer of general annotations that are less practically useful. Additionally, the use of a single threshold to characterise the relationship between sequence similarity and function seems inappropriate as different functions have evolved at different rates and are subject to varying amounts of selection pressure. The most appropriate way to measure and compare methods for annotation transfer performance using sequence similarity is therefore individually for different categories.

Using local thresholds, several comparisons were made between annotation transfer strategies; intra and inter species transfers, local versus global scoring thresholds, and BLAST compared to PSI-BLAST performance. Application of a global score threshold to sequence relationships above which common function could be assumed showed that the BLAST algorithm outperformed PSI-BLAST in its ability to generate alignment scores that were useful in discriminating function. The power of the iterated search procedure meant that many more sequence relationships could be detected but not correctly distinguished from one another using alignment scores. However, these results were shown to exhibit bias due to the different specificities of annotation classes. Performing a similar comparison locally; between equivalent annotation categories showed that neither algorithm produced superior performance for the majority of functions. Making an informed choice as to which alignment algorithm is most appropriate relies on prior knowledge about the degree of sequence divergence within each category. However, this approach is likely to produce far superior results for annotation transfer.

More sequence relationships and more correct annotations could be recovered from comparisons made between human sequences and sequences from other species. This result is to be expected since inter-species sequence relationships often represent orthologues with common function. Where few sequence representatives exist within a species for a given function, significant advantage is obtained by searching a multi-species database populated by greater numbers of functionally equivalent homologues. In cases where large multi-gene families have evolved independently, paralogous sequences might be more likely to share common function. This property was apparent in the results for Biological Processes where highly similar sequences exhibited a greater degree of function conservation. Using locally defined annotation thresholds for each function category, greater accuracies were reported using intra-species sequence relationships for close to half of all tested categories. This suggests that automated methods should either incorporate additional information regarding the source of the sequence relative or use prior knowledge about the annotation category to achieve better performance.

Whilst approaches using prior information about function categories to determine annotation specific scoring thresholds increased accuracy, there are several advantages to using a simpler global threshold applied to all sequence relationships. The BLAST top hit match criterion using a global threshold provided the best overall sequence and annotation category coverage. The PSI-BLAST algorithm could then be used only in cases where BLAST failed to identify a sequence relationship of sufficient strength to infer function. However, results from this approach should always be verified independently using other automated techniques or experimental information where possible. Where deep annotation coverage is required, annotation category specific similarity thresholds applied to sequence relationships produce superior results and allow for tight control of false positive rates.

In addition to exploring the evolutionary relationships between sequence and function, this study exposed several practical limitations of homology based annotation transfer methods. For example, the maximum coverage of annotations obtained using these methods was 70.7% and 43.2% for Molecular Functions and Biological Processes respectively. Because sequence similarity methods are ubiquitous among function prediction approaches and dominate as the source of most known available annotations, there is limited value in further developing these methods to obtain minor performance increases. The problem of identifying closest ancestral sequences can be addressed by phylogenetic methods, and the problem of detecting distant relationships can be addressed by profile based domain or sequence family specific approaches. A more pressing need is to develop alternative approaches to determine sequence relationships that are inaccessible to homology based methods.

Chapter 4

Feature based function prediction

4.1 Chapter aims

This Chapter describes the design, implementation and benchmark results for a feature based function prediction system (FFPred) to tackle the annotation of distant homologues, non-homologous and orphan human protein sequences. This work builds on a previous work, the ProtFun method (Jensen et al. 2003) that models broad function categories using ensembles of neural networks. The networks are trained to recognise patterns of features that correlate with function. The features consist of global features describing whole sequence attributes, for example, iso-electric point, average hydrophobicity and localisation propensities that weakly correlate with function. Combining these weak signals using a set of neural networks achieved classification performances at greater than 50% coverage for a rate of 10% of false positives for 14 GO classes. These comprised broad first level Enzyme Classifications and transcription related categories.

One limitation of the approach is that for a given sequence of interest, predictions are made by selecting the most probable annotation category from a set of candidates. However, in characterising the Gene Ontology annotation system (Chapter 1, Section 1.4), it was noted that human sequences participate in multiple processes and functions, more than 3 on average, rather than being annotated to a single function. Thus one area for improvement in this approach lies in the assignment of several annotations from a set of individual annotation category scores. A second area where the method could be improved is by expansion and incorporation of new features describing functionally relevant attributes of sequence.

A wealth of literature has reported the implicit link between the occurrence of protein disorder and function (Dunker and Obradovic 2001, Dunker et al. 1998, 2008b, Romero et al. 2004, Tompa 2005, Uversky et al. 2005, Vucetic et al. 2007, Xie et al. 2007a,b). Structural disorder in proteins confers flexibility, bypassing constraints imposed by the adoption of regular secondary structure conformations (Iakoucheva and Dunker 2003). Many disordered regions occur at sites of molecular recognition, post-translational modifications, DNA and protein interactions, as well as small molecule-protein interactions (Dunker et al. 1998, Uversky et al. 2005). Upon binding, the disordered regions become ordered, acquiring greater stability than the native state (Zhang et al. 2007). Several experimental studies have shown that the presence of disordered regions in some proteins is essential for their correct functioning (Dunker et al. 2008a, Tompa et al. 2005). At the sequence level, disordered regions are low in complexity; eliciting significant bias towards polar and hydrophilic residues, and away from bulky hydrophobics (Mohan et al. 2006, Vucetic et al. 2003). Consequently they can be successfully predicted from amino acid sequence using machine learning techniques (Kumar and Carugo 2008, Shimizu et al. 2007, Uversky et al. 2007, Ward et al. 2004a). Prediction of disorder across entire proteomes has revealed a correlation with organism complexity; long ($>$ 30 residue) stretches of disorder are predicted to occur in 30-60% of proteins from higher eukaryotes compared to 10-30% in prokaryotes (Jones and Ward 2003, Tompa et al. 2006, Ward et al. 2004a).

Considering the explicit links between disorder and function, and the ability to predict disorder from amino acid sequence, the first part of the chapter examines the design of novel features encoding protein disorder. Predictions of disorder were used rather than experimentally validated definitions of disorder due to the sparse coverage of sequence space provided in curated databases such as DisProt, and high accuracies with which disordered residue assignments could be made from sequence using prediction algorithms. High coverage of sequence space was required to determine statistically valid trends and patterns from the data. Other features in common with the ProtFun method are updated and improved, for example, transmembrane and secondary structure predictions are calculated using evolutionary profile information rather than sequence information. For orphan sequences with no discernible relatives, predictions of secondary structure and transmembrane regions remain unchanged, however, for distantly related sequences, the accuracy of these feature predictions can be greatly improved by incorporating residue conservation information contained within PSI-BLAST profiles (Jones 1999, 2007). In this approach, Support Vector Machine's are trained to distinguish patterns correlating with different function classes independently so that each sequence can be assigned multiple functions. Finally the method is benchmarked against the ProtFun method for a set of equivalent function categories and the merits and limitations of the method discussed.

4.2 Designing features encoding disorder

Since approximately one third of eukaryotic protein sequences contain at least one long ($>$30aa) disordered region, and these regions have been experimentally linked with the correct functioning of the protein, the relationship between disorder and function in the human proteome was characterised. A previous study of the functions of disordered proteins in yeast showed that signalling molecules; kinases, transcription factors and G-proteins were enriched in the set of disordered proteins (Ward et al. 2004b). If these findings are also a feature of disordered human sequences, it is expected that the occurrence of disorder might be useful in function prediction, especially for these categories.

4.2.1 Functional analysis of disordered human sequences

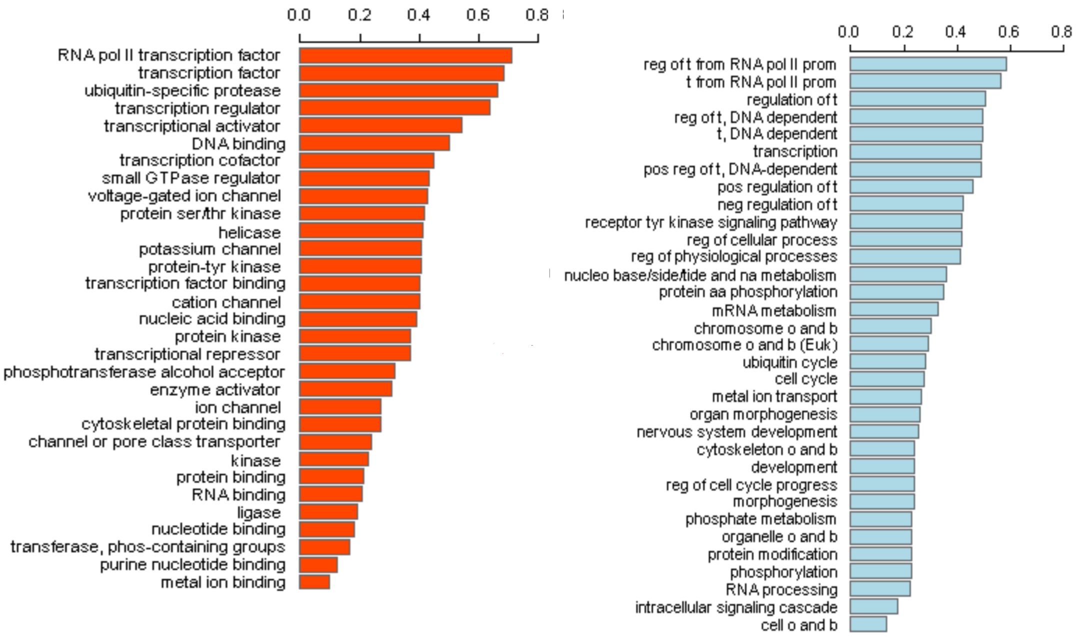
Functional analysis of disordered human sequences was carried out by enrichment statistics (the Fisher Exact test, Fisher, 1954) designed to identify cases of proportional bias between the association of two factors. In this case the factors are the occurrence of disorder and the presence of a particular Gene Ontology annotation. Disordered residues were predicted for the human proteome using the DISOPRED algorithm (Ward et al. 2004a) with default parameters. A sequence was considered disordered if it contained a contiguous stretch of more than 30 amino acids at a per residue false discovery rate of $< 5\%$.

For each individual GO category the Fisher test was performed and a multiple testing correction applied (Bonferroni method). This correction is given as

$$Adjusted_p = pN \tag{4.1}$$

and is designed to minimise the chances of identifying false positive associations by adjusting the resulting p-values proportionally to the number of tests performed N (Bland and Altman 1995). The results (Figure 4.1) show the respective sets of Molecular Function and Biological Process categories that are enriched in disordered sequences.

Many of the function categories that were enriched in disordered human sequences corresponded with the literature and to earlier studies of yeast protein sequences and Gene Ontology categories. In total 31 MF and 33 BP categories were identified that were involved in molecular recognition;



(a) Molecular functions

(b) Biological processes

Figure 4.1: Functional categories enriched in disordered human sequences. The x-axis represents the log10 odds ratio for the degree of over-representation of the proportion of disordered sequences annotated by the category. Several of the category names have been abbreviated for figure clarity, for example, t corresponds to transcription, o and b represent organisation and biogenesis, and regulation has been shortened to reg.

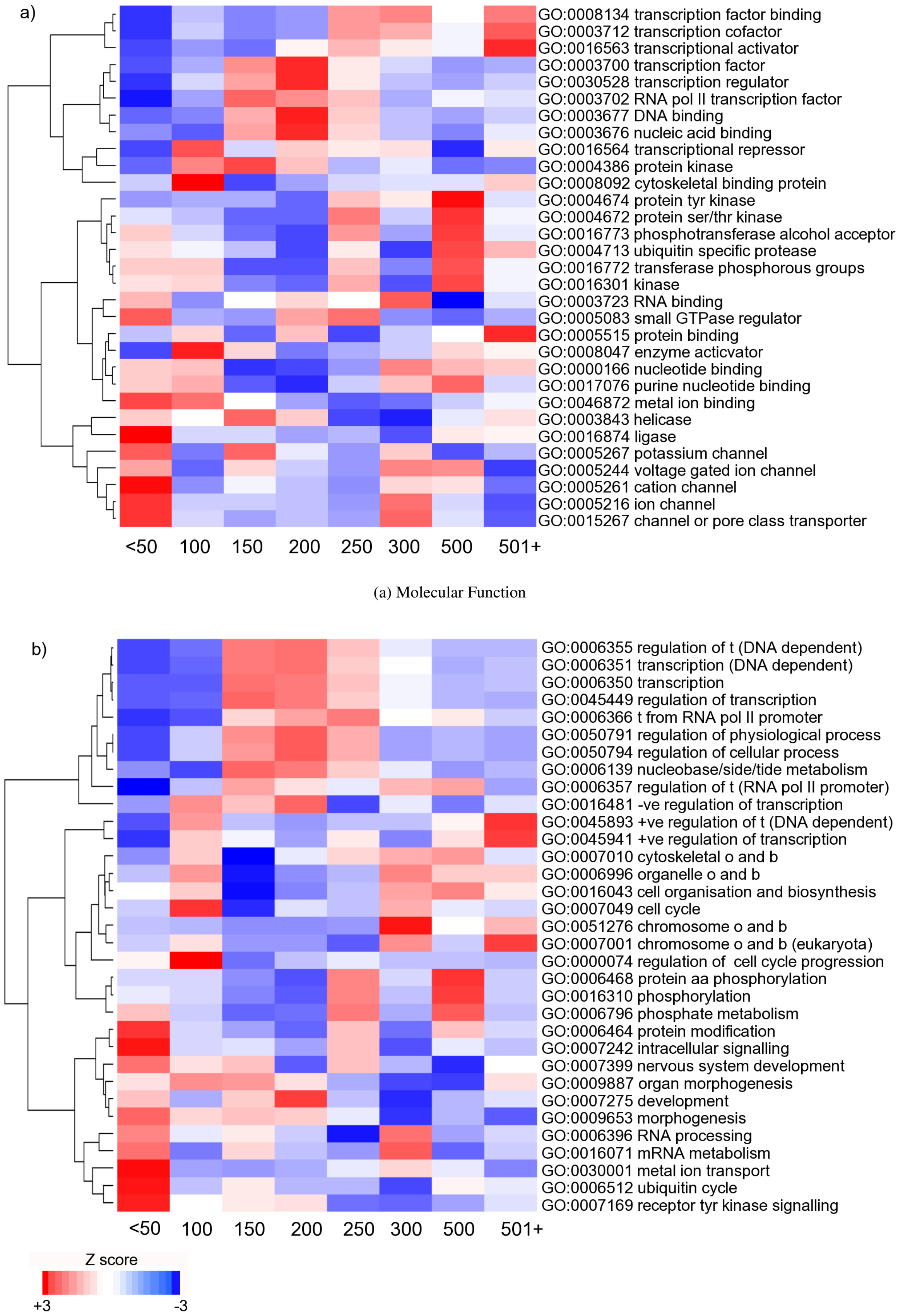
DNA and protein binding as well as transcription and translation (Figure 4.1).

“Transcription factor”, “DNA and protein binding”, “Protein kinase”, and “Ubiquitin protease” MF categories were among those enriched in disordered proteins indicated by the highest log ratios of observed/expected occurrence of disordered proteins (Figure 4.1a). Transcription factor categories were most enriched in disordered proteins, followed by Ion channel and Phosphorylation related functions. Metal-ion and Nucleotide binding functions exhibited smaller yet significant enrichment in disordered proteins. “Transcription regulation”, “Kinase signalling”, “RNA metabolism”, and “Phosphorylation” featured in the set of BP categories that were enriched in disordered proteins (Figure 4.1b). These categories were consistent with those functions reported both experimentally in the literature and in similar analyses of other organisms (Jones and Ward 2003, Tompa et al. 2006).

4.2.2 Encoding strategy for disorder features

To identify aspects of disorder which discriminated between the different function categories, and hence would provide suitable feature descriptors for function prediction, two analyses were performed. The first analysis investigated trends in the distribution of lengths of disordered regions on the basis that the presence of longer disordered stretches might display functional preferences. The second analysis addressed whether the location of the disordered regions within amino acid sequences was statistically associated with function. Since many sequences contained short disordered stretches at either the N or C termini respectively, these distributional aspects aimed to discriminate functions that predominantly or consistently contained disordered residues at either termini from those containing disordered regions throughout the interior of the protein.

First, the entire distribution of disordered region lengths was divided into separate ranges, determined by roughly equal proportioning of the entire length distribution. The proportion of sequences annotated to a particular GO category within each length range was recorded and used to populate a disorder range by GO category matrix. This matrix was then converted to Z scores (normalising by mean and variance within each length range). Proportions of sequences for an annotation category that were significantly greater or less than proportions represented within other categories received a large +ve or large −ve Z score. These results were visualised as heatmaps (Figure 4.2).



(a) Molecular Function

(b) Biological Process

Figure 4.2: Heatmaps showing patterns of disordered region lengths that are associated with function. High Z scores are coloured red whilst low Z scores are coloured blue.

Long regions of more than 500 contiguous disordered residues were over-represented in transcription related function categories. Shorter regions (50 residues or less) were over-represented in proteins performing metal ion binding, ion channel, and GTPase regulatory functions. Proteins annotated with serine/threonine kinase and phosphatase categories were also over-represented with contiguous stretches of disorder 300-500 residues long. Again these findings can be supported by structural evidence. Short disordered regions at the mid to N-terminal regions in small GTPase regulatory proteins mediate a switching mechanism, enabling the protein to interact with multiple binding partners (Menetrey and Cherfils 1999). These correlations were not simply a function of correlations between protein length and GO categories. This is exemplified by considering "Ion Channel" and "Transcription factor binding" categories (Figure 4.2). A statistically significant association between shorter disordered regions and the Ion Channel GO category was observed, yet the average sequence length within this annotation category was more than 900 amino acids. In contrast, for "Transcription factor binding", the opposite trend was observed. The average protein length for this class is closer to 700 amino acids, and an association was reported with long (more than 500 residue) stretches of disorder.

A similar procedure was carried out for location aspects of disorder. Sequences were divided into regions; 50 absolute residues for N and C termini with the interior of the sequence divided into 8 equally proportioned segments. Again a segments by GO category annotation matrix was constructed containing values representing the average proportion of residues that were disordered out of all residues in the segment. These average proportions were converted into Z scores within segment ranges to identify comparatively high or low values with respect to other annotation categories.

A similar visualisation was used to assess the quality of features encoding location aspects for disordered residues (Figure 4.3). The results showed that location patterns corresponding with function displayed less significant associations than length patterns. This is perhaps due to the introduction of noisy signals when defining proportionally equivalent regions to be compared within sequences that vary in length. However, several clear patterns could be observed; "Transcription regulator", "DNA binding", and "RNA pol II Transcription factor" functions were associated with disordered residues in the protein interior, rather than at N and C termini (Figure 4.3b). "Transcription factor activator", "Transcription factor repressor", and "Transcription factor" categories showed significant associations with disordered residues toward the C terminus.

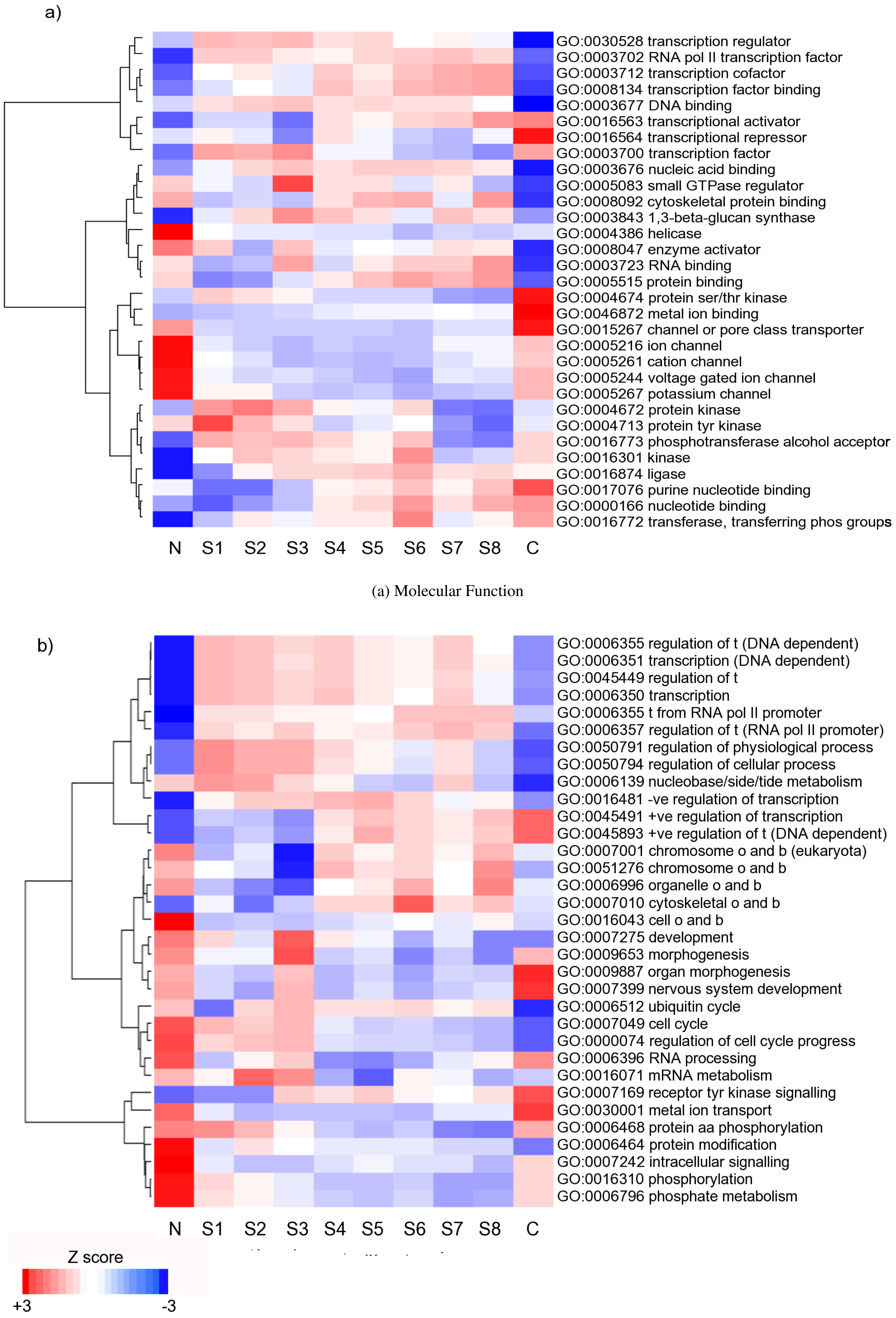
Disordered residues were over-represented at the N terminus within the set of Ion Channel and more specifically potassium channel annotated proteins. A further weak association was observed between disorder at the C terminus and the ion channel categories. These observations can be confirmed by crystal structure information. For example, it has been reported that the majority of voltage-gated potassium channel proteins contain intrinsically disordered residues at their N and C terminus (Sansom 1998). At the N terminus, the residues are responsible for channel inactivation (Magidovich et al. 2006). The disordered residues at the C terminus are adjacent to a PDZ motif mediating binding to scaffold proteins that support the assembly of multiple ion channel subunits into a fully functioning complex (Sansom 1998).

These analyses of length and location dependent patterning of disordered regions with function suggest the appropriate use of these features as part of a function prediction approach, provided that they do not overlap significantly with existing features.

4.2.3 Disorder features in context with other features

The majority of protein features used in the FFPred approach were calculated directly from sequence (Table 4.1). Features were categorised into 14 biological attributes; sequence characteristics, amino acid properties, transmembrane, disorder, secondary structure, PEST (regions of sequences statistically enriched in Proline, Glutamic acid, Serine and Threonine residues thought to be signals for rapidly degradation), low complexity, phosphorylation, N and O glycosylation, signal peptides, protein sorting and coiled coils. Each feature set was either encoded as single values representing the protein sequence as a whole and termed 'global' features, or were 'spatial' (disorder features for example) and described attributes distributed across the length of the sequence.

Topological information from secondary structure and transmembrane residues was encoded in feature vectors similar to those used for disorder features. For regions of secondary structure and predicted transmembrane residues, the choice of segment sizes was consistent with those used for disordered regions. This comprised eight equally proportioned segments with additional N and C terminus regions at 50 residues intervals. Helix, Sheet and Disorder frequency descriptors for different length ranges were also used so that proteins with many short contiguous stretches of helix or sheet or disorder could be separated from those with few longer stretches. For secondary structure, helix frequencies were restricted to greater than 5 contiguous residues and



(a) Molecular Function

(b) Biological Process

Figure 4.3: Heatmaps showing patterns of disordered region location that are associated with function. High Z scores are coloured red whilst low Z scores are coloured blue.

Table 4.1: Feature definitions,types and algorithms used to predict them from sequence.

Feature	Class	Derivation
Global features		
Sequence length	sequence characteristics	Calculated from sequence
Amino acid composition	amino acid composition	20 dimensional vector
Charge	sequence characteristics	Calculated from sequence
Hydrophobicity	sequence characteristics	Calculated from sequence
Iso-electric point	sequence characteristics	Calculated from sequence
Molar extinction coeff.	sequence characteristics	Calculated from sequence
Aliphatic index	sequence characteristics	Calculated from sequence
Molecular weight	sequence characteristics	Calculated from sequence
Signal Peptide	signal peptide	Predicted using SignalP3.0 (Bendtsen et al. 2004)
Localisation	protein sorting	Predicted using PsortII (Nakai and Horton 1999)
Spatial features		
Secondary structure	secondary structure	Predicted using PSIPRED (McGuffin et al. 2000)
Disorder	disorder	Predicted using DISOPRED (Jones and Ward 2003)
Transmembrane regions	transmembrane	Predicted using Memsat (Jones 2007)
Pest regions	PEST	Predicted using pestfind (Rechsteiner and Rogers 1996)
Coiled coils	coiled coils	Predicted using ncoils (Lupas 1997)
Low complexity	low complexity	Predicted using pfilt (Jones and Swindells 2002)
Glycosylation N and O	N and O glycosylation	Predicted using NetNGlyc and NetOGlyc (Hansen et al. 1998)
Phosphorylation	phosphorylation	Predicted using NetPhos (Blom et al. 1999)

sheet frequencies were restricted to greater than 3 residues in order to reduce noise in secondary structure assignments. All features and descriptors are listed in Table 4.2.

Table 4.2: Feature encoding schemes represented by group, description and mathematical transform.

Feature Group	**Index**	**Name**	**Transform**
Amino acids	1 - 20	Percent residue composition	
Sequence Features	21	Sequence Length	
	22	Molecular weight	log(x)
	23	Average hydrophobicity	
	24	Charge	
	25	Molar extinction coeff	log(x)
	26	Iso electric point	
	27	Aliphatic index	
Transmembrane	28	Number of tms	log (1+x)
	29	Percent tm residues	
	30	Nterm tm residues %	
	31	Cterm tm residues %	
	32 - 39	Bins 1-8 tm residues %	
Psipred helices	40	Number of helices	log(1+x)
	41	Percent helical residues	
	42	Nterm helical residues %	
	43	Cterm helical residues %	
	44-51	Bins 1-8 helical residues %	
	52	count helices $<$ 10 residues	log(1+x)
	53	count helices 10-15 residues	log(1+x)
	54	count helices 15-20 residues	log(1+x)
	55	count helices 20-35 residues	log(1+x)
	56	count helices 30-50 residues	log(1+x)
	57	count helices 50-70 residues	log(1+x)
	58	count helices 70-100 residues	log(1+x)
	59	count helices 100+ residues	log(1+x)
Psipred sheets	60	Number of sheets	log(1+x)
	61	Percent sheet residues	
	62	N term sheet residues %	
	63	C term sheet residues %	
	64 - 71	Bins 1-8 sheet residues %	
	72	count sheets $<$ 10 residues	log(1+x)
	73	count sheets 10 - 15 residues	log(1+x)
	74	count sheets 15 - 20 residues	log(1+x)
	75	count sheets 20 - 25 residues	log(1+x)
	76	count sheets 25 - 30 residues	log(1+x)
	77	count sheets 30 - 40 residues	log(1+x)
	78	count sheets 40+ residues	log(1+x)
Psipred - Random coil	79	percent random coil residues	
	80	Nterm random coils %	
	81	Cterm random coils %	
Coiled coils	82	Number of coiled coils	log(1+x)
	83	Percent coiled coil residues	
	84	Nterm coiled coil residues	
	85	Cterm coiled coil residues	
	86 - 93	Bins 1-8 coiled coil residues	

... Continued on next page

Table 4.2 – continued from previous page

Feature Group	Index	Name	Transform
Disorder	94	Number of disordered regions	log(1+x)
	95	Percent disordered residues	
	96	Nterm disordered residues	
	97	Cterm disordered residues	
	98 - 105	Bins 1-8 disordered residues	
	106	count disorder regions < 50	log(1+x)
	107	count disorder regions 50 - 100	log(1+x)
	108	count disorder regions 100 - 150	log(1+x)
	109	count disorder regions 150 - 200	log(1+x)
	110	count disorder regions 200 - 300	log(1+x)
	111	count disorder regions 300 - 500	log(1+x)
	112	count disorder regions 500+	log(1+x)
Pest regions	113	Number of pest regions	
	114	Percent pest regions	
	115	Nterm pest residues %	
	116	Cterm pest residues %	
	117 - 124	Bins 1-8 pest residues	
Low complexity	125	Number of low complexity regions	
	126	Percent low complexity regions	
	127	Nterm low compexity residues %	
	128	Cterm low complexity residues %	
	129 - 136	Bins 1-8 low complexity residues	
Phosphorylation	137	Number Ser phosphorylated residues	log(1+x)
	138	Number Thr phosphorylated residues	log(1+x)
	139	Number Tyr phosphorylated residues	log(1+x)
	140 - 147	Bins 1-8 Number of Ser phosphorylated residues	log(1+x)
	148 - 155	Bins 1-8 Number of Thr phosphorylated residues	log(1+x)
	156 - 163	Bins 1-8 Number of Tyr phosphorylated residues	log(1+x)
	164	Nterm Ser phosphorylated residues	log(1+x)
	165	Cterm Ser phosphorylated residues	log(1+x)
	166	Nterm Thr phosphorylated residues	log(1+x)
	167	Cterm Thr phosphorylated residues	log(1+x)
	168	Nterm Tyr phosphorylated residues	log(1+x)
	169	Cterm Tyr phosphorylated residues	log(1+x)
	170	ATM phosphorylated residues	log(1+x)
	171	CKI phosphorylated residues	log(1+x)
	172	CKII phosphorylated residues	log(1+x)
	173	CAMII phosphorylated residues	log(1+x)
	174	DNAPK phosphorylated residues	log(1+x)
	175	38MAPK phosphorylated residues	log(1+x)
	176	EGFR phosphorylated residues	log(1+x)
	177	GSK3 phosphorylated residues	log(1+x)
	178	INSR phosphorylated residues	log(1+x)
	179	PKA phosphorylated residues	log(1+x)
	180	PKB phosphorylated residues	log(1+x)
	181	PKC phosphorylated residues	log(1+x)
	182	PKG phosphorylated residues	log(1+x)
	183	RSK phosphorylated residues	log(1+x)

... Continued on next page

Table 4.2 – continued from previous page

Feature Group	Index	Name	Transform
	184	SRC phosphorylated residues	log(1+x)
	185	cdc2 phosphorylated residues	log(1+x)
	186	38MAPK phosphorylated residues	log(1+x)
O Glycosylation	187	Number of O glycosylated residues	log(1+x)
	188	Nterm O glycosylated residues	log(1+x)
	189	Cterm O glycosylated residues	log(1+x)
	190 - 197	Bins 1-8 Number of O glycosylated residues	log(1+x)
N Glycosylation	198	Number of N glycosylated residues	log(1+x)
	199	Nterm N glycosylated residues	log(1+x)
	200	Cterm N glycosylated residues	log(1+x)
	201 - 203	Bins 1-3 N glycosylated residues	log(1+x)
Localisation (Psort)	204	PsortII nuclear	
	205	PsortII cytoplasmic	
	206	PsortII mitochondrion	
	207	PsortII cytoskeletal	
	208	PsortII peroxisomal	
	209	PsortII secretory vesicles	
	210	PsortII golgi	
	211	PsortII vacuolar	
	212	PsortII plasma membrane	
	213	PsortII extracellular	
	214	PsortII endoplasmic reticulum	
SignalP	215	SignalP length	log(1+x)
	216	SignalP cscore	
	217	SignalP yscore	
	218	SignalP sscore	
	219	SignalP anchor	

Feature similarity was examined by constructing a matrix of Pearson's correlation values between the 14,641 protein representatives. The pairwise coefficients were converted to a distance measure (using 1-pearson correlation) and used as input to classical MDS (Multi-Dimensional Scaling, Togerson 1958). The positions of the features in the first 3 dimensions showed that there were three orthogonal feature axes (Figure 4.4) that comprised disorder, amino acid compositions and localisation characteristics of the proteins respectively. The isolation of disorder features (green highlight) in the total and relative feature space suggested that disorder features were distinct providing new information that could be useful in function category recognition.

Sequence features, hydrophobicity and charge were related to the frequencies of particular amino acids within proteins and consequently occupy a similar region of the plot. Correlations between predicted phosphorylation sites and frequency of Ser, Thr, and Tyr residues (Pearson correlation 0.2) were due to the fact that high frequencies of phosphorylated residues can only be observed when the relevant amino acid types occurred with a high frequency in the protein. Similarly, the frequencies of predicted O and N glycosylation sites displayed correlations with the occurrence of Asn and Ser/Thr residues. The features most closely related to disorder were random coils, PEST, and low-complexity descriptors with correlation values of 0.472, 0.211, and 0.307, respectively, at the residue frequency level.

These correlations, although relatively weak, indicated that some of the information within the disorder features is also encoded by these related descriptors. Disordered regions in proteins frequently contain residues that are also recognised as low sequence complexity (Tompa et al. 2008); however, a region of low complexity does not always imply structural disorder. For example, fibrous proteins such as collagens and silks are rigidly structured in their native state yet contain repetitive regions of low complexity (Perumal et al. 2008).

PEST motifs are degradation motifs present in proteins involved in protein phosphorylation, protein-protein interactions, and cell adhesion (Rogers et al. 1986). These motifs have been shown to be enriched in an experimentally characterised database of disordered proteins, and the residues that characterise the motifs represent a subset of those amino acids known to be disorder-promoting (Tompa et al. 2008). However, the correlations observed here between predicted occurrences of these features were small. The general spatial isolation of disorder descriptors in feature space suggested that they contain unique biological information not represented by the other features previously used in function prediction.

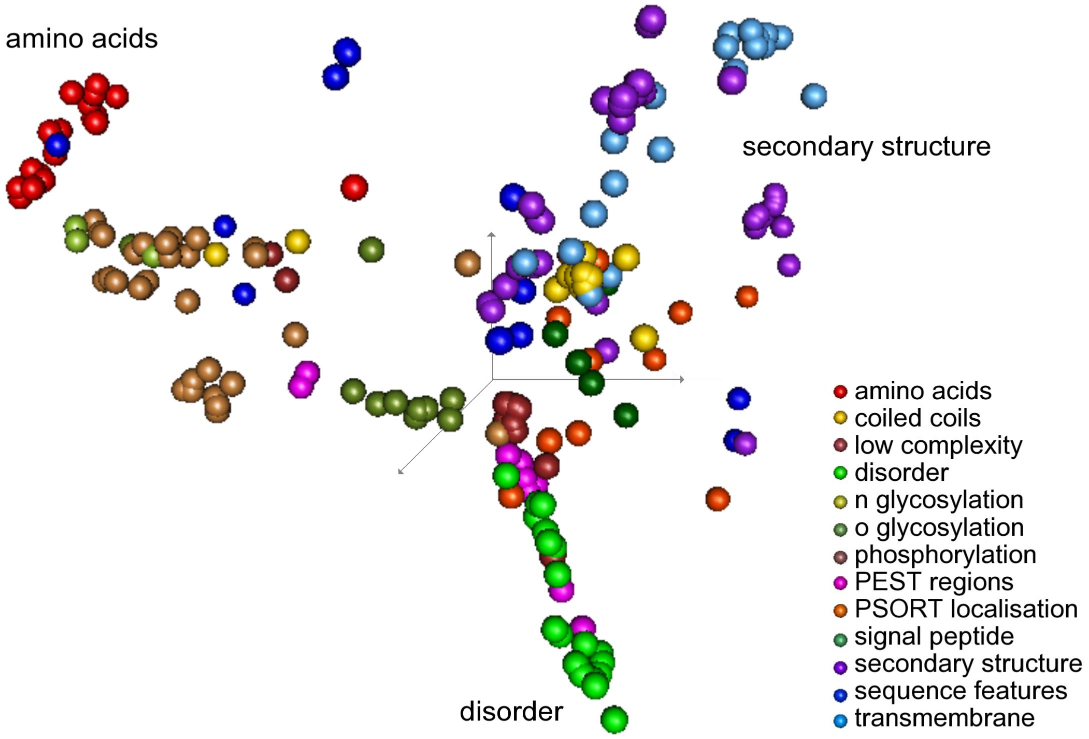


Figure 4.4: MDS plot of feature orthogonality. Similar features lie close together in feature space whilst unrelated features lie far apart.

4.3 Support Vector Classification of Function

121 Gene Ontology function terms and 231 process terms existed with at least 50 example protein sequences in the set of nr60 sequence representatives. It was not anticipated that each function category would relate to the different input features in the same way, however as an initial strategy, SVM's were trained using all features. MF and BP Ontology categories for which first stage training performance was random (achieved a Matthew's correlation of 0 or below) were removed at this stage because it was assumed that further optimisation on these classifiers would not yield sufficient improvement for practical use. For the remaining categories feature elimination (iteratively removing a single feature until a combination was reached that maximised the MCC) and further parameter optimisations were performed.

After this procedure 88 MF and 93 BP GO classifiers remained. A further quality filter of ¿= 50% sensitivity reduced the set to 86 and 91 for MF and BP respectively. The minimum MCC for any classifier was significantly above the random baseline at 0.208. Often classifiers required different inputs. For example, the classifier for cytokines performed best without disorder or transmembrane features, perhaps because sequences belonging to this function class were extracellular proteins that did not contain transmembrane or long disordered regions. Other classifiers, for example, protein phosphatases required all 14 feature inputs for optimal performance.

4.3.1 Training and testing datasets

A positive, negative and ambiguous dataset was constructed for each annotation term. The positive training set comprised sequences that were annotated directly by a GO term or any of its children. The negative training example sets comprised sequences not annotated by the considered GO term or any of its children. Sequences annotated by a parent of the considered GO term were filtered from the negative examples and tagged as ambiguous since their current annotations were incomplete and the example sequence might represent either a future positive or negative example.

The positive and negative example sets for each term were then partitioned into 5 equal sized groups. The groups were constructed using an iterative sequence based partitioning algorithm. Orphan seed proteins were randomly assigned to each group and new proteins sequentially added to the groups with which they shared the most similarity determined by alignment score. The resulting partitions contained maximally similar sequences for the annotation category whilst ensuring between group similarity was minimised.

Each of the 5 SVM's was then trained using a dataset consisting of 4 of the 5 partitions and tested on the remaining maximally dissimilar partition to produce an ensemble of 5 classifiers per annotation category operating on the same input features. This procedure often increases performance over the use of single classifiers by allowing each different classifier to utilise different feature weightings and optimal parameters for prediction. Extra confidence in annotation assignments can then be gained from using multiple independent predictions.

4.3.2 Kernel choice and parameter optimisation

The SVMLight software was used for the training and optimisation (An et al. 1998). The radial basis kernel function was chosen for the feature transformation phase as it is a popular choice in bioinformatics classification problems and has been shown to be most effective over other kernels in protein structure classification, prediction of protein function from microarray data and pattern recognition of DNA sequences (Byvatov and Schneider 2003). The rbf kernel is doubly sigmoidal in shape with a parameter (γ) that controls the width of the sigmoid. A large γ value implies a tall slim functional form whilst a small γ value (close to 0) produces a flat function that approximates the linear kernel. The rbf kernel can approximate both the linear and sigmoid kernels by varying the width parameter, consequently it is frequently the method of choice for generating the kernel matrix (Hsu et al. 2003).

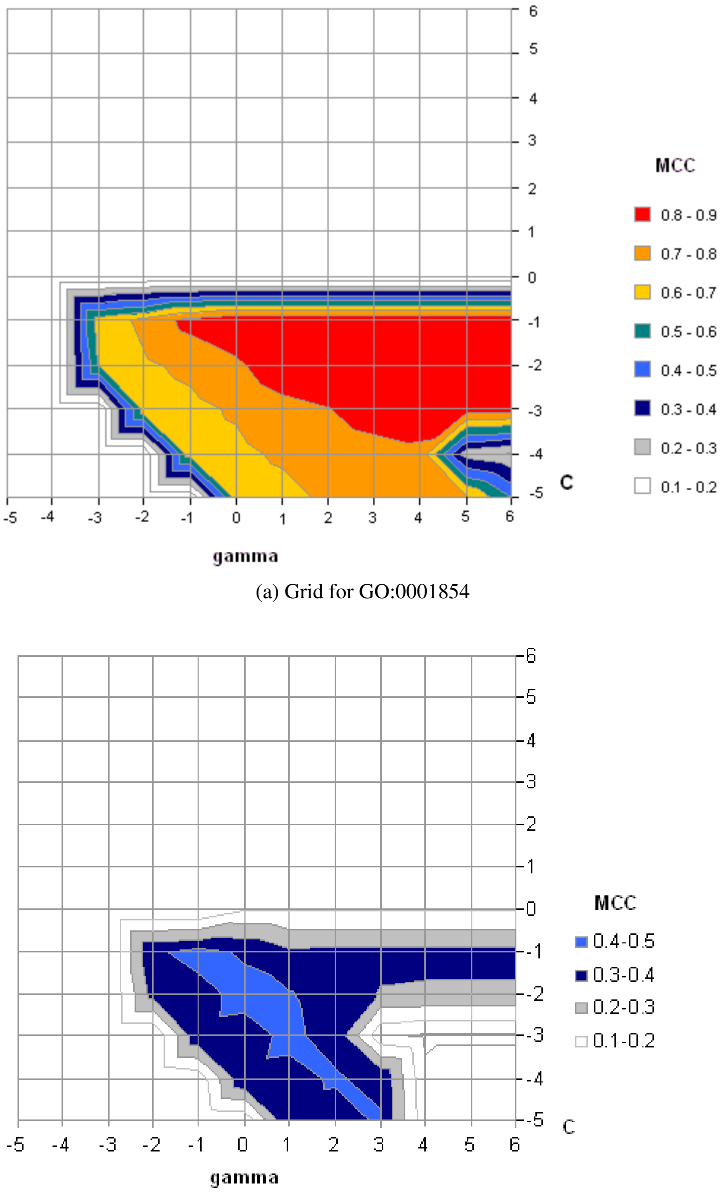
An SVM optimises separation of the feature transformed data by positioning a hyperplane between positive and negative training examples (see Appendix I). The position of the hyperplane is chosen to maximise a margin, the distance between the nearest positive and negative examples, the support vectors. The regularization parameter C controls the trade-off between the cost of errors on the training examples that cannot be perfectly separated by the hyperplane and the complexity of the model (VC dimension). Small values of C allow many classification errors during training (a soft margin), whereas large values of C increase the error penalty so that very few mis-classifications are allowed during training (hard margin) (Bishop 2006).

Both C and γ parameters were optimised by performing a 12x12 grid search (see Figure 4.5 for examples) over the range 1e-6 to 1e+6. For most of the GO term training sets there was a large bias towards the number of negative training examples. This imbalance was controlled by a third parameter J equal to the trade-off between training errors in the positive example set and the negative example set. To simulate training on a balanced dataset J could be set to equal the ratio of negative to positive examples. To bias the performance of each classifier towards low numbers of false positives, the class imbalance was controlled by doubling the cost of an error on a negative example.

4.3.3 Function Category Classification Results

The classification performance was measured for each Gene Ontology category using specificity, precision and Matthew's correlation coefficient as defined in Section 3.2.3 Chapter 3 (Equation 3.4). The number of true positives (tp) represents the number of correctly recognised proteins with a particular function, true negatives (tn) are the number of correctly recognised proteins that do not have the function. False positives (fp) occur when the classifier incorrectly assigns a function to a protein and false negatives (fn) occur when a protein bearing a particular function is missed. The MCC coefficient was used because it is independent of the numbers of positive and negative examples in the test protein sets.

The categories that could be successfully predicted mainly comprised signalling and regulatory functions, and included membrane protein families, transcription factors and other sequences involved in molecular recognition Table 4.3). This result suggests that for sequences that are members of these categories, sequence feature characteristics can be sufficient to infer function. It is also noted that it is difficult to obtain crystal structure information for many of these sequences, which may imply that a lack of rigid conformation is an important determinant of these functions and that this flexibility is encoded in the amino sequence of proteins.



(a) Grid for GO:0001854

(b) Grid for GO:0003676

Figure 4.5: Examples of C and gamma grid search results. Optimal parameters selections for term GO:001854 lie between C 1e-1 and 1e-3 and gamma 100 to 1e+6, whilst the optimal parameter selection for GO:0003676 occupy a much narrower range.

Table 4.3: GO category classifier performance.

