High Specificity Automatic Function Assignment for Enzyme Sequences

Daniel Lee Roden

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of
University College London

Department of Computer Science
University College London

April 2011

I, Daniel Roden, 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.

Acknowledgements

Firstly, I would like to thank my supervisor, David Jones, for providing invaluable support and advice throughout. I would also like to thank all of the members of the Jones lab for providing a great environment in which to work and research.

Love and thanks to my parents, family and friends for all their support and encouragement. Last, but far from least, Justine, for providing me with a wonderful bedrock of love throughout.

Abstract

The number of protein sequences being deposited in databases is currently growing rapidly as a result of large-scale high throughput genome sequencing efforts. A large proportion of these sequences have no experimentally determined structure. Also, relatively few have high quality, specific, experimentally determined functions.

Due to the time, cost and technical complexity of experimental procedures for the determination of protein function this situation is unlikely to change in the near future. Therefore, one of the major challenges for bioinformatics is the ability to automatically assign highly accurate, high-specificity functional information to these unknown protein sequences. As yet this problem has not been successfully solved to a level both acceptable in terms of detailed accuracy and reliability for use as a basis for detailed biological analysis on a genome wide, automated, high-throughput scale.

This research thesis aims to address this shortfall through the provision and benchmarking of methods that can be used towards improving the accuracy of high-specificity protein function prediction from enzyme sequences. The datasets used in these studies are multiple alignments of evolutionarily related protein sequences, identified through the use of BLAST sequence database searches.

Firstly, a number of non-standard amino acid substitution matrices were used to re-score the benchmark multiple sequence alignments. A subset of these matrices were shown to improve the accuracy of specific function annotation, when compared to both the original BLAST sequence similarity ordering and a random sequence selection model.

Following this, two established methods for the identification of functional specificity determining amino acid residues (fSDRs) were used to identify regions within the aligned sequences that are functionally and phylogenetically informative. These localised sequence regions were then used to re-score the aligned sequences and provide an assessment of their ability to improve the specific functional annotation of the benchmark sequence sets.

Finally, a machine learning approach (support vector machines) was followed to evaluate the possibility of identifying fSDRs, which improve the annotation

accuracy, directly from alignments of closely related protein sequences without prior knowledge of their specific functional sub-types. The performance of this SVM based method was then assessed by applying it to the automatic functional assignment of a number of well studied classes of enzymes.

Contents

List of Figures

List of Tables

Abbreviations

AC	Adenylyl Cyclase
ATP	Adenosine-5'-triphosphate
AMP	Adenosine Monophosphate
AUC	Area Under Curve
BETE	Bayesian Evolutionary Tree Estimation
BLAST	Basic Local Alignment Search Tool
CHIEFc	Conservation-controlled HMM Iterative Procedure for Enzyme Family classification
COGs	Clusters of Orthologous Groups
CSA	Catalytic Site Atlas
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic Acid
EFICAz	Enzyme Function Inference by Combined Approach
EC	Enzyme Commission
ET	Evolutionary Trace
FN	False Negative
FP	False Positive
FPR	False Positive Rate
fSDR	function Specificity Determining Residue
GC	Guanylyl Cyclase
GMP	Guanosine Monophosphate
GO	Gene Ontology
GOA	Gene Ontology Annotation
GTP	Guanosine Triphosphate
HMM	Hidden Markov Model
HSP	High Scoring Pair
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LDH	Lactate Dehydrogenase
MCC	Matthews Correlation Coefficient
MDH	Malate Dehydrogenase
MSA	Multiple Sequence Alignment
NAD(+)	Nicotinamide Adenine Dinucleotide
NADH	Nicotinamide Adenine Dinucleotide (reduced form)
NN	Neural Network
PAM	Percent Accepted Mutation
PEDANT	Protein Extraction, Description and Analysis Tool
PSSM	Position Specific Scoring Matrix
PDB	Protein Data Bank
PSI-BLAST	Position Specific Iterated Basic Local Alignment Search Tool
RBF	Radial Basis Function
RE	Relative Entropy

ROC	Receiver Operator Characteristic Curve
SCOP	Structural Classification of Proteins
SE	Standard Error
SeqID	Sequence Identity
SOM	Self Organising Map
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TRE	Total Relative Entropy

Chapter 1 Introduction and Background

1.1 Protein Function

The native state conformation of a protein is essential for its biological activity. Because the structure of the native state is defined by the amino acid sequence, it follows that the precise biological function of a protein is strongly dependant on both sequence and structural properties. Protein function can be a difficult concept to rigorously and unambiguously define and categorise. A general biological description of protein function usually involves a description on three levels:

- **Biological Function**: This describes the effects of the protein on the entire organism;
- **Cellular Function**: This level provides a description of the interactions and pathways that a protein is involved in on a cellular level; and
- **Molecular Function:** Providing a description of the precise biochemical activity of a protein at a molecular level.

A number of functional classification schemes have been proposed towards solving the function categorisation problem, a number of which are described below.

Functional Specificity

The sub-categorisation of function leads to increasingly more detailed, specific descriptions of functions that proteins can perform. Therefore, the concept of functional specificity can be thought of as a hierarchical classification, moving from a general, not very specific description (such as "enzyme"), to a progressively more detailed description of a protein (such as "alcohol dehydrogenase"). It is this detailed form of description and classification that is of major interest in this thesis.

1.1.1 Protein Function Classification Schemes

Several schemes for the description and classification of proteins and their functional properties have been developed (Ouzounis et al., 2003; Whisstock and Lesk, 2003; Riley, 1998). The aim of functional classification schemes is the descriptive categorisation of similar protein functions. There have been attempts which both

concentrate on single organism categorisation (generally associated with a particular genome sequencing project) and also more general classification schemes that either apply to all types of proteins or a particular sub-type such as the enzymes. I will concentrate below on two widely used schemes: the enzyme commission and gene ontology classification schemes.

1.1.1.1 Enzyme Commission Classification Scheme

The enzyme commission (EC) classification is a hierarchical classification scheme for the description of enzyme function and catalysed reactions. This is a well established and widely used scheme, the specific details of which can now be found online (http://www.chem.qmul.ac.uk/iubmb/enzyme/). A database resource, called ENZYME (Bairoch, 1993; Bairoch, 2000), is available, which provides links from the EC descriptions to associated protein sequence databases, such as UniProt (Apweiler et al., 2004).

The structure of the EC naming scheme takes on the form of a four level hierarchy (EC A.B.C.D). The top level (A) consists of six principal enzyme classes, these are: (1) EC 1 – the Oxidoreductases; (2) EC 2 – the Transferases; (3) EC 3 – the Hydrolases; (4) EC 4 – the Lyases; (5) EC 5 – the Isomerases; and (6) EC 6 – the Ligases. The other levels are dependent on the principal class and sub-classify each into progressively more detailed specifics regarding the enzyme reaction catalysed.

The problems associated with this classification scheme, with respect to its use as a description of protein function are well documented (Whisstock and Lesk, 2003; Babbitt, 2003). The main point of caution is that the EC scheme nomenclature was designed as a way of describing the reactions catalysed and not specifically the sequence or structural features of the proteins which catalyse them. A further point of note, especially important in terms of automated function prediction and annotation methods, is the "functional distance" between the specific functional descriptions (Pawlowski et al., 2000). For example, when comparing proteins which have different substrates, it is not always clear from the description the precise degree of difference in the biochemical reactions or the functional properties of the proteins involved. Generally this is overlooked and a simple correlation is assumed between the level of functional specificity and the number of matching values in the four-level EC hierarchy. This problem of functional distance between alternate

protein functions is one that is important when considering the specific accuracy levels and therefore the benchmarking of protein function prediction methods.

1.1.1.2 The Gene Ontology

A more general and detailed classification scheme for all classes of proteins is provided by the gene ontology (GO) project (Ashburner et al., 2000). The gene ontology is designed as a structured ontology with three sections describing the biological processes, cellular components and biological functions of the associated genes or gene products. GO terms are represented by a directed acyclic graph (DAG) in which the level of functional specificity increases as the graph is descended from a more general classification at a 'parent' node to a more specific function at a 'child' node. *Figure 1.1* shows an overview of some of the terms at the top of the GO hierarchy for each of the three main categories. A more detailed view of the ontology can be browsed using the interactive tools available online (www.geneontology.org).

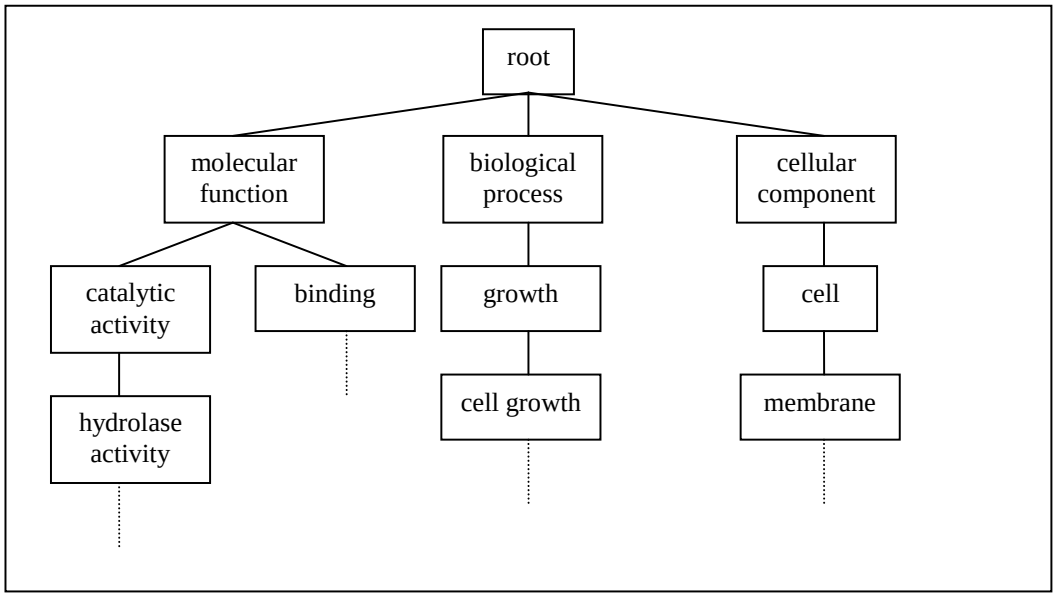

Figure 1.1. Example showing a selection of gene ontology terms. For clarity, not all possible gene annotations are shown at each level.

Concerted efforts are currently underway to provide detailed GO annotations for genes and gene products in major sequence databases and for particular genomes (Gene Ontology Annotation (GOA) project (Camon et al., 2004)). Also, evidence codes are being used in the gene ontology for recording the source of the annotations. This is particularly important for judging the quality and reliability of

the annotated data, especially when benchmarking the reliability of automated methods. There are a number of evidence codes provided for inferring the source, however, the most important distinction is between those that have arisen from expert human manual annotation and those from automated methods.

The gene ontology is currently the most comprehensive general classification available for proteins and is becoming the standard for use in annotation projects and prediction experiments. However, the complexity of the gene ontology requires careful consideration when measuring functional distances, especially with regards to the levels of functional specificity.

1.1.2 Classification of Protein Sequence and Structure

Through evolutionary analysis of the sequence and structural properties of proteins, patterns and relationships become apparent, allowing classification into families of homologous proteins (Orengo and Thornton, 2005). In general it is possible to consider the classification of proteins using clustering algorithms based on sequence or structural similarity measures to define hierarchies. The categories range from general, commonly shared properties at high similarity, to a finer granularity when considering lower levels of similarity. With respect to understanding protein function these classifications can provide important information, as often there is correlation between sequence, structural and functional similarity (Todd et al., 1999). However, the level of sequence and structural similarity is not always a reliable measure of function, meaning more powerful methods of analysis are required, especially when considering specific detailed functional properties.

1.1.3 Evolution and Protein Function

Central to the creation of new protein functions are evolutionary mechanisms and homologous relationships. The continuing accumulation of sequence and structural information is producing significant breakthroughs in the understanding and methods used for analysis of evolutionary aspects of protein sequence, structure and function. Some important concepts relevant to this area are described below.

1.1.3.1 Evolutionary Divergence

During gene replication, mutations can arise in the DNA sequences, producing either synonymous or non-synonymous substitutions. Due to the redundancy of the genetic

code some mutations within codons will produce no change in the translated amino acid sequence (synonymous substitutions). However, others will produce mutations in the translated amino acid sequences (nonsynonymous substitutions). Synonymous substitutions are important when analysing changes in DNA sequences, especially when measuring rates of evolutionary change. The emphasis of this work, however, is on the functions of proteins and therefore nonsynonymous mutations are those of most interest.

The gradual accumulation of mutations from a common ancestor through the process of natural selection is known as divergence. This is the mechanistic basis for both the diversity and similarity seen between groups of homologous proteins when they are classified into sequence, structural and functional families. An understanding of the effects of these mutations is vital for studying changes of functional specificity between homologous proteins and the subsequent development of methods for accurate prediction of function from sequence.

1.1.3.2 Gene Duplication

A key mechanism in the development of new protein functions is that of gene duplication (Ohno, 1970; Taylor and Raes, 2004). Whenever a duplication event occurs, a redundant copy of the gene is created within the organism. Like other mutation events, gene duplication can be advantageous, deleterious or neutral. In general a duplicated gene will be free from evolutionary constraints to undergo divergence, possibly leading to the development of a new specific function without impairing the fitness of the organism. Although the gene pair will be related by a single common ancestor, the two copies may evolve along different pathways creating separation of function, leading to new protein sub-functionalisation.

There are many reported examples of divergent evolution producing changes in the specificity of protein function (Whisstock and Lesk, 2003). A commonly used example is that of the serine proteinases. This is a good example of the possibilities of functional divergence because it shows examples of both the gradual change in specificity through gradual mutational divergence and also large changes in function through point mutations of small numbers of important functional residues (Patthy, 1999).

1.1.3.3 Orthologous and Paralogous Relationships

An important consideration when analysing the evolutionary history of genes, proteins and their functions is the effect of speciation. Two definitions (Fitch, 1970) are required to describe the relationship between genes in different species and gene pairs within the same species:

Orthologs: Are genes in different genomes that have been created by the separation of species, through speciation;

Paralogs: Are genes in the same genome that have been created by gene duplication events.

Identification and discrimination between orthologous and paralogous proteins is an important area for the study and prediction of specific protein functions and also to the field of comparative genomics. The availability of complete genome sequences makes possible attempts to identify and classify orthologous proteins. One approach to this is the clusters of orthologous groups (COGs) method (Tatusov et al., 1997; Tatusov et al., 2003), which uses an all-against-all BLAST based sequence similarity search to identify sets of proteins that occur in at least three different divergent genomes.

Orthologous proteins generally carry out identical or at least very similar functions in their respective genome, because of this, their identification and categorisation can be of particular importance when considering methods for the prediction of function. Accurate differentiation of orthologs and paralogs at different evolutionary distances should provide important information for the separation of specific functional groupings.

1.1.3.4 Sequence Similarity Database Searching

Fast, reliable and efficient solutions are required to identify similarities and possible evolutionary relationships between large numbers of protein sequences. Database search techniques have been developed for this purpose, taking a query sequence as input to provide similarity measures to all other sequences in the search database. The first methods developed for this purpose were FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990), which provided improvements in speed

over dynamic programming methods. The efficiency gains of these methods are provided by the use of heuristic "k-tuple" search techniques which look for matching patterns of consecutive characters of length k in the query and the search database sequences. A local alignment from these seed patterns is then generated to provide similarity scores and identify high scoring pairs (HSPs) of sequences.

An important feature of these methods is their use of a robust statistical framework for calculating the significance of matches between the query and aligned sequences. A value called the expect value (E-value) is used as the basis for this through the use of extreme value statistics. It represents the number of times that you would expect to get the match score observed between a pair of sequences by chance, using a database of known size. Parameters such as the database size and aligned sequence lengths affect this value and should be taken into account when interpreting the output (Jones and Swindells, 2002).

To improve the sensitivity and allow the reliable identification of more distant sequence homologues, powerful profile-based search techniques have been developed. These provide identification of possible homologues at lower values of sequence identity, within a region commonly known as the twilight zone (Feng and Doolittle, 1996). Profile based (Gribskov et al., 1987) and probabilistic methods for sensitive database searching are based on the residue conservation patterns observed from multiple sequence alignments. A widely used extension to the BLAST algorithm is PSI-BLAST (Altschul et al., 1997; Schaffer et al., 2001), which implements an algorithm that carries out iterated database searches using sequence profiles generated from position specific scoring matrices (PSSMs).

Other sensitive search techniques have been developed that use hidden Markov models (HMM) (Eddy, 1996) to generate probabilistic models of residue conservation. Although these methods are more sensitive than the PSSM based profile methods (such as PSI-BLAST) they are also more computationally expensive.

1.1.3.5 Multiple Sequence Alignments

Multiple sequence alignments (MSAs) provide a powerful method for the analysis of evolutionary relationships between families of protein sequences. Columns of

conserved properties within multiple alignments generally indicate structurally and functionally important regions. A number of methods have been developed towards improving the overall sensitivity of multiple alignment approaches (such as: CLUSTALW (Thompson et al., 1994); T-COFFEE (Notredame et al., 2000); and Gotoh, 1999). The most commonly used is progressive alignment, which is based on heuristics that attempt to exploit evolutionary relationships between homologous sequences through the use of a guide tree. The heuristic nature of these algorithms does not guarantee an optimised set of alignments but the advantages of speed and computational efficiency provided compensate for this.

1.2 Automatic Protein Function Prediction

Accurate, reliable and fully automated methods for the prediction of protein function are of major importance in the area of computational biology and bioinformatics analysis. Its importance continues to grow in tandem with the continuing growth of available sequence data from high-throughput genome sequencing projects (Lander et al., 2001; Venter et al., 2001) and structural data from structural genomics projects (see website: http://sg.pdb.org). The difference between available sequence and structural data is significant. As of January 2009, there are 6,964,485 sequences in UniProt release version 39.6 (Apweiler et al., 2004) compared to 55,271 solved structures in the PDB (13-Jan-2009) (Berman et al., 2000). With regards to available functional annotation data, statistics from the gene ontology annotation (GOA UniProt version 67.0) project (Camon et al., 2004) show that there are currently 86,332 distinct proteins that have been manually annotated with GO functional terms. There is clearly a need for automated annotation methods to supplement the data currently available. A number of good reviews are available (Whisstock and Lesk, 2003; Rost, 2003; Valencia, 2005; Watson et al., 2005) covering a range of areas important for prediction and annotation. Here, the aim is to provide a detailed discussion, related to my research, of previous work carried out on sequence based methods for protein function prediction. Particular attention is focussed on methods for accurately discriminating specific functions between homologous groups of proteins.

1.3 Sequence Homology Based Function Prediction Methods

1.3.1 Homology Transfer

The principle method for identifying the function of an unknown protein sequence is through the use of database similarity search techniques such as BLAST (Altschul et al., 1990) or PSI-BLAST (Altschul et al., 1997). A typical approach would be to assign the function of a closely related homolog to that of an unknown query, using a particular threshold of sequence similarity or statistical significance for deciding the reliability of the annotation transfer.

1.3.1.1 Analysis of the Correlation between Sequence, Structure and Functional Similarity

A number of research groups have systematically analysed the correlation between protein sequence similarity and the level of functional conservation. Studies of this kind aim to provide a measure of the accuracy and error associated with using sequence similarity thresholds for the transfer of function. The variation in the analytical methods used has led to discrepancy for specific thresholds between levels of sequence similarity measure and functional conservation (Valencia, 2005). However, a general trend is observed in all the results. As the sequence similarity increases the level of functional conservation also increases, showing a correlation between similarity of sequence and function (Wilson et al., 2000). Although this is also true for differing levels of functional specificity, in general, the more specific the level of function the higher the sequence similarity required for correlation and therefore accurate transfer of function.

1.3.1.2 Analysis of Single-Domain Proteins

An early study by Hegyi and Gerstein (1999) of the relationship between SCOP (Murzin et al., 1995) structural domains and their enzyme function (as specified by the enzyme commission (EC) classification scheme) showed a correlation between major SCOP fold classes and broad functional categories. This analysis was then extended to other structural and functional classification schemes for a detailed analysis of the yeast genome, with an observed fold-function correlation for a number of functional properties. Martin et al. (1998) also investigated the

relationship between general EC class and the CATH fold classification. In this study it was found that the fold was related more closely to the ligand type rather than top level EC number classification.

This work was followed by a number of studies that attempted to get firm threshold values for functional annotation transfer at varying levels of functional specificity. It is difficult to make direct comparisons between all these due to the different methods and functional classification schemes used; however, a summary of these results highlights certain trends:

- Wilson et al. (2000) showed (using a combined ENZYME and FLYBASE (Gelbart et al., 1994) functional classification scheme) that precise function was conserved down to 40% sequence identity and broad functional class down to around 25%.

- Devos and Valencia (2000) used both EC numbers and Swiss-Prot keywords as measures of functional equivalence. Concentrating on the EC conservation results (these are commonly used and therefore more easily comparable between other studies, also the change in level of specificity is easier to see) they state that above 70% sequence identity is required for reliable transfer of all 4 EC numbers, 50-70% for the conservation of the first 3 EC numbers, and that below 30% assignments of function based on sequence identity become problematic.

- Todd et al. (2001) carried out a similar study to Devos and Valencia, using single and multi-domain proteins from CATH (Orengo et al., 1997), with EC numbers as the measure of functional conservation. The results show that the first three EC numbers are conserved with an accuracy of 90% above a 30% sequence identity threshold and that above 40% variation in the fourth EC number becomes rare.

1.3.1.3 Extension of Analysis to Include Multi-Domain Proteins

Due to the importance of multi-domain proteins, especially in eukaryotic genome analysis, some of the above methods have been extended to incorporate multi-domain proteins. Hegyi and Gerstein (2001) extended their earlier work (Hegyi and Gerstein, 1999) and that of Wilson et al. (2000), including both single and multi-domain proteins in a similar analysis. Multi-domain proteins were again taken from

Swiss-Prot and identified as those showing a match to at least two domains of known structure belonging to different SCOP superfamilies. Functional categories were defined using Swiss-Prot keywords. The results showed that there was significantly more conservation of accurate transfer of approximate function for the single (67%) domain proteins compared to the multi-domain (35%), although this value rose to 80% when two domain folds are shared.