GO term	Name	MCC	Sensitivity	Specificity	Precision
Biological Process					
GO:0007608	sensory perception of smell	0.730	0.812	0.995	0.663
GO:0007186	GPCR signaling pathway	0.724	0.660	0.989	0.836
GO:0007156	homophilic cell adhesion	0.714	0.667	0.998	0.769
GO:0007606	sensory perception of chemical stimulus	0.685	0.778	0.993	0.612
GO:0007187	GPCR coupled to cyclic nucleotides	0.669	0.595	0.999	0.759
GO:0006355	regulation of transcription, DNA-dependent	0.578	0.728	0.902	0.584
GO:0016337	cell-cell adhesion	0.578	0.446	0.998	0.761
GO:0019935	cyclic-nucleotide-mediated signaling	0.568	0.579	0.996	0.564
GO:0045449	regulation of transcription	0.563	0.684	0.908	0.600
GO:0006351	transcription DNA dependent	0.559	0.720	0.893	0.566
GO:0006350	transcription	0.556	0.684	0.902	0.597
GO:0006817	phosphate transport	0.539	0.463	0.998	0.633
GO:0007166	cell surface receptor linked signal transduction	0.525	0.511	0.963	0.643
GO:0007200	G-protein signaling, coupled to IP3	0.523	0.447	0.998	0.618
GO:0019932	second-messenger-mediated signaling	0.497	0.433	0.995	0.586
GO:0050794	regulation of cellular processes	0.445	0.605	0.853	0.557
GO:0050791	regulation of physiological process	0.443	0.603	0.854	0.551
GO:0006139	nucleobase, side, tide and nucleic acid metabolism	0.438	0.577	0.858	0.587
GO:0007218	neuropeptide signaling pathway	0.425	0.349	0.998	0.524
GO:0009101	glycoprotein biosynthesis	0.417	0.321	0.997	0.554
GO:0006486	protein amino acid glycosylation	0.414	0.361	0.996	0.488
GO:0008152	metabolism	0.401	0.664	0.747	0.806
GO:0006468	protein amino acid phosphorylation	0.372	0.495	0.950	0.339
GO:0007155	cell adhesion	0.371	0.553	0.938	0.303
GO:0006811	ion transport	0.370	0.447	0.954	0.377
GO:0007600	sensory perception	0.369	0.416	0.974	0.370
GO:0015837	amine transport	0.350	0.508	0.991	0.248
GO:0030001	metal ion transport	0.336	0.438	0.969	0.295
GO:0048015	phosphoinositide-mediated signalling	0.333	0.418	0.992	0.274
GO:0006796	phosphate metabolism	0.331	0.398	0.947	0.364
GO:0006865	amino acid transport	0.320	0.509	0.990	0.209
GO:0007165	signal transduction	0.319	0.290	0.947	0.594
GO:0006812	cation transport	0.315	0.597	0.907	0.215
GO:0006810	transport	0.306	0.339	0.921	0.504
GO:0016310	phosphorylation	0.304	0.453	0.927	0.277
GO:0016567	protein ubiquitination	0.303	0.162	0.997	0.607
GO:0006820	anion transport	0.303	0.351	0.988	0.279
Molecular function					
GO:0001584	rhodopsin-like receptor activity	0.890	0.883	0.996	0.906
GO:0004497	monooxygenase activity	0.890	0.842	0.999	0.941
GO:0030594	neurotransmitter receptor activity	0.763	0.765	0.998	0.765
GO:0004252	serine-type endopeptidase activity	0.719	0.576	0.999	0.905
GO:0008236	serine-type peptidase activity	0.711	0.583	0.999	0.875
GO:0004930	G-protein coupled receptor activity	0.706	0.727	0.984	0.717
GO:0004871	signal transducer activity	0.646	0.716	0.937	0.699
GO:0016757	transferase activity, transferring glycosyl groups	0.591	0.511	0.996	0.697
GO:0030414	protease inhibitor activity	0.568	0.462	0.998	0.706
GO:0004866	endopeptidase inhibitor activity	0.568	0.462	0.998	0.706
GO:0003677	DNA binding	0.568	0.670	0.925	0.596
GO:0005125	cytokine activity	0.558	0.516	0.996	0.615
GO:0004888	transmembrane receptor activity	0.526	0.535	0.967	0.592
GO:0003700	transcription factor activity	0.522	0.405	0.989	0.733
GO:0042626	ATPase activity, coupled to transmembrane movement of substances	0.519	0.480	0.996	0.571
GO:0015293	symporter	0.502	0.588	0.995	0.435
GO:0005275	amine transporter activity	0.498	0.643	0.995	0.391
GO:0008194	UDP-glycosyltransferase activity	0.497	0.500	0.997	0.500
GO:0004713	protein-tyrosine kinase activity	0.488	0.340	0.997	0.720

... Continued on next page

Table 4.3 – continued from previous page

GO term	Name	MCC	Sensitivity	Specificity	Precision
GO:0004674	protein-serine threonine kinase activity	0.430	0.321	0.994	0.610
GO:0004872	receptor activity	0.429	0.511	0.928	0.484
GO:0046943	carboxylic acid transporter activity	0.428	0.647	0.990	0.289
GO:0003824	catalytic activity	0.421	0.620	0.798	0.642
GO:0003676	nucleic acid binding	0.419	0.599	0.841	0.519
GO:0042277	peptide binding	0.412	0.769	0.975	0.230
GO:0005215	transporter activity	0.390	0.491	0.925	0.422
GO:0030528	transcription regulator activity	0.372	0.680	0.853	0.293
GO:0005267	potassium channel activity	0.371	0.174	0.999	0.800
GO:0015276	channel or pore class transporter activity	0.363	0.267	0.999	0.500
GO:0004386	helicase activity	0.350	0.357	0.993	0.357
GO:0000166	nucleotide binding	0.342	0.450	0.903	0.415
GO:0016773	phosphotransferase activity, alcohol group as acceptor	0.331	0.331	0.974	0.395
GO:0016758	transferase activity, transferring hexosyl groups	0.309	0.214	0.997	0.462
GO:0004672	protein kinase activity	0.309	0.239	0.988	0.456
GO:0008233	peptidase activity	0.309	0.675	0.860	0.197
GO:0016740	transferase activity	0.302	0.617	0.793	0.281

Performance of GO classifiers with MCC $\geq$ 0.3 for Molecular Function and Biological Processes. For each ontology term the performance values represent the performance obtained by summing the numbers of true positives, true negatives, false positives and false negatives obtained by testing each of the 5 SVM's on the blind 1/5 partitioned dataset.

4.3.4 Assessing the importance of different features in function prediction

The value of each feature in function classification performance for each Gene Ontology category was assessed by performing a leave one out feature assessment and noting the reduction in classification performance. This strategy is time consuming since it requires re-training of the algorithm after each feature has been omitted, but provides accurate measure of the value of the feature for the particular classifier.

A survey across all Molecular Function and Biological Process categories was performed and summarised by reporting the average percentage MCC loss obtained when removing a feature from all of the classifiers. The results (Figure 4.6) suggest that secondary structure is the most informative feature of all for most of the categories. Secondary structure was less informative for Biological Processes, however was still the most informative feature set. Disorder was strikingly the second most important feature set for prediction of Biological Processes. This finding may be correlated with the differences between the two Ontologies. For example, Molecular Functions deal largely with protein family memberships and binding activities whereas Biological Process annotations describe regulatory and metabolic pathways. It might be that disorder features are relevant to more of the process categories than function categories, giving rise to a greater average value for the BP ontology.

Value of Disorder Features

Since disorder features had not been used previously in function category recognition, individual functions for which these features played an important part were determined. Initially, GO terms for which disorder is expected to contribute were identified using the Fischer Exact test (see Chapter 1 Section 1.2). The method provides a robust measure of over-representation of disorder since the proteins are no more than 60% identical within functional classes. However this value does not account for either feature redundancy or feature interaction effects.

To evaluate the contribution of disorder features for individual categories, the performance loss was measured when disorder features were removed from each classifier using the Matthews Correlation Coefficient (MCC). This measure represents the additional value of disorder features in function prediction, accounting for both interaction and compensatory effects between features. Classifier performance was reported for 26 GO categories (Table 4.4) whose sensitivity

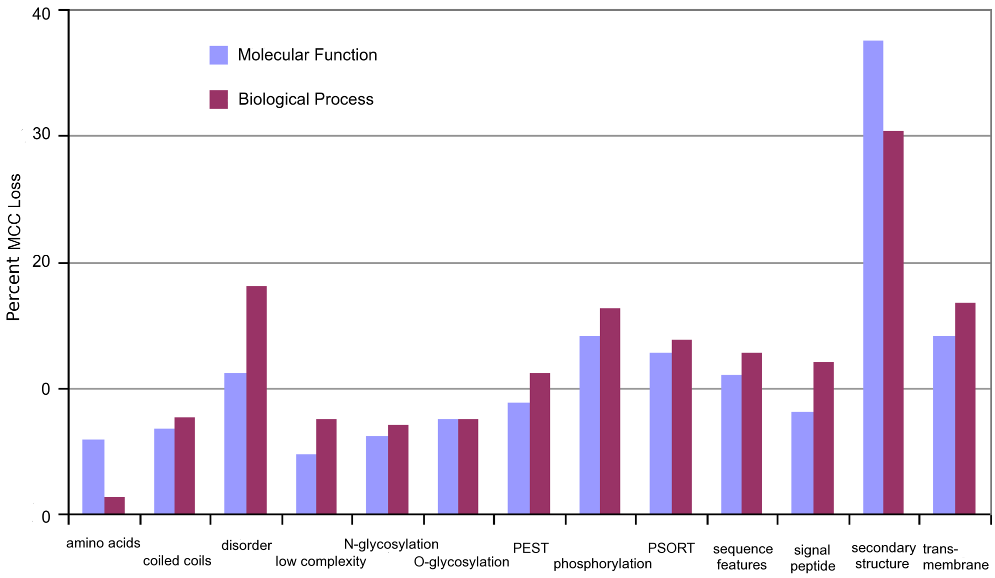


Figure 4.6: Feature importance estimates quantified by the percent loss in classification performance obtained when each feature set is omitted from each classifier.

at a false positive rate of 10% exceeded 50%. The significance of the improvements in correlation coefficients for individual categories were evaluated using Fisher's Z test, which considers both the magnitude of the performance increase and the strength of correlation (Equations 3.5 - 3.7 Chapter 3). The improvements that were significant at the 5% level ($p < 0.05$) were marked in bold (Table 4.4, column MCC+diso).

Classification performance for 11 Biological Process categories and 12 Molecular Function categories that were identified as enriched in disordered proteins were significantly improved when disorder features were added. Several additional GO classes were identified during feature selection that required disorder features for optimal performance that were not identified from the statistical tests (Figure 4.1). These comprised "UDP-glycosyl transferase", "hormone", "growth factors", "transferase", "hydrolase", and "carboxylic acid transporter" MF categories, and "G protein signalling" Biological Process category. The most notable performance gains were observed for "protein tyrosine kinase signalling," "G protein signalling", "ubiquitin specific protease", "transcription", "protein kinase", and "helicase" categories. For some categories; "cation-channel", "ion channel", "metal ion transport", "purine-nucleotide binding", "nucleotide binding", and "DNA binding", little or no performance increase resulted from the addition of disorder features. Particularly for "Ion channels", "Metal Ion transporters", and "Nucleotide binding" categories, other features such as transmembrane regions or secondary structure better characterised the relationship between the primary amino acid sequence of the protein and its function.

The correlation values obtained when classifiers were trained with only disorder features showed that some BP categories relating to transcription, and the "Transcription factor" MF category could be recognised with sensitivities of $> 50\%$ at false positive rates of less than 10%. For these categories, the increased performance resulting from the addition of disorder features was much lower than the correlation obtained from disorder features alone. This result can be explained by the representation of mutual information between random coil, low complexity, or PEST features which reduced the magnitude of the effect of the disorder features. Conversely, for "G protein signalling" and "Receptor tyrosine kinase" BP categories, "Growth factor", "Helicase", "Hydrolase", and "Ubiquitin specific protease" MF categories, the improvement resulting from the addition of disorder features was greater than the correlation obtained using disorder features alone. This finding indicates that disorder features interacted cooperatively with other features in the dataset to achieve a greater performance increase.

Table 4.4: Additional value of disorder features.

GO Category	Description	MCC+diso	MCC-diso	MCC diso
Biological process				
GO:0006139	Nucleo- base/side/tide, nucleic acid metabolism	0.452	0.433	0.233
GO:0006350	Transcription	0.565	0.532	0.333
GO:0006351	Transcription, DNA dependent	0.566	0.546	0.333
GO:0006355	Regulation of transcription, DNA dependent	0.581	0.557	0.353
GO:0006796	Phosphate metabolism	0.348	0.317	0.129
GO:0007169	Receptor tyr kinase signalling	0.343	0.203	0.111
GO:0007200	G protein signalling	0.531	0.404	0.109
GO:0016310	Phosphorylation	0.321	0.299	0.079
GO:0030001	Metal ion transport	0.367	0.367	0.145
GO:0045449	Regulation of transcription	0.572	0.559	0.342
GO:0050791	Regulation of physiological processes	0.455	0.429	0.313
GO:0050794	Regulation of cellular process	0.455	0.435	0.313
Molecular function				
GO:0000166	Nucleotide binding	0.361	0.361	0.107
GO:0003676	Nucleic acid binding	0.486	0.471	0.272
GO:0003677	DNA binding	0.452	0.452	0.293
GO:0003700	Transcription factor	0.538	0.498	0.323
GO:0004386	Helicase	0.362	0.221	0.134
GO:0004553	Hydrolase	0.354	0.200	0.095
GO:0004672	Protein kinase	0.429	0.362	0.142
GO:0004674	Protein serine/threonine kinase	0.479	0.394	0.147
GO:0004713	Protein-tyrosine kinase	0.373	0.304	0.123
GO:0004843	Ubiquitin-specific protease	0.392	0.261	0.098
GO:0005179	Hormone	0.243	0.198	0.103
GO:0005244	Voltage-gated ion channel	0.416	0.416	0.114
GO:0005261	Cation channel	0.447	0.447	0.148
GO:0008083	Growth factor	0.346	0.129	0.133
GO:0008194	UDP glycosyl-transferase	0.500	0.422	0.127
GO:0016740	Transferase	0.316	0.273	0.074
GO:0016773	Phosphotransferase, alcohol group as acceptor	0.339	0.331	0.128
GO:0017076	Purine nucleotide binding	0.365	0.365	0.136
GO:0030528	Transcription regulator	0.371	0.324	0.291
GO:0046943	Carboxylic acid transporter	0.413	0.389	0.140

Classification performance measured by Matthews Correlation Coefficient (MCC) for all features including disorder features (MCC+diso), all features without disorder (MCC-diso) and disorder features alone (MCC diso only).

Throughout this study, classification performance for GO categories has been reported using MCC. This measure accounts for unbalanced training class frequencies encountered for virtually all GO terms, however is sensitive to the different total GO term class sizes. Scoring a single positive or negative result for different terms therefore affects the correlation values to a different degree. For this reason, classification sensitivities obtained at 10%, 5%, and 1% error rates were also reported (Table 4.5). At very low false positive rates (1% FPR) annotation coverage ranged between 0.081 (GO:00050794 regulation of cellular processes) and 0.563 (GO:0008194 UDP glycosyl-transferase) (Table 4.5). At higher false positive rates (10% FPR) coverage was much improved. These statistics are practically useful for whole genome annotation efforts where the number of tested sequences is in the tens of thousands consequently low false positive rates are required.

4.4 Benchmarking against ProtFun method

This method differed from the original ProtFun method in several important ways. Firstly, the predictions for structure, disorder, and transmembrane regions used PSI-BLAST profiles rather than single sequence predictions as feature inputs. Second, additional secondary structure features were encoded that recorded the frequencies of helices and strands of particular length ranges within each protein. Despite these differences, a benchmark comparison between this method and the ProtFun method was attempted. Since ProtFun was not available as a standalone software package, performance of the FFPred method was compared to the predictions made by the ProtFun server for the 14,651 annotated sequences used in this study.

Classifier accuracy was reported for fourteen common categories (Table 4.6). The FFPred method outperformed the ProtFun server for all tested categories using MCC as the performance measure. Improvements were significant at the 95% level using Fisher's Z test for significance of correlation difference, except for the "Ion channel" category. The performance of the FFPred method without the use of disorder feature inputs was also compared with ProtFun server performance so that any improvements in accuracy could be attributed primarily to the inclusion of disorder features or to differences between the other features, the use of different training datasets, and differences between machine learning algorithms.

Four of the function categories; "Ion Channel", "Voltage gated ion channel", "Cation channel", and "Metal ion transport" did not utilise information from disorder features; therefore the perfor-

Table 4.5: Classifier sensitivity obtained at different error rates

GO	Description	Total	Pos	10% FPR	5% FPR	1% FPR
Molecular Function						
GO:0000166	nucleotide binding	3690	1673	0.473	0.285	0.091
GO:0003676	nucleic acid binding	5707	2846	0.576	0.417	0.152
GO:0003677	DNA binding	3018	1723	0.578	0.416	0.147
GO:0003700	transcription factor	1354	789	0.694	0.605	0.359
GO:0004386	helicase	288	132	0.657	0.579	0.279
GO:0004553	hydrolase	184	93	0.596	0.495	0.272
GO:0004672	protein kinase	1096	521	0.590	0.475	0.255
GO:0004674	protein ser/thr kinase	790	374	0.663	0.571	0.342
GO:0004713	protein-tyrosine kinase	549	254	0.670	0.545	0.360
GO:0004843	ubiquitin-specific protease	115	59	0.710	0.645	0.403
GO:0005179	hormone	158	74	0.684	0.557	0.241
GO:0005244	voltage-gated ion channel	281	141	0.602	0.534	0.363
GO:0005261	cation channel	413	218	0.667	0.556	0.391
GO:0008083	growth factor	217	132	0.580	0.449	0.326
GO:0008194	UDP glycosyl-transferase	121	78	0.788	0.788	0.563
GO:0016740	transferase	3018	1492	0.326	0.199	0.046
GO:0016773	phosphotransferase, alcohol acceptor	1305	620	0.529	0.420	0.196
GO:0017076	purine nucleotide binding	3223	1447	0.461	0.299	0.088
GO:0030528	transcription regulator	1856	1121	0.565	0.382	0.134
GO:0046943	carboxylic acid transporter	146	83	0.791	0.721	0.523
Biological Process						
GO:0006139	nucleo- base/side/tide, nucleic acid metabolism	5050	2773	0.522	0.367	0.148
GO:0006350	transcription	3262	1868	0.698	0.535	0.223
GO:0006351	transcription, DNA dependent	2962	1720	0.702	0.538	0.243
GO:0006355	regulation of transcription, DNA dependent	2886	1671	0.734	0.560	0.266
GO:0006796	phosphate metabolism	1519	762	0.469	0.394	0.162
GO:0007169	receptor tyr kinase signalling	146	109	0.500	0.435	0.278
GO:0007200	G protein signalling	81	60	0.638	0.574	0.511
GO:0016310	phosphorylation	1244	609	0.529	0.370	0.117
GO:0030001	metal ion transport	563	305	0.615	0.536	0.135
GO:0045449	regulation of transcription	3155	1783	0.711	0.550	0.269
GO:0050791	regulation of physiological processes	4378	2528	0.524	0.359	0.089
GO:0050794	regulation of cellular process	4442	2590	0.518	0.345	0.081

Each measure in the FPR columns of the table represent specificity or coverage (proportion of true positives) obtained at different false positive rates (FPR).

Table 4.6: Performance comparison with ProtFun.

GO ID	Description	FFPred			ProtFun		
		TPR	FPR	MCC	TPR	FPR	MCC
GO:0030001	Metal ion transport	45.72	3.28	0.37	0.00	0.00	0.00
GO:0005244	Voltage gated ion channel	33.33	0.65	0.36	10.05	1.87	0.08
GO:0005261	Cation ion channel	40.89	0.30	0.45	22.55	0.61	0.22
GO:0006350	Transcription	69.80	9.97	0.57	55.49	7.27	0.43
GO:0006355	Regulation of transcription	70.06	8.32	0.57	43.17	8.80	0.36
GO:0008083	Growth factor	25.36	0.43	0.35	1.40	1.79	-0.01
GO:0005216	Ion channel	44.89	0.90	0.29	33.61	0.95	0.28
GO:0005179	Hormone	40.51	1.60	0.24	18.43	1.96	0.15
GO:0006950	Stress Response	30.28	1.24	0.24	6.92	1.88	0.03
GO:0000955	Immune Response	50.40	6.80	0.43	23.34	0.21	0.06
GO:0005189	Structural Molecule	36.90	0.99	0.24	18.72	1.85	0.10
GO:0004872	Receptor	51.11	0.07	0.43	13.18	9.46	0.18
GO:0004871	Signal transducer	62.11	3.63	0.45	12.56	1.27	0.12
GO:0005215	Transporter	49.12	7.50	0.39	46.22	3.04	0.25

TPR and FPR represent percent true and false positives for the methods, whereas MCC values represent the Matthew's Correlation Coefficient.

mance increase resulted from other methodological differences. For the remaining categories, “Transcription”, “Regulation of transcription”, “Hormone”, and “Growth factors”, the source of performance improvements represented a mix of these effects and the addition of disorder features. The greatest accuracy increase resulting directly from the addition of disorder features was observed for the “Growth factor” category. “Transcription” and “Regulation of transcription” accuracies were improved more by the feature encoding and more recent training datasets used than the addition of disorder features. This result was not surprising considering that the ProtFun features included low complexity, PEST regions, and random coils that overlap considerably with disorder features within these categories.

In this benchmark study, it was difficult to provide an unbiased performance measure that was comparable between the two methods. For ProtFun the assessment was restricted to the use of the server output alone which selects a single most likely GO term assignment per sequence, rather than raw neural network output scores. For the FFPred method, performance measures were derived from the testing procedure described in Section 4.3. The FFPred method permits the assignment of multiple GO terms to a sequence and as such is statistically more likely to outperform a method producing single sequence function assignments. A further problem affecting the validity of the benchmark results was that the ProtFun method was likely to have been trained on at least some of the assessment sequences giving the method an unfair advantage. Despite these concerns, the results suggested that the FFPred method was significantly better overall at recognising function from sequence features.

4.4.1 FFPred Server

The FFPred method was implemented as a public domain server to make the prediction service available to the biological community (Figure 4.7). Since the method was trained and evaluated on human sequences, it was important to investigate its behaviour on other eukaryotic datasets. In order to assess the performance of the method on other organisms, the human classifiers were tested using Gene Ontology Annotations from the GOA project on eukaryotic model organisms zebrafish (*Danio rerio*), mouse (*Mus musculus*), fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*) and yeast (*Saccharomyces cerevisiae*).

Performance statistics; sensitivity, specificity, precision and Matthew’s correlation coefficient (MCC) were reported for classifiers performing better than random. The proteins in each genome that were annotated with one or more GO terms were used as the basis of the benchmark study. A result was considered correct if the server prediction or one of its parent annotations was represented in the GOA annotation. Sequences annotated at less specific GO term levels than the predicted term were omitted from scoring.

As evolutionary distance between the different species and human increased, the overall average classifier accuracy decreased (MCC values in Table 4.7). This could be attributed to a loss in sensitivity across more distantly related species; worm, fly and yeast. The sensitivities obtained for mouse and zebrafish were comparable with human. The average specificities observed for each proteome were consistently high between all organisms. This property is a requirement for predictors applied to whole proteomes to avoid large numbers of false positives where the expected number of GO term annotations (positives) are small compared with the number of sequences that are not annotated with a particular GO term.

The number of classifiers obtaining over 90% specificity at sensitivities of >30% also decreased as evolutionary distance from human increased (Table 4.7). This can be explained by differences in annotation quality between the various proteome annotation efforts, and also as a function of decreasing feature conservation between proteins from distant eukaryotic proteomes. Among 99 categories that were useful in predicting the functions of yeast proteins, the majority were more general annotation terms that had achieved greater accuracies on human proteins. These categories were focused around functions of enzymatic and transmembrane proteins. The majority of terms that performed poorly on the yeast sequences were biological process categories. This observation suggests that the features corresponding with many of these categories in human are not conserved within lower eukaryotes reflecting organism specific biology.

Overall, the benchmark results showed robust classification accuracies across the vertebrate and mammalian proteomes for most annotation categories. The use of this approach is recommended for annotation of vertebrate and mammalian proteomes; however, the benchmark results indicate that when run on proteins from lower eukaryotic organisms, the server is more likely to leave a protein unannotated rather than produce an erroneous annotation. The approach is not recommended for use with proteins from plants or prokaryotic organisms. Key differences in subcellular localisation, signalling pathways and post-translational modification pathways mean that patterns of features corresponding with function are not sufficiently conserved with those from human for effective function prediction.

Table 4.7: Server performance on eukaryotic model organisms

Organism	**Sensitivity**	**Specificity**	**Precision**	**Size**	**Categories**
Human	0.67	0.99	0.68	32 528	197
Mouse	0.48	0.98	0.52	12 684	186
Zebrafish	0.58	0.97	0.64	26 557	186
Fly	0.40	0.98	0.57	13 107	175
Worm	0.47	0.97	0.56	11 770	165
Yeast	0.34	0.97	0.61	5 527	99

Statistics represent sensitivity (proportion of true positives), specificity (proportion of false positives), precision (proportion of predictions that are correct) and proteome size. The number of categories represents the total number of GO terms that could be predicted at a level above random (Matthew's Correlation Value of 0).

FFPred server design and example usage

The FFPred server was designed for ease of use and easy interpretation of prediction results in mind. It accepts single protein sequences as input formatted as plain text or in FASTA format. It is expected that the amino acid sequence of interest represents the entire mature protein product of a gene or at least a genuine transcript. Server results based on sequence fragment inputs may be unreliable as feature information may differ substantially between truncated gene products. Additionally, if the sequence input has been recently processed or is present in the human IPI protein dataset, the user will be immediately directed to a web-page displaying feature information and GO term predictions for the given query sequence.

The server processing model describes the computational steps involved in making a set of GO term predictions from an input sequence (Figure 4.7). A user inputs a sequence, features are calculated using 3rd party software before being passed through the SVM library for prediction. In the case of a typical protein sequence, computation takes 12 to 15 minutes from initial sequence submission to receiving server results via email on an Intel Xeon 3.2 GHz processor running CentOS 4.4. The majority portion of this time is spent screening the GO term SVM library (on average 11 min per sequence).

Server output for sequence submissions are returned to the user by email containing a text summary of GO annotation predictions for an input sequence hyperlinked to a dynamically generated temporary results page (Figure 4.8). The results page details predicted features and GO annotations for the query sequence. The feature predictions are shown in tabular format as well as graphically mapped onto the sequence of interest. This allows for back interpretation of feature patterns responsible for functions.

GO term predictions are represented in hierarchical format or as a single table of individual term results. In the hierarchy view, each GO term is annotated according to whether it was predicted by classifiers present in the library, or whether an annotation was inherited through classifiers representing one or more of the child terms. This view enables the user to contextualise the predictions and derive extra confidence in predictions that are made by both parent and child term classifiers.

The server has two main practical uses; predicting novel annotations for orphan sequences and predicting new annotations for well characterised sequences. A typical example of each sce-

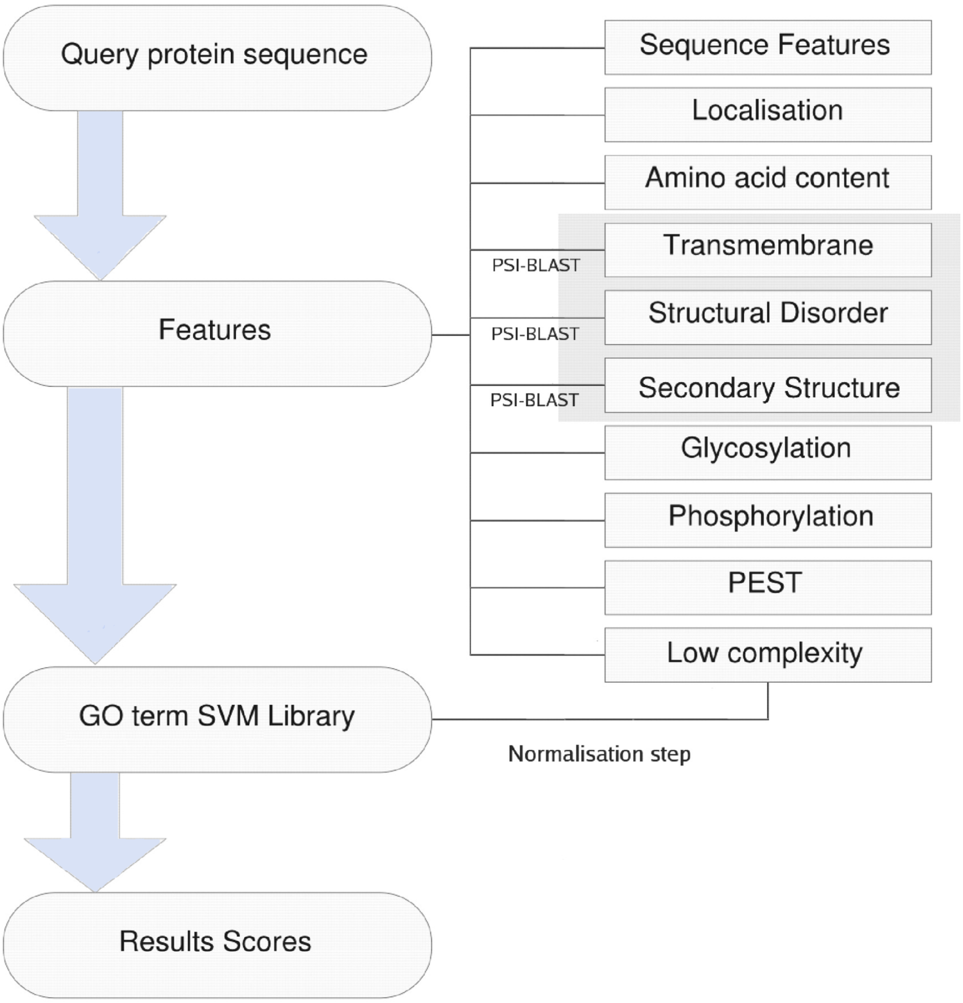
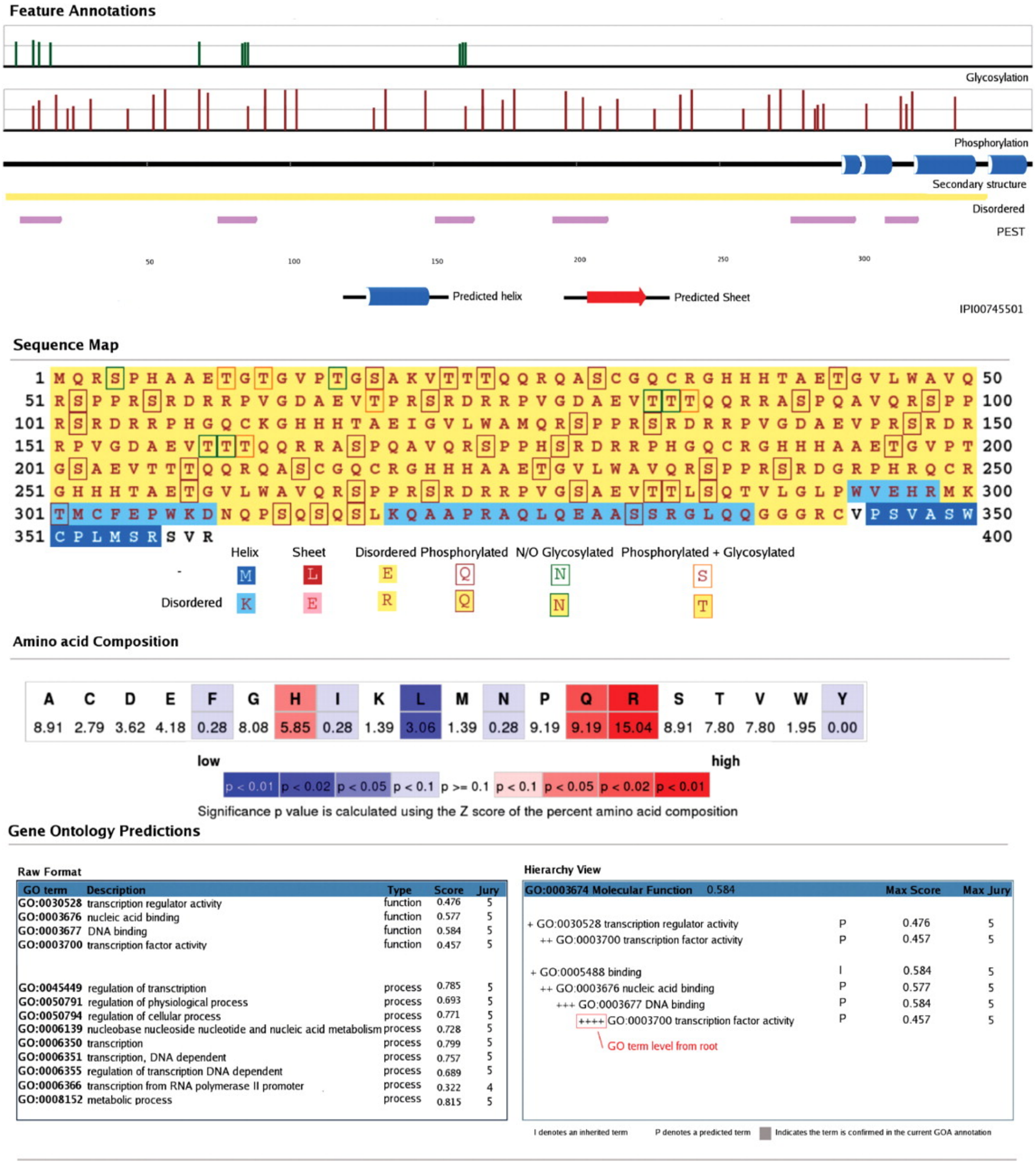


Figure 4.7: Server processing flow chart

nario follows. The first case represents an orphan human sequence IPI00745501 which has no discernible sequence relatives identified by BLAST sequence homology search. FFPred is able to make several predictions for the sequence consisting of several parent-child transcription related function categories. The predictions can be rationalised by analysis of the features that are responsible for function. For example, the sequence is enriched in charged residues, has little secondary structure and is predicted to contain multiple phosphorylation sites. These are all characteristics which frequently occur in DNA binding proteins (Churchill and Travers 1991). Whilst these results are encouraging, they require experimental verification.

A second example involves the re-annotation of a well characterised sequence, lactate dehydrogenase (LDH) known to participate in oxidative phosphorylation through conversion of lactate to pyruvate during metabolism (Markert 1984). Server predictions for this sequence (Listing 4.1) suggest that the enzyme is also responsible for amino acid and nitrogen metabolism, and translation. The amino acid and nitrogen metabolism predictions are false positives that result from similar features obtained for other dehydrogenase enzymes that participate in these processes. These function annotations highlight the fact that this method is sensitive at detecting annotations that cannot be inferred by homology, but lacks resolution where homologues have different functional roles. The annotation "GO:0006412 translation" represents a novel annotation for this sequence that can be supported by literature evidence. One publication "Lactate dehydrogenase is an AU-rich element-binding protein that directly interacts with AUF1" provides *in vivo* direct evidence of its involvement in translation through the observation that LDH is bound to AUF1 on mRNA that is actively translated (Pioli et al. 2002). A second paper "Identification of a nucleic acid helix-destabilising protein from rat liver as lactate dehydrogenase-5" (Williams et al. 1985) shows that *in vitro* lactate dehydrogenase is responsible for DNA helix-destabilisation.

The IPI human sequence dataset contained 2157 examples of poorly characterised or unannotated sequences that could not be related to well characterised sequences by BLAST homology search. FFPred was able to assign function classes to 57% of these sequences. The results have been made publicly available as part of the FFPred database. Each prediction must be considered independently and further available evidence gathered to support server assignments. The approach is not capable of generating highly accurate function assignments, however is ideal for identifying and prioritising a candidate set of functions for a novel sequence.



Sequence Map

```
  1 MQRSPHAAETGTGVPTGSAKVTTTQQRQASCGQCRGHHHTAETGVLWAVQ  50
 51 RSPPRSRDRRPVGDAEVTPRSRDRRPVGDAEVTTTQQRRASPQAVQRSPP 100
101 RSRDRRPHGQCKGHHHTAEIGVLWAMQRSPPRSRDRRPVGDAEVPRSRDR 150
151 RPVGDAEVTTTQQRRASPQAVQRSPPHSRDRRPHGQCRGHHHAAETGVPT 200
201 GSAEVTTTQQRQASCGQCRGHHHAAETGVLWAVQRSPPRSRDGRPHRQCR 250
251 GHHHTAETGVLWAVQRSPPRSRDRRPVGSAEVTTLSQTVLGLPWVEHRMK 300
301 TMCFEPWKDNQPSQSQSLKQAAPRAQLQEAASSRGLQQGGGRCVPSVASW 350
351 CPLMSRSVR                                          400
```

Amino acid Composition

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
8.91	2.79	3.62	4.18	0.28	8.08	5.85	0.28	1.39	3.06	1.39	0.28	9.19	9.19	15.04	8.91	7.80	7.80	1.95	0.00

p < 0.01	p < 0.02	p < 0.05	p < 0.1	p >= 0.1	p < 0.1	p < 0.05	p < 0.02	p < 0.01

Significance p value is calculated using the Z score of the percent amino acid composition

Gene Ontology Predictions

Raw Format

GO term	Description	Type	Score	Jury
GO:0030528	transcription regulator activity	function	0.476	5
GO:0003676	nucleic acid binding	function	0.577	5
GO:0003677	DNA binding	function	0.584	5
GO:0003700	transcription factor activity	function	0.457	5
GO:0045449	regulation of transctription	process	0.785	5
GO:0050791	regulation of physiological process	process	0.693	5
GO:0050794	regulation of cellular process	process	0.771	5
GO:0006139	nucleobase nucleoside nucleotide and nucleic acid metabolism	process	0.728	5
GO:0006350	transcription	process	0.799	5
GO:0006351	transcription, DNA dependent	process	0.757	5
GO:0006355	regulation of transcription DNA dependent	process	0.689	5
GO:0006366	transcription from RNA polymerase II promoter	process	0.322	4
GO:0008152	metabolic process	process	0.815	5

Hierarchy View

GO:0003674 Molecular Function 0.584		Max Score	Max Jury
+ GO:0030528 transcription regulator activity	P	0.476	5
++ GO:0003700 transcription factor activity	P	0.457	5
+ GO:0005488 binding	I	0.584	5
++ GO:0003676 nucleic acid binding	P	0.577	5
+++ GO:0003677 DNA binding	P	0.584	5
++++ GO:0003700 transcription factor activity	P	0.457	5

I denotes an inherited term   P denotes a predicted term   ▨ Indicates the term is confirmed in the current GOA annotation

Figure 4.8: Example server prediction output for IPI00745501. The graphic details feature annotations distributed along the sequence length, for example, secondary structure, phosphorylation and glycosylation residues, PEST and disordered regions. These are highlighted in the sequence map. Amino acid compositional bias is also reported using statistical tests to determine over or under-representation.

Listing 4.1: Lactate Dehydrogenase (LDH) Prediction results

```
 ------------------------------------------------------------------------------
                          JOB ID 5si5sdtgoe73b596
                     Submitted 16-27-3:6-November-2007
 ------------------------------------------------------------------------------

 ---------------------------- GO TERM RESULTS ---------------------------------
 GO term        Description                                      Jury   Score
 ------------------------------------------------------------------------------
 GO:0016491     oxidoreductase activity                           5     0.266
 GO:0008152     metabolic process                                 4     0.915
 GO:0003824     catalytic activity                                4     0.633
 GO:0009058     biosynthetic process                              4     0.627
 GO:0044249     cellular biosynthetic process                     4     0.513
 GO:0005975     carbohydrate metabolic process                    4     0.350
 GO:0006807     nitrogen compound metabolic process               4     0.264
 GO:0006091     generation of precursor metabolites and energy    4     0.148
 GO:0009308     amine metabolic process                           4     0.203
 GO:0019538     protein metabolic process                         3     0.398
 GO:0044267     cellular protein metabolic process                3     0.369
 GO:0019752     carboxylic acid metabolic process                 3     0.236
 GO:0006412     translation                                       3     0.236
 GO:0006519     amino acid and derivative metabolic process       3     0.114
 GO:0006520     amino acid metabolic process                      3     0.021
------------------------------------------------------------------------------
```

4.5 Chapter Summary

Homology based methods for function prediction represent high specificity, low sensitivity methods that can be used to annotate a set of proteins. In contrast, feature based methods such as FFPred are comparatively low in specificity, yet obtain higher coverage for broad function categories. This makes them well suited to drug target prioritisation where a set of candidate functional roles can be suggested for novel proteins. By incorporating more feature characteristics and expanding the training sets to include sequences from other closely related species, the approach might become accurate enough to be incorporated into a function assignment pipeline.

The use of a machine learning approach and creation of function category specific classifiers for this approach was computationally intensive. Initially 5 classifiers for each of 752 and 859 Molecular Function and Biological Process categories were created. Subsequently a smaller set of models that performed better than random were optimised using grid parameter searches equating to 144 training runs per classifier. This procedure amounted to more than 220,000 training runs. In total this took approximately 3 months of cpu time on a compute cluster with 200 2GHz Dual Xeon processors. The results show that these classifiers significantly outperformed those used in the ProtFun method, and covered a much larger set of GO categories. However, maintenance of an up-to-date set of classifiers for GO term prediction is costly in terms of CPU consumption and is labour intensive.

The FFPred approach used only human sequences for training and testing which restricted the applicability of the method to broad function categories for which there were sufficient example sequences. This reduced the size of the training datasets and tuned performance towards function category recognition from human sequences. However, the method could not be applied to sequences from lower eukaryotes. Considering that some functions were better conserved within than between species, it seems likely that addition of sequences from other species might be appropriate for some functions on a case by case basis.

Feature based methods extend homology based methods by allowing the identification of function to relate to conservation of biological characteristics rather than conservation of sequence. In doing so they lack resolution where these characteristics are conserved between similar sequences but function has diverged. This was evident in the annotation results for lactate dehydrogenase where features describing guanylate dehydrogenase and dehydrogenase were similar.

The result was that the LDH sequence received two incorrect function assignments from the FF-Pred server, yet was able to recover an additional annotation. In cases of convergent evolution, where common function is observed but sequence, and sequence derived features are not conserved, extra information from expression or protein interaction characteristics are required to recognise function.

Chapter 5

Designing pairwise features for function prediction

5.1 Introduction and aims

Much of the work in predicting function from sequence information uses a single data source to classify equivalent functions. The most common data source comprises amino acid sequence information represented by pairwise sequence alignments. Feature based methods also use sequence information. The FFPred method in the previous Chapter employed feature sets derived from sequence to classify function. However, the determinants of function are not always captured in sequence information. In some cases, the behaviour of a sequence is governed by the type of cells and tissues it is expressed in, or the cellular compartment that it occupies, or a combination of these characteristics (Eisenberg et al. 2000, Joshi et al. 2004, Ofran et al. 2005, Rost et al. 2003). Several high throughput data sources are available that convey this information, for example microarray expression or protein interaction information. Previous studies have shown that function prediction methods combining information from multiple data sources outperform those that use single sources (Karaoz et al. 2004, Lanckriet et al. 2004a,b, Noble and Ben-Hur 2008).

The ideal function prediction method should be applicable to any sequence, and be capable of annotating highly specific function classes regardless of the homology status of a sequence. Sequence feature based methods that build models of function for each annotation category are applicable to all sequences. The sets of classifiers for individual annotation categories enable tight control of performance since the balance between coverage and error rates can be fine tuned. However, a trade-off is that they are restricted to more general annotations for which there are many available sequence examples. In contrast, methods that use neighbouring relationships (homology-based annotation transfer) are capable of making highly specific annotation assignments because a single example is sufficient to identify all members of a function class providing the relationships can be detected.

Desirable properties for a function prediction method are evident in both neighbourhood and category specific model approaches. The design and implementation of such a method forms the basis of the following chapter. To exploit the value of the FFPred features in specific function category recognition, the sequence feature characteristics have been transformed into measures between pairs of sequences. These can be combined with sequence similarity measures to produce a method that in theory should outperform either of the individual approaches. For further improvements, the use of feature information from diverse and independent data sources is investigated.

Previous function prediction approaches ultimately assign function to sequences in a binary manner. Either a sequence has a function or does not. A lack of completed annotations, and different specificities of annotation categories mean that closely related functions are discounted, or are penalised as false positive assignments. Instead of modelling function in this limited binary capacity, the problem can be posed as a regression between feature characteristics and function similarity. Conceptually, this approach should outperform methods performing binary classification since the degree of function specificity is accounted for in the model. Practically this means that any sequence can be assigned a nearest functional neighbour at any degree of function specificity.

The first part of the chapter explores different semantic similarity measures as suitable candidates for function similarity. Each subsequent section introduces a different data source and its relationship with function. The datasets comprise sequence similarity measures, protein-protein interactions, topology strings, microarray expressions, localisation, domain content and domain fusions. The generation of pairwise features representing each data source is described. The resulting feature matrix is then characterised and the overall importance of the different feature sets estimated by performing correlation analyses.

5.2 Defining a Function Similarity Measure

The concept and common measures of functional similarity are defined in Chapter 1. Four of the most popular methods are evaluated, Resnik (Resnik 1995), Lin (Lin 1998), SimRel (Schlicker et al. 2006) and GFSST (Zhang et al. 2006). The methods differ subtly according to scale and resolution. The first two methods were developed originally as part of the WordNet project (Sigman and Cecchi 2002), however can be applied to any Ontology. The SimRel method defined as

$$-ln(pca) \times \frac{ln(pca)}{ln(A) + ln(B)} \tag{5.1}$$

combines different aspects of both Resnik and Lin measures. GFSST is similar to Resnik but uses a different weighting scheme to score annotation terms. Two aspects of these measures are important, the scale occupied by each similarity or difference measure and the robustness of the measures when annotations are updated.

5.2.1 Selection of a semantic similarity measure

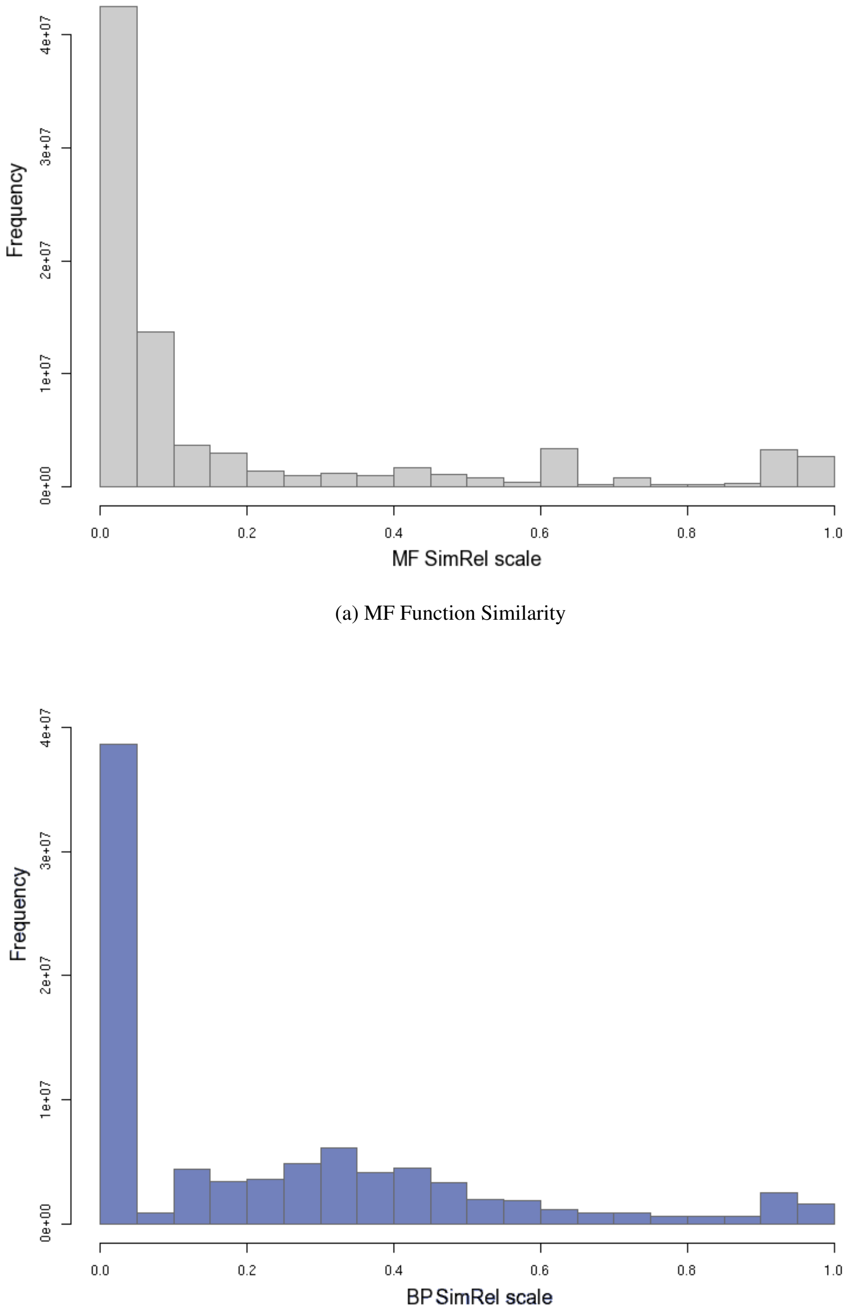
Before selecting a semantic similarity measure to score function similarity, the properties of each distribution were considered. Ideally, a function similarity measure should possess a defined minimum and maximum, and the intervening values should have meaning. In the literature, the Resnik method has frequently been adopted as the measure of choice for computing function similarity since it has the greatest correlation with sequence similarity measures (Lord et al. 2003). However, some circularity exists in this selection criteria as annotations sourced from computational sequence similarity are the primary determinant of most function annotations (see Chapter 3 Section 2.5).

The SimRel measure (Figure 5.1) was selected as the semantic measure used to calculate function similarity. This measure had the advantage of a defined minimum 0 (indicating the common parent is the root node of the graph) and a maximum value at 1 (indicating the annotations are identical and are leaf terms). The greater the similarity value, the closer the common parent to the compared terms. A comparison between the different measures of semantic similarity showed that they were highly correlated (Table 5.1). Thus the consequences of selecting one measure over another would not dramatically impact the regression modelling approach.

Table 5.1: Comparison between semantic similarity measures

Method	**Resnik**	**Lin**	**GFSST**	**SimRel**
Resnik	-	0.961	0.632	0.979
Lin	0.969	-	0.618	0.998
GFSST	0.849	0.824	-	0.634
SimRel	0.977	0.998	0.832	-

The upper triangle values represent correlation between semantic similarities obtained from the MF Ontology whilst lower triangle values represent correlations between semantic similarities obtained from the BP Ontology. All function similarity values that were 0 were omitted from the comparisons since they were common to all methods.



(a) MF Function Similarity

(b) BP Function Similarity

Figure 5.1: SimRel semantic similarity distribution

5.2.2 Function similarity measures

Function similarity can be calculated between sequence pairs using semantic similarities to score pairs of annotations. There are several strategies that can be used to determine function similarity from semantic since most sequences are annotated with more than one GO term. For any pair of sequences, a matrix of semantic similarities can be constructed between GO term pairs (see Table 5.2). The GO term pairs can be combined into a final function similarity score that is either local or global, and uses the maximum or average semantic similarity between GO terms.