Rost (2002) approached an analysis of sequence similarity and conservation of EC numbers in the Swiss-Prot database (Apweiler et al., 2004) with the aim of reducing the effect of the inherent bias in the sequence databases. This bias is proposed to arise from experimental bias in the type of sequence data deposited and also high levels of sequence redundancy. The results obtained by Rost show a clear difference to those of earlier studies, suggesting that the sequence identity threshold required for accurate functional annotation transfer is higher than previously reported. With more than 70% sequence identity required for accurate transfer of all four levels of EC numbers.

Tian and Skolnick (2003) followed this study, also using enzymes, taking into account bias in both functional and sequence properties. This method proposed that a further bias exists in terms of the represented enzyme functional groupings in Swiss-Prot. The figures they obtained were not as pessimistic as those of Rost (2002), but still showed less conservation than most of the other studies previously discussed - suggesting a 60% sequence identity threshold for accurate transfer of all four EC number levels.

The studies of Rost (2002) and Tian and Skolnick (2003) both also looked at the correlation of BLAST and PSI-BLAST E-values with enzyme functional conservation. These show the same general trend seen in the correlation with sequence identity. As statistical significance of the matches decreases, the reliability of specific functional prediction also decreases, and even at particularly significant (low) E-values there are still examples that show incorrect functional conservation. These findings are particularly important because they show that even statistically very significant matches, obtained from powerful homology recognition techniques, can produce incorrect functional assignment.

Although arguments relating to the best datasets to use and the corresponding correct process for removal of bias will most probably continue, the general conclusion is clear. Sequence similarity methods are generally a good indicator of general function, however, they become less reliable when either the level of specificity required is increased or the similarity is reduced. Caution is therefore required when using simple transfer of homology techniques for functional annotation, especially when considering high specificity applications (Devos and Valencia, 2001).

1.3.2 Sources and Extent of Database Annotation Errors

A major concern of automatic annotation efforts is the proliferation of erroneous functional database annotations (Brenner, 1999; Devos and Valencia, 2001; Iliopoulis et al., 2003). Possible reasons proposed for the source of the mistakes in annotation include: insufficient level of sequence similarity used for the annotation; typographical errors; and use of previous incorrect annotations for new annotation. An analysis of the propagation of database errors has been carried out using mathematical modelling techniques which suggests that the annotation errors may grow at an exponential rate with the growth of database sizes (Gilks et al., 2002). Guidelines for successful annotation strategies are described by Iliopoulis et al. (2003). Probably the most important of these is the clear indication and reliability of the source annotation, which constitutes an important part of the GO annotation project and also the detailed information fields in Swiss-Prot. Levels of reliability for automated annotation results can be given depending on whether the source annotation is from a manual expert annotation or a previous automated annotation. A further source of improvement to the quality of annotations, discussed by Ouzounis and Karp (2002), is the regular re-annotation of databases. The time consuming nature of this type of procedure necessitates full automation providing further weight to the need for high-quality automated tools for functional annotation.

1.3.3 Low Specificity Automated Function Prediction

A number of methods have been proposed for automated high throughput annotation of genome sequences. Generally these are based on the understanding that there are problems with the homology based approaches, however, for a large number of annotations, especially when considering more general, lower functional specificity, the accuracy is acceptable.

1.3.3.1 GeneQuiz

One of the earliest automated functional annotation systems was *GeneQuiz* (Andrade et al., 1999) and consists of a combination of sequence similarity (BLAST and FASTA) and rule-based processing algorithms for annotation of both general and more specific functional class. A web-site with details of full genome analyses is available (http://jura.ebi.ac.uk:8765/ext-genequiz/). An overview of the system shows a general methodology common in many of the automated systems:

- A sequence similarity threshold is initially applied to select the most similar sequence pairs to the unknown query sequence;
- Analysis of existing functional annotations (in the case of *GeneQuiz* this is through rule based lexical analysis of functional keywords and EC numbers) is carried out to obtain a consensus result of the most reliable function descriptions to apply to the query sequence;
- Application of annotation to query sequence, sometimes with an indication to the level of reliability of the assignment;
- Option for further manual analysis and editing of the result through use of additional "support methods", such as, multiple sequence alignments and motif database searches.

Assignment of function using a method like *GeneQuiz* shows some improvement over "top-hit" homology transfer because the derived functional annotations are based on a combination of sequence similarity, database quality and source annotation quality.

1.3.3.2 Automatic Annotation of TrEMBL Database

An important system to consider is one underpinning the automated annotation of the TrEMBL (Apweiler et al., 2004) section of the UniProt protein database resource. The algorithmic details and information flow of the system are described in detail elsewhere (Moller et al., 1999; Fleischmann et al., 1999; Kretschmann, 2001), but a look at the overview of the methods used shows a similar (but more complex) integrated rule-based processing approach to that of *GeneQuiz*. An important consideration in the design of this system was that the aim of TrEMBL is

to eventually move all sequences into the related Swiss-Prot database; therefore the rules used for automated annotation are used to help inform the manual annotation procedure.

1.3.3.3 PEDANT (Protein Extraction, Description and Analysis Tool)

The aim of PEDANT (Riley et al., 2005 – http://pedant.gsf.de) is to produce a software system capable of a number of genome scale sequence analysis tasks. This includes automated analysis of protein function based on high-stringency BLAST sequence similarity searches to identify manually annotated homologous proteins for function transfer. A number of different functional classification schemes are used, including EC numbers. The system also assigns sequences according to COGs (Tatusov et al., 2003) and carries out sequence motif and pattern detection searches against a number of sequence motif databases. Although the system provides methods to prevent proliferation of potentially incorrect automatic annotations it is still based on fairly simple sequence similarity based search techniques, and will therefore suffer from the problems already discussed when considering high-specificity predictions.

Recent efforts towards large-scale protein sequence annotation have concentrated on the gene ontology (GO) framework as the basis for the functional classification (for example: Xie et al., 2002; Martin et al., 2004). Again, the main basis for these methods is the use of similarity based search techniques with additional filters to refine the predictions. Xie et al. (2002) describe a method that incorporates a clustering algorithm based on the sequence identity and BLAST E-values to group proteins with potentially similar GO terms. The reliance of this method on text-parsing of annotation literature sources means that it is limited by the quality of the text processing engine and the availability of good literature sources.

The GOtcha method (Martin et al., 2004) is compared to the top-scoring BLAST hit for each input sequence. The key factors related to this method are the accuracy and confidence estimates provided for each annotation. Overall though, the method is aimed at providing greater annotation coverage rather than a major improvement in the level of specificity of predictions.

1.3.3.4 General Limitations

Most of the methods described above suffer from a number of limitations, especially when considering their application to high-quality, reliable and high specificity function annotations. For some of the methods this is down to the fact that the inherent design is for increased coverage of annotations, at a cost of a fairly general level of specific functional classification.

A number of approaches have been proposed for a more detailed analysis of protein function allowing the identification of specific functional sub-types from groups of closely related proteins. Many of these methods aim to take advantage of information describing evolutionary relationships within protein families. These will be the focus of the next section and are of most interest for this thesis.

1.3.4 High-Specificity Phylogenetic Approaches to Protein Function Prediction

One of the main limitations of sequence homology transfer methods, for function prediction, is their performance at identifying specific functional subfamilies in closely related families of sequences. It has been shown that both phylogenetic reconstruction (Eisen, 1998; Eisen and Wu, 2002; Sjolander, 2004; Johnson and Church, 2000) and the identification of functionally determining residues (Livingstone et al., 1993; Casari et al., 1995; Hannenhalli and Russell, 2000; del sol Mesa, 2003; Lichtarge et al., 1996) in functionally related protein families, through the use of multiple sequence alignment (MSAs), can help towards improving the specificity of protein function predictions. With the continued increase in available protein sequences and full genome sequence sets these evolutionary methods are becoming more powerful and important for function analysis.

1.3.4.1 Phylogenetic Reconstruction Methods

Increased sequence information has led to an increase in the use of molecular phylogenetic techniques for analysis and prediction of protein function from sequence. There are three reviews of particular importance in this area (Eisen, 1998; Eisen and Wu, 2002; Sjolander, 2004), describing ways in which phylogeny can be most effectively combined with sequence analysis to improve methods for automated function prediction. A closely related area of research, which is discussed in the next

section, uses phylogenetic information to identify functionally important amino acid residues.

The review by Sjolander (2004) provides an overview and discussion (see *figure 1.2)* of the important stages for the prediction of function using molecular phylogeny. This methodology is an expanded form of that originally proposed by Eisen (1998). Not all stages in this methodology are investigated during the experiments in this thesis, however, it provides a good basis for discussion of some key areas and previous work, related to the prediction of protein function using molecular phylogeny based techniques.

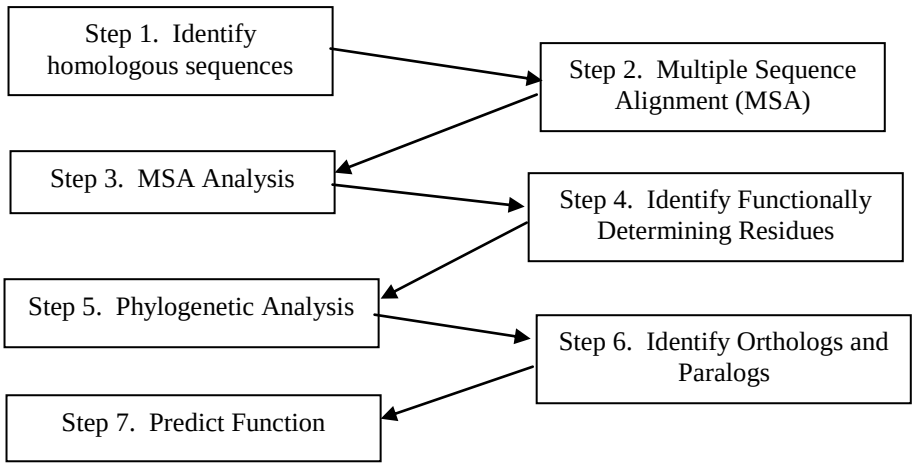

Figure 1.2. Flowchart showing the key stages in molecular phylogenetic analysis of protein function. Adapted from Sjolander (2004).

1.3.4.2 Identification of Homologous Sequences

The first stage is the collection of sequences homologous to the unknown query protein. Three potential limitations are highlighted when using homolog detection for phylogenetic analysis; these are: (i) analysis of protein domains; (ii) possible inclusion of false positives (non-homologs); and (iii) profile drift due to iterated searches. The effects of the second and third problems can be reduced by a number of means, the most obvious being the use of more conservative parameters when including related sequences in the iterated homology search.

1.3.4.3 Multiple Sequence Alignment

High quality MSAs are essential for the accurate and reliable algorithmic reconstruction of phylogenetic trees. A number of applications are available for

multiple sequence alignment; some commonly used ones are: CLUSTAL-W (Thompson et al., 1994); T-COFFEE (Notredame et al., 2000); MAFFT (Katoh et al., 2002); and MUSCLE (Edgar, 2004). When considering automated approaches, a compromise must be reached between the quality of the alignments and the computational efficiency. A further, computationally less demanding source of multiple sequence alignments is from the output of PSI-BLAST through use of the *–m 6* output parameter. These are essentially a concatenation of the multiple pair-wise sequence alignments identified by the sequence database search.