In the local maximum method, the function similarity score is the maximum between all semantic similarity pair scores. Practically, this means that a case where multiple annotations are identical between sequences would obtain the same score as a case where a single annotation was shared between two sequences. The global average score was the first function similarity measure to be used (Lord et al. 2003) and combines information from all pair scores into an average value. Function similarity measures can also be asymmetric (Pesquita et al. 2008). Here, directional bias is introduced by considering pairwise scores for every annotation to one of the sequences (sequence A only), or a single pairwise score for every annotation for sequence B. The score differs from the global average only when different numbers of annotations exist for each sequence.

An asymmetric maximum averaged score was computed using the maximum pair score between a GO term from sequence A and any of the GO terms from sequence B for all GO annotations belonging to sequence A as in Table 5.2. This strategy removed redundancy for multiple similar annotations that might exist for a sequence whilst trying to ensure that the most appropriate annotations were compared. The directionality of the score also accounted for the lack of complete annotations by making each score conditional on the annotation status of the query sequence A. The highest scores for a query sequence could only be obtained when all GO term pairs were matched.

5.3 Feature design for heterogeneous data

Appropriate feature design and encoding methods can produce significant improvements in the performance of machine learning methods beyond those that can be achieved by parameter optimisation. Consequently, the following sections focus on appropriate translation of biological

Table 5.2: Example function similarity calculation between two sequences A and B.

Sequence B/Sequence A	**GO:0000166**	**GO:00016301**	**GO:0016740**	**Ave.**	**Max.**
GO:0003700	0.562	0.000	0.000	0.187	0.562
GO:0000166	1.000	0.379	0.000	0.460	1.000
Ave.	0.781	0.189	0.000		
Max.	1.000	0.379	0.000		

An example of function similarity scoring between two sequences A and B that are annotated with different numbers of GO terms. Using the maximum semantic similarity value, these sequences would be identical (pair score 1). The average semantic similarity between pairs is 0.324. The asymmetric average values are 0.323 when sequence A is the query sequence and 0.394 when sequence B is the query sequence compared to the asymmetric maximum scores of 0.460 and 0.781.

attributes into meaningful numeric feature vectors. To estimate the contribution of a particular feature in predicting function similarity, correlation analysis was performed between individual descriptors and function similarity. The degree of feature redundancy was determined by inter-descriptor correlation analysis to avoid unnecessary calculation of attributes that would contribute little additional value to the final method.

5.3.1 Sequence Similarity

Sequence similarity searches were performed using the SSEARCH algorithm which is part of the FASTA suite of software tools (Lipman and Pearson 1985). The SSEARCH algorithm is an implementation of the Smith Waterman local alignment algorithm (Smith and Waterman 1981). This method produces more accurate pairwise alignments and scores than BLAST, at the compromise of speed. However, the luxury of the use of a modern computer cluster with more than 200 CPu nodes made these computations feasible in just a few hours.

Each sequence was aligned to all other sequences and alignment statistics retained as features (Table 5.3). Pearson's correlation between each attribute and function similarity (SimRel method) was calculated to provide an estimate of how useful each feature might be to infer function. Logistic transforms given by the function

$$\frac{1}{1+exp(^{-t})}\text{where t is of the form } ax+b. \tag{5.2}$$

were used to scale bit score and length features between 0 and 1. Parameters a and b for each transform were determined by optimising a linear regression to the target distribution of function similarities.

Correlations between MF similarity and feature similarity (Table 5.3) were much greater than BP similarity. This is in agreement with the finding reported in Chapter 3, that sequence similarity was a better indicator of shared MF than BP. The correlation between normalised bit score and sequence identity was 0.561, indicating that each measure encoded some mutually exclusive aspects of sequence relationships that might be useful in function prediction. Similarly, query and target coverage features were correlated at 0.520 suggesting that whilst these attributes contained similar information, they also provided novel information for modelling function similarity.

Table 5.3: Sequence similarity features.

Index	**Parameter**	**Transform**	**Cor MF**	**Cor BP**
1	Identity	$\frac{x}{100}$	0.20	-0.00
2	Bit score	$\frac{1}{1+exp-0.01(x-50)}$	0.30	0.00
3	Query coverage	$\frac{x}{100}$	0.24	-0.00
4	Hit coverage	$\frac{x}{100}$	0.21	0.01
5	Length	$\frac{1}{1+exp-0.01(x-100)}$	-0.09	0.01
6	Hit Length	$\frac{1}{1+exp-0.01(x-100)}$	-0.14	0.01

Correlation scores represent individual correlations between transformed feature values and SimRel semantic similarity.

5.3.2 Protein-protein Interactions

Protein-protein interaction information is independent of homology information and can assist in the annotation of non-homologous proteins. These data can potentially be useful for recognising binding functions, and can define relationships that may indicate common BP. The value of the annotations has recently been realised using experimentally derived protein interactions (GOA evidence code IPI) to assign MFs to sequences (See Chapter 2, Section 3). Protein interaction information was sourced from the IntAct database (Hermjakob et al. 2004). Evidence for interactions is compiled from high and low throughput experiments and literature information (see Table 5.4). 24,712 unique interactions were available for the human proteome with 4026 and 6855 sequences possessing either MF or BP annotations.

Initially, a simple feature encoding strategy was employed where each interaction was encoded as a binary evidence vector. The values of each vector denoted presence or absence of experimental interaction information. Several interaction types were sparsely represented within human sequences and were merged into a single evidence category 'other'. A second weighted encoding strategy was investigated to account for important aspects of data quality. For example, some proteins might be more likely to attract other proteins under experimental conditions due to amino acid compositional bias or stickiness (Ispolatov et al. 2005). This suggests that not all reported interactions from the same experiment are equally reliable. To control for this effect, each interaction score was normalised proportionally to the number of interactions observed for each partner sequence. The normalisation is given by

$$Score_{A,B} = I \times \left(\frac{\log \frac{f(A)}{|N|} + \log \frac{f(B)}{|N|}}{2} \right) \tag{5.3}$$

where A and B are two partner sequences and $\frac{f(A)}{|N|}$ and $\frac{f(B)}{|N|}$ are the proportions of total interactions that each sequence participates in. I represents the raw interaction score.

The value of experimentally derived protein interaction feature information in function prediction was determined by correlation coefficients measured between each feature and function similarity (Table 5.4). The measures were calculated for the normalised weighted PPI score and the simple binary score. The feature correlation values were greater between BP similarity and interaction score than MF similarity and interaction score suggesting that overall, experimentally derived interaction information was more valuable for prediction of BPs than MFs.

Overall, correlations between protein interactions and function similarity were low (< 0.1). Co-immunoprecipitation, pull down, protein array and X-ray structure derived interaction features were more highly correlated with function similarity whereas yeast two hybrid, peptide array, imaging and molecular sieving features exhibited marginal or no correlation with function similarity. These findings might relate to aspects of data quality since yeast two hybrid assays are typically noisy and often report associations that cannot be confirmed *in vivo* (Deng et al. 2003). Low correlations also result from cases where the occurrence of the feature in the dataset was rare. However, this did not affect the resulting values from the yeast two hybrid data or peptide array data. All attributes were retained as features despite their low correlation scores as it was expected that when combined together, more weight would be given to those interactions reported in multiple methods thus correcting for low quality annotations. On average, the weighted PPI feature scores were more highly correlated with function similarity. Consequently, this feature representation was adopted as the best representation of feature attributes for the data.

5.3.3 Topology

Topology features were used to represent aspects of sequence that were spatially distributed; secondary structure, transmembrane and disordered regions. These aspects of sequences are 2 dimensional representations of 3 dimensional information, however high confidence structure information is only available for around 3000 human sequences. Consequently, the topology strings present a way of comparing sequences using predictions of structural features. To generate meaningful measures to describe topological similarity between sequences, each sequence was converted to a restricted alphabet, D and X for disordered regions, T and X for transmembrane regions and H, E and C to represent secondary structure. These topology strings were then locally aligned using the Smith-Waterman algorithm from the Align package (Tosatto et al. 2006) with customised substitution matrices (see below).

Table 5.4: Correlation of protein interaction features with function similarity.

Index	Feature	Description	Cor MF	wCor MF	Cor BP	wCor BP
1	affinity chromatography	affinity resins used as tools to purify molecules of interest and their binding partners	0.08	0.07	0.08	0.07
2	anti bait co-immunoprecipitation	(chromatography) specific antibody for the protein of interest is used to generate resin to capture bait	-0.05	-0.04	-0.04	-0.02
3	anti tag co-immunoprecipitation	(chromatography) bait protein is expressed as hybrid protein fused to a tag or peptide for which specific antibodies can be raised	0.08	0.09	0.16	0.16
4	two hybrid	system using transcriptional activity to measure protein interactions. Binding proteins activate transcription by mediating non-covalent interactions. Both proteins must be expressed in the host cell	0.07	0.07	0.05	0.05
5	co-immunoprecipitation	antibody or tag is used to separate bait from prey and simultaneously capture its ligand	0.07	0.07	0.11	0.11
6	far western blotting	a mixture of protein is submitted to electrophoresis then transferred to a membrane. The membrane is incubated with a primary antibody specific for a given protein. A second labelled antibody targets the first and allows visualisation of the protein band	-0.00	-0.00	0.05	0.05
7	molecular sieving	gel filtration method according to molecular weight. Two proteins can be said to interact if they elute in a column fraction with a molecular weight larger than either single molecule.	0.06	0.07	0.07	0.07
8	peptide array	screening of large arrays of synthetic peptides on planar cellulose supports.	-0.07	-0.06	-0.02	-0.02
9	protein array	simultaneous screening of biochemical activities of proteins by washing protein samples over known proteins printed on the array	0.09	0.09	0.15	0.16
10	pull down	affinity chromatography method where the protein of interest is tagged and fixed to resin. Purified proteins are adsorbed to the resin and retained binding proteins identified	0.05	0.07	0.09	0.10
11	surface plasmon resonance	gold coated microarray where small molecule, peptide, protein, sugar, lipid and fragments can be spotted. The ordered protein array is then probed by unlabelled small molecule, peptide, sugar and lipid molecules in real time to establish on and off rates of complex affinities.	0.03	0.03	0.04	0.03
12	x-ray diffraction	complexes captured by crystal structure.	0.04	0.04	0.10	0.10
13	ion exchange chromatography	chromatography method where component proteins are separated by pH and media appropriate to the charge density of the complex.	0.01	0.02	0.01	0.01
14	inferred by curator	evidence based on a curator assumption, either when complete experimental support is not available or when the results are extended by homology to closely related orthologous sequences.	0.12	0.12	0.09	0.09
15	two hybrid pooling	sets of bait and prey hybrid vectors are randomly mated and the positive double hybrid clones sequenced to identify interacting partners.	0.01	0.01	0.00	0.01
16	two hybrid fragment pooling	screening of large numbers of individual proteins against a comprehensive library of randomly generated fragments. Fragments comprise multiple overlapping clones so that the minimal region for binding can be identified.	-0.04	-0.04	-0.13	-0.12
17	electric mobility supershift	gel electrophoresis assay (PAGE or 2D gel) to identify and separate protein molecules according to size.	-0.00	-0.00	0.05	0.05
18	fluorescence imaging	fluorescence microscopy using high intensity light to illuminate the samples. Fluorescent species are then emit light at longer wavelengths.	-0.01	-0.02	0.05	0.05
19	footprinting	interacting proteins are deduced by patterns of degradation or modification that are protected by complex formation.	0.02	0.02	0.06	0.06
20	imaging techniques	microscopy techniques that permit the imaging of molecules at various resolutions depending on the technology used.	-0.00	-0.00	0.04	0.04
21	one hybrid	protein-DNA complementation assay where single promoter acts as bait and is screened against a library of transcription factors.	0.02	0.02	0.04	0.04
22	tandem affinity purification	rapid purification under native complex conditions. TAP method requires fusion of a multiple tag N or C terminally to the target protein of interest. The method is highly selective.	0.03	0.03	0.01	0.01
23	antibody array	microarray consisting of antibodies spotted onto a solid support and incubated with biological sample. Proteins are identified by prior labelling.	0.00	0.00	-0.04	-0.04
24	others	including NMR, circular dichroism, co-sedimentation, cross-linking, fret and facs, kinase and enzymatic assays, lambda phage display and confocal microscopy	0.07	0.08	0.12	0.13

Topology scoring matrices

$$Disorder = \left[\begin{array}{lcc} & \mathbf{D} & \mathbf{X} \\ \mathbf{D} & +3 & 0 \\ \mathbf{X} & 0 & +2 \end{array}\right]$$

$$Transmembrane = \left[\begin{array}{lcc} & \mathbf{T} & \mathbf{X} \\ \mathbf{T} & +3 & 0 \\ \mathbf{X} & 0 & +2 \end{array}\right]$$

$$SSEA = \left[\begin{array}{lccc} & \mathbf{H} & \mathbf{E} & \mathbf{C} \\ \mathbf{H} & +2 & 0 & +1 \\ \mathbf{E} & 0 & +2 & +1 \\ \mathbf{C} & +1 & +1 & +2 \end{array}\right]$$

The secondary structure scoring matrix was derived from the original matrix used in the SSEA algorithm implementation (Fontana et al. 2005). For disordered and transmembrane residues +3 was used to score correctly aligned D residues whilst +2 was used to score pairs of other residues. This resulted in longer alignments between sequences where the highest total alignment score represented topological equivalence (see Listing 5.1 for example alignments). The alignment algorithm computes topological similarity by reducing the topology strings for query and target sequences to block representations. For example, a sequence of three transmembrane regions would be represented as CTCTCT where T denotes a transmembrane region of varying length. Each alignment is computed between block level representations of sequences and scored by mapping the regions back to the full residue alignment. Matches for T-T are scored by the length of the shorter region and mismatches do not contribute to the score. The final normalised score ranges between 0 and 100.

Each alignment produced a normalized score, pairs of from and to regions representing alignment boundaries and coverage statistics which were then scaled to produce feature descriptors ranging between 0 and 1 using different transforms (Table 5.5). Additionally a symmetric measure of similarity between the number of predicted transmembrane regions was included. The score was designed to group sequences with equivalent predicted numbers of transmembrane regions to compliment the alignment scores.

Disordered residues, transmembrane topology and transmembrane coverage features were most highly correlated with function similarity. The number of transmembrane regions was most

Listing 5.1: Example alignments between topology strings

```
QUERY
CCCCCCCCCCTTTTTTTTTTTTTTTTTTTTTTTTTCCTTTTTTTTTTTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTTTTTTTTTTTTTTTTTTTTTTTTTCCCCCCCCCCCCTTTTTTTTTTTTTTT
TTTTTTTTTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCC
SUBJECT
CCCCCTTTTTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
Normalized score: 55.3012
Block alignment: 265-496, 1-334
CTC
CTC
Residue alignment:
CCCCCCCCCCCCTTTTTTTTTTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC--------------------------------------------------------------
----------------------------------------------------

CCCCC-------TTTTTTTTTTTTTTTTTTT-----CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x

QUERY
CEEEEEECCCCCCCCCCCCCCCHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEECCCCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCHHHHHCCCCCCCCCCCCCHHHHHHHHHHHC
CCCCCEEEEECCCCCCCCEEEEEEEECCCCCCCCCCCCCHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCHHHHHHHH
HHHCCCCCCEEEEEECCCCCCCCEEEEEEEECCEEEEEEECCHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCHH
HHHHHHHHHCCCEEEEEECCCCCCCCEEEEEEECCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCC
SUBJECT
CCCCCCCCHHHHHHHHHHHHHHHHHCCHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCC
CCCCCCCCCCCCCCCCCCCCCCCEEEEEEEECCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEECCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEECCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEE
EECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEECCCCCCCCCHHHHHHHHHHHHHCCCCCCHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCEEEEECCHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
HHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHH
HHHHHHHHHHHHHHHCCCCCCCCC
Normalized score: 51.5783
Block alignment:
EC-HCHCECHCHCHCECECH--CHCHCECECE-CH----CHCHCECECHC
CHCHCHCECHCECHCHCECHCHCHCHCECE-CHCHCECHCHCHCHCHCHC
Experimental residue alignment: 2-577, 1-1073

EEEEEE--CCCCCCCCCCCCCCC-----HHHHHHHHH--CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC--HHHHHHHH----CCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEECCCCCCCCCCCC--------HHHHHHHHHHHHHHHHHCCCC----------
--------HHHHH----CCCCCCCCCCCCC-----------------------------HHHHHHHHHHHCCCCCC-EEEEE--------CCCCCCCC-
---------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------EEEEEEE
ECCCCCCCCCCCCC-----------------HHHHHHHHHHHHHHHH-----------------------CCCCC------------------------
-----------------------------------------------HHHHHHHHHHHCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCC-----
--------------------EEEEEE-----CCCCCCCC-----------------------------------------------------EEEEEEE
ECCEEEEEEE---------------CC----HHHHHHHHHHHHHHHHHHH-----------------------------------CCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCC
CCCCCCCCCCCCCCCCCCCC---------------HHHHHHHHHHHCCCEEEEEE----CCCCCCCC--------------------------------
---------EEEEEEECCCCCCCCCCCCCC------------HHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCC----------

CCCCCCCCHHHHHHHHHHHHHHHHHCCHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCC
CCCCCCCCCCCCCCCCCCCCCC-----------------EEEEEEEE-CCCCCCCCCCCCCCCCCCCCHHHHHHHHHH-------CCCCCCCCCCCCCC
CCCCCCCCEEEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHH-CCCCCCCHHHHHHHHHHHHHCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEE
-CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHH------CCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHH---CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCEEEEEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEE--
---CCCCCCCCCHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCEEEEECCHHHHHHHHHCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC-------------------------------------------------HHHHH-----CCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHH--CC-HHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCHHHHH--CCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHH--------CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-
```

Table 5.5: Topology feature listing

Index	Feature	Transform	Cor MF	Cor BP
Disorder				
1	Align score	$\frac{x}{100}$	0.11	0.02
3	Query coverage	$\frac{f(Q)}{length}$	0.14	0.03
4	Hit coverage	$\frac{f(H)}{length}$	0.14	0.02
5	Query disorder coverage	$\frac{f(D_q)}{D_q}$	0.07	0.12
6	Hit disorder coverage	$\frac{f(D_h)}{D_h}$	0.07	0.13
Transmembrane				
1	Align score	$\frac{score}{100}$	0.12	0.08
2	Query coverage	$\frac{f(Q)}{length}$	0.02	0.05
3	Hit coverage	$\frac{f(H)}{length}$	0.04	0.04
4	Transmembrane coverage	$\frac{f(T_q)}{T_q}$	0.16	0.05
5	Coil coverage	$\frac{f(C_h)}{C_h}$	0.13	0.06
6	Transmembrane region score	$\frac{min(R_q, R_h)}{max(R_q, R_h)}$	-0.01	0.14
Secondary Structure				
1	Align score	$\frac{score}{100}$	0.03	0.00
2	Query coverage	$\frac{f(Q)}{length}$	0.00	-0.02
3	Hit coverage	$\frac{f(H)}{length}$	0.01	-0.01
4	Query secondary structure coverage	$\frac{f(SS_q)}{SS_q}$	0.03	-0.06
5	Hit secondary structure coverage	$\frac{f(SS_h)}{SS_h}$	0.03	-0.06
6	Query helix coverage	$\frac{f(helix_q)}{helix_q}$	-0.00	-0.05
7	Query sheet coverage	$\frac{f(sheet_q)}{sheet_q}$	-0.00	-0.03
8	Hit helix coverage	$\frac{f(helix_h)}{helix_h}$	-0.00	-0.05
9	Hit sheet coverage	$\frac{f(sheet_h)}{sheet_h}$	0.01	-0.04

In the Transform column, f denotes a residue frequency value, Q and q represent the query sequence and H and h represent the target or hit sequence from an aligned pair of sequences. SS refers to secondary structure, either a helix or strand residue, and R refers to an entire transmembrane region and T, D and C are transmembrane, disordered and coiled coil residues respectively.

highly correlated with BP similarity. The majority of other features were better indicators of MF than BP suggesting that overall these features might be more useful in MF category recognition.

5.3.4 Cellular Localisation

Subcellular localisations describe the compartmentalisation of a sequence in which its function is performed. In the FFPred method, this information was shown to be of value in function prediction, particularly for BP categories (see Chapter 4 section 3.4). Another recent study demonstrated the value of cellular localisation information in combination with sequence similarity for improving the detection of remote homologues (Shah et al. 2007).

Localisation information was predicted from amino acid sequence using PSORTII (Nakai and Horton 1999), SignalP (Brameier et al. 2007), NucPred (Bendtsen et al. 2004) and MitoPred (Guda et al. 2004) algorithms. The raw motif scores from PSORTII were used to derive feature inputs rather than the final nearest neighbour algorithm probabilities. This was because the PSORT method assumes that each sequence has a single subcellular localisation which is reflected in the algorithm output. This assumption is often violated for proteins that shuttle between nucleus and cytoplasm, or can be found both intracellularly and extracellularly. A good example of this is the nuclear hormone receptors which bind DNA yet cycle between the nucleus and cytoplasm (Krasowski et al. 2005). The use of the raw subcellular localisation information allows multiple amino acid motif signals for cell sorting to be represented as features better reflecting the biological properties of sequences.

SignalP, NucPred and MitoProt offer more recent and accurate localisation predictions than PSORTII. Each algorithm generates a probabilistic score for a single localisation. The scores were combined with the PSORTII features and converted into pairwise localisation similarity measures. The similarity measures for each localisation represented the product of the two probabilities for each sequence pair. The effect of this multiplication meant that sequences both attaining probability 1 of a particular localisation achieved maximum similarity score of 1. PSORT localisations that annotated < 20 sequences, CAAX motifs, bacterial DNA binding motifs and ER arginine motifs provided limited information for function recognition and were excluded from the pairwise feature sets.

Individually, the features were not well correlated with SimRel function similarity as defined in

Table 5.6: Localisation features

Index	Feature	Description	MF Cor	BP Cor
1	PSORT alm	transmembrane discriminant score	-0.09	0.05
2	PSORT bac	bacterial motifs for DNA binding proteins present within $<$ 1% eukaryotic sequences	0.00	-0.00
3	PSORT dna	proportion of matches to 63 DNA binding motifs from PROSITE	0.00	0.01
4	PSORT erl	KDEL/HDEL C terminal motif detection for Endoplasmic Reticulum	0.05	0.07
5	PSORT erm	score for presence of arginines in the first four residues of a signal peptides indicating ER membrane proteins	-	-
6	PSORT gpi	GPI anchor signals	-0.01	-0.02
7	PSORT gvh	signal sequence cleavage site predictions	0.01	-0.00
8	PSORT leu	di-leucine motif important for inclusion in clathrin coated vesicles and lysosomal targeting	0.03	-0.05
9	PSORT m1a	membrane protein with type 1a topology (1 transmembrane region with signal sequence)	-0.00	-0.00
10	PSORT m1b	membrane protein with type 1b topology	0.01	0.02
11	PSORT m2	membrane protein with type 2 topology	-0.01	-0.01
12	PSORT mNt	membrane protein with N tail topology (C terminal region and no signal peptide)	-0.00	-0.01
13	PSORT mip	mitochondrial targeting signal cleavage site	-0.00	0.01
14	PSORT mit	mitochondrial targeting signal at N terminus	-0.01	-0.03
15	PSORT myr	myristoylation/palmitoylation sites from PROSITE	-0.02	-0.05
16	PSORT nuc	discriminant score for nuclear proteins	0.00	0.00
17	PSORT pox	PTS 1 C terminal peroxisomal sorting signal	0.02	0.00
18	PSORT psg	signal peptide score	-0.00	0.07
19	PSORT px2	second weak signal associated with peroxisomal targeting	-0.00	-0.00
20	PSORT rib	ribosomal proteins based on 71 PROSITE regular expressions	0.01	-0.02
21	PSORT rnp	RNA binding motifs from PROSITE	0.00	0.01
22	PSORT tms	number of transmembrane segments	0.01	0.01
23	PSORT top	topology discrimination score	0.16	0.04
24	PSORT tyr	tyrosine motifs in cytoplasmic tail of membrane proteins for lysosomal targeting	-0.00	0.03
25	PSORT myr	N myristolated and palmitoylated proteins using regular expressions	0.00	-0.00
26	PSORT vac	vacuolar targeting signals	0.01	0.01
27	PSORT yqr	tyrosine based pattern for trans-Golgi localization signal	-0.01	-0.00
28	SignalP anchor	probability of signal anchor	0.06	-0.00
29	SignalP signal	probability of signal peptide	0.06	0.02
30	NucPred	probability of nuclear localisation	0.05	0.19
31	MitoProt	probability of mitochondrial localisation	-0.03	-0.07

Equation 5.1 and Table 5.2. However this was to be expected since shared localisation tend to be observed for a large proportion of sequences. For example, 2081 sequences were predicted to localise to the nucleus with probabilities greater than 0.95. Hence these features provide imprecise information for function prediction. Nuclear localisation, endoplasmic reticulum signals and signal peptides were among the most highly correlated with BP similarity, whilst transmembrane topology, signal peptides and signal anchor feature scores exhibited greater correlation with MF similarity.

5.3.5 Domain content and domain fusions

Historically, the structural domain has been defined as the primary unit of functional inheritance (Lee et al. 2007, Moult and Melamud 2000, Todd et al. 1999). In many cases, knowledge of the three dimensional structure of the protein sequence is sufficient to infer some aspect of its function. This is particularly true where structural genomics initiatives have been employed to determine function. However, the proportion of available genome sequences whose structures have been solved is small due to experimental difficulties in obtaining crystal structures (Burley 2000). Fortunately, the value of structural domain annotations can be realised by sequence based methods, since similar fold and function(s) are predominantly a feature of homologous sequences. Several biological knowledge bases contain domain information, for example PFAM (Sonnhammer et al. 1998) and CATH (Orengo et al. 1997). The value of this domain information for function annotation has been demonstrated by direct mappings between domains and function. More than 40% of current human annotations can be determined using domain annotations that coincide with function.

The CATH and SCOP databases are hierarchical classifications of structure arranged in a manner similar to the Enzyme database. CATH groups structures into common classes, architectures, topologies and families, whilst SCOP groups structures by common class, fold, superfamily and family. PFAM domain definitions comprise groupings of homologous sequence families often coinciding with CATH and SCOP annotations for sequences, but extend beyond structural definitions for sequences without any structural homologues. Annotations for PFAM, CATH and SCOP domains are the result of profile-sequence comparisons, where the profile represents an alignment of sequence regions that constitute a domain or family. The alignment profiles are either stored as PSI-BLAST profiles or as Hidden Markov Models (HMMs). Each sequence is typically scanned against a library of domain profiles using HMM search algorithms or the

PSI-BLAST algorithm.

Domain mapping information for sequences has been widely exploited in function prediction approaches beyond the use of direct mappings that are currently supplied for PFAM or CATH via the InterPRO2GO specification. Most methods for inferring function using domain information have employed techniques similar to phylogenetic profiling (Cokus et al. 2007, Marcotte et al. 1999, Melvin et al. 2007, Ramani and Marcotte 2003). Here protein pairs are scored according to a summary of their evolutionary history obtained using automated sequence similarity searches to identify orthologues across a set of reference genomes. Profiles can comprise presence or absence of a domain in the different species or record the number of domain relatives in the species. Sequences sharing consistent profiles frequently possess similar functionality (Ranea et al. 2007).

In this case phylogenetic profiling has not been used. This is partly due to taxonomic bias in fully sequenced genomes which means that some eukaryotic domains are only sparsely represented (Loganantharaj and Atwi 2007, Snitkin et al. 2006). One study of DNA repair proteins in human, yeast and *E. coli* (Galperin et al. 1998) demonstrated that no common domain architectures existed between kingdoms. The WD40 provides another example of a small mobile domain that is not found in bacteria (Doolittle 1995). A second problem arises from the existence of functionally promiscuous domains that have acquired diverse functional roles in different organisms. Evolutionarily, they arise through modification and substitution of parts of domains which act as independent functional units, such as the active sites of enzymes (Todd et al. 1999). The TIM barrels for example, act as a structural scaffold accommodating at least 15 different enzyme families (Nagano et al. 2002). Ferrodoxins are found fused to variety of enzymes of different functions and act as sensory modulators triggering signals in response to intracellular environment (Anantharaman et al. 2001). Phylogenetic profiles built from promiscuous domain homologues often result in an expanded repertoire of possible protein functions for a query rather than focusing in on one or more likely candidate functions. However, annotation errors resulting from automated phylogenetic profile analysis can frequently be corrected by inclusion of domain context information (Forslund and Sonnhammer 2008).

The goal of this approach is to measure similarity between human sequences based on their domain architecture (see Figure 5.2). By using domain information in the context of a single organism, the problems encountered using phylogenetic profiles should be avoided. The method scores entire domain architectures and is based on the method of Hayete and Bienkowska (2005). Domain fusion events are also used to predict and score the architectures of hypothetical complexes. The complexes are determined by linking sequences containing domains that are present as a complete product in other species. In Section 5.2, features were derived from experimental information describing complexes, however this data is only available for a small proportion of human sequences. Hence there is ample scope to represent potential interactions using high coverage low specificity information derived from *in silico* predictions of complexes.

Generating domain and complex architectures

Instead of using the occurrence of a single domains for the purpose of function prediction, each sequence was represented using its entire domain compliment, its multi-domain architecture. Complex architectures refer to the entire multi-domain content resulting from the joining of two sequences. Domain and complex architectures were generated using both CATH and PFAM domain definitions. The CATH database describes domains in terms of independent structural units that can occur alone as autonomously folding units, or as part of a larger structural ensemble (Orengo et al. 1997). PFAM entries represent homologous sequence families that can correspond with CATH domains, or comprise sequence families that have no structural representatives (Sonnhammer et al. 1998). Using the two definitions ensures that information is present for a greater coverage of sequences, and that complementarity between the definitions, the PFAM family or CATH structural domain can be exploited in the prediction approach.

Assignments of domains to sequences were computed from scratch rather than using publicly available mappings from the InterPRO database (Apweiler et al. 2000). This ensured that PFAM and CATH annotations were comprehensive for both the IPI human sequence dataset and the most up to date version of the UniRef database. Most publicly available annotation datasets use profile to sequence alignment methods to assign domains, however, it is well known that these methods can be improved using sensitive profile-profile comparison techniques (Reid et al. 2007, Soding 2005). Because annotation coverage of sequences was considered extremely important, a threading method (pDomTHREADER) based on the mGenTHREADER algorithm (Jones et al. 1999, McGuffin and Jones 2003) was developed for the purpose that required sequence profiles as input (Lobley et al. 2009). Here, the luxury of the Legion compute cluster was available to compute sequence profiles for more than 5 million known sequences in just over 60 days, making whole genome threading at this scale computationally feasible.

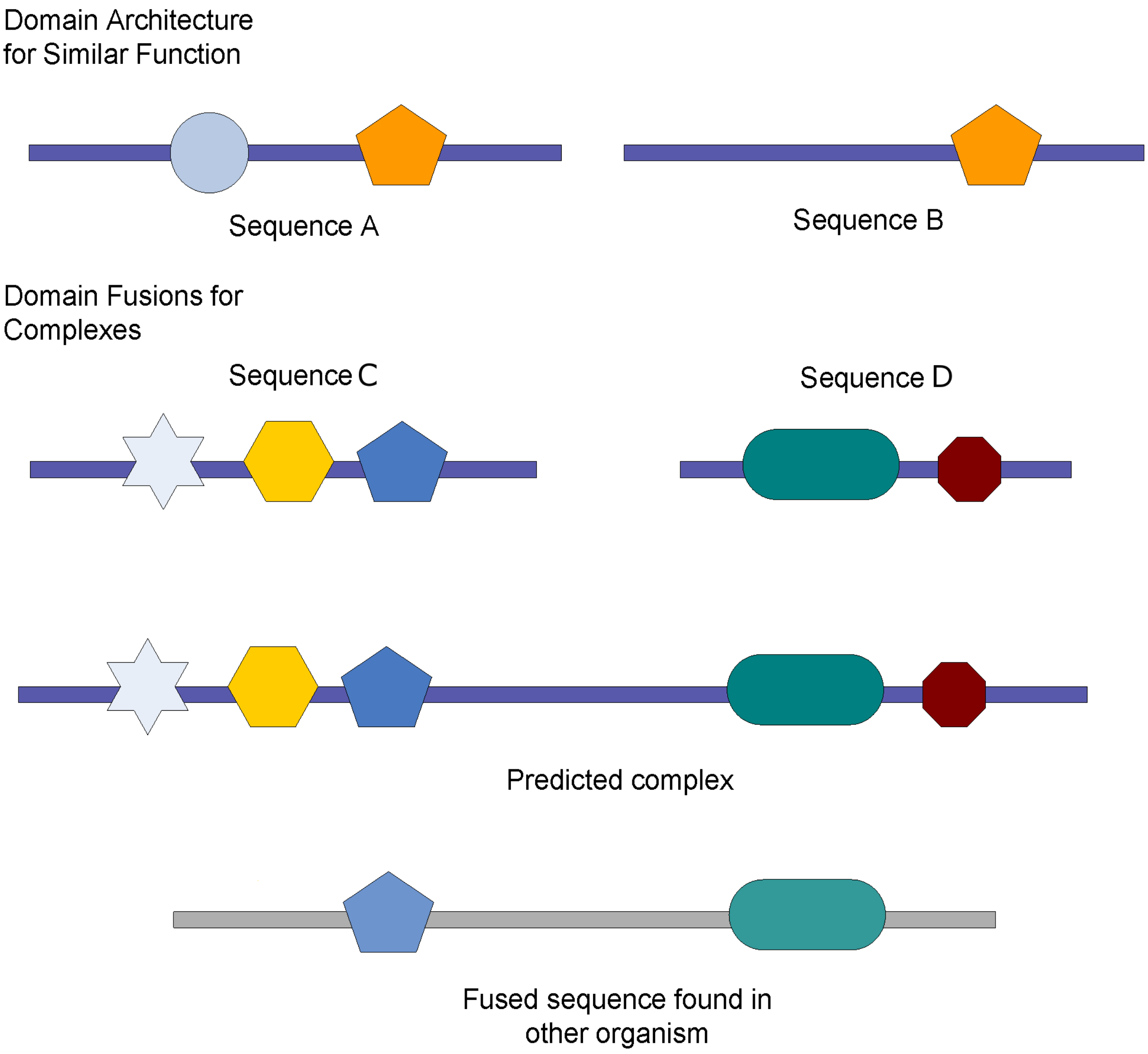


Figure 5.2: Domain architectures for function prediction. In the example above Sequence A and B share a common domain (orange) that is used to score architectural similarity. In the fusion example, Sequence C and D can be linked by the existence of a fused sequence containing a copy of the blue and turquoise domains. The joining of sequence C and D generates a predicted complex that may be used to infer similar functionality.

CATH superfamily identification

CATH annotations were made at the superfamily level. The threading template library was created using the procedure outlined in Figure 5.3. Multiple sequence alignments for each CATH superfamily were kindly supplied by the CATH group (Dr Ollie Redfern). The sequence alignments were then used to seed PSI-BLAST searches for each of the superfamily representatives. The Protein Structure Databank (PDB) sequences were searched first to identify any potential superfamily members that had not yet been classified in the CATH release. Subsequently profiles were constructed by searching a masked version of the UniRef90 database. Masking was carried out using the pfilt algorithm (Jones and Swindells 2002). Search iterations were terminated after at least 3000 sequence relatives had been identified. This procedure ensured that a sufficient diversity of sequences were present in each profile. The profiles were then supplemented with structural information from DSSP (Kabsch and Sander 1983).

CATH superfamilies were assigned to sequences using the pDomTHREADER algorithm. The approach used the same threading and profile alignment algorithm as mGenTHREADER (Jones et al. 1999) but adopted an alternative scoring method that was optimised for recognition of superfamilies rather than folds. A further difference was that mGenTHREADER template libraries were constructed from whole PDB chains whereas pDomTHREADER libraries relied on domain superfamily templates. This improved the accuracy of domain boundary recognition and improved the accuracy of profile templates since they were less likely to drift and accommodate sequences containing other domains.

pDomTHREADER scoring method

Several of the original mGenTHREADER inputs were not relevant for scoring using the superfamily based method. The differences in scoring methods between mGenTHREADER and pDomTHREADER algorithm are shown in Table 5.7. Most of the raw values were scaled using a logistic transform function (Equation 5.2). Target length and z-scores for pairwise energies were not included in the pDomTHREADER score since they offered no improvement to the accuracy of the score. In place, superfamily coverage was introduced. Additionally the energies and alignment scores were re-scaled to ensure maximum resolution when measured over short regions. The scaling parameters were determined by performing regression analysis with different combinations. Parameters were chosen that provided the best resolution between medium to

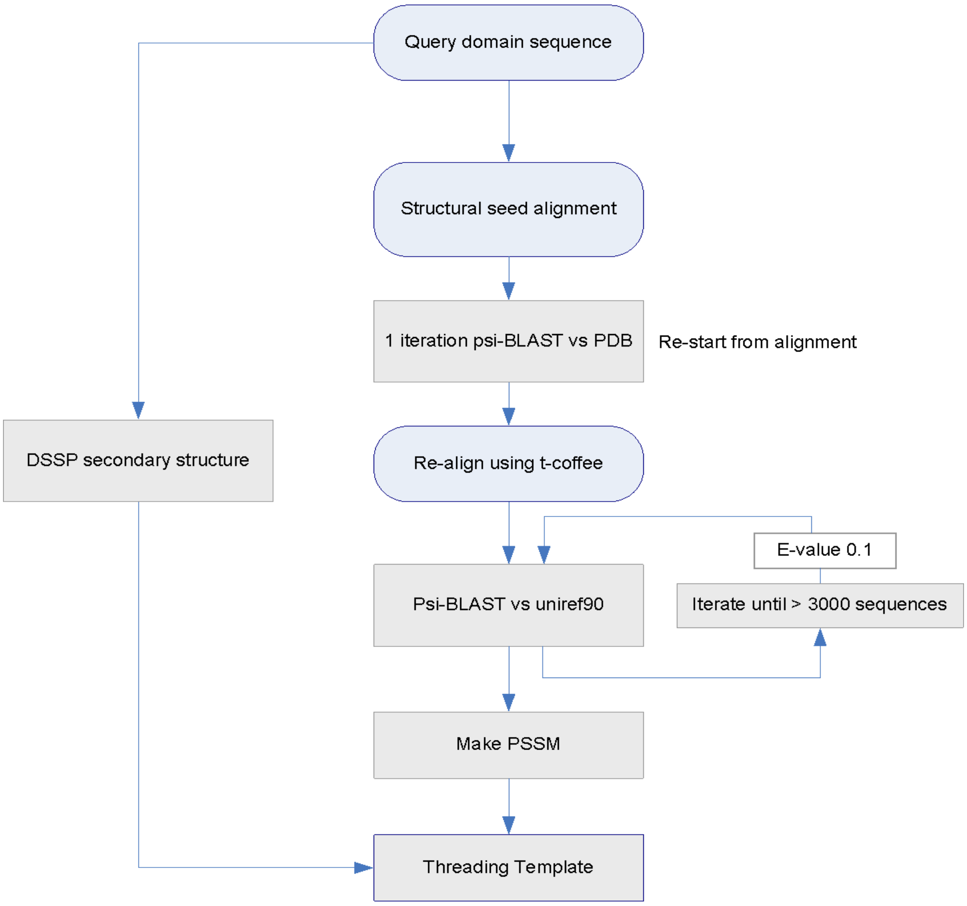


Figure 5.3: Flow chart of the steps involved in creating the structural template library.

high scoring values.

pDomTHREADER was trained in classification mode to provide a clearer distinction between separate homologous superfamilies that could be aligned with high scores. The classification target comprised pairs of CATH S35 representative sequences and 5-fold cross validation experiments were carried out to establish the best parameters for the linear SVM model. The bias parameter j was set to equal the ratio of number of training negative examples to number of training positive examples to simulate training on a balanced class dataset. The cost parameter C was selected by optimising the precision-recall break even point over coarse and subsequently fine grid searches ranging between 1e-3 and 1e+6. The final C parameter was 1.76.

Performance of pDomTHREADER algorithm

The performance of pDomTHREADER was compared to mGenTHREADER, HHPred (Soding 2005), PRC (Madera 2008) and PSI-BLAST (Altschul et al. 1997) for recognising superfamilies. A benchmark dataset comprising whole chains for CATH S35 representatives (Superfamily sequences filtered at 35% identity) was prepared. Third iteration sequence profiles were generated for the full chain sequence for each S35 representative. Each algorithm required subtly different input formats. HHPred required profile HMMs that were produced from PSI-BLAST profiles. PRC used binary checkpoint PSI-BLAST profiles, and the threading methods required matrix input files generated using the makemat programme (Altschul et al. 1997).

Profile-profile comparisons were performed for PSI-BLAST, HHPred and PRC against S35 superfamily delineated profiles. Sequence profiles were aligned to the S35 template library to generate pDomTHREADER results, and S35 whole chain templates were aligned to S35 whole chain sequences for mGenTHREADER. Each algorithm was run using default parameters. Expectation values (E-values) were used to score alignments between sequence profiles, and prediction scores (SVM output and NN output scores) were used for pDomTHREADER and pGenTHREADER respectively.

To compare performance between the methods, the top hits only were considered since this method reflects common practise in whole genome annotation efforts. It also reflects the intended usage of the pDomTHREADER method for superfamily identification. True positive superfamily assignments were scored only for the selected S35 chain regions. Predictions made

Table 5.7: Features used in the mGenTHREADER and pDomTHREADER scores

Input	**mGenTHREADER**	**pGenTHREADER**
Alignment score	a = 0.01, b = 150	a = 0.01, b = -50
Alignment length	a = 0.01, b = 150	–
Coverage	–	$\frac{alignlength}{templatelength}$
Pairwise energy	a = 1, b = 100	a = 0.1, b = 50
Solvation energy	a = 1, b = 10	a = 1, b = 1
Z score energy	Z score	–
Z score solvation	Z score	–
Query length	a = 0.01, b = 150	–
Template length	a = 0.01, b = 150	–

Values termed a and b represent parameters of the exponential function of the form $\frac{1}{1+exp-ax+b}$. Z score energy terms were dropped from the pDomTHREADER score since they added little value when combined with the other features to classification accuracies.

for the other sequence regions were omitted. Additionally, several matches between different superfamilies were not scored as incorrect matches because they corresponded to SAS8 exceptions defined in Reid et al. (2007). These represent cases of structural similarity where genuine homology exists between superfamilies. Genuinely high scoring alignments can therefore be generated between these superfamilies.

The resulting scores from the different algorithms were compared using Errors Per Query (EPQ) plots (Figure 5.4) which report the frequency of errors as a function of the number of predictions performed. EPQ is given by

$$EPQ = \frac{FP}{(TP + FP)} \tag{5.4}$$

and is derived from TP (True Positive) and FP (False Positive) values obtained at a particular score threshold.

Results were averaged by jack-knifing performance, leaving out a single superfamily at a time. At low error rates pDomTHREADER outperformed all other methods. However, at error rates approaching 0.05, PRC recognised more true positives. The poorer performance of the mGenTHREADER algorithm could be explained by the use of whole PDB chains for the threading template library. Many false positive assignments resulted from incorrect whole chain alignments in the pGenTHREADER method that were anchored by other common domains between two chains. Frequently these alignments had over-extended so that different superfamilies were matched. However, the advantage of using whole chain templates is evident in fold recognition where maintenance of an up to date template library is a key determinant of the accuracy of the method. Despite recent improvements in both CATH and SCOP databases, there inevitably exists a significant lag period between the release of a PDB structure and its classification into domains.

The pDomThreader method was then used to predict superfamilies for 5.5 million UniRef sequences. The algorithm was capable of identifying domain annotations for 3.3 million sequences using a stringent filter at a coverage threshold of 50% of the domain template. 44,632 human sequences (68.42% of the IPI human dataset) were included in this set. This figure represents a significant improvement compared to sequence annotation coverage using CATH domains listed in Gene3D database of structural annotations (Yeats et al. 2008) at 59% coverage, and in the current version of the Integr8 genome annotations for IPI human sequences of which 22,651 sequences (31.05%) are annotated.

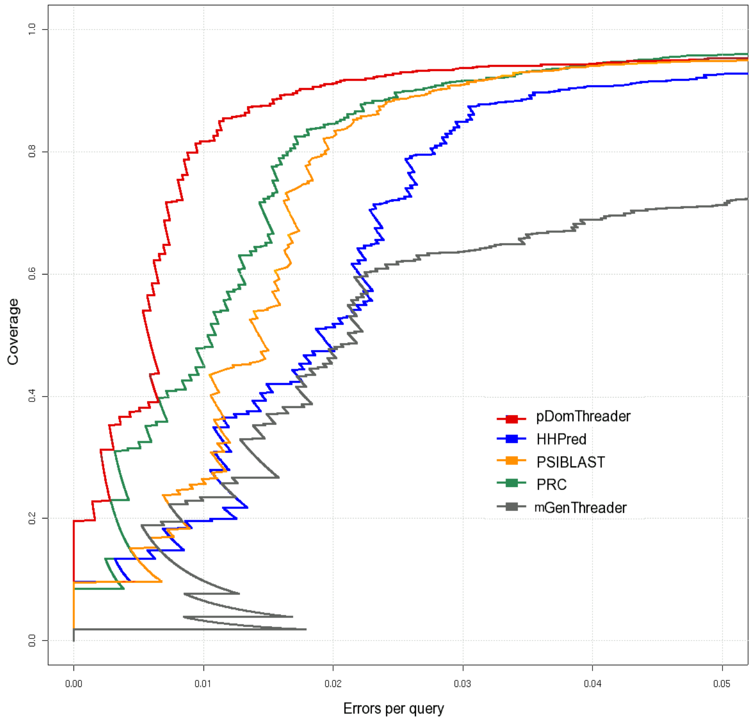


Figure 5.4: Benchmark performance for pDomTHREADER compared to other profile-profile methods.

PFAM family identification

A library of PFAM profile HMM's were downloaded from the HHSearch website (ftp://toolkit.lmb.uni muenchen.de/HHsearch/databases 2008) in the a3m format. In this format sequence alignments are supplemented with predicted and actual secondary structure information. To generate PFAM assignments to sequences, profile-profile matching was carried out using the HHPred algorithm (Soding 2005). The same query sequence profiles were used that were generated for the UniRef sequences for threading. Default parameter settings were used to compare checkpoint profiles for the query sequence against the library of PFAM profile HMM's. Using this approach 63700 human, and 3425730 UniRef sequences could be annotated with at least one PFAM family using an E-value cut-off of 100. This threshold was intentionally permissive to ensure high coverage PFAM assignments and was incorporated into the feature scoring method as a measure of prediction quality.

Scoring domain and complex architectures

Domain architectures were scored by considering each domain occurrence as a single feature. Repeat occurrences of a domain were scored as separate features so that sequences with single domain copies could be discriminated from those containing multiple copies. Rather than using a binary score denoting domain presence or absence, the prediction quality for the domain was used as the feature value (Table 5.8).

Complex architectures were generated by combining the predicted domains from a pair of sequences into a single architecture. Similar to the domain architecture score each occurrence of each domain constituted a feature, and each feature value was comprised of two weights. The first weight represented the prediction quality for the domain whilst the second weight was related to the frequency of occurrence of the sequence in the total set of fused architectures (Table 5.8). This second weight was designed to differentiate between promiscuous sequences and those that occurred rarely in the set of fused architectures.

Using this complex architecture scoring method, sets of sequence pairs with common domain components had identical features. To provide an extra level of resolution between these domain fusions, additional similarity scores were determined for each sequence pair. The scores characterised the relationship between each sequence and the set of parental fusion sequence(s).