Methods for assessing the quality and reliability of regions within multiple sequence alignments have been proposed (e.g. Tress et al., 2003). This type of reliability analysis is important for the accurate detection of conserved functionally determining residues, which is discussed in detail below. There have also been studies that look at reducing the level of sequence redundancy in multiple sequence alignments. An interesting method based on the multi-dimensional QR factorisation of multiple sequence alignments has been proposed by Sethi et al. (2005). This algorithm is specifically designed to reduce evolutionary redundancy in groups of homologous sequences to produce evolutionary optimal sequence sets for phylogenetic analysis.

1.3.4.4 Phylogenetic Analysis and Tree Construction

Algorithmic methods for phylogenetic tree construction are well studied. Sjolander (2004) concludes that the computational efficiencies of distance based reconstruction algorithms (such as neighbour joining) compared to character based (such as maximum parsimony) make them more widely used and applicable to high-throughput computational analysis. A number of other factors can be highlighted regarding the problems in assessing the performance of different tree reconstruction methods, such as, PHYLIP (http://evolution.genetics.washington.edu/phylip.html). The main limitations are: (i) lack of non-simulated test data and (ii) the necessary trade-off that is required between fast efficient computational methods and robustness for high-throughput automated applications. It is concluded that none of the methods show any particular advantage in all cases, with the use of phylogenetic bootstrap analysis (Felsenstein, 1985) combined with a number of multiple alignment and tree construction methods recommended.

An important step towards the inference of function from molecular phylogenetics is the overlay of existing experimental information onto the reconstructed phylogenetic tree containing query and related sequences. A crucial factor is the use of good quality, manually verified, annotation data of the type available from databases like Swiss-Prot. Introduction of evidence tags in the gene ontology to track the source of annotations is also an important development for these types of studies.

1.3.4.5 Identifying Orthologous Relationships

Eisen (1998), Eisen and Wu (2002) and Sjolander (2004) highlight the importance of distinguishing orthologs and paralogs in phylogenetic studies of protein function. This is an important task when considering high-specificity functional properties, because if an ortholog to the query function can be identified then it is likely that they will share identical (or at least very similar) specific functions. The clusters of orthologous groupings (COGs) method is a resource of orthologous relationships between proteins. Other methods developed for the identification of orthologs use phylogenetic reconstruction methods rather than the sequence similarity of COGs (Storm and Sonnhammer, 2002). These methods are likely to give more specific functional information but may be limited for high-throughput methods by increased computational costs.

1.3.4.6 Prediction of Function

The final stage in the analysis process is the actual prediction of likely function for the unknown query protein sequence. Information gained from the earlier stages of analysis should provide a culmination of evidence on which to base a reliable prediction of the unknown protein function. The best way in which to reliably combine this information, to produce accurate high-specificity predictions, will form one of the main research topics of this thesis.

A number of methods have been developed towards improving the level of automation and level of prediction specificity. An early method - Bayesian Evolutionary Tree Estimation (BETE) (Sjolander, 1998) - was applied to SH2 protein domains. The method creates profiles of each sequence in a multiple alignment; an iterative partition algorithm then computes the total relative entropy (TRE) between each profile, progressively grouping together the pairs with the lowest TRE. The aim of this is to find an optimal partition of the phylogenetic tree

of sequences, with the final sequence groupings corresponding to subfamily specific functional profiles. A change in the subfamily annotation of Swiss-Prot for the SRC2_DROME protein was prompted by the analysis results from this method. The BETE method has also been successfully used by Celera Genomics for annotation of functional subfamilies (Sjolander, 2004).

Johnson and Church (2000) investigated the use of phylogenetic analysis to improve the identification of specific ligand-binding functions in two related protein families with similar folds but different binding site specificities. This method was then applied to other unknown sequences to try to identify specific ligand-binding functions. An interesting feature of this method is that the predictive power of the phylogenetic trees for the whole domain sequences and those of just the binding cleft were compared. Analysis of the results showed that whole domain sequence similarity was not a good indicator of binding-site specificity. In contrast the phylogenetic groupings from the binding-site sequence subset showed good differentiation of the different binding specificities. A limitation of this method, especially in terms of extending it to a more general automated approach, is that prior knowledge of binding site locations is required for successful implementation. One way in which this information could be obtained is through the use of automated algorithms for the detection of functionally important residues. These are discussed in detail below and form an integral part of this thesis.

1.3.5 Identification of Function Determining Residues

During evolutionary divergence of protein sequences functionally important residues are conserved due to the pressures of natural selection. Methods for the identification and analysis of the particular amino acids and their physico-chemical properties within these conserved regions are particularly important for prediction of specific protein functions (Valdar, 2002).

1.3.5.1 Entropy

An important concept for the analysis of the level of conservation within regions of aligned sequence residues is entropy. This is commonly defined using a measure of the average uncertainty of an outcome, from information theory, called the Shannon Entropy (Durbin et al., (1998)) (see *equation 1.1*).

$$H(X) = -\sum_i P(x_i) \log P(x_i) \qquad \textit{(equation 1.1)}$$

Where: $P(x_i)$ is the probability of observing the event, x_i , in a discrete set of *k* events. In the context of amino acid conservation, where *k* is commonly taken to be 20 (the number of standard amino acids), there is complete conservation of one amino acid type when the entropy is 0 and outcome, *k*, is certain. Conversely, the entropy is maximised when all amino acids are equally likely and the outcome is maximally uncertain.

1.3.5.2 Sequence Based Methods

Early work in the analysis of residue conservation was carried out by Livingstone and Barton (1993). This method carries out hierarchical clustering of MSAs into sequence subsets, based on criteria such as sequence identity and functional similarity. Conservation scores for residues at each alignment position are then calculated through a simple analysis of the physico-chemical properties (Taylor, 1986) of each of the residues. The method was applied to an alignment of 67 SH2 domains, which led to correct identification of phosphotyrosine-binding residues and also conserved secondary structure elements.

A novel method - *"SequenceSpace"* - developed by Casari et al. (1995) represents each sequence in a multiple alignment as a vector in a multi-dimensional "sequence space". The key feature of this method is the use of principal component analysis to identify the characteristic residues and positions that define the functional specificities of each protein subfamily. Projection of the conserved residues onto lower dimension clusters allows the degree of conservation of the residues to be visualised and measured by the distance of the sequence clusters (vector lengths) from the centre of the space of principle components. An analysis of the Ras-Rab-Rho superfamily is used as an example, showing how the direction of the vectors can be used to define the specific residues of importance for the function of each subfamily. Also, an application of the method to the reduction of phylogenetic tree complexity by using only the identified subset of specific functional residues is shown. The *SequenceSpace* method identified both the highly conserved phosphotyrosine binding residues and more specific peptide binding residues, therefore showing an increased specificity over that of Livingstone et al. (1993).

1.3.5.3 Comparative Analysis of Methods

A study by Pazos et al. (1997) compared four methods for calculating tree-determinant residues: (i) *SequenceSpace*; (ii) evolutionary trace (ET) (Lichtarge et al., 1996); (iii) a method for comparing subfamily conservation (Dorit et al., 1995); and (iv) a method using self-organising maps (SOM) of sequence clusters (Andrade et al., 1997). *SequenceSpace* was shown to be the most effective for the determination of specific functional residues and the *SequenceSpace* and SOM methods were shown to be most stable to the inclusion of distantly related sequences within the multiple alignment.

A more recent study (del sol Mesa et al., 2003) implemented three automatic methods for the prediction of functionally important residues from protein sequences. The primary goal of this study was a systematic, statistical assessment, of the role that conserved "tree-determinant" residues can play in identifying functional specificity. This type of analysis is of particular relevance because it concentrates on methods for automated high-specificity functional analysis. The three implemented methods are:

- *"The Level Entropy Method" (S-method)* – The main aim of this method is to study the conserved residues acting as specific functional tree-determinants using a phylogenetic tree of the protein family. Different partitions of the tree are investigated and the relative entropy is measured to find the most stable tree-level, which produces the most informative separation of sub-families. The physico-chemical properties of the amino acids are not explicitly taken into account in this method;

- *"The Mutational Behaviour Method" (MB-method)* – The aim of this method is to calculate the mutational behaviour of potential tree-determinant positions and compare them to that of the whole sequence family. Mutational behaviour is determined by evolutionary constraints and assessed using correlation matrices and rank correlation criteria. The aim of this study was to identify and separate functional families using conserved residues. The hypothesis is that the mutational behaviour of the tree-determinant residues will be the same as the whole set of family sequences; and

- *"SequenceSpace Automization Method" (SS-method)* – This is an automated implementation of the *SequenceSpace* method of Casari et al (1995). A geometric clustering algorithm calculates an optimal number of clusters from the initial PCA analysis and then attempts to identify positions relating to conserved residues between subfamilies.

Each method was tested on two sets of non-redundant sequence families that have known, single chain, representative structures in the PDB. One set contained 191 families (binding sites associated with various heteroatoms) while the other contains 112 (associated with annotated PDB SITE records). With regards to the coverage of the three methods, it is noted that there are some constraints dictated by the number and level of conservation of the sequences representing each family grouping. The *MB-method* is unaffected by this and will always be able to predict some tree-determinants, whereas the *SS-method* and *S-method* are more sensitive to these factors. The results of this study are judged on the proximity of the identified functional residues to either those heteroatoms deemed functionally important or PDB sites. The results do not clearly stake a claim for any of the three methods over the other. In-fact, as a general rule, it was found that the intersection of prediction results for two, or all three methods, increased the quality of the results. The results were also complicated by their dependency on the type and size of the functional heteroatoms.

A more recent study by Pazos et al. (2006) explores the extension of the *MB-method* to incorporate a functional similarity matrix into the correlation calculation of mutational behaviour of sequences. This is essentially a supervised form of the *MB-method*, with prior functional grouping, and is discussed in more detail in *chapter 4* of this thesis.

The *ConSeq* method of Berezin et al. (2004) identifies functionally important sequence residues through the incorporation of the *"Rate4Site"* algorithm. This algorithm uses the Maximum Likelihood method for phylogenetic tree reconstruction, which, unlike the neighbour-joining methods of phylogeny, takes into account the rate of evolutionary divergence at particular residue positions. This is, however, quite a computationally expensive algorithm when compared to some the other methods previously discussed.

1.3.6 Profile-Based Methods for Identification of Functional Specificity

A group of related methods are those that attempt to construct sensitive profiles for the specific identification of particular functional sub-types. An early study on the use and generation of sequence profiles was published by Gribskov et al. (1987). Following on from this work a number of profile-based and HMM-based methods have been developed to assist the general functional annotation of protein sequences. These include the HMM-based approach of PFAM (Bateman et al., 2004), the profile-based motif approach of PRINTS (Attwood et al., 2003), and the integrated database of resources provided by tools such as InterPro (Hunter et al., 2009).

These methods and their associated database resources are commonly used to help determine the function of unknown protein sequences. However, due to the nature of these methods, they are usually more suited to the annotation of general protein function and care should be taken when annotating a more detailed, specific, level of function (Whisstock and Lesk, 2003; Friedberg, 2006). The main considerations when using these types of approaches are the level of coverage that they provide when annotating function and also the number of sequence representatives used to generate the profiles or HMMs.