Table 5.8: Example domain architecture score

Architecture score			
Sequence	**Sequence A**	**Sequence B**	
Domains	Domain X	Domain X	Domain Y
Prediction	235	128	174
Binary	1	1	1
Binary architecture pair score			1
Weighted architecture score			235+128 * 0.5

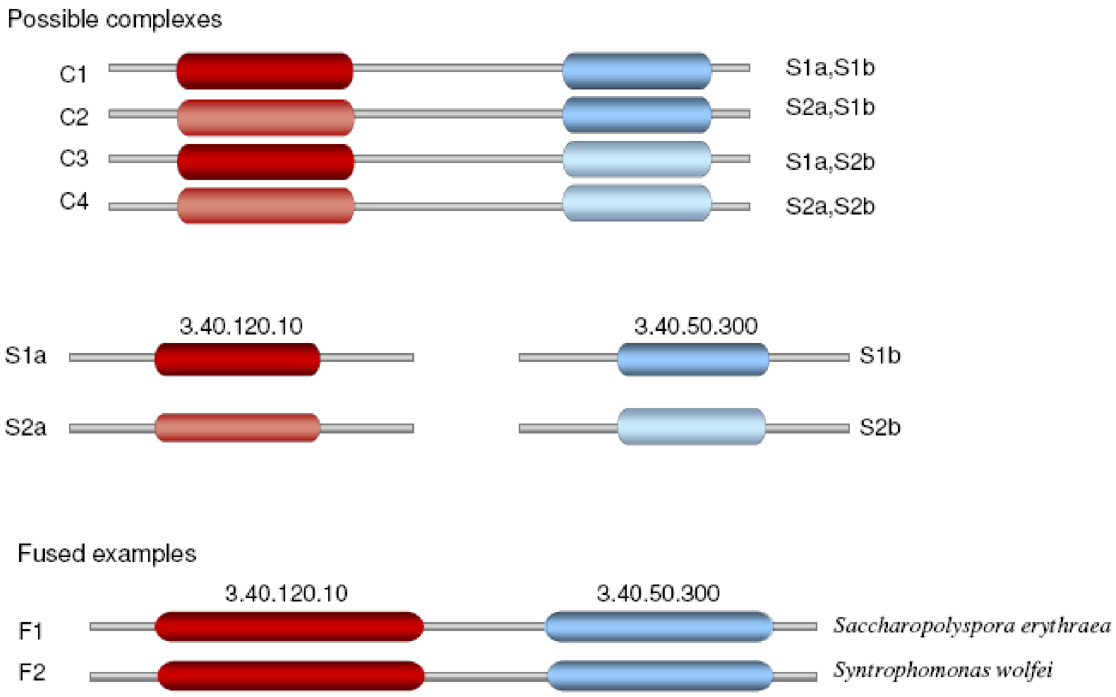
Fusion score			
Sequence	**Sequence A**	**Sequence B**	
Frequency	112	47	
Domains	Domain R	Domain S	Domain T
Prediction	235	128	174
Binary complex	1	1	1
Binary complex score		3	
Weighted complex score		235 + 128 + 174	
Double weighted complex score		log(24238/112)*(235) + log(24238/47)*(128+174)	

The architecture score is composed of a single value for each common domain between two sequences whereas the weighted score is composed of the averaged prediction scores for each of the domain copies. Each single domain that is part of the fused complex is scored separately either using a binary value or the prediction score for each domain. The second weight is derived from the frequency of each sequence in the population of predicted complexes. This weight accounts for the promiscuity of the sequence so that rarely occurring high confidence domain predictions achieve a greater complex architecture score. Alignment bitscores were normalised using the logistic transform (Equation 5.2) and complex sequence frequency weights were downscaled by a factor of 10 to produce values between the interval 0 and 1.

Considering the hypothetical example in Figure 5.5, four human sequences can be linked by the domain fusion 3.40.120.10 - 3.40.50.300. Each pair of sequences shares at least one domain in common with the fusion sequence(s) and can therefore be aligned to the fusion sequence. The set of alignment scores between human sequences S1 .. Sn, and parent fusions F1 .. Fn can be exploited in order to resolve which sequence pair is most closely related to each fusion sequence.

A comparison between two vectors of alignment scores V1 and V2 provided a similarity measure between all fusions F1 .. Fn and human sequences. The rationale behind this approach was that if the sequences S1 and S2 have evolved from a particular parent fusion sequence, it is expected that both sequences might share similar relationships with the set of fusions (F1 .. Fn). To score the similarity of relationships between sequences S1 .. Sn and F1 .. Fn, euclidean distances were calculated between the alignment score vectors V1 and V2. The measures were converted to a similarity measure by a scaling operation (between 0 and 1) followed by an inversion operation (subtract from 1). This co-related or co-evolution score was added to the fused architecture features.

To determine the value of this new score in relation to function similarity, correlation analysis was performed between groups of sequence pairs with common complex architectures and function similarity. A Pearson correlation co-efficient was determined between sequence pairs with a common complex architecture and the set of function similarity values. This analysis reflected the intended usage of the score, since it is designed to provide a ranking of function similarity between sequence pairs with common complex architectures. 48% and 52% of the fusion co-relationship values were positively correlated with function similarity using MF and BP measures respectively. Thus the value of this score in function prediction is questionable. However these results may be symptomatic of incomplete annotations for the sequence pairs, or reflect biological aspects of the data. In cases where all pairs of sequences with a common complex architecture share the same function similarity, the score has no practical utility. In total, 627244 and 837839 human sequences could be linked by domain fusions that were annotated to MF or BP GO terms respectively using CATH domain annotations. Using PFAM families, a much greater number of links could be made (213,538,655). This finding is attributed in part to the wider coverage obtained by PFAM annotations for genome sequences. In addition, some PFAM families represent short repeat regions of sequences or motifs that occur with high frequencies giving rise to greater numbers of PFAM combinations that can be used to make functional linkages.



Align scores

	V1		V2
S1a,F1	300	S1b,F1	242
S1a,F2	124	S1b,F2	136
	V1		V2
S1a,F1	300	S2b,F1	242
S1a,F2	125	S2b,F2	136

Figure 5.5: Example of the complex domain pair score. This score provides additional resolution between sequence pairs that have identical complex architectures. The score comprises a measure of similarity between the degree of relatedness between each sequence and each fusion sequence.

Generating novel biological insight from domains

In generating domain architecture and domain fusion feature scores, novel biological insight could be gained by post examination of the data. For example, several domains of unknown function from PFAM could now be associated with particular functions (see Table 5.9). For example, PF004750 PFAM family occurs in 3 human sequences that receive the GO annotation “GO:0000902 cell morphogenesis”. Other example linkages include PF06582 and “GO:0043565 sequence-specific DNA binding” and PF006352 “GO:0006814 cation transport”.

The putative associations between domain and function may be as a consequence of partial annotation, or be restricted to human sequences only. Additionally, this evidence is not sufficient to propose that these domains are responsible for a particular function since other sequence characteristics might determine function. However, this information is useful in function prediction either directly or indirectly, and highlights the value that can be obtained from the domain architecture approach over using direct PFAM to GO mappings.

Novel biological insights could be generated through examination of the domain fusion information. The majority of the data fell into two distinct cases. The first case showed the value of domain fusions in generating well characterised function linkages. A bi-functional enzyme present in 3 strains of *Mycobacterium* contained two domains, beta lactamase (3.90.850.10) and fumarylaceto acetase (3.60.15.10). The function class hydrolase can be inherited through direct mappings for each domain, so the functional linkage between sequences does not present any new information. However, the sequence pairs that are annotated by each domain cannot be linked by homology-based methods and may indicate some shared pathway or process that are currently unannotated. In total, the domain fusion between 3.90.850.10 and 3.60.15.10 could be used to make 1881 sequence pair links (171 by 11 domain representatives respectively).

The second case for the fusion data constitutes a case of novel structural domain annotation that generates novel functional insight. Evidence for fusions between CATH domains 3.40.120.10 and 3.40.50.300 are observed in 3 sequences from *Saccharopolyspora erythraea, Syntrophomonas wolfei* and *Rhodobacterium bacteriales*. Following the Rosetta stone hypothesis, it is proposed that these sequences share an evolutionary relationship and are likely to share a physical interface, particularly where the distance between two domains in the fused sequence is small *in vivo*. Sequences that bind to one another frequently share an aspect of common

Table 5.9: Putative functional linkages between PFAM and GO terms

PFAM identifier	GO term	Description	N
PF04750	GO:0000902	cell motility	3
PF06844	GO:0008168	methyltransferase	4
PF09317	GO:0050660	FAD binding	2
PF04646	GO:0004672/GO:0005524	protein kinase/ATP binding	4
PF01926	GO:0005525	GTP binding	19
PF08953	GO:0003779	actin binding	5
PF05696	GO:0004984	olfactory receptor	7
PF04515	GO:0015220	choline transmembrane transporter	2
PF06571	GO:0003394/GO:0051539	aconitase/4 iron, 4 sulphur cluster binding	2
PF05638	GO:0016301/GO:0016308	1-phosphatidylinositol-4-phosphate 5-kinase/nucleotide binding	3
PF05614	GO:0005524/GO:0016820/GO:0015662	ATP binding/ATPase activity/hydrolase activity, acting on acid anhydrides	3
PF08983	GO:0004930	G-Protein coupled receptor activity	5
PF06544	GO:0019787	small conjugating protein ligase	3
PF04076	GO:0000287/GO:0004012/GO:0005524	magnesium ion binding/ATP binding/phospholipid-translocating ATPase	2
PF07289	GO:0004437	inositol or phosphatidylinositol phosphatase	5
PF05695	GO:0005524/GO:0016887	ATP binding/ATPase activity	4
PF04844	GO:0005097	Rab GTPase activator	2
PF09324	GO:0015450	P-P-bond-hydrolysis-driven protein transmembrane transporter	2
PF09314	GO:0008138	protein tyrosine/serine/threonine phosphatase	2
PF09687	GO:0005509	calcium ion binding	2
PF08987	GO:0043565	sequence-specific DNA binding	2
PF03781	GO:0005509	calcium ion binding	2
PF04418	GO:0005097	Rab GTPase activator	4
PF04515	GO:0015220	choline transmembrane transporter	2
PF06571	GO:0003994	aconitase/4 iron, 4 sulphur cluster binding	2
PF06352	GO:0015075/GO:0015101	ion transmembrane transporter	6
PF04217	GO:0008271	secondary active sulphate transmembrane transporter	6
PF05638	GO:0016301/GO:0016308	1-phosphatidylinositol-4-phosphate 5-kinase/nucleotide binding	3
PF08424	GO:0006396/GO:0006350	RNA processing/transcription	5
PF04570	GO:0000902	cell motility	3
PF08969	GO:0006512	ubiquitin cycle	3
PF07274	GO:0007242	intracellular signalling cascade	2
PF05908	GO:0006355	regulation of transcription	3
PF05657	GO:0006412	translation	2
PF08401	GO:0006355	regulation of transcription	13
PF04308	GO:0006412/GO:0042254	translation/ribosome biogenesis	2
PF06352	GO:0015695	organic cation transport	5
PF07098	GO:0006857	oligopeptide transport	2
PF06198	GO:0015914	phospholipid transport	9
PF08987	GO:0006355	regulation of transcription	2
PF08648	GO:0006397	mRNA processing	4
PF09314	GO:0006470	protein amino acid phosphorylation	2
PF05590	GO:0006508	proteolysis	13
PF04844	GO:0032313	regulation of Rab GTPase	2
PF07289	GO:0046839	phospholipid dephosphorylation	7
PF08983	GO:007186	GPCR signalling pathway	5
PF07381	GO:0007242	intracellular signalling cascade	2
PF04646	GO:0006468	protein amino acid phosphorylation	4
PF06544	GO:0006464/GO:0006512	protein modification/ubiquitin cycle	2
PF09667	GO:0006355	regulation of transcription	4
PF05614	GO:0006812	cation transport	3
PF04515	GO:0015871	choline transport	2
PF07080	GO:0006814	sodium ion transport	4
PF07469	GO:0006355	regulation of transcription	2
PF09234	GO:0006915	apoptosis	3
PF06571	GO:0006099	tri-carboxylic acid cycle	2
PF09320	GO:0006457	protein folding process	4

N represents the total number of human sequences that were predicted to contain the pfam domain. In each case 100% of these sequences bore the represented annotation.

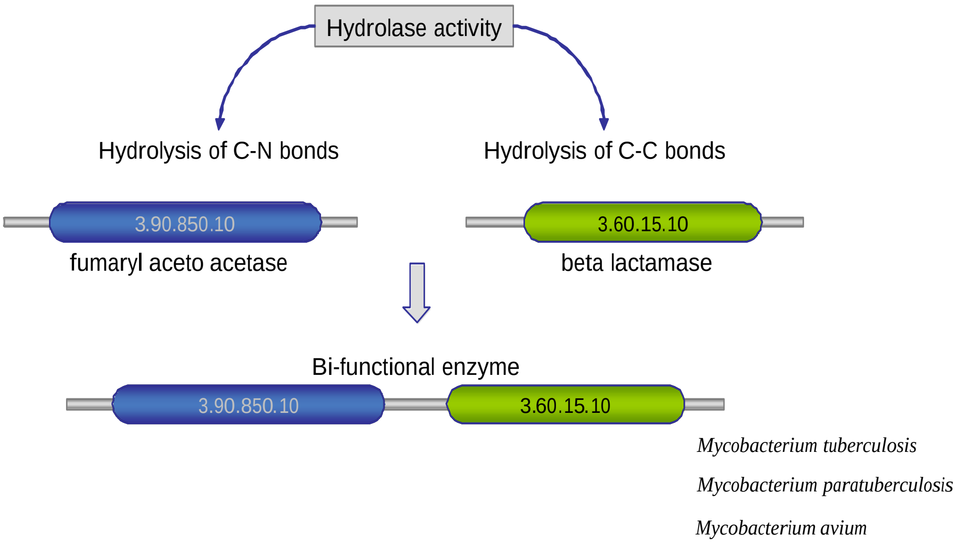


Figure 5.6: Example function linkage using domain fusion information. Hydrolase annotations can be provided for non-homologous sequences using the fusion information.

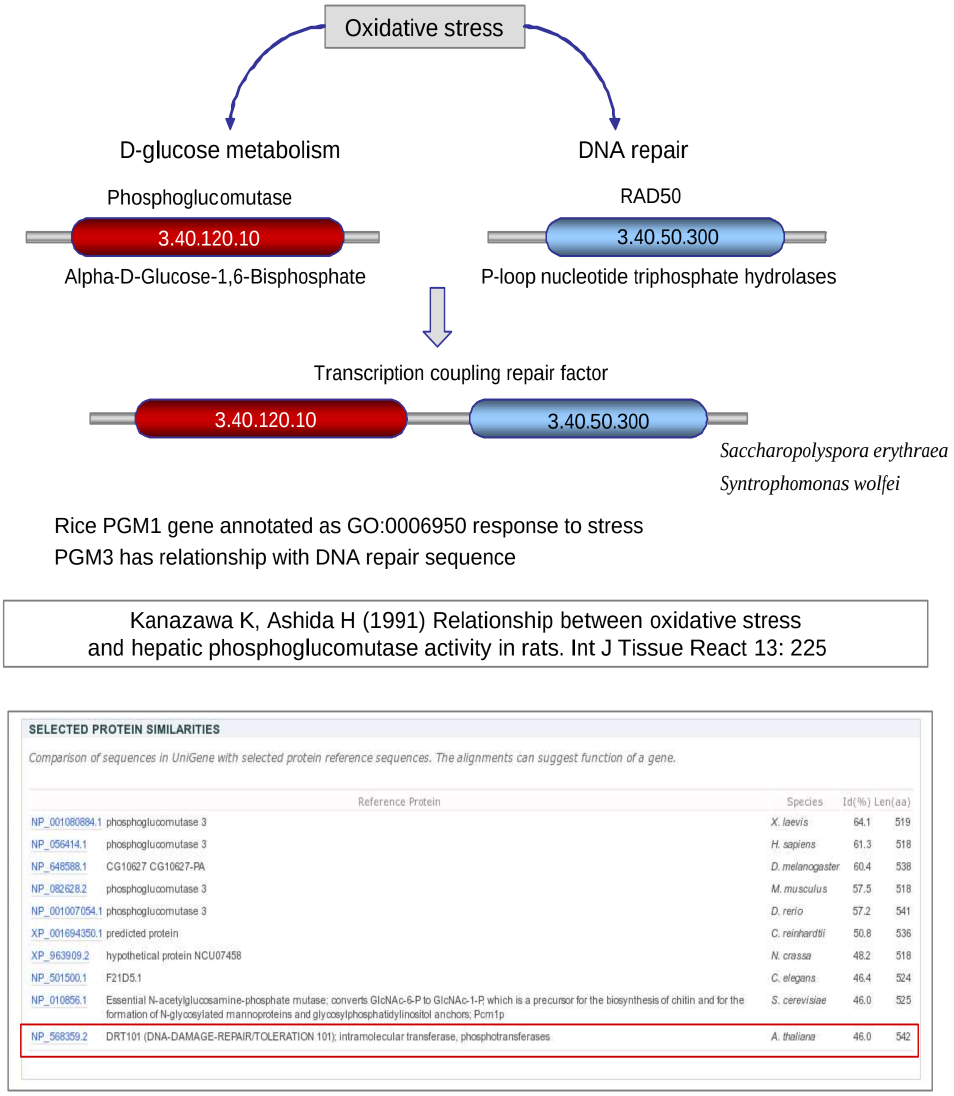
function or participate in similar biological pathways (Date 2007, 2008). The current GO annotations for the human phosphoglucomutases include glycolysis and energy metabolism, whilst for RAD50 include nucleotide binding and DNA binding functions. However, a literature search revealed that rat PGM3 enzyme is related to a DNA repair sequence from *Arabidopsis thaliana*, and PGM1 in the rice genome is annotated with BP term "Response to Stress". Phosphoglucomutases are often used as markers of oxidative stress (Kanazawa and Ashida 1991) and RAD50 is involved in DNA repair in response to oxidative stress. Consequently, it is proposed that the computational fusion method has identified a genuine functional linkage between sequences containing the two domains.

Merely linking all sequences containing either of these domain results in 2136 putative linkages since the P-loop hydrolases are frequently occurring. However it is not proposed that all of these sequence pairs can be annotated with the GO term "Oxidative stress". The complex architecture scoring scheme permits resolution between the pairings because the pairings receive different features where the domain components of the complex differ. In the example case, PGM1 has three copies of the domain 3.40.120.10 and a single copy of 3.30.310.50. RAD50 has predicted architecture 1.20.58.70 - 3.40.310.50. Just 40 sequence pairs share this complex architecture and can be further differentiated by the pair fusion relationship scores.

In the public domain, the assignment of domain 3.40.120.10 to the fusion sequences does not exist, however it can be assigned using the pDomTHREADER method. This putative annotation illustrates the power of the threading approach in domain annotation and the potential of the fusion method in identifying function linkages. Because the available annotations are incomplete for sequences, the value from these fusions cannot be realised computationally, however they present an important dataset of putative functional linkages that warrant further investigation.

5.3.6 Microarray expression information

Microarrays enable simultaneous monitoring of the gene expression levels of thousands of transcripts in different biological contexts. Throughout the past decade, they have been used routinely and extensively in the laboratory as tools for exploring expression changes in disease states in response to cellular stimulus, or to monitor tissue level expressions in normal conditions. The results of these large scale experiments have been data-warehoused in several publicly available repositories. For example the Gene Expression Omnibus, GEO (Barrett and Edgar 2006), Ar-



Oxidative stress
D-glucose metabolism
DNA repair
Phosphoglucomutase
3.40.120.10
Alpha-D-Glucose-1,6-Bisphosphate
RAD50
3.40.50.300
P-loop nucleotide triphosphate hydrolases
Transcription coupling repair factor
3.40.120.10
3.40.50.300
Saccharopolyspora erythraea
Syntrophomonas wolfei

Rice PGM1 gene annotated as GO:0006950 response to stress
PGM3 has relationship with DNA repair sequence

Kanazawa K, Ashida H (1991) Relationship between oxidative stress and hepatic phosphoglucomutase activity in rats. Int J Tissue React 13: 225

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with selected protein reference sequences. The alignments can suggest function of a gene.

Reference Protein		Species	Id(%)	Len(aa)
NP_001080884.1	phosphoglucomutase 3	*X. laevis*	64.1	519
NP_056414.1	phosphoglucomutase 3	*H. sapiens*	61.3	518
NP_648588.1	CG10627 CG10627-PA	*D. melanogaster*	60.4	538
NP_082628.2	phosphoglucomutase 3	*M. musculus*	57.5	518
NP_001007054.1	phosphoglucomutase 3	*D. rerio*	57.2	541
XP_001694350.1	predicted protein	*C. reinhardtii*	50.8	536
XP_963909.2	hypothetical protein NCU07458	*N. crassa*	48.2	518
NP_501500.1	F21D5.1	*C. elegans*	46.4	524
NP_010856.1	Essential N-acetylglucosamine-phosphate mutase; converts GlcNAc-6-P to GlcNAc-1-P, which is a precursor for the biosynthesis of chitin and for the formation of N-glycosylated mannoproteins and glycosylphosphatidylinositol anchors; Pcm1p	*S. cerevisiae*	46.0	525
NP_568359.2	DRT101 (DNA-DAMAGE-REPAIR/TOLERATION 101); intramolecular transferase, phosphotransferases	*A. thaliana*	46.0	542

Figure 5.7: Example novel function annotation “Oxidative stress” is suggested by the fusion between phosphoglyceratemutase enzymes and DNA repair sequences.

ray Express (Brazma et al. 2003), the Stanford Genome Microarray Database (Ball et al. 2005) and the RNA Abundance Database (Manduchi et al. 2004) (RAD). The task of appropriate integration and analysis of these valuable datasets presents exciting opportunities for functional discoveries.

Integrating experiments from different laboratories and platform technologies represents a difficult task due to systematic and non-systematic variability that is embedded within the data. Variability arises from different laboratory bench protocols, alternative probe sequences for representing similar genes and accuracy of measurement bias for the different platforms (Ahmed 2006, Lee and Saeed 2007). Several studies report that it is only possible to integrate information from different experiments at the meta-information level, by comparing outcomes of analyses mapped to pathway or functional categories (Cahan et al. 2007). In contrast, Stevens and Doerge (2005) showed that the accuracy of detection of differentially expressed genes was greater when multiple Affymetrix studies were combined than when single studies were used. Careful and quantitative integration of these experiments, considering the different sources of variance can lead to improved statistical power in biological hypothesis testing (Hu et al. 2005).

Here the integration of diverse experiments performed on the Affymetrix U133A chip has been attempted to determine co-expression relationships for the human transcriptome. Different experiments are combined using low-level (probe-wise) integration in order to merge information from the different experiments. Popular methods of inferring co-expression between transcript pairs including 1st order correlations were investigated. Additionally feature information was generated using bi-clustering techniques. The value of these features was compared with standard co-expression measures by correlation analysis with function similarity.

Datasets and pre-processing

81 Gene Expression Omnibus experiments were used as the source of human microarray datasets all performed on the U133A 3'IVT array (Table 5.10). Any experiment was considered relevant, including those performed using disease or cancer samples since many transcription factors or apoptotic pathways are only activated under these conditions. A single platform was chosen for study for two reasons. First to avoid difficulties surrounding differences between expression measures obtained from different platforms. For example, the Stanford Array and Agilent whole genome arrays use two color reference samples per array to generate relative measures

of expression, whereas Affymetrix and Nimblegen technologies used a single channel per array to measure transcript abundance levels. Second, information loss was avoided that occurs when mapping between probesets for human transcripts.

Obtaining transcript level expression measures from microarray data typically involves three distinct steps; pre-processing, normalisation and summarisation. The pre-processing stage involves background and or probe affinity adjustments. Normalisation scales sample distributions such that means, medians and or variances are equal between different arrays and summarisation steps estimate single transcript values from a group of representative probes. For this analysis a background and probe GC content correction were applied to all arrays on a per experiment basis before carrying out quantile normalisation (Figure 5.8).

The GCRMA background correction step was used to adjust the raw fluorescence data for non-specific binding effects. This step was carried out independently for each experimental dataset since the probe specific affinities can vary according to lab protocols, the quantity of RNA hybridised to the array and the chip type.

A single target distribution was used for quantile normalisation and constructed from 70 maximally varying array samples that had been corrected for background. These samples were chosen such that no two sample distributions were highly correlated (using a sample cut-off of 0.7). This type of rank-based normalisation eliminated systematic differences between samples hybridised to different arrays in different laboratories by enforcing equality between sample distributions.

Summarisation

The quantile normalised probe match values were summarised into transcript level abundances using the median polish algorithm (Irizarry et al. 2003). In order to obtain a high quality and robust representation of transcript abundance only transcript consistent probes were used in summarisation specified using an alternative probe-transcript mapping provided by AffyProbeMiner (Liu et al. 2007). Transcript consistent probes represent those that uniquely and consistently map to a single transcript. In total 26,688 transcripts were represented in each microarray sample that could be mapped to a corresponding protein entry. Mapping was performed conservatively using the RefSeq and NCBI protein annotations for the U133A chip. Sequences that differed by a maximum of 3 amino acids between IPI and proteins represented on the chip were considered

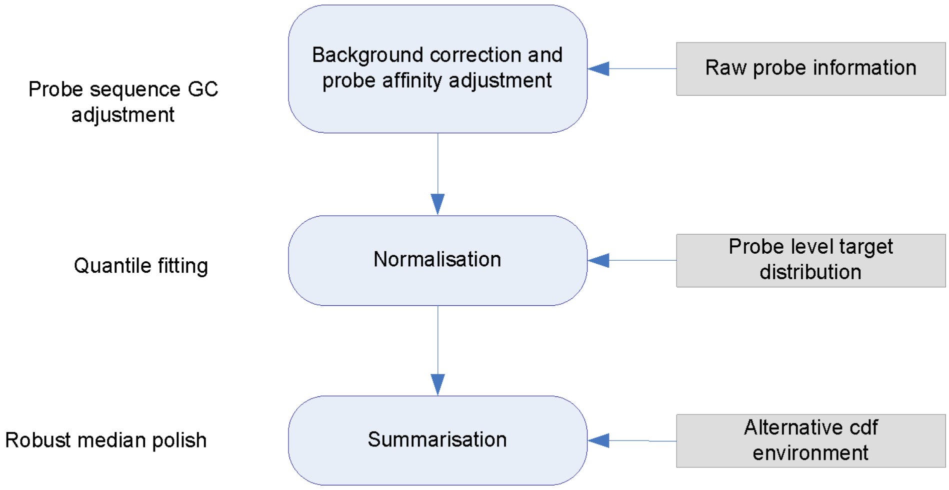

Probe sequence GC adjustment
Background correction and probe affinity adjustment
Raw probe information
Quantile fitting
Normalisation
Probe level target distribution
Robust median polish
Summarisation
Alternative cdf environment

Figure 5.8: Pre-processing flow chart

Table 5.10: Microarray datasets that were combined and integrated for the analysis

Dataset	Source	Description	Samples
GSE1000	Laboratory for Biomedical Materials Research, Boston	osteosarcoma study	10
GSE1133	RNA Profiling Group Novartis	normal tissues	158
GSE1140	University of California, Irvine	PBMC exercise study	14
GSE1295	Research Center for Genetic Medicine, Washington	skeletal muscle metabolic syndrome	24
GSE1297	Landfield, Kentucky	Alzheimers disease study	31
GSE1323	Laboratory of Metabolomics and Systems Biology, Trento	colorectal cancer study	6
GSE1420	Radiation and Cellular Oncology, Chicago	oesophageal cancer progression	24
GSE1462	Department of Neuroscience, Milan	mitochondrial disease & skeletal muscle	15
GSE1561	Bute Medical School, St Andrews	breast cancer study	49
GSE1577	Mount Sinai School of Medicine, New York	TALL BALL and TLL leukaemia study	29
GSE1615	Reproduction and Women's Health, Pennsylvania	PCOS study on theca cells	12
GSE1648	Setton Lab Duke University, Durham	osmotic stress on invertebral disc cells	11
GSE1650	Pulmonary and Critical Care Medicine, Boston	lung from smokers and non smokers	30
GSE1657	Boston University School of Medicine, Boston	Fat deposits from skin regions	24
GSE1729	Haematology University Hospital, Salamanca	Acute myeloid leukaemia subtypes	43
GSE1786	Paediatric Exercise Research Center, Irvine	Effect of exercise on aged muscle	24
GSE1869	Internal Medicine, Division of Cardiology, Carnegie Baltimore	Ischemic and nonischemic heart samples	25
GSE2004	NIH Neuroscience Microarray Consortium, Phoenix	kidney, liver and spleen samples	25
GSE2018	Medicine University of Minnesota, Minneapolis	lung transplant biopsies	34
GSE2113	Ospedale Maggiore IRCCS, Milan	plasma cells from leukaemia patients	52
GSE2144	Institute of Molecular Medicine, Dublin	Oesophageal cell treatments	6
GSE2152	Lauri A. Aaltonen Lab, Helsinki	Uterine fibroids with FH mutations	22
GSE2175	Endocrine Oncology Bart's, London	Pituitary adenoma subtypes	5
GSE2189	Joseph G. Hacia Laboratory, California	Lung cancer drug treatment	18
GSE2225	City of Hope Surgical Research, California	MCF-7 cells before and after treatment	18
GSE2240	Department of Cardiology Grosshardern, Munich	fibrillated atrium	35
GSE2248	Genomics Core Laboratory, New York	comparison of mesenchymal stem cells	6
GSE2280	Children's Hospital of Philadelphia Muschel, Philadelphia	squamous cell carcinoma	27
GSE2361	ENH Research Institute Evanston, Illinois	normal tissue study	36
GSE2395	Research Center for Genetic Medicine, Washington	cystic fibrosis patients	20
GSE2443	Molecular Therapeutics Program, Bethesda	prostate cancer progression	20
GSE2450	University of Arkansas for Medical Sciences, Arkansas	endothelial cells response to drug	16
GSE2485	Molecular Neuro-Oncology lab, Massachusetts	glioblastoma tumours	18
GSE2513	Singapore Eye Research Institute, Singapore	pterygium tissues	12
GSE2531	Wisconsin National Primate Research Center, Wisconsin	placental trophoblasts	7
GSE2638	Experimental Dermatology & Paediatrics, Muenster Germany	TNF stimulated microvascular endothelium	6
GSE2665	Medical Centre Mannheim, University of Heidelberg	lymphnode and tonsil comparison	20
GSE2666	Stem Cell Institute Minnesota, Minneapolis	umbilical cord & bone marrow stem cells	18
GSE2724	Lauri A. Aaltonen Lab, Helsinki	uterine fibroids with FH mutations	11
GSE2725	Lauri A. Aaltonen Lab, Helsinki	uterine fibroids with FH mutations	10
GSE2742	RNA Profiling Group Novartis	colon adenocarcinoma study	27
GSE2815	Molecular Genetics & Microbiology, Albuquerque	MCF7 cells infected with adenovirus	8
GSE3167	Molecular Diagnostic Laboratory, Aarhus Denmark	bladder biopsies	60
GSE3284	Information Dissemination & Coordination, Massachusetts	endotoxin effect on leukocytes	46
GSE3307	Research Center for Genetic Medicine, Washington	muscle biopsys various diseases	121
GSE3356	University of Tuebingen, Germany	smooth muscle response to drug	9
GSE3407	Alan Weiner University of Washington, Seattle	cockayne syndrome	8
GSE3419	National Cancer Institute NIH, Bethesda	keratinocyte stem cells	16
GSE3524	Center for Applied Genomics, New Jersey	Oral squamous cell carcinoma	20
GSE3585	Expression Profiling, Heidelberg	Dilated cardiomegaly heart samples	12
GSE3737	Cell Growth VA Medical Center, San Francisco	PC3 prostate cancer	8
GSE3846	Pulmonary Gene Research, Basel Switzerland	blood after wine water & grape juice	108
GSE3860	Children's Hospital, Boston	Hutchinsons patient fibroblast cell lines	18
GSE4045	Tumorigenesis group, Helsinki	adenocarcinoma subtypes	37
GSE4127	Nippon Medical School, Tokyo	lung cancer cell lines	29
GSE4176	Functional Genomics Experimental Oncology, Switzerland	mantle cell lymphomas	5
GSE4271	Genentech, Inc., California	astrocyte tumour progression	100
GSE4412	NIH Neuroscience Microarray Consortium, Phoenix Arizona	Normal placenta	85
GSE4636	McDonnell Duke University, Durham	LNCaP cells stimulated with Androgen	18
GSE4646	Molecular Biology of Bacterial Pathogens, Prague	Umbilical vein before and after infection	12
GSE473	Research Center for Genetic Medicine, Washington	CD4+ lymphocytes w/wo asthma	88
GSE475	Research Center for Genetic Medicine, Washington	Diaphragm muscle from COPD patients	7
GSE4817	Brain Tumour Research Neurobiology, Chicago	glioblastoma cells	6
GSE4917	Medicine University of Chicago, Illinois	MCF10 breast cells stress response	24
GSE5090	Instituto de investigaciones biomedicas CSIC-UAM, Madrid	omental adipose tissues (obesity)	15
GSE5370	Research Center for Genetic Medicine, Washington	Dematomyositis muscle samples	5
GSE5388	Centre for Neuropsychiatric Research, Cambridge UK	Bipolar samples from 30 adults	61
GSE5389	Centre for Neuropsychiatric Research, Cambridge UK	Bipolar samples from 10 adults	21
GSE5418	Laboratory of Functional Genomics Rockville, MD	PBMCs from patients with malaria	71
GSE5667	Dermatology Mayo Clinic Rochester, MN	Non lesional & lesional atopic dermatitis	17
GSE620	Johns Hopkins Medical Institutions, Baltimore	Cystic fibrosis bronchial epithelium	11
GSE6236	Molecular Biology and Genetics Section, Bethesda	Adult and fetal reticulocytes	28
GSE6691	Haematology University Hospital, Salamanca	Waldenstrom's macroglobulinemia, Bcells and plasma	56
GSE6740	DerOstrowski University of Toronto	HIV infected CD4 and CD8 Tcells	40
GSE6783	Weizmann Institute of Science Rehovot, Israel	EGF effect on HeLa cells	7
GSE6883	OncoMed Pharmaceuticals Inc, California	Tumorigenic breast cancer cells	22
GSE7035	Dana-Farber Cancer Institute Cancer Biology, Boston	PParγ treatment of adenocarcinoma cells	14
GSE781	Genetics & Genomics Boston University, Masachusetts	renal carcinoma vs normal kidney	17
GSE873	Khurana Lab Pennsylvania Muscle Institute, Philadelphia	Extraoculur muscle and limb comparison	5
GSE974	Hall Cardiology Lillehei Heart Institute, Minneapolis	Myocardial remodelling after implant	38
GSE994	Pulmonary and Critical Care Medicine, Boston	Smoking induced changes in lung	75
Total	**81**		**2346**

equivalent.

By performing normalisation and summarisation steps using a common reference sample, variation between experiments was greatly reduced, however could not be eliminated. This was demonstrated by eigen value decomposition of the inter-sample correlation matrix to visualise sample variance in fewer dimensions (see Figure 5.9). Proper batch adjustments could not be performed on these data without the risk of over-correction leading to the loss of biological information. Estimating batch variation is a well documented problem when combining microarray data between studies. Several replicates between the different groups were required for reliable variance estimation, however in this case different numbers of samples from different experimental laboratories are present. Additionally, biological sample equivalence is difficult to determine due to limited descriptive information supplied with the data. Consequently no inter-study batch correction was performed.

The case of the reticulocyte experiment (Figure 5.9) is a good example of the problem. Reticulocytes are immature red blood cells containing residual amounts of RNA (Goh et al. 2007). The goal of the experiment was to determine the transcripts that are present following differentiation from erythroid cells. Consequently, these samples should be outliers from the rest of the expression data, however without the presence of another similar experiment performed in a different laboratory to normalise against, a batch correction would destroy the distribution of expressions within these samples by up-weighting expression values such that they are aligned with other experiments. The resulting dataset would be artificial and conclusions drawn from any subsequent analysis likely to be false.

The global Pearson correlation coefficient between all transcript pairs was used to compare co-expressions between transcripts, assuming that the effect of experimental batch was constant across for different transcripts expressed at any intensity. This correlation measure was therefore used to judge the relationship between co-expression and function similarity (Figure 5.10). A weighted version of Pearson's correlation measure was used to determine co-expression between transcript pairs. The equation is given by

$$R_w = \frac{\Sigma w_i x_i y_i - \Sigma w_i x_i \cdot \Sigma w_i y_i}{\left(\frac{\Sigma w_i x_i^2 - \Sigma w_i x_i^2}{\Sigma w_i}\right) \cdot \left(\frac{\Sigma w_i y_i^2 - \Sigma w_i y_i^2}{\Sigma w_i}\right)} \tag{5.5}$$

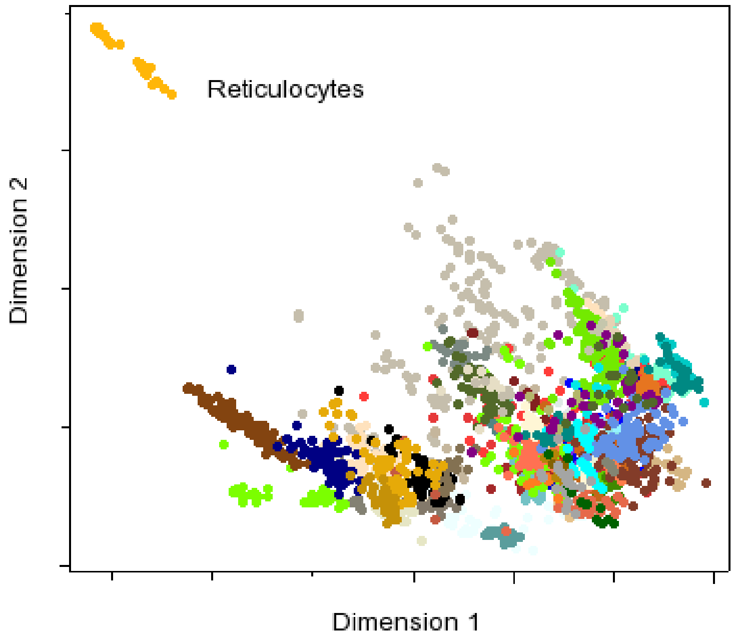


Figure 5.9: Visualisation of inter-study experimental variation. Each data point represents a different biological sample color coded according to the study from which the dataset was sourced. Samples generally cluster by experiment type, however the variation between different experiments is entwined with inter-study variation that cannot be reliably measured or removed without compromising biological sample differences. The Reticulocyte experiment shows up as an outlier which is consistent with the fact that this cell type contain relatively small amounts of DNA.

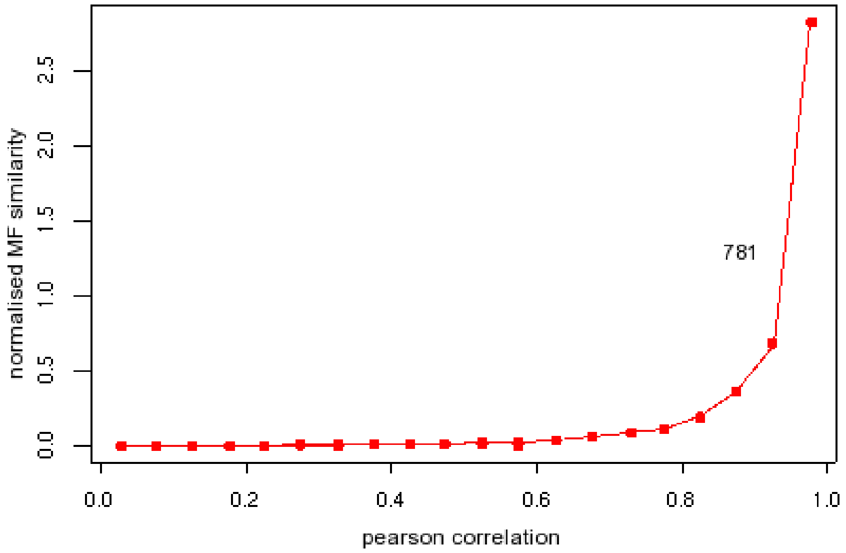
where the weights usually sum to 1. The weighted correlation reduced bias in the magnitude of the co-expression correlation coefficients by down-weighting the contribution of similar samples. The weights were determined by subtracting inter-sample correlation coefficients from 1.

Transcripts that were highly correlated were more likely to share similar functionality. This trend was more evident for BP similarity than MF similarity. However, the frequency of transcripts with correlation values of ≥ 0.9 was small (781 and 802 respectively for MFs and BPs) demonstrating limited applicability of these data alone in function prediction. This finding was consistent with another study which investigated the power of co-expression measures in determining functional relationships (Daub and Sonnhammer 2008).

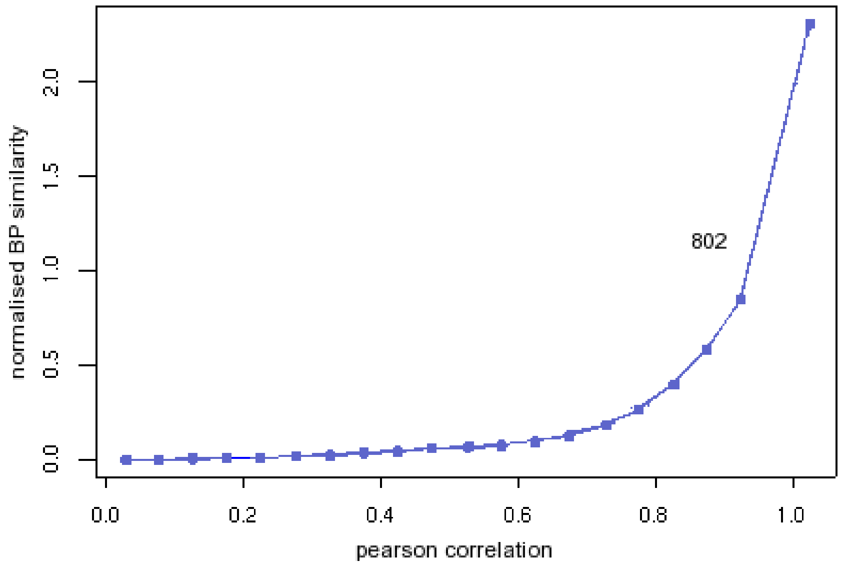
The result may be a consequence of data quality. For example, including expression intensities from just a few noisy samples is sufficient to lower the correlation coefficient between co-regulated sequences. Additionally co-expression across all of the sample conditions might not be required to indicate common functionality. Supporting evidence for this claim comes from previous work that showed that temporal patterns from subsets of sample conditions were sufficient to indicate a set of common functions (Brown et al. 2000, Iyer et al. 1999). To extract temporal expression patterns from the data that might be useful in function prediction, bi-clustering was carried out. Several publicly available bi-clustering software packages are available (Madeira and Oliveira 2004), however due to the size of the transcript to sample conditions matrix (26688 x 2342), bi-clustering was feasible only using binary data.

Sample Discretization

Discretization of the expression matrix into binary values allowed for further control of batch experimental variance from diverse samples as well as efficient computation of bi-clusters. Before discretization was carried out, replicate sample conditions were merged into single values using the one-step Tukey average given by



(a) MF co-expression correlation and function similarity



(b) BP co-expression correlation and function similarity

Figure 5.10: The relationship between co-expression and function similarity determined by correlation analysis. Each data point represents an averaged function similarity value over a range (interval 0.1) of co-expression correlation coefficients. Function similarity measures have been scaled using Z scores to emphasize the trends.

$$\begin{aligned}T &= \frac{\sum wx}{\sum w}\\ w &= (1-u^2)^2\\ u &= \frac{x-m}{(c\cdot s)-\epsilon}\\ s &= median(|x-median(x)|)\end{aligned} \tag{5.6}$$

where c controls the degree of smoothing and ϵ avoids division by zero. w are the resulting weights proportional to the degree of deviation of each item from the median.

Cases where multiple individual sample donors had been profiled in an experiment, for example, as part of a disease comparison study, were also merged into a single value. This ensured a minimum of sample redundancy despite some inevitable information loss. After replicates were merged, 432 distinct biological samples resulted. In order to discretize the data, properties of the sample intensity distributions were used. After normalisation using either GCRMA or RMA algorithms, expression intensities within a sample are bi-modally distributed (Figure 5.11). The first peak likely represents a mixture of intensities for transcripts that are not expressed or lowly expressed so that the signal is close to noise. The second peak comprises intensities that are clearly differentiated from noise or that represent abundant expression. These assumptions about the two distributions were exploited in order to define a probabilistic sample specific cut-off for discretization.

Generalised lambda distributions were used to model each sample distribution. The generalised lambda distribution is well suited to the task since it is flexible and can adopt many different distributional shapes. The distribution is specified by four λ parameters

$$F^{-1}(\mu) = \lambda_1 + \frac{\mu^{\lambda_3} - (1-\mu)^{\lambda_4}}{\lambda_2} \tag{5.7}$$

which specify location, inverse scale parameter and left and right hand skew of the distribution tails respectively. Mixtures of generalised λ distributions were estimated and fit to each sample distribution using the R GLDEX function to estimate the lambdas. Probability distributions were empirically determined for 'expressed' and 'not expressed' distributions using the lambda

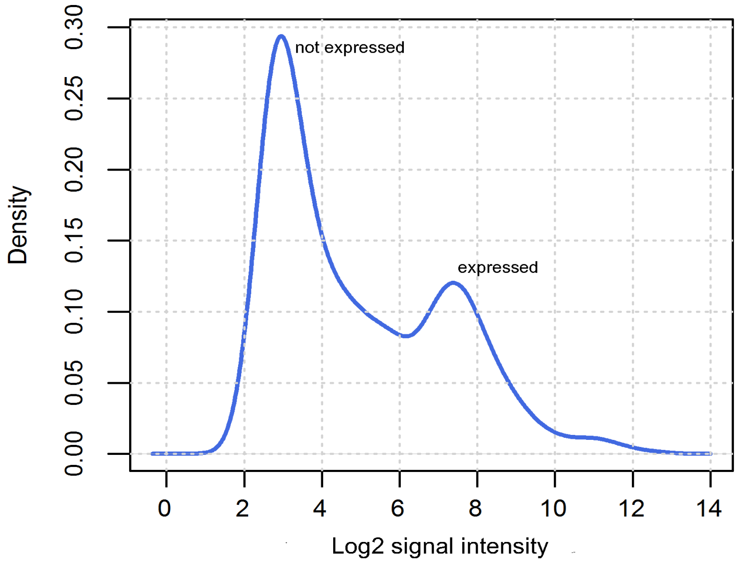


Figure 5.11: Sample distribution for GSM showing 2 distinct modes of expression

values. The expression value at which the probability of the 'not expressed' distribution reached 0.001 was used to determine each sample specific cut-offs to transform expressions into binary values.

An example of a model fit to the experimental sample GSM15875 (Figure 5.12) demonstrates the appropriate use of the lambda distribution. The distribution parameters (2.64 5.89e-5 7.15e-5 5.93e-5) yielded a fit to the observed data with log odds likelihood ratio of 967 and Kolmogorov-Smirnov p-value of 1 indicating that observed and expected distributions came from the same population. These statistics indicated high quality function estimation using the lambda distributions. At probability < 0.001, a sample specific expression value cut-off of 6.245 was determined above which the likelihood of an expression intensity belonging to the 'not expressed' distribution was small. This procedure was carried out individually for each sample.

Bi-Clustering

Bi-clustering is an unsupervised learning algorithm with the objective of grouping rows and columns of an underlying matrix into maximally correlated sub-clusters (see (Madeira and Oliveira 2004) for a review). In the context of gene expression, the bi-clusters correspond to sub matrices of co-regulated genes over a subset of conditions. Since clusters may overlap, the approach is well suited to the task of determining functionally relevant signatures in microarray data where multiple groupings between tissues or patient samples may indicate a particular function. The task of finding all significant bi-clusters in a given expression matrix is NP-hard, therefore many different flavours of bi-clustering have been developed making different assumptions and optimizations in order to converge on a final solution.

The BIMAX algorithm (Prelic et al. 2006) was used to generate bi-clusters using a minimum of 15 transcripts by 5 sample conditions for each cluster. It was not possible to compute all bi-clusters in one run due to memory requirements exceeding 64G RAM. To exhaustively sample all possible bi-clusters from the data matrix, more than 43 trillion unique samplings of conditions are required ($\frac{432!}{(432-5)!}$). However, repeated permutations of the row and column ordering of the binary expression matrix were carried out (100 times) to generate a diverse set of clusters. Each permutation iteration produced different sets of 80,000 unique clusters that were subsequently merged into a final dataset. Constraints were imposed on the set of bi-clusters so that they contained no more than 60% of common transcripts. This avoided redundant information since

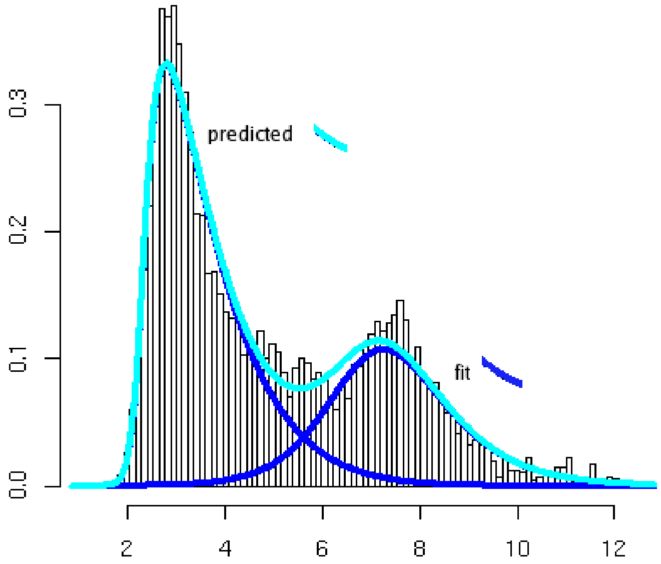


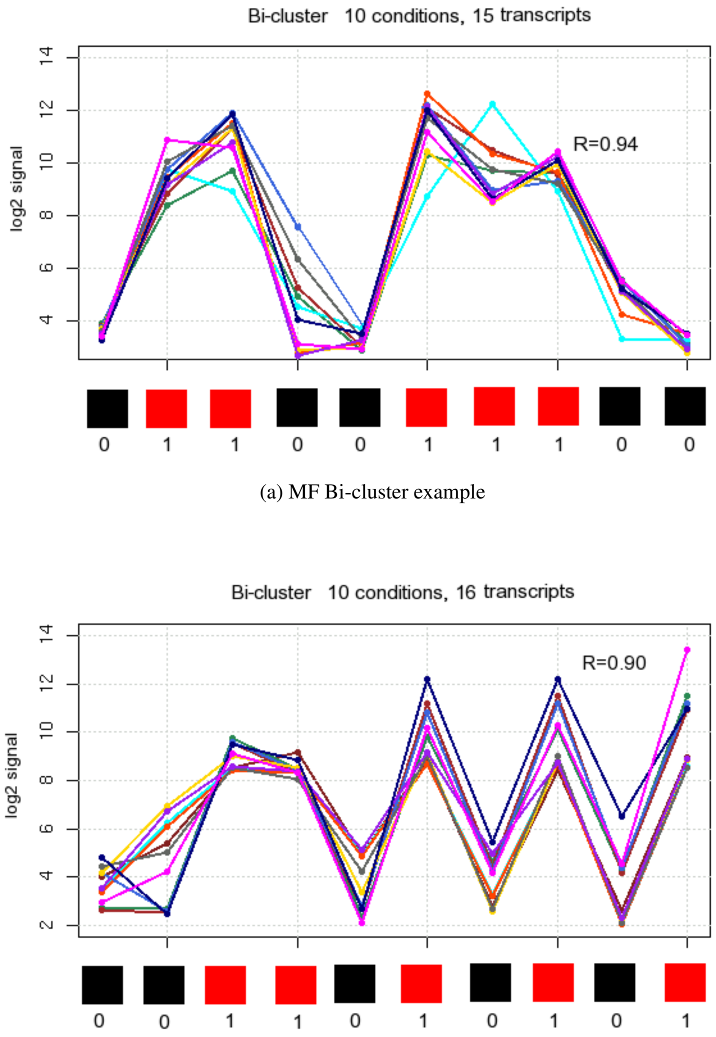
Figure 5.12: Example of the generalised lambda distribution model fit.

the feature representation did not use biological sample information. Whilst it could not be certain that this procedure sampled the entire set of bi-clusters, each transcript was at least a member of a single bi-cluster, and every sample condition appeared in at least one bi-cluster. In total 23,912 bi-clusters were produced.

One advantage of such a transparent approach in obtaining the bi-clusters is that novel biological knowledge could be generated. The two example bi-clusters (Figure 5.13) are highly correlated in just 10 sample conditions. The MF cluster comprises zinc binding sequences whilst the BP cluster contains transcription regulators. Both zinc binding and transcription regulation are popular annotation terms, however, the likelihood of observing these terms compared to a random model is significant at p-values 8.8e-137 and 1.80e-72 respectively based on the hyper-geometric distribution with p-value adjusted for 80,000 repeated samplings of the data. This demonstrates the power and efficiency of the bi-cluster approach in detecting functionally co-regulated modules. Additional knowledge can be gained by the fact that these sequences could be differentiated from other sequences with similar function by the pattern of experimental conditions which specified the bi-cluster.

The bi-clusters would not be detected using the global correlation co-efficient. Average correlations of 0.42 and 0.48 resulted from the transcript pairs measured over all sample conditions. This demonstrated the power of the bi-clustering technique in picking out conditions that determined functionally relevant clusters. Mapping the sample conditions back to the original data matrix identified the following conditions for the MF zinc binding cluster, colorectal adenocarcinoma, adipocyte,B-cell lymphoblasts, chronic lymphocytic leukaemia, coronary smooth muscle, blood post alcohol, umbilical vein, non-ischemic heart, skeletal muscle and peripheral blood dendritic cells. This result demonstrates the use of diverse experimental samples from both disease and normal tissues in generating functionally relevant patterns of expression. The sample conditions for the second cluster comprised PBMC's, bipolar brain sample, dorsal root ganglion, skeletal muscle from a patient with fascicular muscular dystrophy, blood post grape juice, liver, non-functioning pituitary, Rholo cord blood, right frontal lobe from brain tissue and T-Cell lymphocytic lymphoma. The biological significance and accuracy of these results warrants further investigation.

To prepare the bi-clusters for function prediction, pairs of transcripts were represented as a feature matrix where each feature corresponded to the correlation coefficient calculated using ex-



(a) MF Bi-cluster example

(b) BP Bi-cluster example

Figure 5.13: Co-expression profiles for sample bi-clusters

pression intensities from the sample conditions common to a particular bi-cluster. As additional features, the global correlation coefficient and Euclidean distance between pairs of transcripts measured over all sample conditions were determined.

To determine the value of these features in function prediction, Pearson correlation between the sum of bi-cluster correlations and function similarity was measured within the nr80 dataset. Values of 0.242 and 0.211 were obtained with MF and BP similarity respectively using the bi-cluster features. These measures represented a significant improvement over the relationship that could obtained with function similarity using the global correlation (0.101, 0.046) and Euclidean distance (0.083, 0.024) measures. This suggests the effective use of bi-cluster generated patterns for function prediction. However how much of this information is not represented by other feature sets remains to be determined.

5.3.7 Characterising feature relationships

In total, more than 49,000 feature vectors were computed for the regression modelling approach. The degree of overlap between features derived from different datasets was investigated, since combinations of correlated, overlapping features offers little additional value to machine learning approaches than using single data sources alone.

Multi-dimensional scaling (MDS) was performed using inter-feature distances to enable feature relationships to be visualised. The feature distances were taken as the absolute Pearson correlation between the sum of feature pair scores representing a data source and subtracting from 1. The resulting eigen-vectors were used to transform the feature distances to approximate feature relationships in three dimensions (Figure 5.11).

The features from the different data sources were spread out in the plot except for Localisation (LOC) and sequence similarity information (SW) which showed some correlation (0.211). This result was expected because highly similar sequences frequently co-localise, and the majority of localisation features comprise sequence motifs that are present among close homologues. All other features were far apart in the plot suggesting that they contained unique information that could be effectively combined to achieve greater accuracy in function prediction.

The position of each feature set in the plot was also influenced by the degree of overlap between

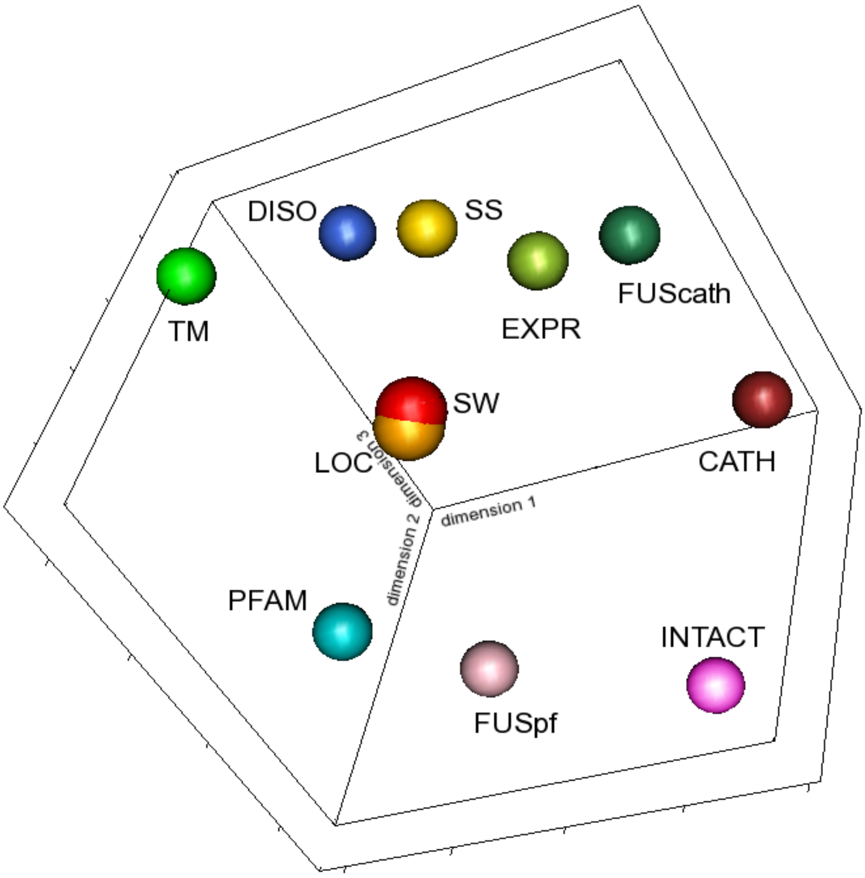


Figure 5.14: Visualisation of feature relationships in 3 dimensions. Features that are close together in the plot are relatively more highly correlated within a population of sequence pairs than those that are not. Features representing different data sources are represented as an averaged data point. the corresponding data source labels have been abbreviated to Sequence Similarity (SW), Localisation (LOC), Secondary Structure (SS), Transmembrane regions (TM), Disorder (DISO), PFAM family fusions (PFAMfus), CATH superfamily fusions (CATHfus), Expression (EXPR) and Protein Interactions (INTACT).

feature pairs from different data sources (Table 5.11). Sequence pairs represented by domain architecture and domain fusion features shared marginal similarity due to a population of sequence pairs that did not possess domain annotations. In the calculation of correlation between features, these values were assigned 0. It should also be noted that the use of Pearson correlation measures the linearity of relationships between feature sets. It is highly likely that in some cases, the relationships between features are non-linear, however, no single similarity measure between features is capable of capturing all the desired characteristics of the relationships which include ranking similarity, variability and non-linearity between feature scores, consequently linearity is assumed and the correlation measure reflects the magnitude of non-linear deviations between the feature scores discounting the absolute value of each feature.

Experimentally determined interaction (INTACT) and expression derived features (EXPR) supplied the fewest functional linkages whilst Localisation (LOC), secondary structure (SS) and disorder features (DISO) contributed the most. This is because the majority of sequences contain disordered stretches at either N or C termini and can therefore be aligned using the topology strings method. Sequence similarity features covered just one quarter of the total dataset using an E-value cut-off of 1000. The maximum overlap between feature pairs occurred between topology features whilst PFAM architecture and fusion data sources, and CATH and CATH fusion data sources were mutually exclusive. Just 23 sequence pairs were represented by every data source.

5.4 Chapter summary

With the aim of developing a machine learning approach that combines multiple sequence characteristics, for example, homologous relationships, domain architectures, expression patterns and protein interactions, sets of pairwise sequence features have been designed. Each dataset conveyed different information that could be used to determine function. Individual descriptors for localisation, experimentally determined protein interactions and disorder were barely linearly correlated with function similarity, however this does not preclude that when combined these features might interact co-operatively with one another to produce a greater overall correlation.

The Pearson correlation measures between individual features and function provide a rough guide to the strength of the feature in modelling function similarity. However, these values are not comparable between the different data source features because each feature set covers differ-

Table 5.11: Overlap between feature pairs.

Dataset	**SW**	**LOC**	**SS**	**TM**	**DISO**	**CATH**	**CATH$_{fus}$**	**PFAM**	**PFAM$_{fus}$**	**EXPR**	**INT**
SW	5,678,450										
LOC	5,080,934	23,946,660									
SS	5,654,927	23,854,091	26,701,482								
TM	1,185,179	4,660,987	5,158,493	5,165,493							
DISO	5,037,943	21,039,854	23,486,065	4,608,029	23,566,068						
CATH	198,809	222,748	243,066	113,320	193,905	243,136					
CATH$_{fus}$	204,717	256,373	287,989	46,032	257,545	0	288,071				
PFAM	948,019	4,102,761	4,551,402	929,599	3,934,392	93,916	93,916	4,558,248			
PFAM$_{fus}$	265,795	1,208,784	1,354,624	202,550	1,205,964	3,268	5,188	0	1,355,659		
EXPR	27,676	113,666	128,353	24,946	118,313	977	1,362	20,061	7,193	128,778	
INT	1,424	3,768	4,350	352	4,095	245	313	954	295	27	4357

The frequencies across the diagonal represent the number of sequence pairs that are represented by each data source in the nr80 dataset. The lower triangle details pair overlap between row and column data sources. Feature datasets have been abbreviated to Sequence Similarity (SW), Localisation (LOC), Secondary Structure (SS), Transmembrane regions (TM), Disorder (DISO), PFAM family fusions (PFAMfus), CATH superfamily fusions (CATHfus), Expression (EXPR) and Interaction (INT).

ent numbers of sequences. Additionally these values are biased when calculated over example annotation sets because most annotations have been made by homology methods. This is emphasize by the high correlation values obtained for CATH and PFAM domain architecture scores compared to known function similarity values. More than 40% of current human sequence annotations can be made using direct PFAM2GO mappings even where the evidence source of the original annotation is not a homology based method. Subject to these effects, the correlation values were not used to eliminate features, because combining multiple weak features might result in superior overall performance in the final prediction method.

The next challenge is to integrate all of the different features together to produce a high coverage, high accuracy function prediction method. The unique aspect of this study is the sheer volume of features to be combined, and optimisation of feature representation for the task of function prediction.

Chapter 6

Combining features for function prediction: Performance evaluation and benchmark

6.1 Introduction and aims

Few function prediction methods employ diverse data sources in order to make function assignments to sequences. Consequently, their applicability is restricted to a subset of sequences that can be represented by a particular data source. The problem is made more complex by the problem of semantic inequality between function descriptions. For example, functions that are generally descriptive such as “Receptor activity”, are not as important (that is semantically informative) as specific functions which might describe a detailed activity. To circumvent these problems, a method has been developed that integrates both sequence-based feature information from experimental and non-experimental data sources to predict the degree of function similarity between pairs of sequences. The function similarity measure accounts for this semantic inequality by weighting annotations according to their level of descriptive detail. The combination of diverse data sources using homology independent and homology-dependent information provides the means to infer functional relationships for any sequence regardless of its homology status. Thus the ultimate goal of the approach is to outperform annotation transfer made by homology searches.

In the previous Chapter (5) a function similarity measure was characterised for determining the degree of annotation similarity between sequence pairs. Features from different information sources, protein interactions, co-expression, and localization were designed for input to the method, and their relationships with function similarity characterised. Many of the features were shown individually to correlate weakly with function similarity. The effective combination of these features to predict function similarity therefore constitutes a particular challenge to be addressed in this chapter. Once combined, the performance of the method must be assessed both technically, to determine how well the models can reproduce known function similarities, and practically, with the goal of function annotation transfer accuracy in mind.

The major obstacle when combining the different features lies in combining effectively the weak features with stronger ones so that noise is not introduced into the system. There are several ways to perform feature integration including vector space integration, classifier integration and kernel methods (Noble and Ben-Hur 2008). In each approach, the integration step is performed using different information. For example, vector space integration operates at the raw feature level combining all the descriptors into a single model. Classifier integration approaches combine the results of different models to make a final prediction. Kernel methods (see Appendix I for more details) represent a more flexible approach for integrating data sources that do not require numeric vector representations.

Classifier integration techniques include committee approaches and boosting techniques as well as ensemble methods. Each classifier can represent a different model obtained using the same feature data, or use different features to model the same output. These integration techniques are appropriate for regression problems such as modelling function similarity as well as classification problems. The FFPred method described in Chapter 4 employed both vector space integration and classifier integration techniques. First the feature descriptors or vectors were used to generate a single classifier (vector space integration). Subsequently different classifiers trained using the same features (ensemble classifiers) were combined to produce a final result using a majority voting scheme (committee approach).

Two of the strategies are outlined in a flow chart (Figure 6.1). The first approach (1. vector space integration) combines all of the features generated from the different data sources into a single model. In the second model integration approach (2. regression model integration), a single regression model is developed specifically for each data source, and the models combined to learn the function similarity measure. By comparing the two approaches the best integration technique for the final method can be adopted. This comparison is performed by evaluating the ability of each model to reproduce predicted similarities similar to the known degree of function similarity between pairs of human sequences. The performance of each method is also judged using test datasets filtered at different homology levels (nr80 and nr35). The merits and complexity of each approach are discussed and the practical utility of the method is tested by making predictions for a set of uncharacterised human sequences.

Using the best prediction approach determined by evaluating the different integration strategies, a final function prediction method is produced. This method is then applied in a practical setting to the provide predicted annotations for a set of uncharacterised human sequences. An attempt to validate a set of high scoring predictions is made by reviewing the original feature information that forms the basis of the approach and by consulting independent information resources, Bioinformatic knowledge-bases and literature resources where possible.

6.2 Methods

The procedures for generating vector space (Figure 6.2a) and regression model integration (Figure 6.2b) approaches are outlined. In the vector space approach, all of the features were combined into a single regression model using function similarities as the response variable. A training example dataset was prepared comprising a representative sub-sample of sequence pairs. ϵ-sensitive Support Vector Regression (SVR) was used to build models from the training data. In the regression model integration approach a separate model of function similarity was produced describing the relationship between the feature descriptions and function similarity. Each training data set represented a sub-sample of the sequence pairs that could be represented by each data source. The sampling procedures, training and testing strategies are described in detail for the two approaches below.

6.2.1 Vector space integration

In total, 49,231 feature descriptions were produced for the vector integration technique. Sampling procedures were carried out on the pairwise sequence feature matrix for two reasons. First, the number of possible feature pairs for human sequences was so large (more than 4.2 billion pairs) that practical difficulties when handling all of the information were encountered. Second, the sampling permitted controlled reduction in bias resulting from the skewed distributions of function similarities (see Chapter 5 Section 2). The regression models could then be computed using a small subset of the data provided that it was representative of the entire sequence pair feature matrix.

The sampling procedure was carried out in two steps. Sequence pairs were randomly sampled

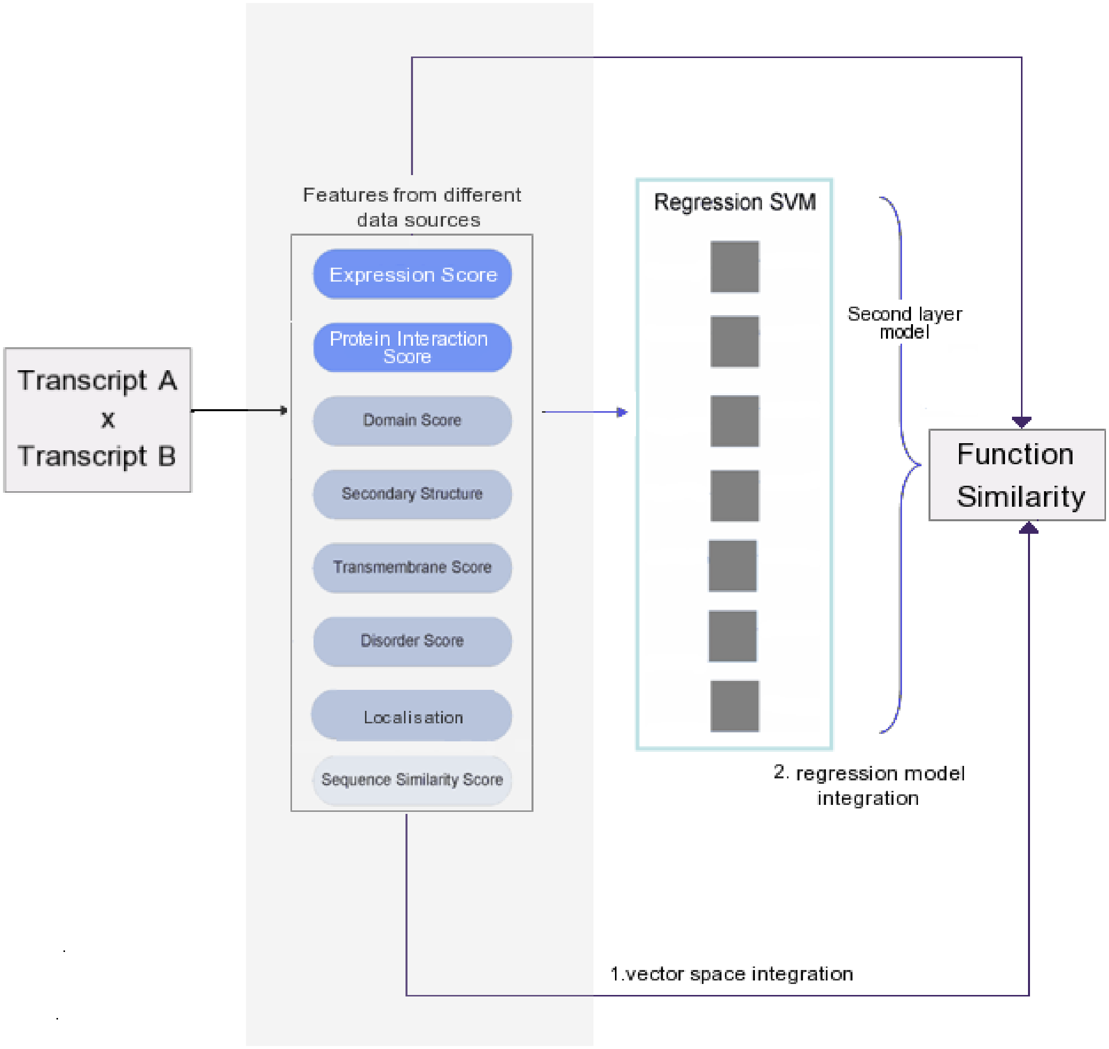
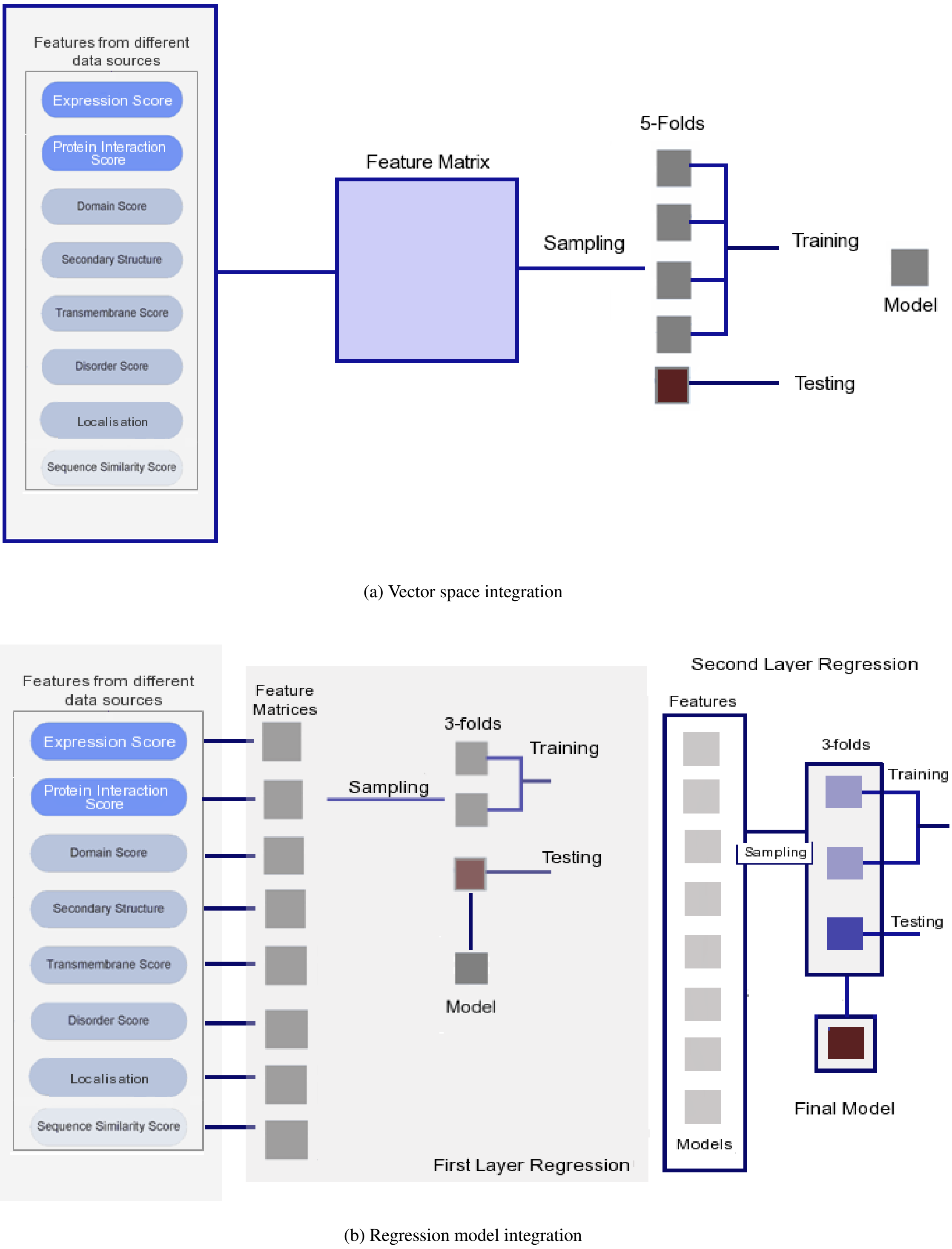


Figure 6.1: Integration strategies for function prediction. Both vector space integration and regression model integration paths through the chart are numbered. The vector space integration technique involves combining all features in one step from the different data sources. The regression model integration approach involves combining the outputs of different models; one specific to each data source into a second layer model.



(a) Vector space integration

(b) Regression model integration

Figure 6.2: In the vector space approach, the feature matrix comprises all of the feature descriptions from each data source. The matrix undergoes sampling and subsequently 5-fold cross validation is performed to produce a final model. In the regression model integration approach, a feature matrix is constructed for each data source. Sampling is carried out on each matrix and 3-fold cross validation performed to produce each data source model. The model outputs are then used as feature inputs to a second round of regression modelling to produce the final model.

so that there were no more than 10,000 pairs represented by identical features regardless of their value. Secondly, the pairs were filtered according to Manhatten distance ($d_{XY} = \sum(|x - y|)$) between feature values with a threshold of 0.02. This threshold was deemed sufficient to maintain feature diversity within the training sets whilst significantly reducing their size. For the calculation sparse matrix elements were replaced with 0 values. The absolute difference between function similarity values was constrained to 0.05 for values < 0.8 and 0.01 for values ≥ 0.8. These steps ensured that too many highly similar sequence pairs were not discarded since these data represented the most important part of the regression.

Sequence pairs from the nr80 evidence source balanced data set were used for training. After sampling, this training matrix was further divided into 5 independent subsets by randomly selecting sequence pairs. The resulting unique data partitions (Table 6.1) were further filtered to ensure that feature pairs representing high function similarity values were over-represented compared to those that represented low function similarity values (Figure 6.3). Specifically, the frequency of values in the 0-0.05 range of function similarities could be no more than 80% of the frequency of the 0.95-1.0 range of function similarities. This ensured that the regression fits were deliberately skewed towards high function similarity values which represented the most important parts of the distribution as well as being the most reliable data for model fitting.

6.2.2 Data source integration method

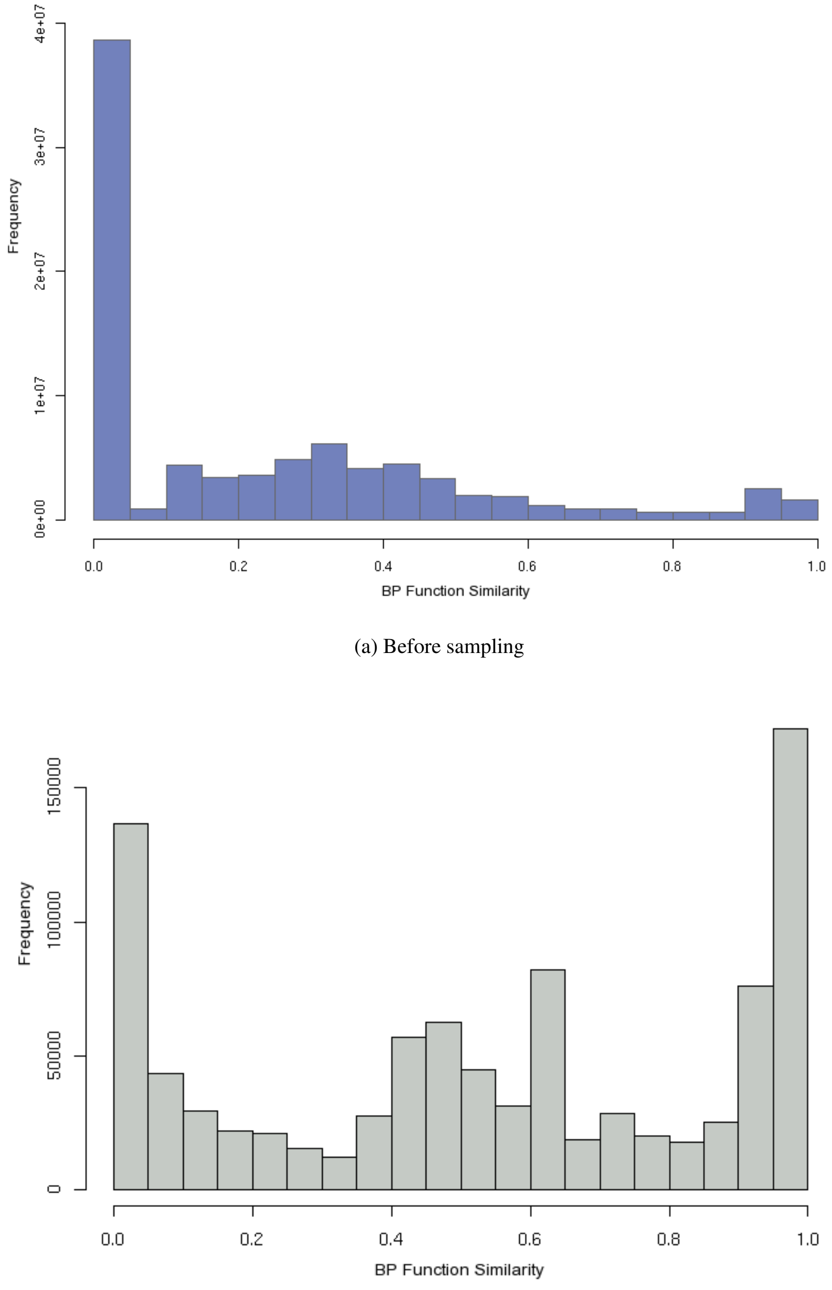
In this approach, an independent regression model was created using the descriptive features in Chapter 5 that were designed from the raw data representing each data source. A second level modelling step was then carried out to combine the model outputs. This approach is more flexible than the vector space integration because new data sources can be added without perturbing the existing regression models. The technique also permits different kernel functions to be used for each data source.

Again sampling was carried out on each independent data source feature matrix using two separate criteria. For the sparse matrices representing CATH superfamilies, PFAM families, protein interactions and fusions, random sampling was carried out on each feature matrix of sequence pairs ensuring that no more than 1,000 pairs were represented by common features. This figure was reduced to 500 for the protein interaction dataset due to its small size. Full matrices were sampled using the Manhatten distance method with the same thresholds of 0.01 and 0.05 applied

Table 6.1: Resulting data set sizes after the sampling procedure was carried out.

Dataset	**MF Size**	**BP Size**
1	361,856	425,901
2	361,939	442,227
3	361,906	440,936
4	361,873	414,623
5	361,928	425,948

Vector integration: training dataset sizes. Each count represents the number of sequence pairs present in each of 5 folds after sampling. The 5 folds were used for training and cross-validation.



(a) Before sampling

(b) After sampling

Figure 6.3: The distributions of BP similarity before and after the sampling procedure was applied.

to the difference between function similarity values above and below 0.8 respectively. These data were then partitioned into 3 independent training datasets per data source and the resulting function similarity distributions filtered (Table 6.2).

6.2.3 Support Vector Regression training

Epsilon sensitive Support Vector Regression (SVR) was used to create models combining the different feature inputs. Two regressions were carried out per training dataset for MF and BP function similarity measures respectively. Fold cross validation experiments were performed to select the best cost parameter (C) and width parameter (ϵ) for regression. These parameters influence the number of support vectors in different ways. The width parameter is related to the degree of noise in the relationships between features and regression target. A smaller epsilon value implies a tighter fit to the data, typically resulting in a solution with more support vectors. The C parameter controls the cost of errors on the examples. Generally, higher C values result in fewer support vectors. The optimal model was considered the one which produced the greatest correlation with function similarity in the test dataset. The test data sets comprised the fold partitions that were not used in the training procedure.

The linear kernel was used to train the vector space models. For the data source regression models, three different kernels were used, the linear kernel, spline kernel and radial basis function kernel. Training runs using the RBF (Radial Basis Function) kernel required an extra parameter γ to be tuned whereas the spline and linear kernels required only the width and cost parameters to be optimised. Different kernels performed best on different data sources (Table 6.3). For very large and sparse feature matrices with little overlap between features, the linear kernel only was used because the effect of feature interactions would be minimal, and the kernel generalised on a solution in minutes to hours rather than days.

Model quality was evaluated using the correlation coefficient between predicted function similarity and actual function similarity using all pairs of sequences from each data source. The better performing kernels were frequently RBF or linear kernels (bold highlight Table 6.3) selected by their greater correlation with known function similarities. To determine the strength of relationship between each data source and function similarity, smoothed plots were made using comparing predicted to known function similarities (Figures 6.4 and 6.5). The degree of noise present in each model fit is also represented by the shaded area. These fits could be used to infer

Table 6.2: Training dataset sizes for different data sources.

Dataset	**MF Size**			**BP Size**		
SW	27,631	28,221	27,167	33,179	33,468	32,898
LOC	367,832	368,211	322,491	407,821	407,029	407,141
SS	469,340	468,298	469,911	510,000	508,264	510,151
DISO	431,830	430,791	430,206	451,438	449,367	445,367
TM	416,858	415,899	415,823	83,716	83,721	83,416
PFAM	30,010	30,192	29,864	31,321	32,616	31,995
PFAMfus	2,389,558	2,367,451	2,372,906	3,127,963	3,124,913	3,124,186
CATH	33,413	32,981	33,544	33,397	34,291	32,656
CATHfus	627,244	626,131	627,453	837,839	836,992	837,147
EXPRS	210,869	201,197	211,679	264,318	272,888	268,888
INTACT	511	507	520	695	675	681

Data sources have been abbreviated to SW (sequence similarity), LOC (localisation), SS (secondary structure), DISO (disorder), TM (transmembrane), PFAMfus (PFAM fusion information), CATHfus (CATH fusion information), EXPRS (expression information) and INTACT for protein-protein interactions. PFAM and CATH represent protein family and structural domain features. Each count represents the total number of sequence pairs that were used for training in each of the 3 training folds.

Table 6.3: Training results for independent data source models, and vector space model.

Dataset	Kernel	Parameters	MFcor	Stdev.	Parameters	BPcor	Stdev.
SW	linear	C 80, ϵ 0.05	**0.432**	0.03	C 80, ϵ 0.10	**0.112**	0.06
	spline	C 1, ϵ 0.001	0.412	0.05	C 2, ϵ 0.05	0.061	0.01
	rbf	C 12, ϵ 0.1, γ 2000	0.431	0.06	C 10, ϵ 0.05, γ 150	0.093	0.01
LOC	linear	C 4, ϵ 0.05	0.174	0.01	C 110, ϵ 0.2	0.132	0.01
	spline	C 1, ϵ 0.05	0.171	0.02	C 2, ϵ 0.1	0.129	0.02
	rbf	C 10, ϵ 0.1, γ 8	**0.175**	0.01	C 0.1, ϵ 0.1 γ, 12	**0.133**	0.01
SS	linear	C 1e-4, ϵ 0.1	0.211	0.02	C 1, ϵ 0.2	**0.121**	0.02
	spline	C 1, ϵ 0.05	0.224	0.01	C 0.001, ϵ 0.3	0.101	0.01
	rbf	C 0.5, ϵ 0.05, γ 10	**0.267**	0.03	C 0.5, ϵ, 0.2 γ 100	0.114	0.02
DISO	linear	C 15, ϵ 0.1	0.265	0.03	C 0.15, ϵ 0.3	0.177	0.03
	spline	C 1, ϵ 0.5	**0.281**	0.02	C 0.01, ϵ 0.2	**0.241**	0.02
	rbf	C 200, ϵ 0.1, γ 100	0.198	0.02	C 1000, ϵ 0.1, γ 0.5	0.156	0.02
TM	linear	C 0.1, ϵ	0.278	0.02	C 50, ϵ 0.1	0.198	0.01
	spline	C 0.01, ϵ 0.05	**0.314**	0.02	C 0.1, ϵ 0.4	0.243	0.02
	rbf	C 10, ϵ 0.10, γ 0.7	0.309	0.01	C 10, ϵ 0.10, γ 0.7	**0.243**	0.01
PFAM	linear	C 15, ϵ 0.1	0.779	0.02	C 12, ϵ 0.1	0.757	0.02
PFAM fus	linear	C 0.01, ϵ 0.3	0.459	0.01	C 0.01, ϵ 0.1	0.112	0.01
CATH	linear	C 100, ϵ 0.2	0.666	0.01	C 10, ϵ 0.2	0.671	0.01
CATH fus	linear	C 0.1, ϵ 0.2	0.421	0.01	C 0.1, ϵ 0.1	0.370	0.02
EXPRS	linear	C 1.34, ϵ 0.1	0.327	0.01	C 0.01, ϵ 0.1	0.367	0.01
INTACT	linear	C 1.5, ϵ 0.1	0.217	0.01	C 1.6, ϵ 0.1	0.051	0.01
ALL MODEL	linear	C 0.1,ϵ 0.2	0.368	0.03	C 0.1, ϵ 0.1	0.248	0.02

The correlation performance represents the average Pearson's correlation values for different training folds tested on nr80 sequence pairs present in each data source. The ALLMODEL results were computed on the entire nr80 sequence pair matrix. Note that the different sized test sets means that these values are only comparable between different kernels within the same dataset. The best result for each dataset is highlighted in bold. Data sources have been abbreviated to SW (sequence similarity), LOC (localisation), SS (secondary structure), DISO (disorder), TM (transmembrane), PFAMfus (PFAM fusion information), CATHfus (CATH fusion information), EXPRS (expression information) and INTACT for protein-protein interactions.

model quality and differentiate between stronger and weaker data sources in the approach.

All of the relationships between observed and predicted MF similarity exhibited non-linearity apart from those representing Disorder and Expression features. The strongest data sources were Sequence Similarity (SW), Localisation (LOC) and PFAM families (PFAM) based on small standard deviations around the central relationship and a greater density of predicted high function similarity values compared to other data source (Figure 6.4). The relationships between observed and predicted BP similarity were also non-linear for all data source models except sequence similarity (SW), secondary structure (SecStr), transmembrane (Transmem), Disorder and expression (EXPR) information (Figure 6.5). This observation suggests the limited use of general homology-based features in predicting BP similarity. Prediction models for Localisation (LOC), PFAM family (PFAM), and PFAM and CATH fusion data sources (PFAM_FUS and CATH_FUS) produced the most informative BP similarity scores with the smallest standard deviations to the fits.

Compared to other features, Disorder features seemed of little value in predicting MF and BP similarities producing a flat, linear relationship with known function similarities. However, this result could be obtained because the predicted function similarity for Disorder features are only useful in a very small proportion of cases and the signal becomes lost when averaging over the entire nr80 dataset. Additionally, these fits are made to observed function similarities which despite being carefully balanced for evidence sources and filtered for specific annotations, are still heavily populated by incomplete annotations.

The predicted function similarities from each of the 11 data source models were then used as 'complex feature' inputs to a second layer regression integration.

6.2.4 Integrating complex features

The complex feature pair matrix was sampled to produce 3 independent partitions for training and cross-validation. Single features representing each data source were added successively one at a time until all data source features were combined into a single model. The integration was performed sequentially using the sequence similarity score as the first feature in order to demonstrate the performance improvement that could be obtained beyond this baseline by including different features. To assess model quality the degree of successful annotation transfers obtained

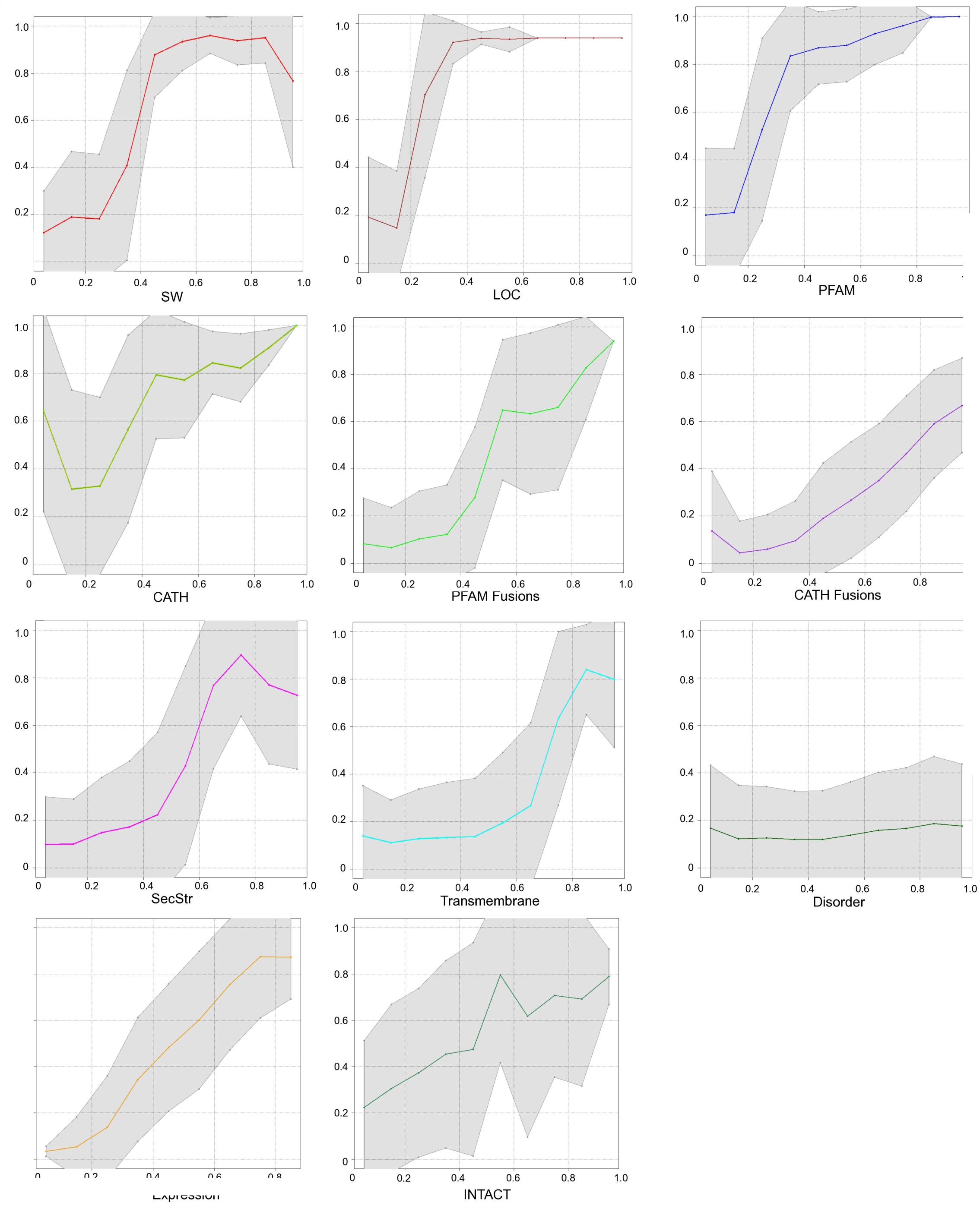


Figure 6.4: Relationships between predicted function similarity (y-axis) and observed MF similarity (x-axis). Each data point represents the average observed function similarity over successive intervals (0.1) of MF similarity. The shaded area of each plot represents the region bounded by the standard deviation at each point. Note that using expression similarity measures, the number of data items present in the last interval was < 5, consequently this category was merged with values from the previous interval.

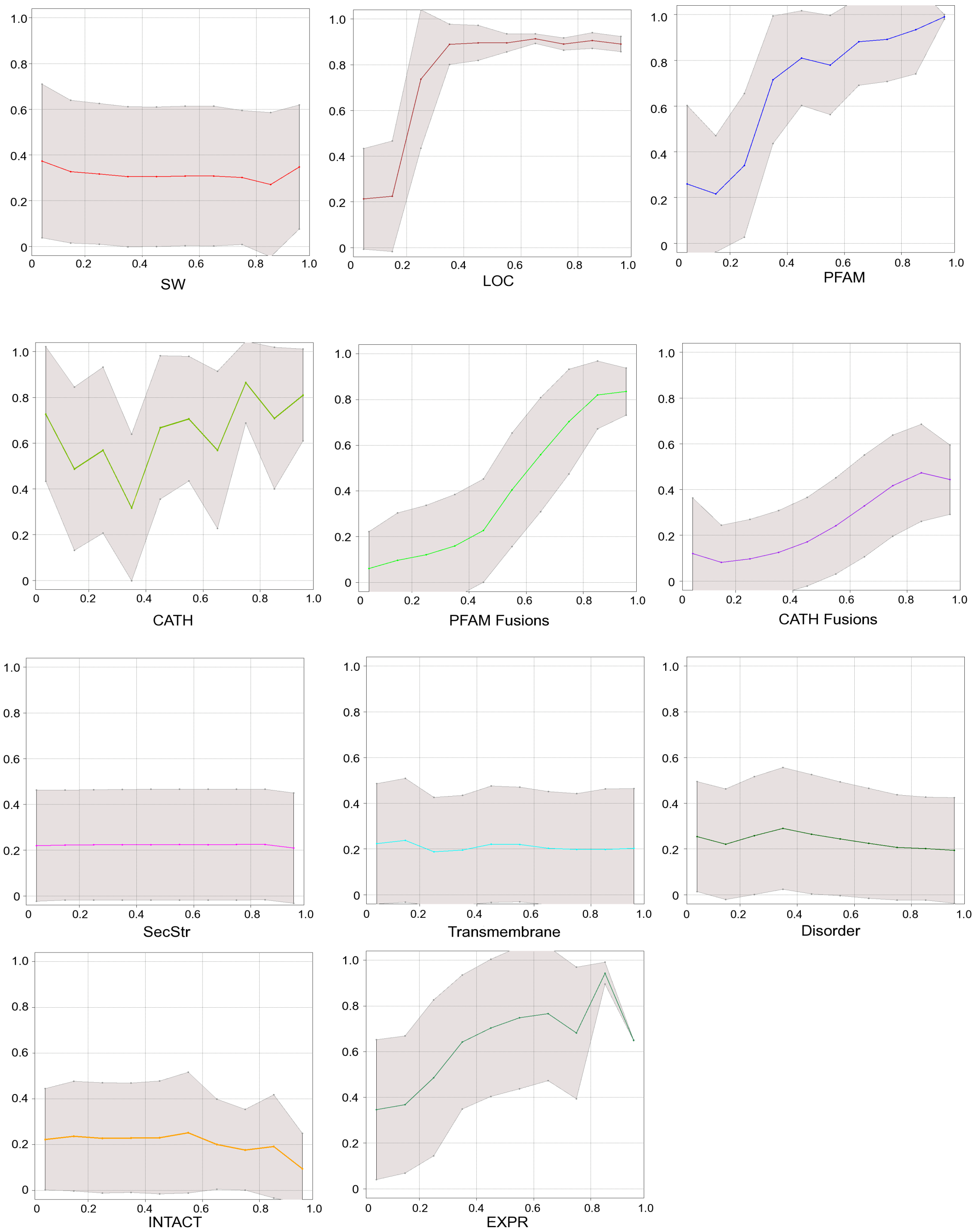


Figure 6.5: Relationships between predicted function similarity (y-axis) and observed BP similarity (x-axis). Each data point represents the average observed BP similarity over successive intervals (0.1) of predicted function similarity. The shaded region represents the area bounded by the standard deviation associated with each data point.

by sliding a threshold over predicted function similarities was measured using the nr80 dataset sequence pairs. Each GO annotation transfer that could be made between a query and target sequence was assigned a match state (1) if the GO terms were equivalent or the target GO term was a parent of the query term. Otherwise the target annotation was considered a potential false positive (0). Annotation transfer was propagated to the root annotation term in each GO graph during scoring so that all possible annotation matches were considered. Model quality was judged using aAUC (actual Area Under Curve) measured between prediction output and function similarity (Table 6.4). The actual area under the curve measures the number of data items captured under the curve in the ROC like plot. It is more appropriate as an unbiased performance statistic than the standard AUC measure, which is a proportion, to compare between classification methods when tests are conducted using datasets of varied sizes. Larger values indicate better performance, and perfect classification is indicated by an aAUC value equal to the product of positive and negative test cases. Group-wise Pearson's correlation was also computed between predicted and actual function similarities for each set of sequence pairs grouped by common sequence query. ROC like curves were also plotted reflecting actual true and false positives (Figure 6.6) to show the nature of performance improvement when data sources were included in the models. This was necessary since AUC measures reflect overall model quality at all frequencies of true and false positives whereas visualisation of the ROC like curves highlighted cases where addition of a particular data source to a model might increase performance specifically at either low or high false positive rates.