For example, in the case of PFAM, the HMMs contained in the database are generated at a protein domain level and are clustered into PFAM families using homology based measures, rather than specific functional class. Therefore, it is possible for single families of PFAM HMMs to contain sequences of different specific functional sub-classes. The consequence of this, when using PFAM to assign specific enzyme function, is that although the number of false positive annotations at a more general level of enzyme classification should be reduced due to the increased sequence coverage, they are more likely to be unsuitable for determining more specific enzyme classes.

It is these potential limitations of the general profile and HMM based approaches that led to the development of the BLAST-based methods of specific enzyme annotation investigated in this thesis. They also led to the development of other more sophisticated profile and HMM based methods for the specific purpose of functional annotation that are discussed below.

Three particularly important approaches, with regard to protein function prediction and subsequent application to the improvement of the accuracy and level of specificity, are those of Hannenhalli and Russell (2000), Tian and Skolnick (2004), and Pazos and Sternberg (2004). Each of these methods is quite distinct and has been applied to different datasets and functional classification schemes.

Hannenhalli and Russell (2000) describe a method for the identification of functional sub-types and also functionally specific residue positions. Given a multiple sequence alignment and information regarding the specific functional properties of each sequence a set of hidden Markov model (HMM) profiles can be constructed to represent each specific function. Potential functional specificity determining residues are then identified using a relative entropy based measure, which takes into account the likelihood that particular amino acids will be specifically associated with one functional sub-type over the others. A protein sequence of unknown specific function can then be compared to the specific profiles to identify the most probable specific function. Four large enzyme families (nucleotidyl cyclases, eukaryotic protein kinases, lactate/malate dehydrogenases and trypsin-like proteases) with good experimental information, regarding the specific functional properties, were used to test the method. Examples were chosen that could not be separated by simple sequence comparisons or phylogenetic tree comparison to demonstrate the power of the method, with accuracies (for the four enzyme families listed above) of 96% compared to 80% and 74% for sequence similarity and BLAST respectively. This analysis was then extended to include 42 PFAM (Bateman et al., 2004) alignments and was also shown to outperform both BLAST searching and sequence similarities when identifying most of the specific functional subtypes.

The method of Tian and Skolnick (2004) uses a combined system – EFICAz (Enzyme Function Inference by Combined Approach) - of four recognition methods to improve the accuracy of enzyme function predictions, they are:

1. *CHIEFc (Conservation-controlled HMM Iterative Procedure for Enzyme Family classification):* This procedure consists of carefully built HMMs from multiple sequence alignments of each enzyme family. A method, based on information theory, is then used to identify functionally discriminating residues (FDRs) for each enzyme family HMM derived by CHIEFc.

2. *Pairwise Sequence Identity:* A specific reliability threshold is used for each enzyme family.

3. *Recognition of FDRs in Multiple Pfam enzyme families:* This uses the same Shannon entropy measure to identify FDRs as method (1) but PFAM alignments are used in place of the CHIEFc generated HMMs.

4. Recognition of multiple high specificity PROSITE (Hulo et al., 2004) Patterns

One of the main outcomes of this study is the importance, of the CHIEFc family FDR recognition method, to the high accuracy recognition results that are obtained. This is perhaps unsurprising as the CHIEFc method is purposely designed for the accurate recognition of specific enzyme functions, defined by their annotated EC numbers. As a result of this and the added effects of the other three methods, the combined EFICAz approach shows high accuracy and high sensitivity during testing on enzyme sequences in Swiss-Prot and also when applied to automatic annotation of the *E. coli* K12 proteome. A comparison of enzyme function annotations made by EFICAz and KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000) for this genome showed that EFICAz predicted 114 more potential enzyme coding genes at the specificity level of four EC numbers. The majority of these in KEGG are either partially annotated (with 54 out of 69 showing correlation with the partial annotations provided by EFICAz) or are marked as hypothetical proteins and did not have any annotation. These results suggest that EFFICAz is applicable to automated genome annotation and able to make novel specific enzyme predictions.

The approach developed by Pazos and Sternberg (2004), known as PHUNCTIONER, varies from the other two methods in that it uses multiple structural alignments with the resultant profiles as the basis of its predictions and recognition of functionally important areas. Also, the profiles used in this study are based on the GO functional classification scheme. Starting from a structural alignment, proteins with the same annotated GO terms are extracted and grouped together. Profiles of the functionally conserved residues for each GO term are identified using a conservation score and high entropy positions are filtered out.

Position specific scoring matrices (PSSMs) are then created for each profile and the performance for prediction of GO terms is compared to the use of sequence identity. PHUNCTIONER is found to perform better than the sequence homology based method in most cases. This is especially true in cases of low (generally less than 20%) sequence identity. A further application of PHUNCTIONER was a comparison to the *SequenceSpace* and the Mutational Behaviour (del sol Mesa et al., 2003) methods for identification of functionally determining residues. The findings indicate that the PHUNCTIONER method is able to identify residues that are related to more general lower-specificity GO functional classification, whereas *SequenceSpace* and the mutational behaviour method identify residues that are related to more specific functional properties.

Each of these approaches show good application to the prediction of protein function and the identification of functionally determining residues for specific functional subtypes. These methods all share a common limitation, which is their reliance on pre-determined functional sub-groups. The implementation of all three of these methods depends on a prior knowledge and availability of a sufficient amount of annotated sequence or structural representatives, with the same function, on which to base the specific functional profiles.

1.3.7 Sequence and Structure Based Methods

The use of structural information in addition to sequence can provide added insight into the determination of specific functional residues and protein interfaces (Watson et al., 2005; Lichtarge and Sowa, 2002; Filizola and Weinstein, 2005). These methods generally share similar features to the sequence based approaches, with the main difference being the requirement of representative, three-dimensional protein structures, for the final analysis of the results. This is especially true for those methods that rely on the spatial clustering of residues (Lichtarge et al., 1996; Landgraf et al., 2001; Glaser et al., 2003) to assess the accuracy of predicted functional residues within biochemically active sites.

1.3.7.1 Evolutionary Trace Method

The evolutionary trace method (ET) (Lichtarge et al., 1996) uses evolutionary information available from multiple sequence alignments to map predicted,

functionally important residues, onto proteins of known structure. Through use of a set of sequence percentage identity thresholds a multiple sequence alignment can be partitioned into clusters to form a dendrogram (phylogenetic tree). A consensus sequence can be obtained for the set of sequences either between or within each cluster. This identifies those residues that are indicative of either the general functional class (and therefore conserved in a larger number of proteins), or those that are only conserved within a subfamily (and therefore relate to the specific function of the subfamily cluster). A number of extensions to the method have been proposed that provide more robust statistical analyses of the results and also improved levels of automation (Madabushi et al., 2002; Aloy et al., 2001).

There are a number of other approaches which have looked at utilising structural information to improve the quality and specificity of functional site identification and protein function prediction (Watson et al., 2005). The method of Landgraf et al. (2001) is described as an extension of the evolutionary trace, with one of the main differences being that phylogenetic relationships are not used as input. The theoretical basis for not using phylogenetic information is that proteins with multiple functional clusters could be averaged out in the phylogenetic tree, or highly conserved residues associated with one function could overshadow those of a secondary function. The evolutionary trace method is not designed to detect secondary functional clusters; therefore the authors use a form of correlated mutation analysis to highlight conserved clusters through regional similarity relationships. This 3D cluster analysis technique has structural information at the core of the functional analysis and is not of direct interest with regards to predicting function from sequence information. However, the correlated mutation analysis that is part of this method is of interest and (as we have seen above) has been shown by del sol Mesa et al. (2003) and Pazos and Valencia (2006) to be successfully applicable to sequence based studies of functional specificity.

Finally, an important consideration when attempting to identify functionally active conserved residues, from both sequence and structure, is the differentiation between those that are structurally and functionally important (Chelliah et al., 2004). A method incorporated into *ConSeq* and *ConSurf* (Armon et al., 2001), which uses neural network predictions to differentiate between buried and exposed residues in

globular proteins, is one proposed solution to this. However, this assumes that functional residues are always solvent accessible and all buried residues are associated with structurally conserved regions. This is generally a difficult problem to solve, due to the unavoidable ambiguity in classification of residues that are responsible for structural or functional protein properties.

1.4 Non-Homology Based Methods for Function Prediction

It is worth briefly mentioning some methods for protein function prediction which are not based primarily on sequence homology detection. One approach is that of Jensen et al. (2003), which uses derived physico-chemical sequence properties instead of sequence similarity. These sequence features are then used as input to a system of neural networks for the prediction of GO classifications. The advantages of this method are that it can predict functions for sequences with no known homologous relationships (orphan sequences); however, the limitation is that the predictions obtained are mostly low specificity general classifications. Other approaches to non-homology based prediction of function through the use of protein-protein interaction data have also been described (Marcotte, 2000). A further method of interest in relation to sequence based homology prediction is that of Espadaler et al. (2005). This method investigates a combined approach to the combination of sequence homology and protein-protein interaction data for use in improving structural and functional annotation.

1.5 Overall Conclusions and Summary

The comparison of the many different approaches to automated function prediction, especially those aimed towards improving the overall accuracy and specificity of the functional annotations is an inherently difficult task. This is due to a number of contributing factors:

- The lack of an unambiguous description of protein function, especially when trying to compare levels of specificity; and
- The lack of benchmark datasets that can be used as a clear way to distinguish, compare and judge the performance of newly developed prediction methods (Tetko et al., 2005).

The efforts of the gene ontology consortium and the annotation projects such as GOA are making important contributions to the standardisation of how protein functions are described and annotated within sequence databases. However, problems still remain, even with this scheme, as to how best to compare and measure the specific functional distance between two predicted functional terms. For example, if the actual function of a protein is *x* and the predicted function is *y*, how should the resulting accuracy of this prediction be measured? As has been discussed earlier, the EC scheme gives a widely used way of estimating this by treating the number of correctly predicted EC numbers as roughly comparable to levels of functional specificity. This has a number of problems, (i) It is only applicable to enzymes and (ii) it is possibly too simplistic and will cause valuable information to be lost and not considered when assessing the results. The problem is possibly more difficult when considering gene ontology terms. Due to the graph-based architecture of the GO hierarchy an intuitive way of measuring functional distance may be to count the number of edges between terms, or possibly for comparing levels of specificity, the depth of the term-node in the graph could be used as a measurement. The subjective nature of defining protein function makes this a problem that may not be solvable in an exact way.

As we have seen in studies on the level of sequence similarity required for the simple transfer of function via homology, clear levels of sequence similarity required for specific levels of functional inference are difficult to agree upon. These problems of firm comparisons are increased when comparing the many different techniques for improving the prediction of functional specificity or identifying functionally important residues. This is particularly problematic when looking at ways to incorporate these techniques into an automated high-throughput approach to high-specificity function prediction. Mainly because the question of which methods to include to best achieve these aims is difficult to definitively answer.

It has been shown that the incorporation of evolutionary analysis of protein families, through phylogeny, improves the accuracy of high-specificity function prediction in comparison to simple homology transfer methods (Eisen, 1998; Eisen and Wu, 2002; Johnson and Church, 2000; and del sol Mesa, 2003). These methods also aid in the identification of functionally important amino acid residues. However, there are

many difficulties still to be overcome for the development of methods and their integration into a fully automated solution to the problem of reliable, accurate, high-specificity protein function prediction from sequence.