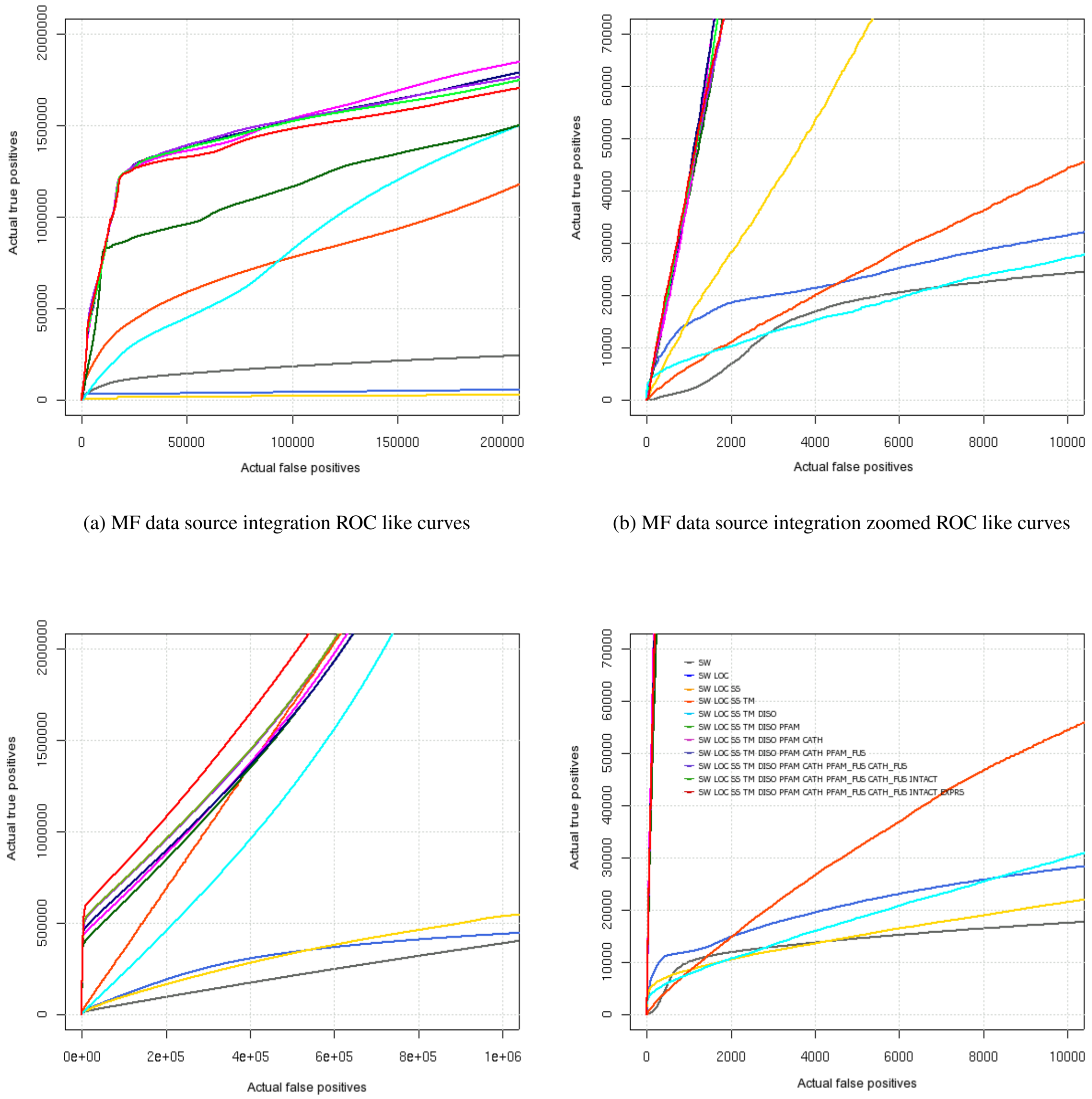
Addition of different features affected the models in different ways. For example, a five feature model of MF similarity (SW+LOC+SS+TM+DISO) was more sensitive at high false positive rates than the four feature model (Figure 6.4a). At low false positive rates the opposite was observed and the four feature model outperformed the five feature model. This effect could partly be explained by the resolution and range of the particular feature score which presented the most useful information for function prediction. The Disorder feature score varied between 0.3 and 0.75 and did not provide good resolution between highly functionally similar sequences compared to other feature scores. Clearly by comparison to annotation transfers made using Smith Waterman sequence similarity measures (SW), using an approach combining all features together was superior.

For the purpose of examining the value of different feature sets, this assessment provided useful information, but by no means was exhaustive. However, producing models for all combinations

Table 6.4: Model prediction performance

Model	MF aAUC	MFcor	BP aAUC	BPcor
All feat				
SW	4.69e+11	-0.122	6.42e+11	0.112
SW+LOC	2.36e+12	-0.112	6.58e+11	0.129
SW+LOC+SS	2.38e+12	0.080	6.32e+12	0.103
SW+LOC+SS+TM	3.36e+12	0.155	6.39e+12	0.105
SW+LOC+SS+TM+DISO	3.58e+12	0.229	6.43e+12	0.124
SW+LOC+SS+TM+DISO+PFAM	3.62e+12	0.282	6.51e+12	0.126
SW+LOC+SS+TM+DISO+PFAM+CATH	3.64e+12	0.285	6.55e+12	0.145
SW+LOC+SS+TM+DISO+PFAM+CATH+PFAMfus	3.61e+12	0.289	6.62e+12	0.150
SW+LOC+SS+TM+DISO+PFAM+CATH+PFAMfus+CATHfus	**3.63e+12**	**0.289**	6.69e+12	0.151
SW+LOC+SS+TM+DISO+PFAM+CATH+PFAMfus+CATHfus+INTACT	3.45e+12	0.279	6.74e+12	0.153
SW+LOC+SS+TM+DISO+PFAM+CATH+PFAMfus+CATHfus+INTACT+EXPRS	3.57e+12	0.269	**2.35e+13**	**0.155**

aAUC values represent the absolute area under curve calculated using the actual true and false positive values to a threshold of 1e+6 false positives. This statistic permits comparison of the different models with different sized test datasets. The best performing models are highlighted in bold. Data sources have been abbreviated to SW (sequence similarity), LOC (localisation), SS (secondary structure), DISO (disorder), TM (transmembrane), PFAMfus (PFAM fusion information), CATHfus (CATH fusion information), EXPRS (expression information) and INTACT for protein-protein interactions.



(a) MF data source integration ROC like curves

(b) MF data source integration zoomed ROC like curves

(c) BP data source integration ROC like curves

(d) BP data source integration zoomed ROC curves

Figure 6.6: Integration performance adding a single complex feature data source at a time. The data sources have been abbreviated to SW (Sequence similarity), LOC (Localisation), SS (Secondary structure), TM (Transmembrane), DISO (Disorder), PFAM, CATH, PFAMfus (PFAM fusion information), CATHfus (CATH fusion information), EXPRS (Expression information) and INTACT (Interaction information).

of 11 features was impractical in computational terms given time constraints and necessary parameter optimisation. The ordering of features combinations was selected to incorporate the simplest homology related features first, followed by orthogonal, experimentally derived information. This strategy ensured that the added value of a feature set beyond performance obtained by sequence similarity annotation transfer methods could be realised.

The best predictions of BP similarity could be obtained using all features, whereas for MF similarity, there was little to be gained by adding in domain fusion (PFAM_FUS and CATH_FUS) and expression (EXPRS) and interaction information (INTACT). To determine the significance of the differences between group-wise correlation coefficients, Fisher's significance of correlation difference test (Equation 3.7) was applied to the correlation values using the number of groups (5694) as the sample size. The results of these tests suggest that overall, the difference between the 9 feature MF model (0.289) and the 11 feature model (0.269) was not significant (z=1.13, p=0.130) therefore all features were retained for the combined data source modelling approach.

Five integration methods were subsequently performed on the data matrix comprising 11 complex features. A simple maximum feature pair score rule (MAXfus), the average majority pair score rule (AVEfus), the sum of pairwise feature scores (SUMfus), linear SVR model and RBF SVR models. The MAXfus, AVEfus and SUMfus rules represent simple methods of combining features to produce a single function similarity score assuming an equal contribution between each feature set and the similarity score. Theoretically, the linear and RBF kernel methods should outperform the simpler score combinations as they permit different weights to be assigned to different features. In the case of the RBF regression, the non-linear relationships between features and function could be correctly handled.

6.3 Results and model application

Several performance comparisons were made between the different models to illustrate both technical and practical aspects of model quality in successful annotation transfer. For the technical assessment, the agreement of the predicted function similarities was compared to known function similarity between pairs of sequences. In the practical assessment, the degree of successful annotation transfers was determined using the predicted function similarity score. The distinctions between technical and practical qualities were necessary because a technically competent model can perform poorly in the GO annotations transfer test due to the presence of incomplete annotations in the test dataset.

6.3.1 Technical performance of the different model integration approaches

The goal of the technical model quality assessment was to determine how well each model had learnt the function similarity measure. The ability of each model to produce values closest to known function similarities was determined using rank correlation statistics for sequence pairs from nr80 and nr35 datasets. The number of functionally nearest neighbours recovered in the top sequence hit, top 5, and top 10 neighbours was also reported (Table 6.5). The expanded neighbourhood analysis was important to determine how well the method performed in ranking an unknown sequence against multiple relatives with common function.

The distinction between rank correlation measure and number of correct nearest neighbours was necessary because the number of functionally identical neighbour sequences is related to the popularity of each annotation class. Frequently occurring annotations could potentially yield at least 100 neighbour terms that were functionally equivalent (ranked equal 1st) by observed function similarity, but not by predicted function similarity. This technique can result in low correlation coefficients when the method has performed well.

Overall, the results indicate that the predicted function similarities are good indicators of known function similarity. In 72.12% and 84.04% of sequence pairs annotated by MF and BP terms respectively, the nearest neighbour sequence determined by predicted function similarity is a nearest neighbour by observed function similarity (Table 6.5). In the nr35 dataset, the equivalent measures were 60.69% and 74.46% suggesting that each method performs less well when sequences are not closely related.

When the neighbourhood size was increased from a single neighbour to 5 or 10 members, the proportion of nearest functional neighbours decreased where MF annotations were considered, yet increased where BP annotations were considered. This result may be influenced by the fact that MF annotations tended to be more specific and complete than BP annotations, therefore the total population of sequences sharing identical annotations is smaller than for BP annotations. Equally, this may indicate that the method is better suited to grouping sets of sequences by common BP annotation category than by MF annotation category. This finding is consistent with results in Chapter 3 where annotation transfer by sequence similarity appeared to be better conserved for BP categories within species than for MF's.

The correlation between observed and predicted MF similarity measures (Table 6.5) was higher than for BP similarity. This suggests that the method was more successful, that is more information was present in the features that reflected the properties of MF similarity than BP similarity.

The RBF data source integration model produced a greater proportion of top hit annotations that were nearest neighbours than other methods. It also produced the best correlation between predicted and observed function similarities. The FUSsum method performed similarly, though not as well as the linear data source integration method indicating that the assignment of weights to the different features was an effective method of predicting function similarity. The vector space model outperformed the linear (LIN) and RBF kernels when more distantly related sequences were used. This might be a result of the compression of more than 49000 features into 11 features in the data source integration approach. However, it might also be related to the different sampling steps producing fewer training examples in the data source models than the vector space model. Performance on the nr80 dataset however, was better for the RBF data source model.

Increasing the number of nearest neighbours did not improve the recognition of MF annotations, perhaps due to the smaller numbers of sequence pairs with identical functions. Equally it may be that the method identifies nearest neighbour sequences that are correct but are not identified as such because the annotations do not exist in the GOA dataset. The top 5 nearest neighbours provided a greater enrichment of functionally similar sequence neighbours in the BP nr80 dataset. This may reflect the fact that BP annotations are less complete than MF annotations, thus greater numbers of sequence pairs can be recovered with identical function similarity scores. However, in the nr35 dataset this trend was not observed. Overall, the results suggest that to achieve the most accurate automated annotation transfers, the nearest neighbour approach is a good choice since there is a strong likelihood that sequence pairs will share similar functionality.

6.3.2 Practical assessment of model quality in annotation transfer

Whilst the technical assessment was useful in establishing the best feature integration modelling approach, it did not provide useful information regarding the expected error rates when using

Table 6.5: Technical model quality assessment results

MODEL	**nr80 cor**	**nr35 cor**	**top**	**top5**	**top10**
Molecular Function					
Vector space	0.41	0.31	70.87	50.09	49.22
MAXfus	0.39	0.22	66.18	51.34	46.60
SUMfus	0.46	0.29	69.21	54.24	46.32
AVEfus	0.44	0.25	64.21	58.31	44.10
LIN	0.47	0.34	71.04	61.10	56.35
RBF	**0.52**	**0.31**	**72.46**	**63.18**	**56.92**
Biological Process					
Vector space	0.23	0.18	83.61	84.21	84.19
MAXfus	0.19	0.12	79.88	78.40	78.41
SUMfus	0.23	0.17	80.13	85.29	81.54
AVEfus	0.21	0.14	72.21	80.98	72.21
LIN	0.25	0.20	84.04	89.37	81.08
RBF	**0.29**	**0.24**	**85.94**	**90.45**	**82.85**

The vector space model represents the performance of the linear kernel trained using 49231 features. The MAXfus, SUMfus and AVEfus models are data source integration models computed by taking the maximum, sum or average of each of the data source features. LIN and RBF both represent the linear and radial basis kernels trained using individual data source features. The correlation statistic represents Spearman's rank correlation coefficient. Cases where the maximum function similarity score between sequence pairs was 0 were excluded from the comparison to avoid positively skewing the statistics. The left and right hand values in the nearest neighbour columns are the proportion of actual nearest neighbour sequence pairs recovered in the set of predicted top, top5 and top10 ranked sequence pairs.

the predicted function similarities to make annotation assignments in practice. A benchmark performance of the models in annotation transfer was therefore carried out using both nr80 and nr35 datasets. ROC-like curves were produced by recording frequencies of true and potential false positives by applying different thresholds to the predicted function similarity score. In this assessment sequences annotated to general terms (less than 3 levels from the root term) were excluded from scoring, and annotations were only propagated to the third level of each GO graph hierarchy. This strict criterion ensured that the performance curves reflected the ability to transfer detailed annotation descriptions. Similar to Section 6.3.1, true positive assignments were made to annotation transfers where a query annotation was a child term of the target annotation sequence. Potential false positives represented cases where no relationship existed between known and predicted annotations.

Performance was assessed using all sequence pairs and the closest neighbour sequence only. This distinction permitted the performance assessment to reflect common practice in high-throughput automated annotation where as long as annotations can be transferred between closest sequence neighbours, then the method is considered successful. The ROC curve assessments using all sequence pairs provide useful information about the discriminatory capacity of the predicted function similarity scores in making function assignments to sequences. Whilst these performance statistics do not provide a definitive assessment of the methods in annotation performance because false positives may represent novel and correct annotations, they represent the ability of each method to reproduce known assignments in a standard and controlled test environment.

The ROC curves (see Figures 6.7 and 6.8) showed that the RBF kernel outperformed all other methods (magenta and blue series) when performance was assessed using the nr80 dataset, although the improvement was not significant when transferring MF annotations (Figure 6.7a)). However, when a similar assessment was made using nr35 sequence pairs, the vector space integration method outperformed the other methods when MF annotations were transferred. This result might be a consequence of information loss resulting from compression of the features into single values in the data source models, however the same trend was not observed when the equivalent comparison was made between BP similarity models.

The MAX fus, AVE fus and SUM fus approaches reflect performance that could be effectively obtained by unsupervised clustering algorithms to combine data sources. These rules produced similar quality MF predictions to the SVR methods when using the nr80 but not nr35 dataset for testing. In both cases, lower quality predictions than the SVR method were made when considering BP annotations. This result justifies the use of the supervised approach to obtain weights to optimally combine the different data sources.

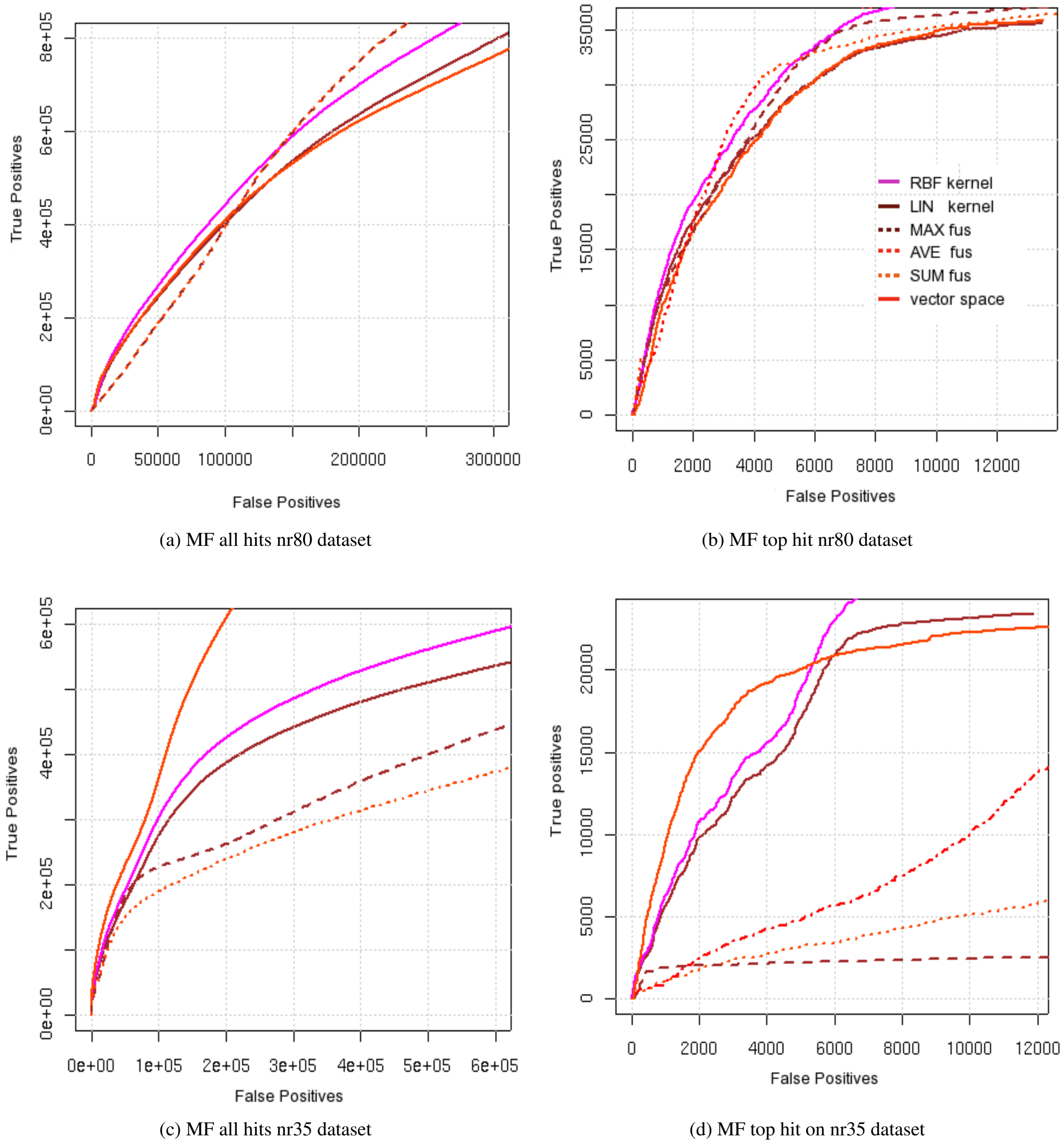
Overall, MF category assignments were more accurate than BP category assignments, however the drop in performance observed between nr80 and nr35 datasets was emphasized when MF categories rather than BP categories were transferred. This observation is in agreement with the hypothesis that the determinants of BPs are not as heavily biased towards homology features as MFs.

Despite high confidence that the closest neighbour sequence will be the most functionally similar using both MF and BP similarity models, performance in annotation transfer using the similarity scores did not reflect this. For example, using nearest neighbour sequences only, 15000 MF annotation assignments and 9052 BP assignments could be made at an error rate of 10%. This result was not unexpected since so few of the annotations are complete for human sequences, and highlights the problems of functional multiplicity and true negative assignments in function classification. As an example, performance of and annotations could be obtained at 10% error rate using the same score criteria. This sub-optimal performance results from the transfer of multiple annotations to sequences using a single score. Because the score is a composite measure of different GO annotation similarities, it cannot applied to transfer all annotations between sequences in the same way. To overcome this problem it might be more appropriate to adjust the score according to the annotation assignment in question.

The RBF models produced the best annotation transfer results between pairs of sequences according to the ROC curves. These models also produced the best estimates of function similarity. A possible reason for this result is that the RBF kernel is non-linear and permits modelling of feature interaction effects that cannot be represented using a linear sum (linear kernel). Practically, this means that co-occurring weaker features can be strengthened in combination with one another, rather than treating each data source as an independent predictor of function similarity.

6.3.3 Establishing the value of different data sources

Useful information regarding the value of different biological features from each data source could be determined by investigating the optimised RBF regression models. For example, little



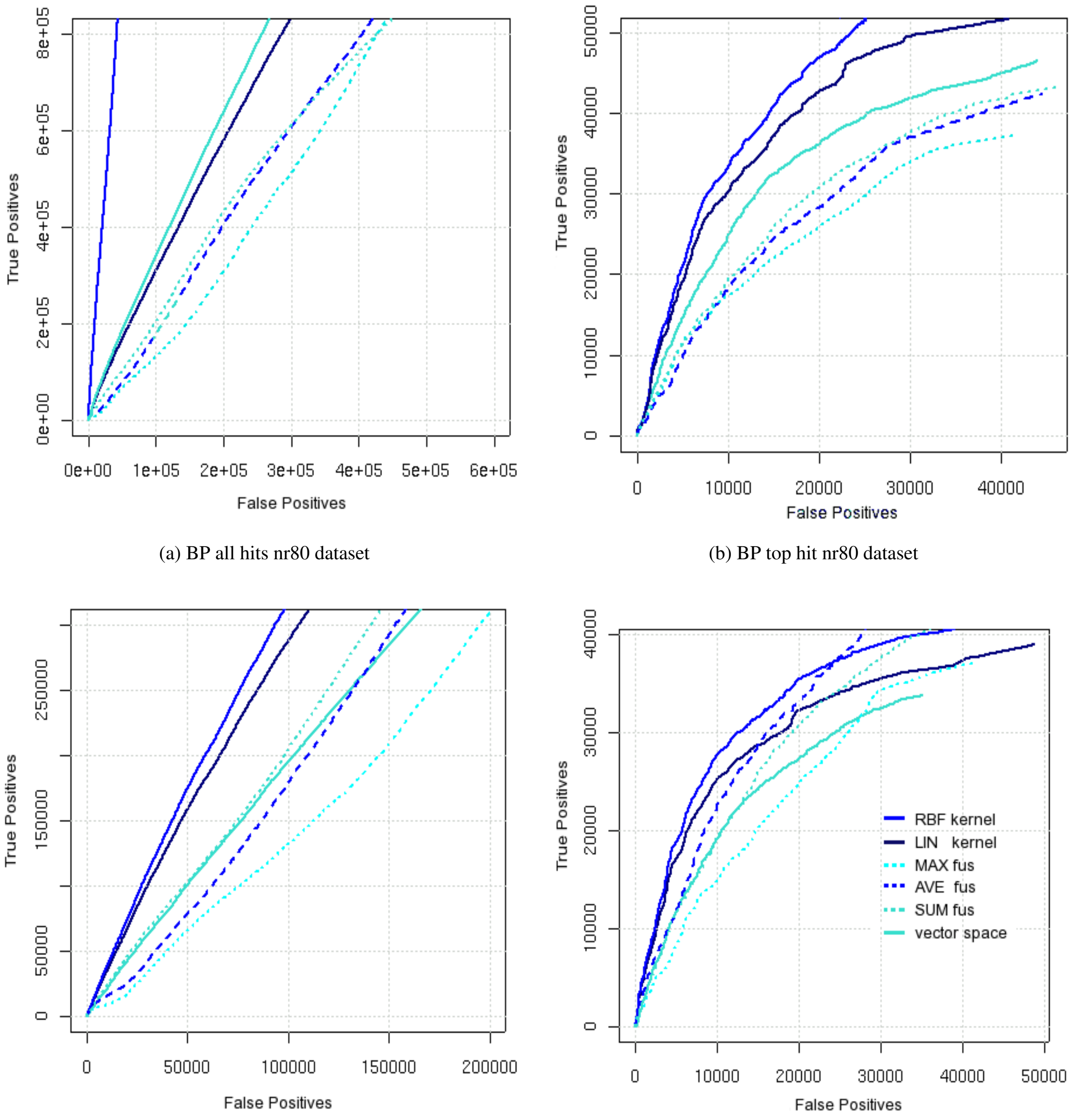
(a) MF all hits nr80 dataset

(b) MF top hit nr80 dataset

(c) MF all hits nr35 dataset

(d) MF top hit on nr35 dataset

Figure 6.7: MF similarity model performance.



(a) BP all hits nr80 dataset

(b) BP top hit nr80 dataset

(c) BP all hits nr35 dataset

(d) BP top hit on nr35 dataset

Figure 6.8: BP similarity model performance

information is gained by incorporating information from data source features that are numerically unique, but present no unique information to the final prediction approach.

The value of each data source in function similarity prediction was determined by iteratively retraining the model removing a single feature at a time. The reduction in Spearman's correlation co-efficient between predicted and actual function similarity was reported during the feature elimination test to estimate the value of the data sources to the model.

The proportion of the maximum correlation between MF and BP similarity achieved using all features represented the unique contribution of each data source to the data source integration model (Figure 6.9). This feature elimination technique accounts for feature interaction effects as well as redundant contributions between similar features.

The removal of PFAM, CATH, sequence similarity (SW) and secondary structure features (SS) resulted in the greatest performance loss between predicted and actual MF similarity. These results correlated with the value of single features since they also provided the strongest relationships alone with MF similarity (Pearson correlation 0.64). Localisation (LOC), CATH fusion features (CATH_FUS) and interaction features (INTACT) resulted in the greatest performance loss when predicting BP similarity. This result is in contrast to the single feature contributions which suggest that PFAM, CATH and fusion information was the most valuable. The correlation between single feature performance and feature loss between the BP data sources was 0.31. This disparity between single feature strength alone and when omitted from the best model suggests that there is more feature interaction in the BP model than in the MF model of function similarity. The weaker features from the topology information and expression information seem to be more powerful when combined with other features resulting in a greater loss when removed from the RBF model, whereas the contributions from the stronger features (PFAM) could be compensated by other features in the model.

These results seem to reflect a common finding throughout this thesis; that MFs tend to be better recognised by homologous relationships between sequences, whereas BPs require conservation of a diverse range of weak signals that are less frequently represented by sequence similarity relationships.

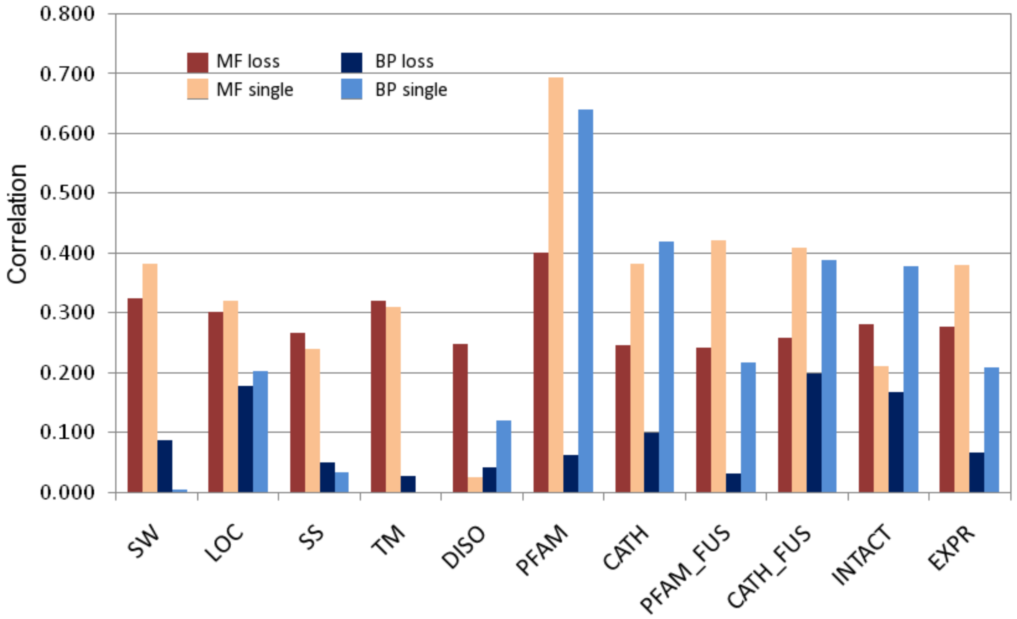


Figure 6.9: Contribution of data sources in function similarity prediction. The dark bars represent the proportional loss in correlation between predicted and MF similarity. The light bars represent the correlation performance of single features. Data sources have been abbreviated to SW (sequence similarity), LOC (localisation), SS (secondary structure), TM (transmembrane topology), DISO (disorder), PFAM_FUS (PFAM fusions), CATH_FUS (CATH fusions), INTACT (interactions) and EXPR (expression information).

6.3.4 Annotation of uncharacterised human sequences

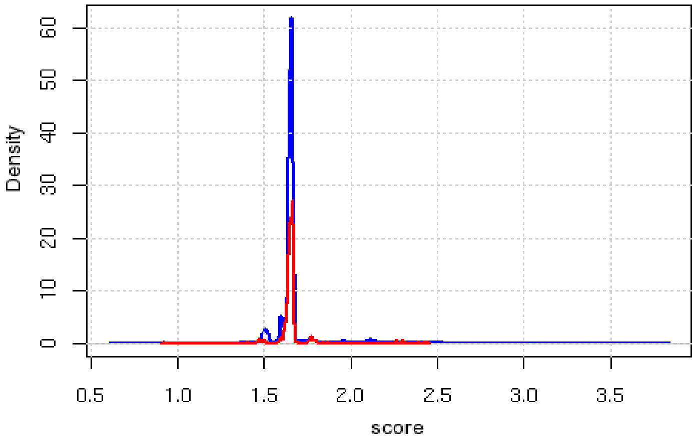
The RBF models of function similarity were used to annotate a set of 11669 uncharacterised IPI sequences. These comprised a mixture of close homologues whose functions could be inferred by standard annotation transfer practices, sequences with homologous relationships to annotated sequences but whose annotation might require manual intervention, and harder targets that displayed no significant sequence similarity with annotated sequences in the GOA UniProt set. Annotations for many members of these sets could be found in other bioinformatic resources, for example SwissProt or Ensembl records, but not in the IPI version used here.

Before making predictions using the function similarity models, posterior probabilities were determined for transfer of individual GO classes given a particular score. This step was considered necessary due to a lack of heterogeneity between score distributions obtained for different functions (see Figure 6.10). For example, GO term GO:0006355, "DNA-dependent regulation of transcription", the positive examples heavily overlapped with the negative examples, whereas the distributions were cleanly separated for the MF term GO:0001584 "Rhodopsin-like GPCR". This justifies the need for annotation class dependent probability distributions specific to each GO term to make the highest quality annotation assignments.

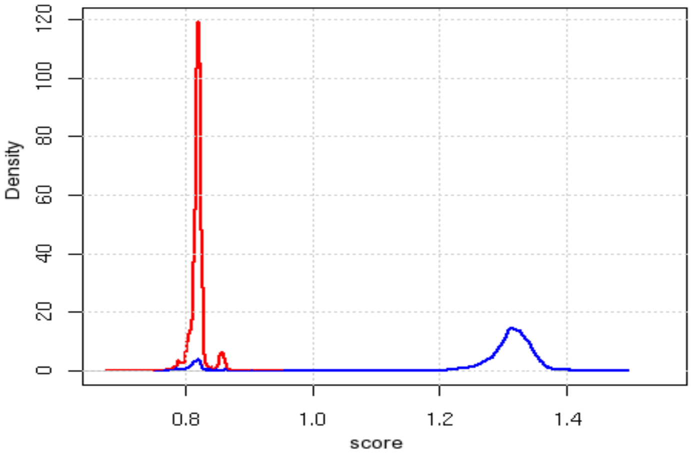
The probabilities were calculated from a modified score incorporating a weight for the neighbour rank of the score. This adjustment was introduced to up-weight the strength of sequence relationships that were closest by rank to the query sequence based on the results in Section 6.3.1 showing that more than 70% of nearest neighbour sequences were the actual closest relative in terms of function annotation. The rank weighted scores corresponded to the exponential of the negative log of the scaled rank ($w = \exp(-srank)$). The scaled rank was fixed so that ranks of more than 50 were fixed at 50, and so that ranks of 1 became 1e-6.

Individual parameters were estimated separately for each GO term using non-linear least squares fitting (nls function) in the R stats package (R Development Core Team 2009). In cases where fewer than 10 positive or negative assignments were made to a given GO term in the nr80 dataset, these fits could not be confidently computed. For such GO terms, information was borrowed from other terms by pooling a sample of 5 GO terms that produced a conservative probability distribution.

The scoring method was applied to the dataset of uncharacterised IPI human sequences. Pairwise



(a) GO:0006355 Regulation of transcription



(b) GO:0001584 Rhodopsin-like GPCR

Figure 6.10: Estimating probability distributions for GO term annotation transfer. The score distribution of known positive annotations is shown in blue for comparison with the potentially negative examples shown in red.

sequence relationships were computed between the unannotated sequences against a library comprising 22307 specific MF and 21177 BP annotated sequences. The novel sequence annotations were further sub-classified into 3 groups, easy, medium and hard. The easy group corresponds to a mixture of sequences for which an annotated homologue could be found either in human or other species at $\geq$ 80% identity. These cases are considered trivial and may represent splice variants, incomplete gene sequences or cases where the GOA annotations are lagging behind information contained within other annotation datasets. The medium set contains sequences that can be annotated using homologous relationships at E-values < 0.001. In these cases, annotation by homology transfer is questionable, however, some functional knowledge can be obtained by the use of homologous sequence relationships. The classification of 'hard' refers to sequences for which no closely related annotated homologues can be determined, thus automated annotation transfer becomes difficult without manual intervention. This class contained 4543 sequences.

Applying a probability threshold of 0.001 to the annotation transfers and excluding general annotations (those 1 or two levels from the root annotation classes) produced predicted annotations for 20,667 uncharacterised sequences. The annotations could be subdivided into the three different classes, easy medium and hard (Figure 6.11). MF annotations could be predicted for a large proportion of the easy cases (79.86%) and for smaller proportions of medium and easy targets (82.24% and 63.91% respectively). The proportion of annotation assignments that could be made to BP classes was much lower across all target classes and coverage was similar regardless of the target sequence status. This result corresponds with previous observations in Chapter 3 that conservation of BPs in human sequences is not highly correlated with sequence similarity.

A subset of the novel function predictions were selected for validation using literature searching and manual curation. The set comprised a sample of 10 annotations with high scores and annotation probabilities of < 0.001 from the MF and BP hard targets sets. Sequences that were annotated as fragments or were less than 30 amino acids long were excluded from this set. Supporting evidence for each prediction was compiled by consulting the original feature information and conducting searches of bioinformatic databases for additional evidence. Some of the cases were trivial. The annotation could be confirmed by consulting the equivalent entry in one of the bioinformatic knowledge bases, Swissprot or Ensembl for example. However these annotations were not present in the IPI database version used in this study. In other cases, the annotations were completely novel and required experimental validation. For 7 out of the 10 sequences, a case could be made to support the GO term predictions using external reference information. In

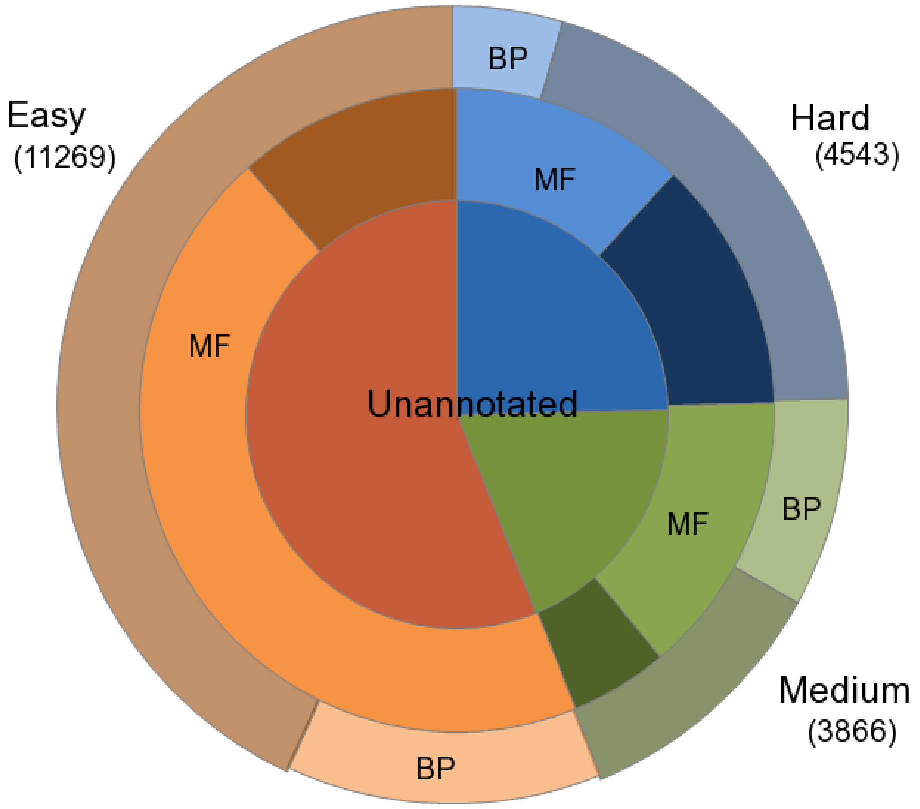


Figure 6.11: Annotation coverage pie chart. The smaller central pie chart reflects the numbers of sequences belonging to each target class. The dark areas of the outer pies represent the proportion of unannotated sequences whilst the lighter portions reflect the proportion of annotated sequences using a p-value threshold of 0.001.

the remaining 3 cases, no additional data could be obtained to support the predictions.

In order to provide a balanced view of the method some potential false positive assignments are highlighted. These occurrences present a significant challenge for automated annotation transfer methods using any data source and should be avoided where possible because a single wrong assignment can lead to rapid propagation of errors across all biological knowledge bases. There are two particular types of error encountered using this approach, the first where the original annotation assignment is a rare instance of an incorrect annotation for a particular GO term class. This leads to high probabilities of correct annotation transfers when the predicted function similarity to the sequence is high. A second type of false positive, and less dangerous, is the case where feature annotations in the primary data are incorrect from a data source that is important (has a high contribution) to the final function similarity.

One such example is present in Table 6.6 where an uncharacterised sequence IPI00514093 receives a weakly predicted PFAM domain (PF00859 'PseudoUSynth2' at p-value 0.0042). Sequences containing this domain modify uracil in RNA molecules and receive the GO annotation term GO:0001522 "pseudouridine synthesis" . Because PFAM families are strong indicators of function in the model, the uncharacterised sequence is linked to the synthase enzymes with high scores, and a high probability of annotation assignment ($p < 1.28e-05$). Close inspection of this low complexity sequence suggests the annotation may have occurred by chance since the synthase sequences also contain low complexity regions. These types of error can be avoided by manual inspection of the primary features responsible for the function linkages, or by applying a more conservative filter to the primary feature information at source.

6.4 Chapter Discussion

Tackling the challenge of integrating diverse feature characteristics from different data sources has proved fruitful in developing a method to produce high quality function assignments for the human proteome. The approach developed here is more accurate than the FFPred feature based method, and than predictions that can be made by using sequence similarity methods. This new method is also more flexible than most competitor function prediction methods since it is designed in a modular manner and can be easily updated through the addition of new data sources or re-evaluated as new function annotations become available.

Table 6.6: Example sequence annotations

Accession	Class	Description	Prediction	Evidence
IPI00418595	Medium	Transmembrane protein	GO:0016627 oxidoreductase activity, acting on the CH-CH donors GO:0046872 metal ion binding GO:0043169 cation binding GO:0016125 sterol metabolism GO:0006629 lipid metabolism	Annotation confirmed in Swissprot: Q6ZNB7 fatty acid hydroxylase, 7 transmembrane regions
IPI00444938	Easy	Disks homologue	GO:0016301 kinase activity GO:0016773 phosphotransferase, contains PDX domain known to bind ATP GO:0017076 purine nucleotide binding GO:0019205 nucleobase, nucleoside,nucleotide kinase GO:0019261 1,4-dichlorobenzene catabolic process GO:0055088 lipid homeostasis GO:0055092 sterol homeostasis	Guanylate kinase annotation in Swissprot: Q5H9Q4 Annotation confirmed in reference PMID: 8909548 Swissprot: Q15700
IPI00395779	Hard	C20orf27	GO:0016301 kinase activity GO:0017076 purine nucleotide binding	No supporting evidence
IPI00375576	Hard	FLJ42001	GO:0016472 metal ion binding GO:0003700 transcription factor GO:0008270 zinc ion binding	No supporting evidence
IPI00747209	-		GO:0017171 serine hydrolase GO:0004175 endopeptidase activity GO:0004252 serine type peptidase	Annotation confirmed in Swissprot: Q59H04
IPI00400897	Hard	FLJ27365	GO:0016866 intramolecular transferase GO:0016616 oxidoreductase, acting on the CH-OH group of donors, NAD or NADP as acceptor	weak sequence similarity to penicillin acylase (BLAST)
IPI00783233	Easy	RING finger protein	GO:0046872 metal ion binding GO:0008270 zinc ion binding	Annotation confirmed in Swissprot: Q969Q1
IPI00514093	Hard	SLAIN1 homologue novel stem cell gene	GO:0016866 intramolecular transferase GO:0003676 nucleic acid binding GO:0001522 pseudouridine synthase	No supporting evidence
IPI00644456	Easy	FGD4 protein	GO:0046578 regulation of Ras protein signal transduction GO:0030695 small GTPase regulator GO:0005085 guanyl-nucleotide exchange factor	Annotation confirmed in Swissprot Q49A55
IPI00657778	Hard	- 28 kDa protein	GO:0008624 induction of apoptosis by extracellular signals GO:0016481 oxidoreductase activity GO:0046872 metal ion binding	No supporting evidence

Overall, the most successful annotations were produced using RBF kernel regression to combine predicted function similarity scores representing each data source. The results from this regression comprised the final predicted MF and BP function similarities. Comparing between all sequence pairs, the method was able to reproduce function similarities with correlation values of 0.38 and 0.29 for MF and BP Ontologies respectively. These values suggest that significant improvements can be made to the method perhaps through the use of more and diverse features. However, some limitations to the method are inevitably imposed by data quality and the ability to evaluate novel predictions.

The addition of new data sources can be achieved independently by creating a model specifically describing the relationship between new features and function similarity. As new high throughout genomic and proteomic technologies produce a wealth of data describing transcription factor binding sites and protein-DNA and protein-protein interactions, the information can be easily converted into features for inclusion into the model. In the current implementation, many sequence pairs can only be sparsely represented by secondary structure and/or localisation information, a proportion of which display common function. These relationships can be strengthened by including these new and homology independent data sources into the approach.

Overall, the performance of the method is encouraging. Specific annotation predictions could be made for a large proportion of unannotated sequences with and without annotated homologues in other species. The approach can therefore be applied to any sequence regardless of its homology status. Confidence in the ability of the model to reproduce known sequence pair rankings according to well established function similarity measures is high since more that 78% of functionally nearest neighbours could be recovered in the benchmark assessment. However, the successful transfer of annotations between these functionally related sequences remains challenging. The most straightforward approach to solving this problem was adopted here by using prior information to compute posterior probabilities of annotation transfer conditional on each annotation category. However, in the light of the fact that available function annotations are incomplete, this solution is not ideal since overly conservative or tolerant estimates can lead to the dangerous situation of error-ridden annotation propagation. Extra resolution between functionally similar sequences will increasingly become apparent as new annotations become available, and the quality of existing annotations increases.

This method currently operates on a single species (human) which is restrictive where functions are not well represented in the human genome. By statistical chance alone, the likelihood of detecting a same function relationship is greatly improved where the proportion of sequences displaying that function are high. Although some functions are better transferred by within-species comparisons, a further extension to the work might include generating features using relationships between sequences from different species. This task represents a significant undertaking considering the volume of annotated sequences in other species, and the required data pre-processing and feature design steps. In particular the integration of experimental information between species, where the use of different platform technologies are used in different experimental set-ups can produce highly variable data. In addition, the definition of functionally equivalent sequences between species can be problematic, introducing inaccuracies that might lead to inappropriate information transfer.

The potential for improving prediction accuracies by borrowing information from orthologues is undoubtedly great since extra confidence can be obtained from the existence of multiple highly similar sequence pairings with consistent annotation. However, a further concern is that the evolutionary distance at which same functions between species are preserved is unlikely to be a constant. In particular Jensen et al. (2006) have shown that expression behaviour for sequences with regulatory functions is only conserved within primates. This is in contrast to core homeostatic functions, which tend to be highly conserved in sequence, and in experimentally defined behaviour across most eukaryotes. As a machine learning problem, this might be well posed as an adaptive, or on-line approach, where the information used in modelling, or the type of modelling approach is adapted according to the nature of the test case.

Chapter 7

Discussion

Function prediction perspective: current status and future prospects

In this post-genomic era most of the genome sequences of the component parts of model organisms, their genes and proteins are known. The major focus both in biology and in the development of computational methods is in understanding the complex and subtle interplay between components that govern cellular responses and ultimately an organism's behaviour. A first task in this challenge requires a catalogue of the functions of the component genes and proteins in order to better understand how the pieces of the puzzle might co-exist to elicit physiological responses. In the computational world this translates to a major focus in the area of function prediction. The complex nature of this challenge cannot be underestimated since for many uncharacterised genes and proteins, the only available information is protein sequence or regulatory signals in DNA.

There are two fundamental problems in computational function prediction, the more obvious task of making accurate assignments of function to sequences, and uncovering possible causative mechanisms for particular functionalities. Most automated function prediction methods attempt to provide probable answers to the first of these questions but frequently generate associations that might provide useful insight into the second. However, even in the light of current biological knowledge, the task of predicting function from sequence remains extremely difficult. The major bottleneck centres around the acquisition of current function annotations. These are predominantly sourced from homology-based annotation transfers used in an automated fashion to assign functions to uncharacterised genome sequences. Whilst these methods are successful at providing some degree of specific annotation to sequences, they are not universally applicable across all of sequence and function space (see Chapter 3). The majority of these assignments comprise obvious annotation cases, and therefore recycle current knowledge rather than generating new information that can be used to further our understanding of the relationship between sequence and function.

The effects of this recycling are evident in the curation process for function annotation schemes. At present more than 70% of human GO annotations are made using homology based annotation assignments, demonstrated in Chapter 2. This becomes a problem for function prediction methods which attempt to learn the complex nature of relationships between sequence and function because the information that remains when sequence similarity is effectively masked out is a rather sparse representation of functionally diverse sequences. Pattern recognition algorithms used for prediction then struggle because the patterns are rarely present in sufficient quantities to be recognised. In fact, it is difficult for any function prediction approach to yield significant improvements over simple homology-based methods because the signal from homologues is over-represented in any sizeable sample of annotated sequences. Although other accurate methods of assigning function to sequences exist beyond homology based annotation transfer, they predominantly comprise low throughput experimental methods that cannot produce similar volumes of information at controlled accuracies. Until progress is made in developing accurate high throughput experimental technologies, the current computation challenge remains the integration of weak and noisy information to predict function.

The definitions of function have been effectively captured and organised in the form of machine readable ontologies. These ontologies permit multiple levels of description to be assigned to a given sequence, thus unifying different biological concepts of function in a single data structure. The flexibility of inheritance supported by interlinking the annotations in this system is desirable, however it creates problems for function prediction methods concerning the definition of specific annotation descriptions and their equivalence. For example, any function prediction can be considered correct at some level of specificity if the logic of an Ontology is followed by propagating annotation categories to their highest common ancestor. For example, the annotation “Molecular Function” may be inherited from any low level enzyme or binding annotation. It is therefore impossible to compare the quality of published annotation methods whose goal is to make Gene Ontology category assignments to sequences. Favourable performance statistics are frequently quoted over a test dataset without detailing the nature of the annotation assignment, whether specific or more general. The majority of measured statistics comprise averages computed over a whole range of annotation categories of different specificities. Thus the true accuracy of these methods is concealed since the value of a correct assignment to a specific annotation category is clearly greater than that of a rare annotation category.

Despite the problems encountered when trying to predict annotation classes within ontologies, there are clear advantages in the use of ontologies. They facilitate a cross disciplinary merging of language used to describe function. For example, an enzyme's precise role may be ultimately defined by the chemistry involved in a specific catalytic reaction. In developmental biology, the role of a sequence might be sufficiently described by the term "limb generation". These annotation descriptions, and therefore the two fields are unified by a common unit, the gene product or protein component.

The move away from well structured static hierarchies for function definitions towards flexible graphs of annotation relationships has been well received throughout the biological and biomedical communities. Indeed this is evident in the number of new descriptive ontologies that have appeared, largely through the collaborative efforts of community wide researchers (Bodenreider and Stevens 2006). Example ontologies include the Disease Ontology (Du et al. 2009), the Human and Mammalian Phenotype ontologies (Robinson et al. 2008, Smith et al. 2005, Tasan et al. 2008), the Cardiovascular Gene Ontology (Lovering et al. 2008) and Plant-Associated Microbe Gene Ontology (PAMGO) (Torto-Alalibo et al. 2009). The emergence of these systems re-iterates a pressing need to develop fast, generic solutions for computational prediction of annotations from sequence. The methodology employed to predict GO annotations here is both applicable to any sequence and any Ontology provided that sufficient information exists for machine learning to be carried out.

In Chapter 2, the structure of GO and current status of the human GO annotation assignments was reviewed. This study revealed that although annotation assignments are made to more than 58% of available sequences, just 2% and 8% of the GO class assignments were at their most specific. This suggests that numerous estimates of the annotation status of the human genome as 'approaching completeness' are enthusiastic, and that our ability to make performance assessments of GO class prediction tools is limited by this lack of information. This problem will undoubtedly be encountered when undertaking predictive modelling of any of the ontologies mentioned above. It is therefore important by design that prediction methods should be responsive to new information with minimal tuning for updates, and that criteria for performance testing clearly account for the specificity of annotation assignments that can be made.

Supervised machine learning methods to tackle this problem have been the focus in this work. In particular, SVMs were selected because they are adept at handling noisy and high dimensional pattern recognition problems. However, their performance is determined by the quality of information used to learn these patterns. Practically this is a consequence of the amount of consistent and diverse example annotations made to sequences. Common to most bioinformatic prediction problems, and especially true in function prediction, is the fact that the assignment of a true negative is always ambiguous. The result that a protein does not bind ligand in a functional assay may not be interpreted as evidence supporting the absence of this role. It is more likely that the experimental conditions under which such an interaction might occur have not been encountered. The implications for supervised machine learning approaches are that supposedly negative example cases are penalised during the learning phase, reducing the ability of the method to detect patterns. If the conditions of learning are relaxed such that these ambiguous cases can be tolerated, an approach may be viewed as inaccurate when assessing performance on labelled test cases. The consequences for regression models are similar since the ill-defined negatives simply add a significant noise component to the fitting procedure.

Techniques to avoid this problem include transductive or semi-supervised learning where the labelling of an example, positive or negative, is carried out during the learning phase. In function prediction, this would require significant manual intervention to determine the confidence level of true negative assignments. Too many unlabelled training examples could result in a loss of sensitivity of the prediction approach if functionally similar sequences became labelled as functional equivalents. Fundamentally, the problem of assessing model quality in the light of current knowledge remains unsolved because novel predictions cannot be independently evaluated.

Unsupervised techniques, for example clustering approaches, present an interesting alternative because the underlying structure of the biological information used to make a prediction remains unperturbed. As new function information is acquired, groupings of sequences according to biological features can simply be revised rather than having to rebuild and retrain complex models. However, a significant downside to the unsupervised methodology is a lack of power to differentiate functionally useful information from noise contained within descriptions of biological characteristics. For example, in the experimentally determined protein interaction data included in the SVR feature based approach, weights were assigned to each experimental method that could be interpreted as the reliability of this information in predicting function. In fact protein interactions sourced from yeast-two-hybrid data, a notoriously noisy protocol, barely contributed to the relationship between protein interactions and function. Unsupervised clustering approaches do not make use of this information. However, the trade-off is that more false positives are likely to be encountered and error rates cannot be tightly controlled.

In spite of the problems resulting from sparsely available knowledge of function, a successful function prediction method was developed using supervised machine learning. The use of ϵ sensitive regression support vector machines seemed particularly appropriate for handling large amounts of noisy feature information that could be obtained from various experimental data sources. Non-linear kernel functions (RBF's) were adept at defining and handling complex inter- and intra-feature relationships. Ultimately the method was capable of identifying functionally similar sequences with good accuracy, which inspires confidence in the approach. It is believed that significant improvements to the method can only be made through the addition of new and function information-rich biological data sources.

As elegantly pointed out in Sadowski and Jones (2009), the future of function prediction depends on improvements in function definitions, identification of positional and non-positional indicators of function and the ability to provide a definitive dataset of completed function annotations. This first point implies clarity and stability within the GO graph structure, together with consistent annotation assignments to sequences between species and across bioinformatic knowledge-bases. Currently, annotation terms may be nominated by the scientific community providing their existence can be justified and approved by expert curators. Between major releases, these changes coupled with the retirement or merging of existing annotation classes can dramatically alter the GO Graphs and cause problems for those function prediction approaches that implement inheritance to infer higher GO class memberships. They also affect those using function similarity measures that exploit positional information from the GO categories in the graph to define a local common ancestor term with which to score semantic similarity.

The capacity to make GO term function predictions in the FFPred approach and the ability to computationally model function similarity were limited by the underlying data. In fact, the models of function similarity were capable of identifying the correct functionally nearest sequence neighbour in more than 80% of test cases. This suggests that appropriate algorithms exist with which to build effective function prediction approaches with, despite the lack of available and high quality training example information. It is the ability to determine accurate estimates of model quality, and assess their performance on new data that is lacking.

The problem of identifying positional and non-positional indicators of function follows the theme of modularity in protein function, an increasingly popular concept arising in systems biology (Dani and Sainis 2007). These modules can be thought of as units of functional inheritance. Specifically, a functional module is considered a single, or set of characteristics of genes and proteins that are sufficient and necessary for function. Historically, the structural domain has been coined a unit of functional inheritance (Lee et al. 2007, Moult and Melamud 2000, Todd et al. 1999). More recently, the entire domain architecture of a sequence has been shown to correspond more closely with function (Krishnamurthy et al. 2007, Lee et al. 2005). However, these concepts concentrate on the presence of structure whereas the absence of clearly defined secondary structure within sequences, the presence of disorder, has also been inextricably linked to the correct functioning of some proteins (Dunker et al. 2008a, Tompa et al. 2005). Thus it seems reasonable to assume that the modular units of functional inheritance are non-uniform throughout sequences.

Perhaps during evolution, the properties of sequences which are retained are a mixture of convenience coupled with selection of those necessary elements with which to perform function. These may be specified according to the available materials (amino acids), or adapted from source materials through mutation. This hypothesis is supported by the fact that functions can arise from the existence of a few catalytic residues (a positional indicator of function) within a structural scaffold that supports an appropriate interaction with substrate. The presence of catalytic residues alone is not sufficient to infer function, it is these residues coupled with a particular scaffold that permits the enzyme to perform its function. In other cases, the modular unit of functional inheritance can comprise the expression behaviour of a sequence under some experimental conditions (a non-positional indicator of function) together with its cellular localisation.

The identification of such modules is perhaps more difficult than predicting the function of a given sequence of interest. Modules comprising both positional and non-positional features can be inferred from associations made between characteristic properties of sequences and function. For example, the back interpretation of microarray and protein interaction information in Chapter 5 permits valuable and novel biological insights to be made. However, these associations can occur coincidentally rather than for the necessary preservation of function. Even if attempts to experimentally verify these trends are made, for example, using site-directed mutagenesis to remove amino acid side chains or delete regions from proteins whilst quantifying the effect on some functional behaviour, it is difficult to interpret these results without the use of crystal structure data, or other molecular visualisation techniques. Other regions or residues within a sequence may compensate for the changes such that the overall function of the protein is retained. In this case the true importance of a site in specifying function may be lost. Despite these difficulties, once unveiled, discovery of these modules provides new functional knowledge that can feed back into function prediction approaches that are ready to reap the rewards.

In the regression-based function similarity prediction approach developed here, little positional information was used. This is partly because the availability of such information is sparse for human sequences and also because this information is often captured by domain or sequence family information. One of the more accurate sources of functional site information is from crystal structures where conservation of side chain positions or regions of consistent backbone conformation may be sufficient to infer function (see Bandyopadhyay et al. (2009) for a review). However, short of producing homology models for all human sequences for which close templates could be identified, there is a lack of available crystal structure information for the human genome that can be used in such an approach. Even where homology models could be produced, there exist questions regarding the degree of accuracy of side-chain placements (Eyal et al. 2005), which would seem to be a critical aspect of structure-based function site prediction methods.

The major contribution of this work is in presenting a flexible framework within which the assignment of any labels from an ontological structure to sequences can be predictively modelled. Currently this method superceeds other methods because annotation specificities are accounted for in the approach. The method requires only a small fraction of sequences to be annotated to fulfil minimum data source modelling requirements because features describing sequence pair relationships are used rather than features describing characteristics of single sequences. As our current knowledge of function improves and new annotation assignments are made to sequences, the method will auto-update to an extent. This is because the relationships between sequences can remain constant whilst the scoring of these relationships can be adjusted as sequences acquire new function annotations. Larger changes resulting from the addition of thousands of new sequence annotations, or a change in ontology structure might necessitate a complete re-build of the approach.

Finally, the production of a gold standard dataset of sequences with fully completed annotations is a necessary and community-wide requirement for the successful development and testing of any function prediction method. Such attempts are being made using mouse knock-out data to decipher phenotype (Gondo et al. 2009, Shaw 2009). However, the interpretation of this information with respect to biological molecules is again complex since elements of different biological pathways may adapt or compensate for gene loss or gene mutation effects. At the current time this appears to be the most important and challenging bottleneck that must be overcome in order to advance the field.

Machine Learning in Bioinformatics

Machine learning algorithms are a branch of artificial intelligence concerned with enabling a computer to learn patterns and rules. Learning can be inductive (a reasoning process to support but not guarantee a conclusion) or deductive (the reasoning process guarantees the conclusion). Machine learning techniques are used widely in search engines, medical diagnosis, bioinformatics and cheminformatics. Popular Bioinformatics applications include secondary structure prediction, detecting promoter regions in DNA sequences, classification of protein families, domains and functions and class discovery using microrarray data.

Types of machine learning algorithm

Machine learning approaches can be differentiated by the their learning style. Most algorithms can either be classed as unsupervised or supervised. Unsupervised algorithms assume no prior knowledge and detect naturally occurring patterns within data. In contrast supervised approaches extract rules or patterns that are indicative of some known features of the data, for example, class membership or a continuous variable. Common types of machine learning algorithm are listed in Table A-1.

Table A-1: Types of machine learning algorithms

Type	**Definition**
supervised learning	algorithm generates a function that maps inputs (a numeric vector) to outputs (numeric vector or class label). Examples of desired inputs and outputs are used to the behaviour of the function
unsupervised learning	a set of inputs are modelled without specifying desired input-outputs
semi-supervised learning	function is learnt using combinations of labelled and unlabelled examples
reinforcement learning	algorithm learns a policy given an observed fact. The policy has an impact on an environment and the response feedsback to guide the learning algorithm
transduction	similar to supervised learning except no function is constructed. New outputs are predicted based on training inputs, training outputs and new inputs.
learning to learn	algorithm learns its own inductive bias based on previous experience

Unsupervised learning algorithms

In unsupervised learning a model is fitted to a set of observations without assuming a particular outcome. Unsupervised learning algorithms can be divided into data compression algorithms that rely on probability distributions over sets of inputs, and clustering algorithms that are not probabilistic.

Clustering algorithms attempt to group similar objects. These are defined according to a distance or similarity measure applied to some observed characteristics of the data. Clusters can be formed hierarchically or by data partitioning. Hierarchical algorithms establish an initial cluster and successively generate additional clusters by adding new objects and merging existing clusters. Partitioning algorithms such as K-means and self organising maps determine all clusters simultaneously.

One of the main bioinformatic application areas for clustering methods is microarray analysis. Frequently researchers identify co-regulated genes or transcripts by grouping them into clusters according to some similarity or distance metric. Subsequently meta-information such as pathway data, functional categories or family memberships are overlaid onto the clusters to draw inference from these data.

Supervised and semi-supervised learning algorithms

Semi-supervised and supervised learning algorithms exploit prior knowledge to build a predictive model. Labelled examples are used to optimise a function that maps between sets of known inputs and outputs. The inputs usually comprise a vector of characteristics describing data items of interest. Outputs can be class membership (classification models) or a continuous variable (regression models). Common to semi and fully supervised algorithms are a training phase during which labelled examples of inputs and outputs are presented to the algorithms to learn the parameters of a function mapping between input and output spaces (Figure A-1).

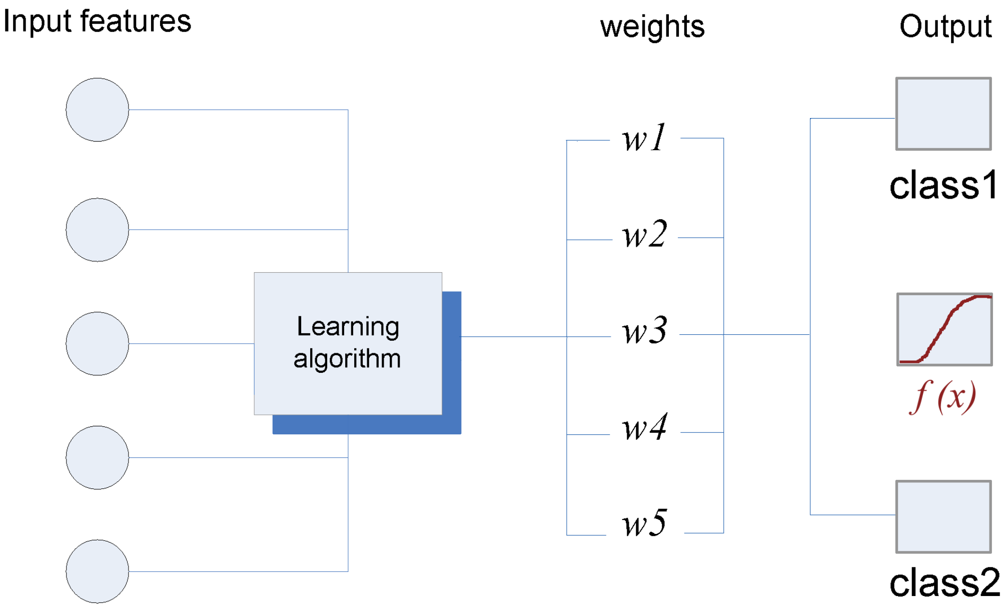


Figure A-1: Supervised learning

In semi-supervised algorithms the training data comprises a mixture of labelled and unlabelled items. Labels can be acquired for the unlabelled data during training (transductive learning), or can be estimated using separate independent models trained on labelled data (co-training). Once trained, models are validated using information criteria concerning pre-labelled test data.

Neural networks and Support Vector Machines (SVMs) represent two of the most popular algorithms for pattern recognition. In binary classification, SVMs provide highly effective and accurate solutions. They are suited to tackling both noisy and high dimensional problems where the number of feature characteristics is high compared to the number of training examples. This scenario is often termed small n (examples), large p (features). In contrast, neural networks operate efficiently when n is much larger than p.

Neural Networks (NN) are a branch of artificial intelligence comprising layers of nodes (neurons) to transfer information between input and output layers. During the training phase example inputs and outputs are passed through the network in order to optimise a set of weights. The number of nodes and topology of the network can vary according to the problem specification. Too many nodes can lead to overfitting resulting in poor performance on new test cases. The use of too few nodes may result in a poor solution. In contrast, the Support Vector Machine (SVM) is by design more resistant to these problems.

Support Vector Machines

In classification mode SVM's optimise the position of a linearly separating hyperplane to assign class membership. The input feature space is transformed by a kernel function ϕ which can be thought of as a similarity matrix describing the relationship between features. This 'feature space' is the transformed space in which positioning of the hyperplane is performed.

The Support Vectors (SV's) are those objects lying closest to the separating hyperplane and consitute the decision boundary. The optimal separating hyperplane is found by maximizing the distance (margin) between support vectors either side of the hyperplane (Figure A-2). This strategy avoids overfitting and is well suited to problems of high dimensionality because the goal of the algorithm is simply to maximize the margin in the feature space.

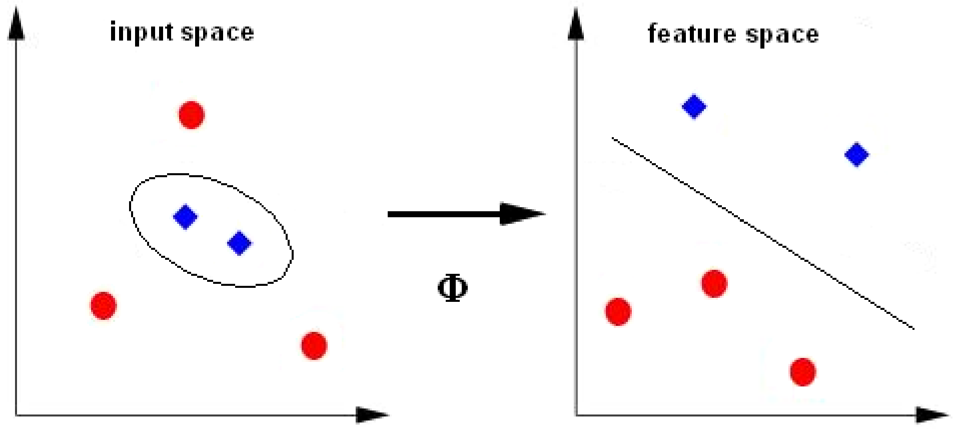


Figure A-2: Schematic of SVM algorithm

The problem of maximising the margin is posed as finding the solution to a set of quadratic inequalities. Considering a dataset with inputs $x_i ... x_n$ and known outputs $y_i ... y_n$ which either belong to a class ($y_i = +1$) or do not belong to a class ($y_i = -1$), the problem of finding the separating hyperplane is defined in Equations A-1 and A-2, providing the data are linearly separable.

$$x_i \cdot w + b \geq +1 \text{ for } y_i = +1 \tag{A-1}$$

$$x_i \cdot w + b \leq -1 \text{ for } y_i = -1 \tag{A-2}$$

The equations can be combined into a set of inequalities:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \; \forall_i \tag{A-3}$$

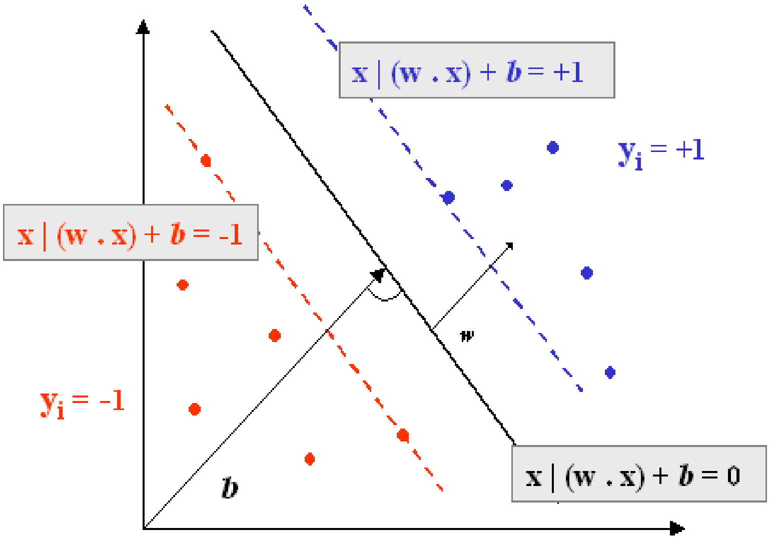


Figure A-3: Separating hyperplane in transformed feature space. The hyperplane is shown in black and the respective margins in red and blue.

The points that determine the upper margin H_1 and lower margin H_2 lie on hyperlanes with normal w and perpendicular distance from the origin $|1-b|/||w||$. In the perfect example separation no points lie between the two parallel margins and the pair of hyperplanes that maximise the margin are given by minimizing $||w||^2$. This minimisation is carried out using a Lagrangian formulation of the problem (Equation A-4). Lagragian multipliers enable the problem to be reformulated only using the dot product between the training data items. This allows the algorithm to be generalised to the non-linear case, and replaces consraints on the inequalities with contraints on the multipliers.

$$LP \equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} \alpha_i \tag{A-4}$$

L_P must be minimized with respect to w (Equation A-5) and b subject to contraints that $\alpha_i \geq 0$ and that all α_i vanish (Equation A-6). This is a quadratic programming problem and can be formulated in a dual fashion (L_D in Equation A-7).

$$w = \sum_i \alpha_i y_i x_i \tag{A-5}$$

$$\sum_i \alpha_i y_i = 0 \tag{A-6}$$

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{A-7}$$

In the support vector training process for the linear case L_D is maximised with respect to α_i subject to constraints. The points which are the support vectors (define the margin) have positive α_i. All other training points have $\alpha_i = 0$. This property means that if the training set was reduced in size not compromising any of the support vectors, then the exact same hyperplane would be found. The optimal solution is determined using (KKT) Karush-Kuhn-Tucker theory (Karush 1939, Kuhn and Tucker 1950).

Non-linear SVMs

Practically, classification problems tend not to be linearly separable in the input space and require transformation into a higher dimensional feature space to acheive linear separation. Since the training data is only ever present as a dot product in the Lagragian form, the higher dimensional feature space need not be explicitly defined or the transform function known. SVM's use kernel functions to represent the transformed feature space in a higher dimensional space that can be linearly separated. This is also known as the 'kernel trick' (Bishop 2006).

Mercer's theorem states that any continuous, symmetric, positive semi-definite kernel function K(x, y) can be expressed as a dot product in a high-dimensional space. Commonly used kernels include the sigmoid, spline kernel, anova kernel and radial basis function kernels (Table A-2). As an example, this specification means that over a set of 1000 training examples with 256 features, a 50 fold reduction in computational efficiency can be observed.

Table A-2: Kernel functions

Kernel	**Equation**
linear	$x \cdot x'$
sigmoid	$tanh(a\dot{(}x_i^T x_j) + \text{offset})$
radial basis	$\frac{\lvert x_i - x_j \rvert^2}{\sigma}$
polynomial	$(a\dot{(}x_i^T x_j) + \text{offset})^d$
anova	$\sum_{1<=qi_1 \ldots < i_D <=qN} \prod_{d=1}^{D} k(x_{id}, x'_{id})$
spline	$\prod_{d=1}^{D} 1 + x_i x_j + x_i x_j min(x_i, x_j) - \frac{x_i + x_j}{2} min(x_i, x_j)^2 + \frac{min(x_i, x_j)^3}{3}$

Here x and x' represent feature vectors and x_i and x_j are indexes of feature vector elements.

There exists only a single best hyperplane for cases where the input data is truly and absolutely separable. In most cases the data will not be exactly separable, either due to the choice of kernel function or due to noise in the training data. To handle noisy training data, and avoid laborious searches for more appropriate kernels two slack variables are introduced permitting a soft margin (Equation A-8). The soft margin tolerates "errors" in the training data such that a proportion of data points may lie within the margin of the hyperplane.

$$
\begin{aligned}
x_i \cdot w + b &\geq +1 - \varepsilon \text{ for } y_i = +1 \\
x_i \cdot w + v &\leq -1 - \varepsilon \text{ for } y_i = -1 \\
\varepsilon &\geq 0 \, \forall_i
\end{aligned}
\tag{A-8}
$$

The algorithm assigns extra cost for errors by introducing a tunable parameter C. The maximimal margin minimization is then $||w||^2 + C(\sum_i \varepsilon_i)$ rather than $||w^2||$. The Lagrangian primal L_P and L_D are re-written (Equations A-9 and A-10). μ_i are the extra lagrange multipliers used to enforce positivity of the ε_i.

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{A-9}$$

$$L_P \equiv \frac{1}{2}||w||^2 + C\sum_i \varepsilon_i - \sum_i y_i(x_i \cdot w + b) - 1 + \varepsilon_i - \sum_i \mu_i \varepsilon_i \tag{A-10}$$

Regression Support Vector Machines

In regression mode the support vector machine optimises a function that minimises the error of the desired output function. Examples include predicting a continuous variable such as a property of an object such as length, or cost. The mathematical formulation of the problem and the solution are analagous to the classification SVM. The implementation is similar to regression in a 3 layer neural network except that in the case of Support Vector Regression (SVR), the input weights are pre-determined by the training patterns.

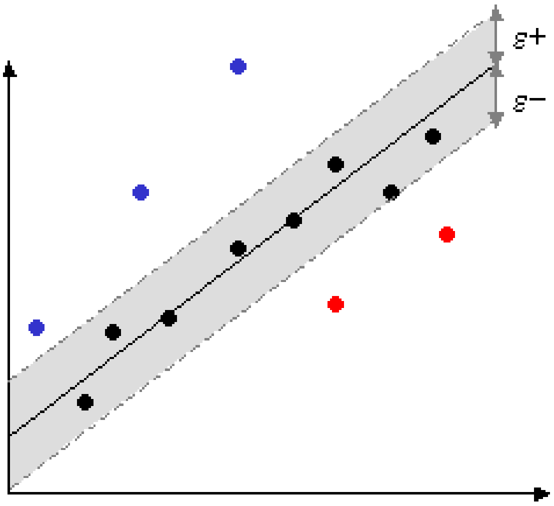


Figure A-4: Epsilon sensitive Support Vector Regression

The SVR algorithm is also referred to as ”shrinking the width of the tube” (see Figure A-4). This is because unlike classic linear regressions a measure (ϵ) is introduced below which errors on the fit are discounted. The C parameter controls the trade off between the flatness of the function and the amount up to which deviations larger than the specified error are tolerated. Like the SVM in classification mode, the SVR only depends on a subset of the data (the support vectors) which line the tube. Kernel functions can be applied to handle non-linearity. The benefits of SVR over classical regression algorithms can be realised for large (high dimensional) and noisy training datasets.

The algorithm can be thought of as a flattening of the function defined in Equation A-11.

$$f(x) = (w \cdot x) + b \text{ with } wX \text{ and } b \in \Re \tag{A-11}$$

The problem can be reformulated as a quadratic minimisation (Equation A-12. Similar to SVM classification, two slack variables ξ_i and ξ^* are introduced (see Figure A-5) to make the solution feasible. Only the points outside the shaded region are subject to cost. The formulation of the problem is then defined as in Equation A-13.

$$\begin{aligned} &\text{minimize } \frac{1}{2}||w||^2 \\ &\text{subject to } y_i - (w, x_i) - b \leq \epsilon \\ &\quad\text{and} \quad (w, x_i) + b - y_i \leq \epsilon \end{aligned} \tag{A-12}$$

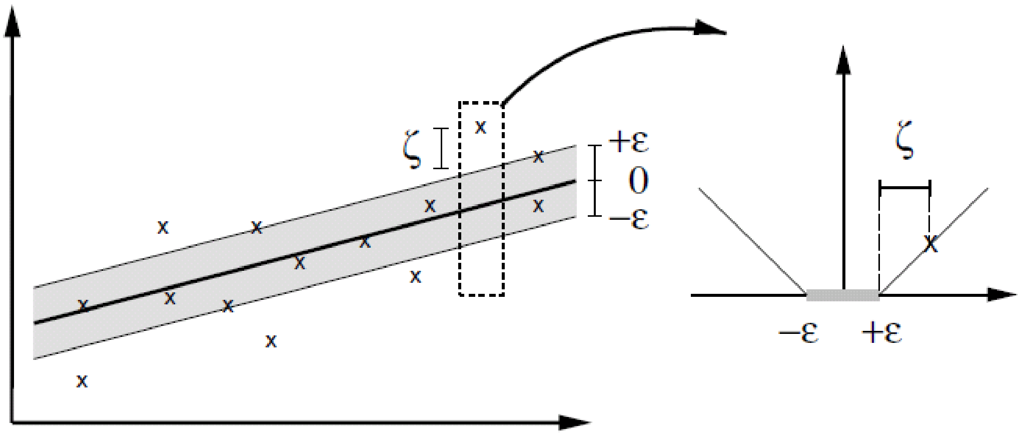


Figure A-5: Epsilon sensitive support vector regression with slack variables ξ adapted from Smola et al. (2003)

$$\begin{aligned} \text{minimize } & \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i*) \\ \text{subject to } & y_i - (w, x_i) - b \leq \epsilon + \xi_i \\ \text{and } & (w, x_i) + b - y_i \leq \epsilon - \xi_i* \\ \text{where } & \xi_i, \xi_i* \geq 0 \end{aligned} \tag{A-13}$$

Similar to the SVM, the dual formulation of this problem (Equation A-14) is produced by introduction of Langrange multipliers and is solved under KKT conditions.

$$\begin{aligned} L|D| =& \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i*) - \\ & \sum_{i=1}^{l}\alpha_i(\epsilon + \xi_i - y_i + (w, x_i) + b) - \\ & \sum_{i=1}^{l}\alpha_i * (\epsilon + \xi_i * + y_i - (w, x_i) - b) - \\ & \sum_{i=1}^{l}(\nu_i\xi_i + \nu_i * \xi_i*) \end{aligned} \qquad \text{(A-14)}$$

Parameter Optimisation for SVM and SVR

Practically, there are several parameters to be tuned during SVM/R training. Initially an appropriate kernel function must be chosen. Secondly the cost parameter (C) must be determined. The value of C controls the number of datapoints tolerated in the margin for classification or the degree to which the deviation from the error margin of regression is tolerated. In cases where the number of class examples are unequal in the training set, a third parameter (bias) can be used to control the trade-off between training errors made on the positive or negative examples in order to simulate training on a balanced dataset. The choice of kernel function may introduce a variable number of extra parameters.

Parameter optimisation strategies include simple grid searches, gradient descent approaches and genetic algorithms. In the grid search method a range is defined for the parameters. The algorithm is trained successively using parameters selected at evenly spaced intervals within the range. The 'best' parameter set are those that optimise a fitness measure. If the parameter range and intervals are sufficiently diverse, this method is guaranteed to find the optimum solution. However where more than one or two parameters must be tuned, the method becomes increasingly computationally inefficient since each range must be cross-trialed dramatically increasing the number of training runs. For example, a coarse grid search of 1e-6 to 1e+6 with 12 evenly spaced intervals over 3 parameters requires 1728 training runs.

Gradient descent algorithms are more efficient at searching large parameter spaces, however, they are not guaranteed to find the best parameter set. In the approach, initial ranges are specified and training runs performed. The gradient of performance improvement is monitored during successive training runs. Adjustments are made to the parameter values either upwards or downwards after each training iteration until no further improvements are made. The approach assumes that the parameter surface is fairly smooth and that there exists a well in which the optimal solution lies (see Figure A-6). Frequently this is not the case, and where the surface is wave-like, the method may converge on a sub-optimal parameter set by reaching a local minimum.

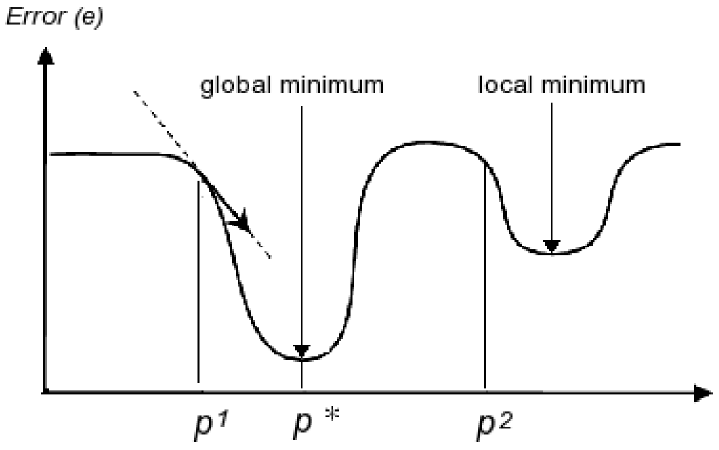


Figure A-6: Example parameter surface. The goal is to reach parameter set *p** which can be achieved by passing through *p1*. However the if a search begins at or passes through *p2*, *p** may never be reached due to the existence of a local minimum close to *p2*.

Genetic Algorithms (GA's) simulate evolutionary events in order to efficiently sample large parameter spaces. These events include mutation, selection, recombination and inheritance. Initially, performance can be measured for each parameter value by varying one parameter and fixing the rest, or by random sampling. From these results a population of values are selected and trialed. The best parameter sets are then selected for reproduction to produce a new population subject to evolutionary events. These parameters are used for training runs and the process repeated until a stable solution is reached, or until a maximum number of generations have been produced. GA's generally settle on good solutions, but like the gradient descent methods, may not produce the optimum solution. One problem lies in generating sufficiently diverse populations that include the parents of potentially good solutions.

SVM/R Training and test strategies

Cross validation strategies are used to assess how well a model might perform in practice as well as for parameter optimisation. For effective learning the training data should be representative of the test case scenario in which the model(s) will be applied. The training and test cases must be mutually exclusive.

Cross validation can be carried out by random repeated training and testing iterations (random sampling approach), or by jack-knifing, leaving out a single training example at a time and repeating for the whole training set. This is known as Leave One Out Cross Validation (LOOCV). This strategy is rigorous for small datasets, however can be computationally impractical on larger ones. In these cases, N-fold cross validation can be performed. This strategy involves partitioning the training data into N separate folds. Training is performed using a single fold and testing carried out on the remaining fold. Cross validation ensures that parameter selections provide realistic performance estimates on unseen data.

Performance measures

Performance measures for machine learning problems emphasize different aspects of quality. Different measures can more or less well suited to different tasks. In classification, AUC (Area Under Curve), MCC Matthew's Correlation Coefficient), precision-recall break even point and F measures are commonly used statistics. AUC represents the area under the Receiver Operating Characteristic curve (ROC). This curve details the proportion of true and false positives obtained at different threshold distances from the SVM hyperplane. Scores greater than zero represent predicted positive classifications whilst those below zero represent predicted negative classifications. These values are compared to the known class assignments in order to obtain performance statistics.

Typically a confusion matrix is constructed. This matrix comprises four values, True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These represent the proportion of predicted positive classifications that are correct, the proportion of predicted negatives that are correct, the proportion of predicted positives that are false, and the proportion of predicted negatives that should be positive.

Table A-3: Confusion matrix

	Actual	
Predicted	**True**	**False**
True	TP	FP
False	FN	TN

AUC measures are well suited to classification tasks performed using balanced training datasets, that is those where the frequency of positive and negative training examples are roughly equal. A value of 0.5 indicates performance obtained when class assignments are made at random. MCC measures are the class based equivalent of Pearson's correlation coefficient. A value of 1 implies perfect classification whilst a value of 0 denotes random performance.

In the MCC calculation, the contribution of positive and negative examples to the score is made equal, thus class imbalance does not affect the resulting values. For instances where false positives can be tolerated providing at least some of the predictions are correct, the precision recall break even point might be used. This measure balances the proportion of true positives (recall) against the likelihood of a positive result being correct (precision). Between MCC measures from different classifiers, performance is only comparable where class sizes are similar due to the different contributions of a single test case to the magnitude of the correlation. To make these comparisons it is appropriate to compare actual true and false positives.

Performance measures for SVR include Pearson's correlation (Equation 1.3) and Euclidean distances (Equation 1.4) where the magnitude of the difference or the similarity of score magnitude is important respectively. Alternatively, Kendall's tau

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{A-15}$$

and Spearman's correlation

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$
$$d_i = x_i - y_i \tag{A-16}$$

may be used where only the rank of the predicted score is important. Sum of squares error given by

$$SS_{err} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{A-17}$$

reflects the variability between the regression target and response variable, and is useful in descriminating between models acheiving similar performance by other measures.

Appendix II: Tables

Table A-4: Gene Ontology categories better predicted either between or within species.

Category	Description	Bw	Pw	Bb	Pb	Size	Bw-Bb	P value	Pw-Pb	P value
Molecular Functions										
GO:0008199	ferric iron binding	3.80	3.80	0.71	0.71	31	8.18	1.11e-16	8.18	1.11e-16
GO:0004674	protein ser/thr kinase	1.06	1.07	0.53	0.52	893	7.85	2.11e-15	8.12	2.22e-16
GO:0004984	olfactory receptor	3.80	3.80	3.25	3.80	618	6.82	4.64e-12	0	5.00e-01
GO:0004571	mannosyl-oligosaccharide 1,2-alpha-mannosidase	3.80	3.80	1.80	1.80	29	5.09	1.75e-07	5.09	1.75e-07
GO:0016524	latrotoxin receptor	2.32	2.00	0.28	0	26	4.89	5.06e-07	4.79	8.14e-07
GO:0008440	inositol trisphosphate 3-kinase	3.80	3.80	0.83	0.83	13	4.70	1.30e-06	4.70	1.30e-06
GO:0008009	chemokine	1.74	1.90	0.58	0.54	69	4.70	1.32e-06	5.50	1.93e-08
GO:0004434	inositol or phosphatidylinositol phosphodiesterase	1.29	1.38	0	0	53	4.57	2.41e-06	4.86	5.75e-07
GO:0004175	endopeptidase	0.98	0.97	0.70	0.70	1045	4.55	2.66e-06	4.29	8.92e-06
GO:0004467	long chain fatty acid CoA ligase	3.80	3.80	0.95	0.95	13	4.51	3.25e-06	4.51	3.25e-06
GO:0004222	metalloendopeptidase	1.32	1.32	0.78	0.77	266	4.39	5.63e-06	4.48	3.65e-06
GO:0004743	pyruvate kinase	3.80	3.80	1.24	1.24	14	4.23	1.08e-05	4.25	1.08e-05
GO:0016773	phosphotransferase activity, alcohol acceptor	0.79	0.76	0.60	0.60	1883	4.15	1.65e-05	3.61	1.54e-04
GO:0004672	protein kinase	0.81	0.78	0.60	0.60	1574	4.14	1.72e-05	3.67	1.22e-04
GO:0003918	DNA topoisomerase (ATP-hydrolyzing)	3.80	3.80	1.14	1.14	12	3.98	3.40e-05	3.98	3.40e-05
GO:0042379	chemokine receptor binding	1.49	1.54	0.53	0.50	71	3.95	3.85e-05	4.30	8.49e-06
GO:0015071	protein ser/thr phosphatase	1.56	1.56	0.24	0	37	3.85	5.87e-05	4.54	2.80e-06
GO:0047760	butyrate-CoA ligase	3.80	3.80	0.66	0.66	9	3.85	5.94e-05	3.85	5.94e-05
GO:0008067	metabotropic glutamate receptor	1.95	1.95	0.97	0.95	64	3.80	7.25e-05	3.90	4.89e-05
GO:0004245	neprilysin activity	3.80	3.80	1.51	1.46	14	3.80	7.31e-05	3.89	5.09e-05
GO:0016165	transcriptional repressor	3.80	3.80	1.32	1.32	12	3.73	9.77e-05	3.73	9.77e-05
GO:0004448	isocitrate dehydrogenase	3.80	3.80	1.32	1.32	12	3.73	9.77e-05	3.73	9.77e-05
GO:0004618	phosphoglycerate kinase	3.80	3.80	0.55	0.37	8	3.63	1.39e-04	3.84	6.24e-05
GO:0008113	peptide-methionine-(S)-S-oxide reductase	3.80	3.80	1.05	0.90	10	3.63	1.39e-04	3.83	6.29e-05
GO:0022844	voltage-gated anion channel	1.76	1.76	0.66	0.69	45	3.59	1.68e-04	3.49	2.40e-04
GO:0004370	glycerol kinase	3.80	3.80	1.27	1.27	11	3.58	1.70e-04	3.58	1.70e-04
GO:0016503	pheromone receptor	3.80	3.80	1.67	3.80	13	3.37	3.69e-04	0	5.00e-01
GO:0004785	superoxide dismutase	3.80	3.80	0	0	6	3.29	4.99e-04	3.29	4.99e-04
GO:0004143	diacylglycerol kinase	1.84	1.84	0.60	0.64	31	3.28	5.21e-04	3.17	7.66e-04
GO:0003847	1-alkyl-2-acetylglycerophosphocholine esterase	3.80	3.80	0.97	0.97	8	3.16	7.86e-04	3.16	7.86e-04
GO:0004926	non-G-protein coupled 7TM receptor	3.80	3.80	2.03	2.03	15	3.07	1.08e-03	3.07	1.08e-03
GO:0004818	glutamate-tRNA ligase	3.80	3.80	1.07	1.07	8	3.05	1.15e-03	3.05	1.15e-03
GO:0004830	tryptophan-tRNA ligase	3.80	3.80	0.78	0.78	7	3.02	1.28e-03	3.02	1.28e-03

... Continued on next page

Category	Description	Bw	Pw	Bb	Pb	Size	Bw-Bb	P value	Pw-Pb	P value
GO:0022836	gated channel	0.70	0.69	0.49	0.50	808	3.01	1.32e-03	2.74	3.05e-03
GO:0004832	valine-tRNA ligase	3.80	3.80	2.00	3.80	14	2.99	1.41e-03	0	5.00e-01
GO:0004657	proline dehydrogenase	3.80	3.80	1.96	1.96	13	2.91	1.80e-03	2.91	1.80e-03
GO:0004329	formate-tetrahydrofolate ligase	3.80	3.80	1.96	1.96	13	2.91	1.80e-03	2.91	1.80e-03
GO:0008518	reduced folate carrier	3.80	3.80	2.00	3.80	13	2.85	2.21e-03	0	5.00e-01
GO:0004499	flavin-containing monooxygenase	3.80	3.80	1.91	3.80	12	2.84	2.28e-03	0	5.00e-01
GO:0003810	proteinglutamine gammaglutamyltransferase	3.80	3.80	1.91	1.91	12	2.84	2.28e-03	2.84	2.28e-03
GO:0001608	nucleotide receptor	1.09	1.12	0.28	0.28	52	2.83	2.33e-03	2.94	1.65e-03
GO:0005102	receptor binding	0.67	0.67	0.81	0.84	1186	-2.36	9.32e-03	-2.94	1.66e-03
GO:0033558	protein deacetylase	0.91	0.95	1.82	1.82	30	-2.36	9.18e-03	-2.27	1.16e-02
GO:0016881	acidamino acid ligase	1.23	1.24	1.58	1.53	193	-2.38	8.76e-03	-1.97	2.45e-02
GO:0004857	enzyme inhibitor	0.94	0.94	1.19	1.19	414	-2.48	6.64e-03	-2.50	6.24e-03
GO:0016757	glycosyl transferase activity	0.93	0.93	1.15	1.16	518	-2.49	6.30e-03	-2.66	3.93e-03
GO:0019213	deacetylase	0.93	0.96	1.87	1.87	33	-2.58	4.88e-03	-2.49	6.32e-03
GO:0015269	calcium activated potassium channel	0.52	0.48	1.44	1.53	35	-2.59	4.74e-03	-2.95	1.59e-03
GO:0004559	alpha mannosidase	1.82	1.82	3.80	3.80	10	-2.62	4.43e-03	-2.62	4.43e-03
GO:0046872	metal ion binding	0.87	0.88	0.93	0.93	7397	-2.66	3.95e-03	-2.49	6.45e-03
GO:0016787	hydrolase activity	0.72	0.71	0.79	0.80	5174	-2.66	3.85e-03	-3.27	5.47e-04
GO:0046873	metal ion transporter	0.58	0.53	0.79	0.80	648	-2.67	3.80e-03	-3.42	3.16e-04
GO:0005261	cation channel	0.57	0.52	0.81	0.82	541	-2.69	3.61e-03	-3.43	3.01e-04
GO:0017176	phosphatidylinositol Nacetylglucosaminyltransferase	0.66	0.66	3.80	3.80	6	-2.72	3.25e-03	-2.72	3.25e-03
GO:0008533	astacin	1.35	1.35	3.80	3.80	8	-2.73	3.12e-03	-2.73	3.12e-03
GO:0043169	cation binding	1.04	1.04	1.10	1.10	6872	-2.74	3.06e-03	-2.64	4.18e-03
GO:0015662	ATPase activity	0.97	0.93	1.43	1.45	159	-2.88	2.00e-03	-3.25	5.83e-04
GO:0003676	nucleic acid binding	0.95	0.96	1.09	1.11	1711	-2.88	1.99e-03	-3.17	7.69e-04
GO:0003735	structural constituent of ribosome	1.74	1.76	1.96	1.99	706	-2.93	1.71e-03	-3.05	1.14e-03
GO:0008889	glycerophosphodiester phosphodiesterase	2.00	2.00	3.80	3.80	14	-2.99	1.41e-03	-2.99	1.41e-03
GO:0008484	sulfuric ester hydrolase	0.85	0.92	1.91	1.76	36	-3.04	1.17e-03	-2.42	7.78e-03
GO:0004597	peptideaspartate betadioxygenase	2.03	2.03	3.80	3.80	15	-3.07	1.08e-03	-3.07	1.08e-03
GO:0046914	transition metal ion binding	1.22	1.22	1.30	1.31	5036	-3.09	9.91e-04	-3.25	5.72e-04
GO:0016832	aldehyde lyase	2.18	2.18	3.80	3.80	20	-3.33	4.34e-04	-3.33	4.34e-04
GO:0016831	carboxy lyase	1.00	0.97	1.85	1.85	64	-3.34	4.12e-04	-3.44	2.86e-04
GO:0004565	beta galactosidase	1.59	1.59	3.80	3.80	13	-3.50	2.36e-04	-3.50	2.36e-04
GO:0005272	sodium channel	0.46	0.39	1.35	1.48	64	-3.50	2.36e-04	-4.24	1.10e-05
GO:0004835	tubulin tyrosine ligase	2.60	2.60	3.80	3.80	45	-3.89	4.95e-05	-3.89	4.95e-05
GO:0001619	lysosphingolipid and lysophosphatidic acid receptor	1.40	2.00	3.80	2.00	14	-3.98	3.51e-05	0	5.00e-01
GO:0004767	sphingomyelin phosphodiesterase	0.84	0.84	3.80	3.80	11	-4.19	1.39e-05	-4.19	1.39e-05
GO:0004629	phospholipase C	0.16	0.16	1.30	1.34	62	-4.38	5.80e-06	-4.54	2.82e-06
GO:0008270	zinc ion binding	1.41	1.42	1.55	1.54	4114	-4.41	5.15e-06	-3.92	4.50e-05
GO:0016772	transferase, transferring phosphate groups	0.73	0.71	0.92	0.90	2841	-4.91	4.47e-07	-5.18	1.09e-07