The key aims of this literature review were: (i) to give a critical discussion of the area relating to automated prediction of protein function, with a concentration on methods that have been used to improve the accuracy and specificity of the prediction results; and (ii) the highlighting of current "state-of-the-art" automated methods for high-specificity function prediction from sequence. The most satisfactory conclusion appears to be that there are a number of different methods that show varying levels of ability to predict specific functional properties. The comparative analysis of Pazos et al. (1997) showed the superiority of the *SequenceSpace* method for determining specific functional subgroups, however, this method suffers from problems associated with the level of automation possible. The later study of del sol Mesa (2003) implemented three automated methods (including a semi-automated form of *SequenceSpace*) for comparison and concluded that the best results are obtained from combinations of the methods. The hidden Markov model based sub-profile method of Hannenhalli and Russell (2000) has also been shown to work well for both identifying specificity determining functional residues and application to functional sub-type prediction. It is these two studies, along with the ideas contained in the sub-alignment phylogenetic reconstruction studies of Johnson and Church (2000) that will form an important part of this thesis.

In conclusion, the best approach for a fully automated approach to high-specificity function prediction from sequence appears to be a combination of the optimal properties of a number of methods. Using evolutionary information relating to the relationships between homologous protein sequences it should be possible to accurately identify specific functional details that have been acquired through the process of evolutionary divergence. Approaches to combining these methods and extracting important algorithmic features in reliable, automated ways, form a major part of the research in this thesis. These ideas and methods are then extended to investigate the feasibility of using machine learning techniques, namely support vector machines (SVMs), to identify the function specificity determining residues

(fSDRs) in a fully automated way, from multiple sequence alignments, without using any prior knowledge of the functional sub-types of the constituent sequences.

1.6 Outline of Research Thesis

The major aim of this research was the development and assessment of methods for use in an integrated and automated system for the prediction of detailed, specific, protein molecular functions, from sequence information. In a review of the literature a number of methods have been described which investigate function prediction, using sequence information and algorithmic techniques, for improving the accuracy of specific functional inference. However, to my knowledge, there are at present no methods that successfully combine these features into one high-throughput, accurate and robust fully automated system for the prediction of specific protein functions.

The overall goal of this research was the development and investigation of methods for re-evaluating the sequence similarity of homologous proteins to generate an improved scoring method for assessing functional similarity. An overview is presented, in *figure 1.3,* of the main stages involved in this process. First, a sequence database homology search is carried out using a query protein sequence of unknown molecular function. An MSA is returned from this along with an associated sequence similarity score (such as a BLAST E-value) for each sequence, which is used to order the sequences by similarity to the query. Using a homology transfer method for function prediction, the query sequence would be assigned the same function as the most significant annotated sequence above a similarity threshold. However, this will lead to incorrect annotations in circumstances where the most significant sequence is not the same specific function as the query. A simplified example of this is shown in *figure 1.3,* where the query sequence (with function = *func_B)* shows a greater degree of sequence similarity to 3 sequences (*seq1, seq2 and seq3*) with function = *func_A.*

In a case such as this, additional properties must be taken into account to provide an improved method for assessing functional similarity between the query and the group of sequences with function = *func_B.* Methods are proposed that aim to automatically identify amino acids that are indicative of evolutionary conservation within groups of functionally specific proteins. This can be thought of as a form of "phylogenetic filtering" of the aligned sequence columns, to create a more relevant,

functionally determining, sub-set of aligned residues. The example in *figure 1.3* shows four aligned columns that have conserved residues within the specific functional groupings and variation between.

It then becomes possible to calculate a new measure of sequence similarity - using only the sub-set of amino acids most likely responsible for determining the specific functional properties - and thus re-order (or cluster) the sequences to provide an improved measure of functional similarity. From the example in *figure 1.3*, it can be seen that when only considering the four aligned columns containing the fSDRs, the query sequence is most closely related to the group of sequences with function = *func_B* and therefore predicted, correctly, to be of that specific function.

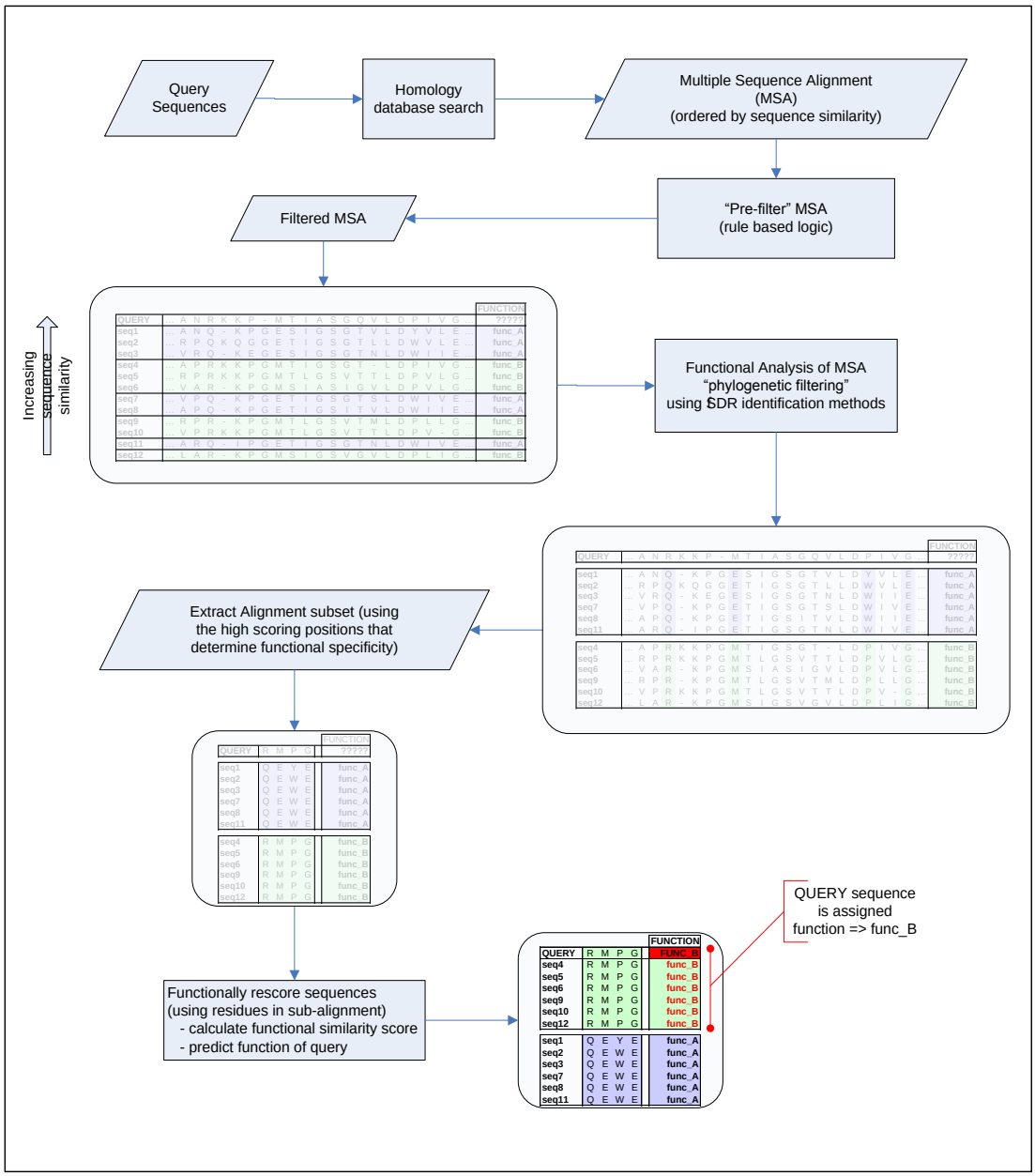

Figure 1.3. Conceptual overview of the proposed methods of analysis and key areas of investigation carried out in this research thesis.

With regards to the automatic identification of functional specificity determining sets of residues, a disadvantage to the methods analysed in *chapter 4*, of this thesis, was their requirement for prior knowledge of the specific functional classifications of the sequences contained within the MSAs. This limits the use of these methods to alignments of functionally well-characterised sequences, thus preventing a more general approach to the classification problem and limiting the possible uses to a much reduced sample space of functionally annotated sequences. To circumvent this requirement it was suggested that machine learning methods, such as support vector machines (SVMs), could be used for the automatic identification of fSDRs in multiple sequence alignments. The analysis, in *chapter 5*, investigates the feasibility of using SVMs towards automatically identifying fSDRs and thus the possibility of incorporating this identification into a fully automated system for improving the specific functional classification of enzyme sequences.

The target audience of the methods analysed in this thesis is expected to be researchers and genome annotators, who are primarily interested in accurate, high specificity, functional genome annotation, when close homologs with differing specific functional properties are available to provide an evolutionary analysis. Although the analysis within this thesis concentrates on the functional classification of enzyme molecular function, it is expected that the methods would be generally applicable to other types of proteins. To test this hypothesis, however, an alternative benchmark set of protein sequences and the use of relevant functional classification schemes would be required.

In summary, the analyses presented in this thesis aim to investigate automatic, computationally efficient methods for the transformation of sequence similarity scores into a measure of functional similarity, which provides a reliable and accurate measure of specific enzyme functional classification.

Chapter 2 Investigation into the Functional Conservation of Enzyme Sequences and Dataset Definitions

2.1 Introduction and Aims

The work of Rost (2002), Tian and Skolnick (2003), and Todd et al. (2001), among others, shows that the level of correlation between protein function and sequence similarity measures follow a common relationship; where the accuracy for functional transfer becomes greater with a higher level of sequence similarity. The work in this section aims to provide an initial investigation into the level of error involved when using homology based sequence similarity measures for the assignment of protein function and provide the source for the benchmark datasets of multiple sequence alignments used within this thesis. An important factor of this work was the investigation of homology transfer when applied to the prediction of high-specificity protein function. The functional classification chosen for this analysis was the Enzyme Classification (EC) scheme. This method of classification was chosen because it has already been widely used with good success in the studies mentioned above (Rost, 2002; Tian and Skolnick, 2003; Todd et al., 2001) and provides a relatively simple and effective way of computationally measuring the level of functional specificity. Through comparison of the number of shared EC numbers between the input query sequence and the homologous sequences obtained from a database similarity search, an understanding of the level of specific function prediction at varying sequence similarity thresholds can be obtained.

Most previous studies of this type have aimed to identify detailed relationships between sequence similarities (such as percentage sequence identity or statistical E-value scores) to obtain definitive threshold values for varying levels of sequence and functional conservation. This study also provides an understanding of these properties but aims to concentrate on the areas of high functional specificity, by looking at the correlation between sequence homologues and the correlation to the conservation of all four numbers in the EC classification hierarchy. A further aim is

to provide a set of benchmark examples where the high-scoring "top-hit", to a "target" sequence, obtained from a PSI-BLAST homology search does not identify a protein sequence with the same specific function as the query sequence.

2.2 Methods

2.2.1 Collection of "target" Enzyme Sequences

The method followed for the collection and identification of enzyme sequences for analysis is based on that of Tian and Skolnick (2003) and Rost (2002). The Swiss-Prot (version 46) section of the UniProt (Apweiler et al., 2004) (version 4.0) sequence database was used as the source of the analysis sequences. From the Swiss-Prot database, which contained 168,297 sequences, a total of 43,572 enzyme sequences with fully annotated EC codes at all 4 levels of the hierarchy were identified. These enzyme sequences in the "target" sequence set were identified in the following way:

- All sequences that have annotated EC numbers in the "Description (DE)" field of their records in the Swiss-Prot database were identified, sequences which fulfil any of the following criteria were then removed from the final target set:

    1. They contain incomplete EC annotations and therefore undetermined numbers (e.g. EC 1.2.3.- would be classed as an incomplete annotation and therefore removed);

    2. They have multiple EC annotations and are therefore defined as multifunctional enzymes;

    3. Contain any of the following keywords in the "Description (DE)" or "Keyword (KW)" field of Swiss-Prot ("probable", "hypothetical", "putative", "by homology", "by similarity");

    4. Are identified as fragments and therefore contain the keyword "fragment" in the Swiss-Prot "Description (DE)" field.

This process identified 45,164 sequences. All 100% identical sequences were then identified and a single, randomly selected, representative of each sequence cluster was kept in the dataset. This reduced the target set by a further 1592 sequences to

produce the final enzyme sequence set of 43,572 sequences. These sequences consist of 1901 distinct enzyme classes measured to all four levels of EC specificity. These were tagged and identified as "target" sequences in the sequence search database used in the next stage and are referred to as "target" sequences at later places in this thesis. These criteria were used to ensure that all of the sequences added to the target set had associated functional annotation data which was complete and most importantly, of a high quality, obtained from the "gold-standard" annotations in the Swiss-Prot database.

2.2.2 Identification of Homologous Sequences

After identification and extraction of the fully annotated enzyme "target" sequence dataset a PSI-BLAST (Altschul et al., 1997) database search was carried out to identify homologues for each of the 43,572 target enzymes. This was so that the level of functional inference from sequence similarity search measures could be assessed. A PSI-BLAST search was carried out for each of the target enzyme sequences against the UniProt (Swiss-Prot + TrEMBL) database (version 4.0), which contained 1,757,967 sequences. To improve database search efficiency and reduce the number of false positives, each input sequence was filtered using the SEG low complexity filter (Wootton and Federhen, 1996) and all of the sequences in the search database were filtered using the low complexity, trans-membrane and coiled-coil filter options of the *pfilt* application (Jones and Swindells, 2002). The sequence database search was carried out using 3 iterations of PSI-BLAST (version 2.2.10), using the default iteration inclusion value (*-h* parameter) of 0.001 and an output E-value threshold of 10. Also, the maximum number of sequences included in the BLAST search output and resultant multiple sequence alignments (MSAs), was set at 5000 using the *–v* and *–b* command line parameters. Finally, with regards to composition-based sequence statistics - which are calculated from the sequence composition of the database sequences (Schaffer et al., 2001) - the default setting, which includes these calculations, was applied through the setting of the *–t* command line parameter (*-t T*). All other search parameters were left unchanged from the default settings of PSI-BLAST (*blastpgp*) version 2.2.10.

The resulting output list of detected homologues was then filtered to remove all sequences not identified as belonging to the functionally annotated "target" enzyme

sequence dataset. This was so that comparisons could be made between the functions of the query sequences and those identified as homologues in the PSI-BLAST search.

2.2.3 Definition of EC Conservation Accuracy

The method used to calculate the accuracy of specific EC functional conservation, with respect to sequence similarity measures, is described in *equation 2.1*. This is based on the method used by Rost (2002), slightly adapted to take into account ranges of similarity thresholds.

$$Accuracy = 100 * \left(\frac{Matching}{All} \right) \qquad \textit{(equation 2.1)}$$

Where: *"Matching"* signifies the number of functionally matching sequence pairs within a defined range of sequence similarity threshold values; and *"All"* signifies the number of all sequence pairs within this same range.

2.2.4 Calculation of Global Sequence Identity

A full Needleman-Wunsch pair-wise sequence alignment algorithm (Needleman and Wunsch, 1970) was used to calculate a global percentage identity score between the query sequence and all "target" sequences identified in the database search. The *needle* application from the EMBOSS (Rice et al., 2000) software suite was used with the default parameters: BLOSUM62 substitution matrix; gap open penalty of 10.0; and gap extension penalty of 0.5.

2.3 Results and Discussion

2.3.1 Level of EC Functional Conservation

A comparison between the level of EC functional conservation and sequence similarity measures (PSI-BLAST E-value and global sequence identity) was carried out to assess threshold levels for reliable, accurate transfer of specific enzyme function by homology. The first step in the analysis of the data was an investigation of the level of functional conservation with respect to the observed PSI-BLAST E-values between each of the identified query-target pairs. The method described in *section 2.2.3* was used to calculate the accuracy of functional transfer, within E-value ranges, which were calculated by taking the minus of the log (to base 10) of

the E-value. The results of this analysis are shown in the graph in *figure 2.1* for the two levels relating to the most specific level of functional correlation available with the EC classification scheme. These are: (i) the first three EC numbers are conserved (EC3: n.n.n.-); and (ii) all four EC numbers are conserved (EC4: n.n.n.n). It can be seen from *figure 2.1* that as the level of functional specificity increases (from EC3: n.n.n.- to EC4: n.n.n.n), the accuracy of functional transfer using the PSI-BLAST E-value decreases. Overall these results seem to agree quite closely with those of Rost (2002) in his study of 1st iteration PSI-BLAST E-values. The results show that even at very statistically significant E-values, commonly used for functional transfer (such as 10^{-50} => -log(E-value)=50), the accuracy of exact specific function prediction (all four EC numbers are conserved) is only just slightly greater than 90%. Similarly, the results comparing EC conservation accuracy to sequence identity, in *figure 2.2* show that even at levels above 50% identical residues, the accuracy of specific functional transfer is less than 100%.

When considering the correlation between sequence identity and functional conservation, these results agree most closely with those of Todd et al. (2001). The results reported by Rost (2002) are much more pessimistic and report that upwards of 70% sequence identity (local sequence identity reported from PSI-BLAST) is needed to transfer all 4 EC numbers with comparable levels of accuracy. A more recent study by Tian and Skolnick (2003) reports yet another different threshold requirement of 60% sequence identity (global sequence identity) for at least 90% accuracy for the same level of specific function transfer between sequence pairs. The main differences between the results of these studies is thought to lie in the disparate way in which the datasets from each have been formed, especially with regards to the particular thresholds that have been applied for sequence and functional redundancy removal.

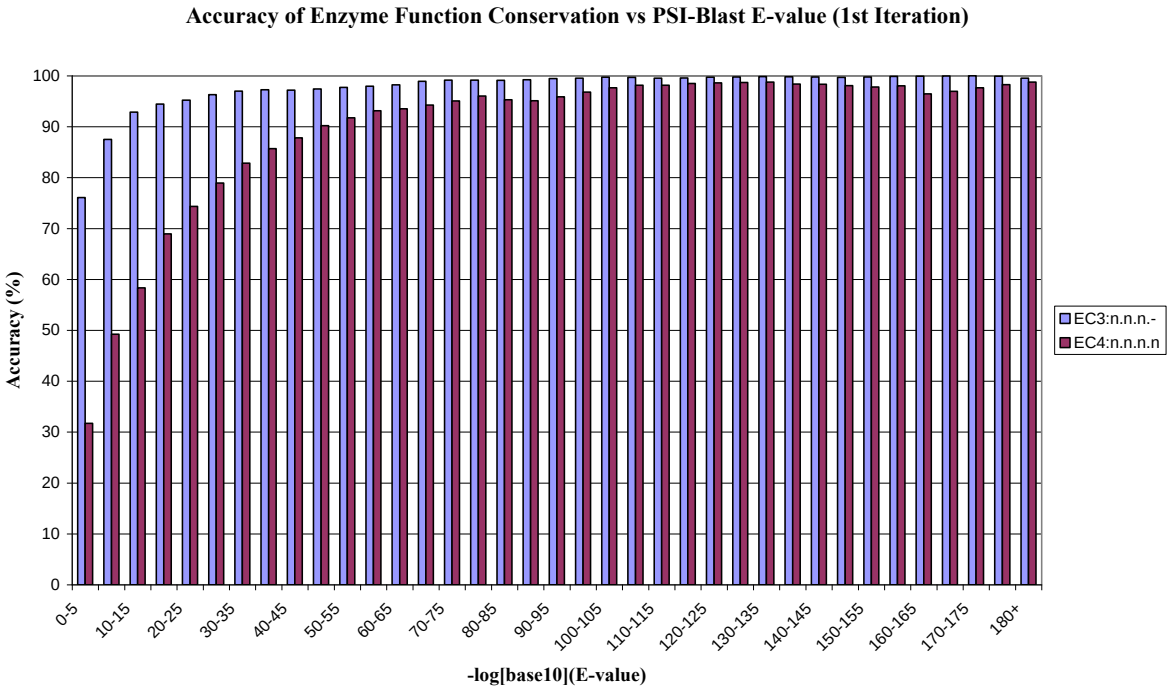

Figure 2.1. Graph showing the accuracy, using equation 2.1, of function prediction using PSI-BLAST E-values, obtained from sequence pairs in the 1st iteration of the database search results. Where, EC3:n.n.n.- are the results for the first three EC numbers predicted correctly; and EC4:n.n.n.n for all four EC numbers correctly predicted.

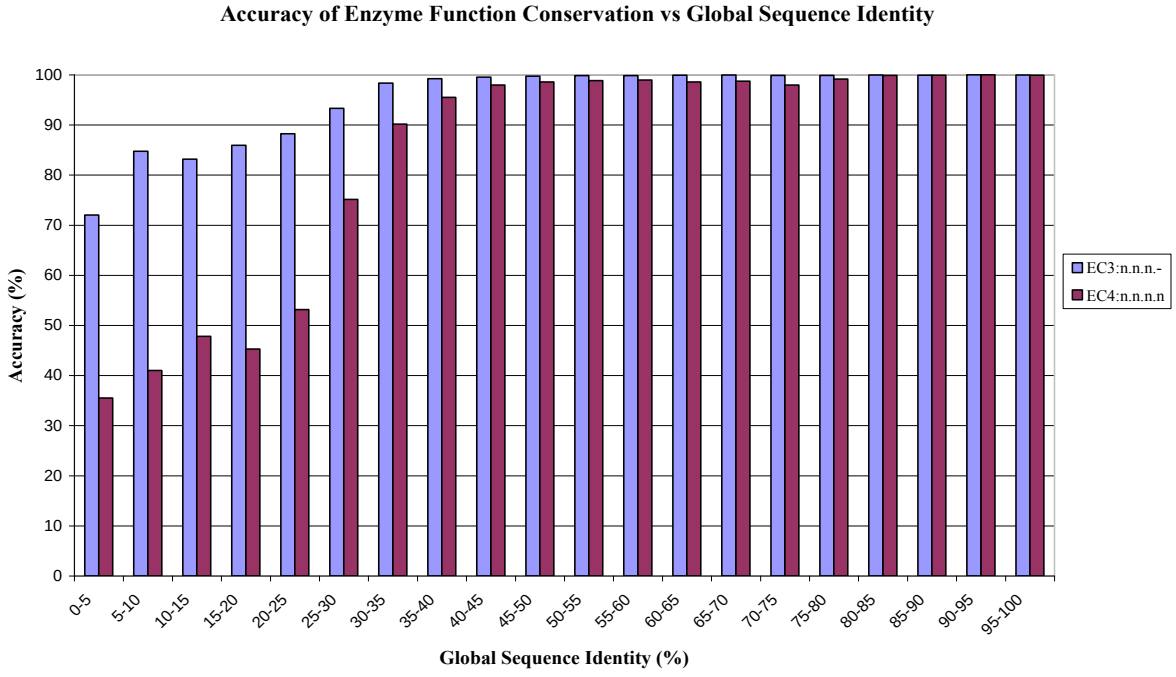

Figure 2.2. Graph showing the accuracy, using equation 2.1, of function prediction using global sequence identity, obtained from sequence pairs in the 1st iteration of the database search results. Where, EC3:n.n.n.- are the results for the first three EC numbers predicted correctly; and EC4:n.n.n.n for all four EC numbers correctly predicted.