... Continued on next page

Category	Description	Bw	Pw	Bb	Pb	Size	Bw-Bb	P value	Pw-Pb	P value
GO:0016740	transferase	0.71	0.69	0.86	0.85	4530	-5.18	1.08e-07	-5.40	3.31e-08
GO:0003824	catalytic activity	0.46	0.46	0.60	0.60	11998	-7.96	8.88e-16	-7.51	2.85e-14
GO:0019787	small conjugating protein ligase	1.15	1.20	3.80	3.80	90	-12.36	0.00e+00	-12.11	0.00e+00
GO:0016301	kinase	0.33	0.32	0.92	0.92	2392	-14.44	0.00e+00	-14.73	0.00e+00
Biological Processes										
GO:0007608	sensory perception of smell	1.44	1.45	0.47	0.33	562	11.53	0	13.25	0
GO:0007606	sensory perception of chemical stiulus	1.34	1.36	0.51	0.38	605	10.21	0	12.01	0
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	1.97	1.95	0.87	0.89	214	8.04	4.44e-16	7.70	6.55e-15
GO:0048002	antigen processing and presentation of peptide antigen	1.83	1.80	0.83	0.84	223	7.46	4.21e-14	7.10	6.12e-13
GO:0019882	antigen processing and presentation	1.65	1.64	0.83	0.85	226	6.10	5.47e-10	5.85	2.52e-09
GO:0006470	NADPH regeneration	1.55	1.51	0.79	0.79	261	6.03	8.26e-10	5.82	2.96e-09
GO:0006003	fructose 2,6-bisphosphate metabolism	3.80	3.80	0.66	0.59	15	5.44	2.63e-08	5.57	1.29e-08
GO:0006313	transposition, DNA-mediated	1.50	1.50	0.42	0.18	91	5.10	1.66e-07	6.20	2.83e-10
GO:0006438	valyl-tRNA aminoacylation	3.80	3.80	0.99	0.99	14	4.67	1.54e-06	4.67	1.54e-06
GO:0007600	sensory perception	0.78	0.77	0.49	0.43	1004	4.61	1.99e-06	5.46	2.38e-08
GO:0050877	neurological system process	0.74	0.74	0.46	0.40	1057	4.61	2.02e-06	5.42	2.94e-08
GO:0007223	Wnt receptor signaling pathway	1.84	1.84	0.64	0.61	41	3.71	1.0e-04	3.79	7.55e-05
GO:0006334	nucleosome modeling	1.67	1.74	1.15	1.18	179	3.40	3.41e-04	3.73	9.49e-05
GO:0032196	transposition	1.13	1.04	0.42	0.18	91	3.33	4.30e-04	4.03	2.76e-05
GO:0006422	aspartyl-tRNA aminoacylation	3.80	3.80	0.00	0.00	6	3.29	4.99e-04	3.29	4.99e-04
GO:0006436	tryptophanyl-tRNA aminoacylation	3.80	3.80	0.78	0.78	7	3.02	1.28e-03	3.02	1.28e-03
GO:0006435	threonyl-tRNA aminoacylation	3.80	3.80	0.43	0.43	6	2.92	1.77e-03	2.92	1.77e-03
GO:0006537	glutamate biosynthetic process	3.80	3.80	1.96	1.61	13	2.91	1.80e-04	3.46	2.65e-04
GO:0006424	glutamyl-tRNA aminoacylation	3.80	3.80	1.32	1.32	8	2.78	2.74e-03	2.78	2.75e-03
GO:0007018	microtubule-based movement	1.37	1.39	1.05	1.07	216	2.38	8.69e-03	2.33	9.83e-03
GO:0006511	ubiquitin-dependent protein catabolic process	0.74	0.74	0.57	0.58	713	2.35	9.42e-03	2.15	1.59e-02
GO:0046116	queuosine metabolic process	3.80	3.80	1.14	1.14	6	2.30	1.07e-02	2.30	1.07e-02
GO:0046114	guanosine biosynthetic process	3.80	3.80	1.14	1.14	6	2.30	1.07e-02	2.30	1.07e-02
GO:0008618	7-methylguanosine metabolic process	3.80	3.80	1.14	1.14	6	2.30	1.07e-02	2.30	1.07e-02
GO:0008616	queuosine biosynthetic process	3.80	3.80	1.14	1.14	6	2.30	1.07e-02	2.30	1.07e-02
GO:0046118	modification-dependent macromolecule catabolic process	3.80	3.80	1.14	1.14	6	2.30	1.07e-02	2.30	0.01
GO:0006259	DNA metabolic process	0.46	0.44	0.57	0.58	1845	-2.21	1.35e-02	-2.99	1.41e-03
GO:0042981	regulation of apoptosis	0.24	0.24	0.39	0.42	891	-2.23	1.3e-02	-2.68	3.6e-03
GO:0006810	transport	0.69	0.67	0.76	0.79	3771	-2.24	1.24e-02	-3.55	1.91e-04
GO:0048523	negative regulation of cellular process	0.23	0.22	0.33	0.35	1970	-2.25	0.012	-2.89	1.93e-03
GO:0031497	chromatin assembly	0.59	0.61	0.91	0.98	205	-2.29	1.11e-02	-2.66	3.88e-03
GO:0006468	protein amino acid phosphorylation	1.28	1.26	1.41	1.41	1246	-2.30	1.07e-02	-2.69	3.57e-03
GO:0006464	protein modification	0.88	0.88	0.96	0.98	3256	-2.35	9.35e-03	-2.98	1.42e-03
GO:0031163	metallo-sulfur cluster assembly	1.70	1.70	3.80	3.80	8	-2.35	9.34e-03	-2.35	9.34e-03

... Continued on next page

Category	Description	Bw	Pw	Bb	Pb	Size	Bw-Bb	P value	Pw-Pb	P value
GO:0016310	phosphorylation	1.21	1.19	1.34	1.35	1307	-2.35	9.32e-03	-2.94	1.66e-03
GO:0051234	establishment of localization	0.67	0.65	0.75	0.77	3843	-2.37	8.87e-03	-3.62	1.46e-04
GO:0008033	tRNA processing	0.74	0.76	1.22	1.31	103	-2.38	8.61e-03	-2.72	3.27e-03
GO:0006066	alcohol metabolic process	0.69	0.68	0.90	0.90	560	-2.41	7.99e-03	-2.66	3.92e-03
GO:0006974	response to DNA damage stimulus	0.56	0.53	0.78	0.81	537	-2.46	7.02e-03	-3.18	7.28e-04
GO:0006505	GPI anchor metabolic process	0.54	0.54	1.52	2.00	28	-2.46	6.93e-03	-3.66	1.28e-04
GO:0015031	protein transport	0.56	0.55	0.75	0.78	714	-2.47	6.82e-03	-3.05	1.13e-02
GO:0007067	mitosis	0.37	0.34	0.76	0.88	169	-2.47	6.74e-03	-3.46	2.69e-04
GO:0022403	cell cycle phase	0.26	0.25	0.55	0.62	290	-2.47	6.73e-03	-3.11	9.24e-04
GO:0043067	regulation of programmed cell death	0.24	0.23	0.41	0.42	906	-2.49	6.35e-03	-2.83	2.35e-03
GO:0008202	steroid metabolic process	0.57	0.57	0.85	0.86	317	-2.51	6.07e-03	-2.52	5.94e-03
GO:0042559	pteridine and derivative biosynthetic process	1.27	1.27	2.32	2.32	26	-2.52	5.94e-03	-2.52	5.94e-03
GO:0006583	melanin biosynthetic process from tyrosine	0.88	0.88	3.80	3.80	6	-2.53	5.74e-03	-2.53	5.74e-03
GO:0046489	phosphoinositide biosynthetic process	0.45	0.44	1.39	1.58	32	-2.53	5.69e-03	-3.09	1.01e-03
GO:0043412	biopolymer modification process	0.83	0.83	0.92	0.93	3396	-2.55	5.36e-03	-2.97	1.48e-03
GO:0006207	*de novo* pyrimidine base biosynthetic process	1.24	1.24	3.80	3.80	7	-2.56	5.21e-03	-2.56	5.21e-03
GO:0022613	ribonucleoprotein complex biogenesis and assembly	0.54	0.59	1.23	1.25	59	-2.57	5.10e-03	-2.45	7.21e-03
GO:0006796	phosphate metabolism	0.83	0.81	0.95	0.95	1651	-2.63	4.28e-03	-2.75	3.01e-03
GO:0006793	phosphorus metabolic process	0.82	0.81	0.95	0.95	1651	-2.70	3.43e-03	-2.87	2.08e-03
GO:0006729	tetrahydrobiopterin biosynthetic process	0.66	0.66	3.80	3.80	6	-2.72	3.25e-03	-2.72	3.25e-03
GO:0018200	peptidyl-glutamic acid modification	0.66	0.66	3.80	3.80	6	-2.72	3.25e-03	-2.72	3.25e-03
GO:0046146	tetrahydrobiopterin metabolic process	0.66	0.66	3.80	3.80	6	-2.72	3.25e-03	-2.72	3.25e-03
GO:0018214	protein amino acid carboxylation	0.66	0.66	3.80	3.80	6	-2.72	3.25e-03	-2.72	3.25e-03
GO:0065004	protein-DNA complex assembly	0.55	0.52	0.86	0.94	301	-2.72	3.24e-03	-3.58	1.74e-04
GO:0015937	coenzyme A biosynthetic process	1.32	1.32	3.80	3.80	8	-2.78	2.75e-03	-2.78	2.75e-03
GO:0006260	DNA replication	0.60	0.60	0.96	1.04	257	-2.85	2.21e-03	-3.49	2.44e-04
GO:0006629	lipid metabolism	0.60	0.59	0.77	0.80	1136	-2.91	1.83e-03	-3.52	2.12e-04
GO:0006415	translation termination	1.42	1.42	3.80	3.80	9	-2.92	1.76e-03	-2.92	1.76e-03
GO:0006397	mRNA processing	0.46	0.46	0.76	0.80	387	-2.94	1.64e-03	-3.36	3.96e-04
GO:0006281	DNA repair	0.64	0.61	0.93	1.00	461	-3.08	1.04e-03	-4.09	2.17e-05
GO:0006396	RNA processing	0.49	0.49	0.69	0.72	935	-3.14	8.35e-04	-3.51	2.27e-04
GO:0045885	positive regulation of survival gene product expression	0.60	0.60	3.80	1.63	7	-3.20	6.77e-04	-1.03	0.151
GO:0006506	GPI anchor biosynthetic process	0.61	0.61	1.96	2.04	26	-3.23	6.09e-04	-3.44	2.92e-04
GO:0007166	cell surface receptor linked signal transduction	0.79	0.79	0.93	0.94	2638	-3.43	2.98e-04	-4.06	2.47e-05
GO:0016070	RNA metabolic process	0.41	0.41	0.59	0.63	1573	-3.59	1.64e-04	-4.45	4.20e-06
GO:0006353	transcription termination	0.19	0.19	3.80	3.80	7	-3.61	1.54e-04	-3.61	1.54e-04
GO:0000087	M phase of mitotic cell cycle	0.55	0.55	3.80	3.80	8	-3.63	1.39e-04	-3.63	1.39e-04
GO:0007186	GPCR signalling pathway	1.32	1.30	1.50	1.52	1674	-3.73	9.71e-05	-4.31	8.04e-06
GO:0032012	regulation of ARF protein signal transduction	2.41	2.41	3.80	3.80	32	-3.74	9.10e-05	-3.74	9.10e-05
GO:0006561	proline biosynthesis	1.70	1.70	3.80	3.80	16	-3.79	7.47e-05	-3.79	7.47e-05

... Continued on next page

Category	Description	Bw	Pw	Bb	Pb	Size	Bw-Bb	P value	Pw-Pb	P value
GO:0010468	regulation of gene expression	0.78	0.79	0.90	0.92	4952	-4.19	1.38e-05	-4.47	3.92e-06
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.81	0.82	0.94	0.96	4730	-4.22	1.25e-05	-4.81	7.56e-07
GO:0006350	transcription	0.58	0.57	0.76	0.81	2154	-4.30	8.42e-06	-5.66	7.70e-09
GO:0012501	programmed cell death	0.35	0.36	0.76	0.81	477	-4.44	4.46e-06	-4.85	6.10e-07
GO:0008219	cell death	0.37	0.37	0.78	0.84	495	-4.64	1.72e-06	-5.15	1.31e-07
GO:0006915	apoptosis	0.35	0.37	0.79	0.84	467	-4.67	1.52e-06	-5.14	1.41e-07
GO:0006412	translation	1.39	1.39	1.73	1.79	778	-4.67	1.48e-06	-5.55	1.45e-08
GO:0045449	regulation of transcription	0.85	0.86	1.00	1.03	4598	-5.02	2.52e-07	-5.75	4.48e-09
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.37	0.37	0.53	0.55	5985	-5.84	2.62e-09	-7.13	4.91e-13
GO:0006512	ubiquitin cycle	0.73	0.73	1.25	1.32	575	-6.27	1.81e-10	-7.10	6.43e-13
GO:0042254	ribosome biogenesis	0.82	0.91	3.80	3.80	36	-8.57	0	-8.31	0

BLAST and PSI-BLAST results have been abbreviated to Bw, Pw, Bb and Pb, for within (w) and between species (b) results respectively. The values of these columns represent the transformed Fisher — statistics. Bw-Bb column contains t-test statistics and Bp, the corresponding P values filtered at $P < 0.01$ for BLAST sequence relationships. Pw-Pb contains t-test statistics for PSI-BLAST comparisons. Pp column contains P-values for the PSI-BLAST t-test statistic. N represents the number of unique sequences that bear each annotation in the human proteome. Positive t-values denote categories better predicted within species whilst negative t-values denote categories better predicted between species.

Glossary of terms and equations

Name	Definition	Equation
Mutual information	Similarity measure between two vectors x and y	$MI(X:Y) = \sum\sum p(x,y) log \left[\frac{p(x,y)}{p_1(x)p_2(y)} \right]$ (1)
Pearson's correlation	Similarity measure between two vectors x and y	$r_{XY} = \frac{\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$ (2)
Euclidean distance	Ruler distance between two vectors x and y of length N	$d_{XY} = \sqrt{\sum (x-y)^2}$ (3)
Neighbourhood chisq	neighborhood chi-squared frequency method	$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)}$ (4)
TPPK kernel	topology pairwise kernel operating between 4 data items $x_1 to x_4$, each representing a feature vector	$K((x_1,x_2),(x_3,x_4)) = K(x_1,x_3) \times K(x_2,x_4) + K(x_1,x_4) \times K(x_2,x_3)$ (5)
MLPK kernel	Metric learning pairwise kernel operating between 4 data items $x_1 to x_4$, each representing a feature vector	$K((x_1,x_2),(x_3,x_4)) = (K(x_1,x_3) + K(x_1,x_4) - K(x_2,x_3) + K(x_2,x_4))^2$ (6)
Manhatten distance	Distance measure between two vectors x and y of length N.	$d_{XY} = \sum(\lvert x-y \rvert)$ (7)
GO term specificity	A specificity measure for Gene Ontology terms where x represents either the number of child nodes of the term or the frequency of annotations in a sequence population	$GO_{spec} = ln(\sum x + 1)$ (8)
E-value	Expectation value representing the likelihood of observing a sequence similarity score $S\prime$ by chance in a database of a particular size n. m represents the length of the query sequence.	$E = mn2^{-S\prime}$ (9)
Bit score	BLAST alignment bit score computed from the raw alignment score S determined by summing aligned amino acid scores from a substitution matrix and adjusted by K and λ	$S\prime = \frac{\lambda S - lnK}{ln2}$ (10)
Identity	Proportion of identical residues measured between a pair of aligned sequences.	$Identity = \sum_i^{i=1} \frac{1}{len}$ (11)
Fisher's r to z	Fisher's transform for the Pearson correlation coefficient r to Z score	$Z = 0.5 \times \frac{ln(1+r)}{ln(1-r)}$ (12)
Fisher's correlation difference	T test to compare significance of difference between two transformed correlation coefficients. The variance is given by σ for sample sizes N_1 and N_2.	$t = \frac{z_1 - z_2}{\sigma_{1,2}}$ (13) $\sigma_{1,2} = \sqrt{\left(\frac{1}{N_1 - 3}\right) + \left(\frac{1}{N_2 - 3}\right)}$ (14)

... Continued on next page

– continued from previous page

Name	Definition	Equation
Bonferroni	Bonferroni multiple testing adjustment. P-values are multiplied by the number of tests performed to stabilise the family-wise error rate.	$Adjusted_P = p \times N$ (15)
Sensitivity	Sensitivity, coverage or true positive rate represent the proportion of true positives tp recovered by a classifier when a positive assignment is made.	$Sensitivity = \frac{TP}{(TP + FN)}$ (16)
Specificity	Proportion of true negatives recovered by a classifier when a negative prediction is made.	$Specificity = \frac{TN}{(TN + FP)}$ (17)
Precision	Likelihood of a prediction being correct when it has been classified as a positive.	$Precision = \frac{TP}{(TP + FP)}$ (18)
MCC	Matthew's correlation coefficient. The class based equivalent of Pearson's correlation used to assess classifer accuracy.	$MCC = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}$ (19)
Linear kernel	Linear sum of weights kernel	$k(X) = x \cdot x$ (20)
Spline kernel	One step linear approximation to the cubic spline kernel	$k(x) = \prod_{d=1}^{D} 1 + x_i x_j + x_i x_j min(x_i, x_j) - \frac{x_i + x_j}{2} min(x_i, x_j)^2 + \frac{min(x_i, x_j)^3}{3}$ (21)
RBF kernel	Radial Basis Function kernel	$k(x) = \frac{\lvert x_i - x_j \rvert^2}{\sigma}$ (22)
PPI score	protein interaction score	$Score_{A,B} = I \times \left(\frac{\log \frac{f(A)}{\lvert N \rvert} + \log \frac{f(B)}{\lvert N \rvert}}{2} \right)$ (23)
EPQ	Errors Per Query, a measure of classification accuracy. FP represents the number of False Positives, whilst TP represents the number of True Positives.	$EPQ = \frac{FP}{(TP + FP)}$ (24)
Rw	weighted Pearson's correlation	$R_w = \frac{\Sigma w_i x_i y_i - \Sigma w_i x_i \cdot \Sigma w_i y_i}{\left(\frac{\Sigma w_i x_i^2 - \Sigma w_i x_i^2}{\Sigma w_i} \right) \cdot \left(\frac{\Sigma w_i y_i^2 - \Sigma w_i y_i^2}{\Sigma w_i} \right)}$ (25)
f Lambda	lambda distribution	$F^{-1}(\mu) = \lambda_1 + \frac{\mu^{\lambda_3} - (1 - \mu)^{\lambda_4}}{\lambda_2}$ (26)

... Continued on next page

– continued from previous page

Name	**Definition**	**Equation**
Tukey bi-weight	One step tukey bi-weight for robust averaging. ϵ is a small positive constant that avoids division by 0, whilst c controls the degree of smoothing as data items become more distant from the median.	$$T = \frac{\sum wx}{\sum w}$$ $$w = (1 - u^2)^2$$ $$u = \frac{x - m}{(c \cdot s) - \epsilon}$$ $$s = median(\lvert x - median(x)\rvert) \quad (27)$$
Kendalls tau	Correlation co-efficient between ranked pairs. n_c and n_d correspond to the number of concordant and discordant pairs of ranks respectively.	$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (28)$$
Spearman's rank	Correlation co-efficient between ranked pairs. n represents the number of data items in each vector and d is the difference between ranks of two vector values X_i and Y_i. The result is equivalent to Pearson's method calculated between ranked data items.	$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$ $$d_i = x_i - y_i \quad (29)$$
Sum of squares	Sum of squares error between two vectors used to indicated variance about a central fit.	$$SS_{err} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad (30)$$

Abbreviations

BP	Biological Process
BLAST	Basic Local Alignment Search Tool
CATHfus	CATH fusion features
DISO	Disorder features
DNA	Deoxyribonucleic acid
EXPRS	Expression features
GO	Gene Ontology
HMM	Hidden Markov Model
INTACT	Protein-Protein interaction features
K-NN	K-Nearest Neighbours
LOC	Localisation features
MF	Molecular Function
MDS	Multi-dimensional Scaling
NN	Neural Network
PEST	Proline Glutamic acid, Serine and Threonine rich sequences
PFAMfus	PFAM fusion features
PPI	Protein-Protein Interaction
PSI-BLAST	Position Specific Iterated Basic Local Alignment Search Tool
PSSM	Position Specific Scoring Matrix
RBF	Radial Basis Function
RNA	Ribonucleic acid
SS	Secondary Structure
SVR	Support Vector Regression
SVM	Support Vector Machine
SW	Smith Waterman sequence similarity
TAP	Tandem Affinity Purification
TM	Transmembrane features
Y2H	Yeast two hybrid

Bibliography

Addou S, Rentzsch R, Lee D, and Orengo CA. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J Mol Biol*, 387(2):416–420, 2009.

Ahmed FE. Microarray RNA transcriptional profiling: part I. Platforms, experimental design and standardization. *Expert Rev Mol Diagn*, 6(4):535–540, 2006.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 215(3): 403–410, 1990.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3392, 1997.

An A, Viii L, Viii L, Viii L, and Joachims T. Making large-scale support vector machine learning practical, 1998.

Anantharaman V, Koonin EV, and Aravind L. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J Mol Biol*, 307(5):1271–1272, 2001.

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, and Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, 36(Database issue):D419–425, 2008.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, 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 NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, and Zdobnov EM. InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, 16(12):1145–1150, 2000.

Aravind L, Watanabe H, Lipman DJ, and Koonin EV. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*, 97(21):11319–24, 2000.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

Bader GD, Betel D, and Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31 (1):248–250, 2003.

Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*, 28(1):304–305, 2000.

Bairoch A and Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res*, 24(1):21–25, 1996.

Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, and Sherlock G. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*, 33(Database issue):D580–582, 2005.

Bandyopadhyay D, Huan J, Prins J, Snoeyink J, Wang W, and Tropsha A. Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: Ii. case studies and applications. *J Comput Aided Mol Des*, Jun 2009. (ENG).

Barrett T and Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*, 411:352–359, 2006.

Basu MK, Carmel L, Rogozin IB, and Koonin EV. Evolution of protein domain promiscuity in eukaryotes. *Genome Res*, 18(3):449–451, 2008.

Baumgartner Jr. WA, Cohen KB, Fox LM, Acquaah-Mensah G, and Hunter L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–48, 2007.

Ben-Hur A and Noble WS. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1: i38–46, 2005.

Bendtsen JD, Nielsen H, von Heijne G, and Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–785, 2004.

Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

Bjorklund AK, Ekman D, Light S, Frey-Skott J, and Elofsson A. Domain rearrangements in protein evolution. *J Mol Biol*, 353(4):911–913, 2005.

Bland JM and Altman DG. Multiple significance tests: the bonferroni method. *BMJ*, 310(6973):170, Jan 1995.

Blom N, Gammeltoft S, and Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 294(5):1351–1352, 1999.

Bodenreider O and Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform*, 7(3):256–264, 2006.

Brameier M, Krings A, and MacCallum RM. NucPred–predicting nuclear localization of proteins. *Bioinformatics*, 23(9):1159–1160, 2007.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, and Sansone SA. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 31(1):68–71, 2003.

Brenner SE. Errors in genome annotation. *Trends Genet*, 15(4):132–133, 1999.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares Jr. M, and Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1): 262–267, 2000.

Burley SK. An overview of structural genomics. *Nat Struct Biol*, 7 Suppl:932–934, 2000.

Byvatov E and Schneider G. Support vector machine applications in bioinformatics. *Appl Bioinformatics*, 2(2):67–77, 2003.

Cahan P, Rovegno F, Mooney D, Newman JC, r. d. St Laurent G , and McCaffrey TA. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, 401(1-2):12–18, 2007.

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, and Versteeg R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 291(5507):1289–1292, 2001.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, and Botstein D. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, 282(5396):2022–2028, 1998.

Chia JM and Kolatkar PR. Implications for domain fusion protein-protein interactions based on structural information. *BMC Bioinformatics*, 5:161, 2004.

Churchill ME and Travers AA. Protein motifs that recognize structural features of DNA. *Trends Biochem Sci*, 16(3): 92–97, 1991.

Cokus S, Mizutani S, and Pellegrini M. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics*, 8 Suppl 4:S7, 2007.

Copley RR, Letunic I, and Bork P. Genome and protein evolution in eukaryotes. *Curr Opin Chem Biol*, 6(1):39–45, 2002.

Corpet F, Gouzy J, and Kahn D. The ProDom database of protein domain families. *Nucleic Acids Res*, 26(1):323–326, 1998.

Dani DN and Sainis JK. Modularity: a new perspective in biology. *Indian J Biochem Biophys*, 44(3):133–139, 2007.

Date SV. Estimating protein function using protein-protein relationships. *Methods Mol Biol*, 408:109–117, 2007.

Date SV. The Rosetta stone method. *Methods Mol Biol*, 453:169–170, 2008.

Daub CO and Sonnhammer EL. Employing conservation of co-expression to improve functional inference. *BMC Syst Biol*, 2:81, 2008.

Deng M, Zhang K, Mehta S, Chen T, and Sun F. Prediction of protein function using protein-protein interaction data. *Proc IEEE Comput Soc Bioinform Conf*, 1:197–206, 2002.

Deng M, Sun F, and Chen T. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, pages 140–141, 2003.

Devos D and Valencia A. Practical limits of function prediction. *Proteins*, 41(1):98–107, 2000.

Devos D and Valencia A. Intrinsic errors in genome annotation. *Trends Genet*, 17(8):429–431, 2001.

Dobson PD and Doig AJ. Predicting enzyme class from protein structure without alignments. *J Mol Biol*, 345(1): 187–189, 2005.

Doolittle RF. The multiplicity of domains in proteins. *Annu Rev Biochem*, 64:287–314, 1995.

Dozmorov I, Knowlton N, Tang Y, Shields A, Pathipvanich P, Jarvis JN, and Centola M. Hypervariable genes–experimental error or hidden dynamics. *Nucleic Acids Res*, 32(19):e147, 2004.

Du P, Feng G, Flatow J, Song J, Holko M, Kibbe WA, and Lin SM. From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, 25(12):i63–68, 2009.

Dunker AK and Obradovic Z. The protein trinity–linking function and disorder. *Nat Biotechnol*, 19(9):805–806, 2001.

Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, and Villafranca JE. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*, pages 473–474, 1998.

Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, and Uversky VN. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 9 Suppl 2:S1, 2008a.

Dunker AK, Silman I, Uversky VN, and Sussman JL. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol*, 18(6):756–764, 2008b.

Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3):163–167, 1998.

Eisen JA and Fraser CM. Phylogenomics: intersection of evolution and genomics. *Science*, 300(5626):1706–1707, 2003.

Eisen JA and Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor Popul Biol*, 61(4):481–487, 2002.

Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

Eisenberg D, Marcotte EM, Xenarios I, and Yeates TO. Protein function in the post-genomic era. *Nature*, 405(6788): 823–826, 2000.

Engelhardt BE, Jordan MI, Muratore KE, and Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol*, 1(5):e45, 2005.

Enright AJ and Ouzounis CA. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, 2(9):RESEARCH0034, 2001.

Enright AJ, Iliopoulos I, Kyrpides NC, and Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, 1999.

Eyal E, Gerzon S, Potapov V, Edelman M, and Sobolev V. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J Mol Biol*, 351(2):431–442, Aug 2005.

Finn R, Griffiths-Jones S, and Bateman A. Identifying protein domains with the Pfam database. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.5, 2003.

Fontana P, Bindewald E, Toppo S, Velasco R, Valle G, and Tosatto SC. The SSEA server for protein secondary structure alignment. *Bioinformatics*, 21(3):393–395, 2005.

Forslund K and Sonnhammer EL. Predicting protein function from domain content. *Bioinformatics*, 24(15):1681–1687, 2008.

Friedberg I. Automated protein function prediction–the genomic challenge. *Brief Bioinform*, 7(3):225–232, 2006.

Frigyesi A, Veerla S, Lindgren D, and Hoglund M. Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics*, 7:290, 2006.

ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt . Blastclust - blast score-based single-linkage clustering. URL, 2007. URL `ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt`.

ftp://toolkit.lmb.uni muenchen.de/HHsearch/databases . Hhsearch hmm library download. URL, August 2008.

Galperin MY, Walker DR, and Koonin EV. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*, 8(8):779–780, 1998.

Ge H, Liu Z, Church GM, and Vidal M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet*, 29(4):482–486, 2001.

Gerlt JA and Babbitt PC. Can sequence determine function? *Genome Biol*, 1(5):REVIEWS0005, 2000.

Gerstein M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des*, 3(6):497–512, 1998.

Goh SH, Josleyn M, Lee YT, Danner RL, Gherman RB, Cam MC, and Miller JL. The human reticulocyte transcriptome. *Physiol Genomics*, 30(2):172–178, 2007.

Gondo Y, Fukumura R, Murata T, and Makino S. Next-generation gene targeting in the mouse for functional genomics. *BMB Rep*, 42(6):315–323, 2009.

Groth D, Lehrach H, and Hennig S. GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res*, 32(Web Server issue):W313–317, 2004.

Guda C, Fahy E, and Subramaniam S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, 20(11):1785–1794, 2004.

Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, and Brunak S. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj J*, 15(2):115–120, 1998.

Hawkins T, Luban S, and Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci*, 15(6):1550–1556, 2006.

Hawkins T, Chitale M, Luban S, and Kihara D. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins*, 74(3):566–572, 2009.

Hayete B and Bienkowska JR. Gotrees: predicting go associations from protein domain composition using decision trees. *Pac Symp Biocomput*, pages 127–128, 2005.

Healy MD. Using BLAST for performing sequence alignment. *Curr Protoc Hum Genet*, Chapter 6:Unit 6.8, 2007.

Hegyi H and Gerstein M. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res*, 11(10):1632–1640, 2001.

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, and Apweiler R. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–455, 2004.

Hishigaki H, Nakai K, Ono T, Tanigami A, and Takagi T. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, 18(6):523–531, 2001.

Hollich V, Storm CE, and Sonnhammer EL. OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics*, 18(9):1272–1273, 2002.

Hsu CW, Chang CC, and Lin CJ. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003. URL `http://www.csie.ntu.edu.tw/~cjlin/papers.html`.

Hu P, Greenwood CM, and Beyene J. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, 6:128, 2005.

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, and Clamp M. The Ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41, 2002.

Iakoucheva LM and Dunker AK. Order, disorder, and flexibility: prediction from protein sequence. *Structure*, 11(11): 1316–1317, 2003.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 31(4):e15, 2003.

Ispolatov I, Yuryev A, Mazo I, and Maslov S. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res*, 33(11):3629–3635, 2005.

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson Jr. J, Boguski MS, Lashkari D, Shalon D, Botstein D, and Brown PO. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83–87, 1999.

Jansen R, Greenbaum D, and Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37–46, 2002.

Jaroszewski L, Rychlewski L, and Godzik A. Improving the quality of twilight-zone alignments. *Protein Sci*, 9(8): 1487–1496, 2000.

Jensen LJ, Gupta R, Staerfeldt HH, and Brunak S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, 19(5):635–642, 2003.

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, and Bork P. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, 443(7111):594–597, Oct 2006.

Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2): 195–202, 1999.

Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, 2007.

Jones DT and Swindells MB. Getting the most from PSI-BLAST. *Trends Biochem Sci*, 27(3):161–164, 2002.

Jones DT and Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, 53 Suppl 6:573–578, 2003.

Jones DT, Tress M, Bryson K, and Hadley C. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins*, Suppl 3:104–111, 1999.

Joshi T and Xu D. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, 8:222, 2007.

Joshi T, Chen Y, Becker JM, Alexandrov N, and Xu D. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast Saccharomyces cerevisiae. *OMICS*, 8(4):322–323, 2004.

Kabsch W and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2577, 1983.

Kanazawa K and Ashida H. Relationship between oxidative stress and hepatic phosphoglucomutase activity in rats. *Int J Tissue React*, 13(5):225–231, 1991.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, and Yamanishi Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue): D480–484, 2008.

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, and Kasif S. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A*, 101(9):2888–2893, 2004.

Karush W. Minima of functions of several variables with inequalities as side constraints. In *M.Sc Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

Kelly W and Stumpf M. Protein-protein interactions: from global to local analyses. *Curr Opin Biotechnol*, 19(4): 396–403, 2008.

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, and Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988, 2004.

Khan S, Situ G, Decker K, and Schmidt CJ. GoFigure: automated Gene Ontology annotation. *Bioinformatics*, 19 (18):2484–2485, 2003.

Kirac M, Ozsoyoglu G, and Yang J. Annotating proteins by mining protein interaction networks. *Bioinformatics*, 22 (14):e260–260, 2006.

Koski LB and Golding GB. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, 2001.

Krasowski MD, Yasuda K, Hagey LR, and Schuetz EG. Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). *Nucl Recept*, 3:2, 2005.

Krishnamurthy N, Brown D, and Sjolander K. Flowerpower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol Biol*, 7 Suppl 1:S12, 2007.

Kuhn HW and Tucker AW. Nonlinear programming. In Neyman J, editor, *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, Berkeley, CA, USA, 1950.

Kumar S and Carugo O. Consensus prediction of protein conformational disorder from amino acidic sequence. *Open Biochem J*, 2:1–5, 2008.

Kyrpides NC. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects worldwide. *Bioinformatics*, 15(9):773–774, 1999.

Lagreid A, Hvidsten TR, Midelfart H, Komorowski J, and Sandvik AK. Predicting gene ontology biological process from temporal gene expression patterns. *Genome Res*, 13(5):965–969, 2003.

Lanckriet GR, De Bie T, Cristianini N, Jordan MI, and Noble WS. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004a.

Lanckriet GR, Deng M, Cristianini N, Jordan MI, and Noble WS. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, pages 300–301, 2004b.

Lee D, Grant A, Marsden RL, and Orengo C. Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins*, 59(3):603–605, 2005.

Lee D, Redfern O, and Orengo C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 8 (12):995–1005, 2007.

Lee NH and Saeed AI. Microarrays: an overview. *Methods Mol Biol*, 353:265–300, 2007.

Lee SI and Batzoglou S. Application of independent component analysis to microarrays. *Genome Biol*, 4(11):R76, 2003.

Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, and Apweiler R. UniProt archive. *Bioinformatics*, 20(17): 3236–3237, 2004.

Lespinet O, Wolf YI, Koonin EV, and Aravind L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*, 12(7):1048–1049, 2002.

Letovsky S and Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19 Suppl 1:i197–204, 2003.

Liebermeister W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

Lin D. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

Lipman DJ and Pearson WR. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, Kahn A, Ryan MC, Nuhanovic A, Munson PJ, Reinhold WC, Kane DW, and Weinstein JN. AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets. *Bioinformatics*, 23(18):2385–2390, 2007.

Lobley A, Sadowski MI, and Jones DT. pGenTHREADER and pDomTHREADER: New Methods For Improved Protein Fold R ecognition and Superfamily Discrimination. *Bioinformatics*, 2009.

Loganantharaj R and Atwi M. Towards validating the hypothesis of phylogenetic profiling. *BMC Bioinformatics*, 8 Suppl 7:S25, 2007.

Lopez-Bigas N, De S, and Teichmann SA. Functional protein divergence in the evolution of Homo sapiens. *Genome Biol*, 9(2):R33, 2008.

Lord PW, Stevens RD, Brass A, and Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

Lovering RC, Dimmer E, Khodiyar VK, Barrell DG, Scambler P, Hubank M, Apweiler R, and Talmud PJ. Cardiovascular go annotation initiative year 1 report: why cardiovascular go? *Proteomics*, 8(10):1950–1953, May 2008.

Lupas A. Predicting coiled-coil regions in proteins. *Curr Opin Struct Biol*, 7(3):388–393, 1997.

Madeira SC and Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform*, 1(1):24–45, 2004.

Madera M. Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, 24 (22):2630–2631, 2008.

Magidovich E, Fleishman SJ, and Yifrach O. Intrinsically disordered C-terminal segments of voltage-activated potassium channels: a possible fishing rod-like mechanism for channel binding to scaffold proteins. *Bioinformatics*, 22 (13):1546–1550, 2006.

Maglott DR, Katz KS, Sicotte H, and Pruitt KD. NCBI's LocusLink and RefSeq. *Nucleic Acids Res*, 28(1):126–128, 2000.

Manduchi E, Grant GR, He H, Liu J, Mailman MD, Pizarro AD, Whetzel PL, and Stoeckert Jr. CJ. RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinformatics*, 20(4):452–459, 2004.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, and Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*, 30(1):281–283, 2002.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, and Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

Mardis ER. Anticipating the 1,000 dollar genome. *Genome Biol*, 7(7):112, 2006.

Markert CL. Lactate dehydrogenase. Biochemistry and function of lactate dehydrogenase. *Cell Biochem Funct*, 2(3): 131–134, 1984.

Marshall E. Human genome project. Emphasis turns from mapping to large-scale sequencing. *Science*, 268(5215): 1270–1271, 1995.

Martin DM, Berriman M, and Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5:178, 2004.

McGuffin LJ and Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, 19(7):874–881, 2003.

McGuffin LJ, Bryson K, and Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4): 404–405, 2000.

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

Menetrey J and Cherfils J. Structure of the small G protein Rap2 in a non-catalytic complex with GTP. *Proteins*, 37 (3):465–473, 1999.

Mika S and Rost B. Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol*, 2(7):e79, 2006.

Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, and Uversky VN. Analysis of molecular recognition features (MoRFs). *J Mol Biol*, 362(5):1043–1049, 2006.

Moult J and Melamud E. From fold to function. *Curr Opin Struct Biol*, 10(3):384–389, 2000.

Murzin AG, Brenner SE, Hubbard T, and Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, 1995.

Nagano N, Orengo CA, and Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol*, 321(5):741–745, 2002.

Nakai K and Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24(1):34–36, 1999.

Nariai N and Kasif S. Context specific protein function prediction. *Genome Inform*, 18:173–182, 2007.

Noble WS and Ben-Hur A. Integrating information for protein function prediction. In Lengauer T, editor, *Bioinformatics-From Genomes to Therapies*, volume 3, pages 1297–1314, 2008.

Ofran Y, Punta M, Schneider R, and Rost B. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov Today*, 10(21):1475–1482, 2005.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, and Thornton JM. CATH -a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1098, 1997.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, and Brazma A. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35(Database issue):D747–750, 2007.

Pennisi E. Human genome. Reaching their goal early, sequencing labs celebrate. *Science*, 300(5618):409, 2003a.

Pennisi E. Human genome. A low number wins the GeneSweep Pool. *Science*, 300(5625):1484, 2003b.

Perumal S, Antipova O, and Orgel JP. Collagen fibril architecture, domain organization, and triple-helical conformation govern its proteolysis. *Proc Natl Acad Sci U S A*, 105(8):2824–2829, 2008.

Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, and Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9 Suppl 5:S4, 2008.

Pioli PA, Hamilton BJ, Connolly JE, Brewer G, and Rigby WF. Lactate dehydrogenase is an AU-rich element-binding protein that directly interacts with AUF1. *J Biol Chem*, 277(38):35738–45, 2002.

Ponting CP and Russell RR. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002.

Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, and Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9): 1122–1129, 2006.

Punta M and Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol*, 4(10):e1000160, 2008.

R Development Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Ramani AK and Marcotte EM. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*, 327(1):273–274, 2003.

Ranea JA, Yeats C, Grant A, and Orengo CA. Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS Comput Biol*, 3(11):e237, 2007.

Rechsteiner M and Rogers SW. PEST sequences and regulation by proteolysis. *Trends Biochem Sci*, 21(7):267–271, 1996.

Reid AJ, Yeats C, and Orengo CA. Methods of remote homology detection can be combined to increase coverage by 10*Bioinformatics*, 23(18):2353–2360, 2007.

Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

Riley M. Functions of the gene products of Escherichia coli. *Microbiol Rev*, 57(4):862–952, 1993.

Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, and Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–615, 2008.

Rogers MF and Ben-Hur A. The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics*, 25(9):1173–1177, 2009.

Rogers S, Wells R, and Rechsteiner M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science*, 234(4774):364–368, 1986.

Romero P, Obradovic Z, and Dunker AK. Natively disordered proteins: functions and predictions. *Appl Bioinformatics*, 3(2-3):105–113, 2004.

Rost B. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94, 1999.

Rost B, Liu J, Nair R, Wrzeszczynski KO, and Ofran Y. Automatic prediction of protein function. *Cell Mol Life Sci*, 60(12):2637–2640, 2003.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, and Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, 32(18):5539–5545, 2004.

Sadowski MI and Jones DT. The sequence-structure relationship and protein function prediction. *Curr Opin Struct Biol*, 19(3):357–362, 2009.

Sangar V, Blankenberg DJ, Altman N, and Lesk AM. Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics*, 8:294, 2007.

Sansom MS. Ion channels: a first view of K+ channels in atomic glory. *Curr Biol*, 8(13):R450–452, 1998.

Schlicker A, Domingues FS, Rahnenfuhrer J, and Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302, 2006.

Schultz J, Milpetz F, Bork P, and Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, 95(11):5857–5864, 1998.

Serres MH and Riley M. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb Comp Genomics*, 5(4):205–212, 2000.

Serres MH and Riley M. Gene fusions and gene duplications: relevance to genomic annotation and functional analysis. *BMC Genomics*, 6(1):33, 2005.

Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, and Rubio A. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 2(4):330–338, 2005.

Shah AR, Oehmen CS, Harper J, and Webb-Robertson BJ. Integrating subcellular location for improving machine learning models of remote homology detection in eukaryotic organisms. *Comput Biol Chem*, 31(2):138–142, 2007.

Shaw DR. Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr Protoc Bioinformatics*, Chapter 1:Unit1.7, 2009.

Shimizu K, Hirose S, and Noguchi T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, 23(17):2337–2338, 2007.

Sigman M and Cecchi GA. Global organization of the Wordnet lexicon. *Proc Natl Acad Sci U S A*, 99(3):1742–1747, 2002.

Singh AK. Querying and mining biological databases. *OMICS*, 7(1):7–8, 2003.

Sjolander K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, 20(2): 170–179, 2004.

Smith CL, Goldsmith CA, and Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6(1):R7, 2005.

Smith TF and Waterman MS. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981.

Smola AJ, Scholkopf B, and Scholkopf B. A tutorial on support vector regression. Technical report, Statistics and Computing, 2003.

Snel B, Bork P, and Huynen M. Genome evolution. Gene fusion versus gene fission. *Trends Genet*, 16(1):9–11, 2000.

Snitkin ES, Gustafson AM, Mellor J, Wu J, and DeLisi C. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, 7:420, 2006.

Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.

Sonnhammer EL, Eddy SR, and Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–410, 1997.

Sonnhammer EL, Eddy SR, Birney E, Bateman A, and Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322, 1998.

Stevens JR and Doerge RW. Combining Affymetrix microarray results. *BMC Bioinformatics*, 6:57, 2005.

Stewart MG, Kuppersmith RB, and Moore AS. Searching the medical literature on the Internet. *Otolaryngol Clin North Am*, 35(6):1163–1164, 2002.

Storm CE and Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, 2002.

Stuart JM, Segal E, Koller D, and Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

Tasan M, Tian W, Hill DP, Gibbons FD, Blake JA, and Roth FP. An en masse phenotype and function prediction system for Mus musculus. *Genome Biol*, 9 Suppl 1:S8, 2008.

Taylor JS and Cristianini N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, June 2004.

Tian W and Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*, 333(4):863–872, 2003.

Titz B, Schlesner M, and Uetz P. What do we learn from high-throughput protein interaction data? *Expert Rev Proteomics*, 1(1):111–111, 2004.

Todd AE, Orengo CA, and Thornton JM. Evolution of protein function, from a structural perspective. *Curr Opin Chem Biol*, 3(5):548–556, 1999.

Tompa P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, 579(15): 3346–3354, 2005.

Tompa P, Szasz C, and Buday L. Structural disorder throws new light on moonlighting. *Trends Biochem Sci*, 30(9): 484–489, 2005.

Tompa P, Dosztanyi Z, and Simon I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J Proteome Res*, 5(8):1996–2000, 2006.

Tompa P, Prilusky J, Silman I, and Sussman JL. Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins*, 71(2):903–909, 2008.

Torto-Alalibo T, Collmer CW, and Gwinn-Giglio M. The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. *BMC Microbiol*, 9 Suppl 1:S1, 2009.

Tosatto SC, Albiero A, Mantovan A, Ferrari C, Bindewald E, and Toppo S. Align: a C++ class library and web server for rapid sequence alignment prototyping. *Curr Drug Discov Technol*, 3(3):167–173, 2006.

Troyanskaya OG, Dolinski K, Owen AB, Altman RB, and Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A*, 100 (14):8348–8353, 2003.

Uversky VN, Oldfield CJ, and Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit*, 18(5):343–344, 2005.

Uversky VN, Radivojac P, Iakoucheva LM, Obradovic Z, and Dunker AK. Prediction of intrinsic disorder and its use in functional proteomics. *Methods Mol Biol*, 408:69–92, 2007.

van Noort V, Snel B, and Huynen MA. Predicting gene function by conserved co-expression. *Trends Genet*, 19(5): 238–242, 2003.

Vazquez A, Flammini A, Maritan A, and Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21(6):697–700, 2003.

Vert JP, Qiu J, and Noble WS. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S8, 2007.

Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S, and Konig R. GOPET: a tool for automated predictions of Gene Ontology terms. *BMC Bioinformatics*, 7:161, 2006.

Vogel C, Bashton M, Kerrison ND, Chothia C, and Teichmann SA. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*, 14(2):208–216, 2004.

Vucetic S, Brown CJ, Dunker AK, and Obradovic Z. Flavors of protein disorder. *Proteins*, 52(4):573–574, 2003.

Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, and Uversky VN. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res*, 6(5):1899–1906, 2007.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, and Feldman MW. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A*, 102(15):5483–5488, 2005.

Wang JZ, Du Z, Payattakool R, Yu PS, and Chen CF. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, 2007.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, and Jones DT. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–2139, 2004a.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, and Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, 337(3):635–635, 2004b.

Williams KR, Reddigari S, and Patel GL. Identification of a nucleic acid helix-destabilizing protein from rat liver as lactate dehydrogenase-5. *Proc Natl Acad Sci U S A*, 82(16):5260–5264, 1985.

Wilson CA, Kreychman J, and Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, 297(1): 233–239, 2000.

Wu J, Hu Z, and DeLisi C. Gene annotation and network inference by phylogenetic profiling. *BMC Bioinformatics*, 7:80, 2006.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, and Eisenberg D. DIP: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289–291, 2000.

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, and Uversky VN. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res*, 6(5):1917–1922, 2007a.

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, and Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res*, 6(5):1882–1888, 2007b.

Yanai I, Wolf YI, and Koonin EV. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*, 3(5):research0024, 2002.

Yao Z and Ruzzo WL. A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics*, 7 Suppl 1:S11, 2006.

Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, and Orengo C. Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res*, 36(Database issue):D414–418, 2008.

Yeung KY, Medvedovic M, and Bumgarner RE. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biol*, 5(7):R48, 2004.

Zehetner G. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res*, 31(13):3799–3803, 2003.

Zhang P, Zhang J, Sheng H, Russo JJ, Osborne B, and Buetow K. Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, 7:135, 2006.

Zhang Y, Stec B, and Godzik A. Between order and disorder in protein structures: analysis of ”dual personality” fragments in proteins. *Structure*, 15(9):1141–1147, 2007.

Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, and Wong WH. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23(2):238–243, 2005.

Zmasek CM and Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, 2002.