The aim of this study was not an exhaustive comparison between the many methods and previous studies carried out in this area because this has been discussed extensively in previous work. However, the results shown in *figure 2.1* and *figure 2.2* do achieve the aim of highlighting the problems, which have been previously reported (Todd et al., 2001; Rost, 2002; Tian and Skolnick, 2003), regarding the use of sequence homology for specific functional inference. These are that it is not a simple matter to make a definitive prediction of enzyme function, based on simple sequence similarity measures and that the disparate nature of the datasets used makes it difficult to even agree on the best thresholds to use (Valencia, 2005).

2.3.2 Functional Analysis of PSI-BLAST "top-hit" Sequences

A common approach to assigning the function of an unknown protein sequence is through the transfer of function from a previously annotated homologous sequence with the most significant, "top-hit", sequence similarity score. This approach was used to assess the number of correct predictions that would be expected when simulating the prediction of the specific function of the "target" sequence set in this way. As expected, the results showed that a majority of cases (42453 (out of 43572) in the first iteration and 41637 (out of 43572) in the final iteration) are examples of a correct prediction from the top PSI-BLAST hit (rank position one). This was expected due to the large amount of potential sequence redundancy within the source Swiss-Prot database. There are however a number of examples where this is not the case and the first correct specific functional sequence result occurs at rank position (ordered with respect to decreasing statistical significance of the sequence alignments) two or lower, with 354 and 1214 examples in the first and final iterations respectively. A third case, which make up the remainder of the examples, is where no correct functional hits are found. These types of examples are not considered further in this work as they are not suitable for use in the discrimination between the specific functional sub-types of sequence homologues.

Interestingly, it is the 1st iteration PSI-BLAST results which give the largest number of correct examples, with respect to a specific functional match at all 4 levels of the EC hierarchy. Also, this means that in a number of cases the iterated PSI-BLAST process actually causes a deterioration of the functional accuracy of the "top-hit". An interesting discussion on the effect of PSI-BLAST iteration on functional transfer

is provided by Tian and Skolnick (2003). They show that the E-values of closely related query-sequence pairs (above 70% identity) tend to increase in later iterations, but decrease for those below 70%. This result suggests that some thought must be made as to whether an iterated database search is the best approach to annotation of specific enzyme function, and if so, the E-values used to interpret the results must be carefully considered in the context of the iteration from which they came.

The relatively low number of *"incorrect"* sequence examples is likely due to the inherent bias within the Swiss-Prot database and the associated "target" enzyme sequences. Both Rost (2002) and Tian and Skolnick (2003) give detailed discussions of these estimated database redundancy issues. For this study I have not pursued the effects of potential bias any further because it is not definitively clear if, or how, any potential sequence redundancy should be removed. This is especially true when considering the use of multiple sequence alignments and the associated evolutionary information in later stages of this work, because the level of evolutionary divergence observed in certain sequence residues can be crucial when determining the specific functional sub-type. Also, it was decided to concentrate on the alignments generated by the 1st iteration of the PSI-BLAST sequence database search. This is because of the results described above, related to the deterioration in functional inference in the later iterations and also because in this work it is the more closely related sequences that are of most interest. Therefore, the use of an iterated search to identify and include more distantly related sequences, in the resulting MSAs, is of lesser importance for this study.

2.4 Collection and Definition of Datasets

One of the main aims of the "top-hit" functional analysis was the identification of a set of data that could be used as an experimental benchmark for comparing the performance of specific function prediction techniques investigated in this thesis. It was decided that this data should consist of examples that show *"incorrect"* specific function prediction when transferring the function from the top "target" sequence hit from PSI-BLAST. This was deemed an appropriate form of benchmark because it simulates real problem cases likely to be encountered by a researcher attempting to determine the specific function of an unknown protein sequence. Therefore, any

automated approach which consistently improves on the accuracy of this simple homology based method should be highlighted by this benchmark.

The approaches to the benchmark dataset collection are described below. Two different methods are described. This is because, due to limitations in the size and quality of the "initial" dataset, it was decided to develop an alternative method to collect a much larger set of "artificial" incorrect benchmark examples. The data content of each of these datasets is a set of MSAs, generated by PSI-BLAST, through the use of the *–m 6* command line parameter. These MSAs were used for the benchmark studies because they are very computationally efficient to generate and are of a good quality.

Unless stated otherwise all MSAs analysed are generated from the 1st iteration of a PSI-BLAST database search (using the *blastpgp* executable - version 2.2.10), which is the same as a gapped-BLAST database search. Therefore the notation: BLAST and PSI-BLAST is used interchangeably.

2.4.1 Collection of the "Initial" Benchmark Dataset

2.4.1.1 Method

The initial approach taken to identify a benchmark dataset for use in testing and validation, focused on a selected subset of the "incorrect", "top-hit" predictions, obtained from the BLAST analysis. It was decided to extract this subset from the examples which showed incorrect "top-hit" prediction results in both the first and final PSI-BLAST iteration results. This restriction was made because it meant that the sequence ranking and associated MSAs for both of the iterations could be compared if required in later studies. Two further criteria were used in an attempt to improve the dataset quality: (i) the removal of all examples that share zero EC numbers between the query and the highest ranked "target" sequence, to remove cases from the dataset which highlight potential problems related to potentially misleading functional distances in the EC nomenclature; and (ii) the removal of examples which had less than 5 sequences with the same specific function as the query in the multiple alignments.

2.4.1.2 Properties of the "Initial" Dataset

The above steps led to a final dataset containing 126 sets of PSI-BLAST multiple sequence alignments. These represent 76 distinct 4 digit EC classes, with all 6 of the general enzyme classes being represented. This dataset will be referred to as the "initial" dataset in any later discussions involving its use.

2.4.2 Collection and Definition of Expanded "Artificial" Benchmark Datasets

2.4.2.1 Overview of "Artificial" Dataset Creation Method

Due to the small size of the "initial" dataset described above, it was decided to create a second, expanded, benchmark dataset from the PSI-BLAST analysis, by using a much larger set of aligned target sequences. The construction of this dataset was done via the post-modification of a subset of MSAs that satisfied particular criteria of the original 43,572 database searches. Again, the main aim of this dataset was the collection of examples which show an incorrect specific functional comparison between the query and the most significant enzyme "target" sequence. It is proposed that this situation can be simulated by removing all of the sequences found in the database search, which have the same specific EC function as the query and are classed as more significant than the first incorrect sequence hit. An overview of the method is shown in *figure 2.3*. After removal of these "correct" sequence hits, a set of examples remain that produce an "incorrect" prediction of function, when using the most significant remaining sequence from the BLAST output. To provide reference to the fact that these datasets consist of ordered multiple sequence alignments - where the top-ranked (1^{st}) sequence is always of a different "*incorrect*" specific function to the query sequence - datasets of this form are described as "*All1stINCORRECT*" throughout the thesis. Although these are not examples of "naturally" occurring incorrect examples, from a protein sequence database search, they should be of a high enough quality to provide an accurate prediction benchmark. Indeed, the nature of the Swiss-Prot database - from which the target enzymes were collected – is itself an "artificial" construct containing numerous biases from historical and research origins (Rost, 2002).

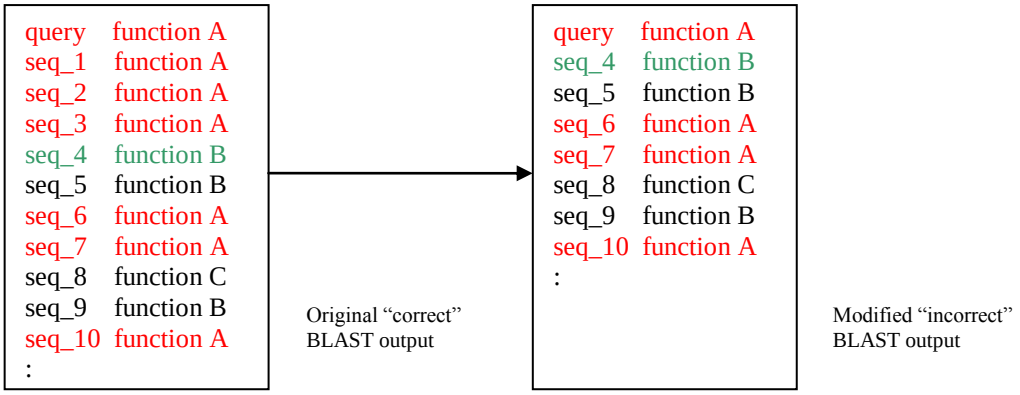

Figure 2.3. Overview of the process used to create the artificial "All1stINCORRECT" dataset examples. The original BLAST output (left) shows an example where the three most significant sequences (seq_1, seq_2, seq_3) have the same function as the query but not as seq_4. Removing these three sequences produces the modified "incorrect" BLAST output (right) where seq_4 is now the most significant, top-ranked, sequence hit.

The source data for this dataset was the 43572 PSI-BLAST searches obtained from the target sequences. All analysis of the output sequence properties is confined to the sequence homologs identified in the 1st PSI-BLAST iteration. The process was as follows: (i) 284 "empty-set" examples were removed (i.e. those that have no target sequences in the output); (ii) 15201 "all-correct" examples were removed (i.e. those that have only target sequences with the same specific function as the query in the output). This identified a reduced set of 28087 examples.

2.4.2.2 Method Used to Ensure a Minimum Level of Functional Diversity in the Benchmark Multiple Sequence Alignments

Two further restrictions for inclusion – the "*MSA functional diversity criteria*" - were then applied: (i) only include examples with at least 10 target sequences with the same specific function as the query and are less significant than the first incorrectly matching target sequence. This reduced the dataset to 6114 examples; and (ii) only include examples with at least 10 target sequences having a different specific function to that of the query. This led to the identification of 4189 "*artificial – All1stINCORRECT*" examples that successfully satisfy all of the criteria set for the inclusion of MSAs within the benchmark datasets of PSI-BLAST generated multiple sequence alignments. The choice of 10 sequence examples was partly arbitrary, but mainly influenced by the fact that it was the number used by Hannenhalli and Russell (2000) when selecting PFAM (Bateman et al., 2004) based MSAs, for a

similar analytical purpose. This is an improved method of ensuring a degree of functional diversity within the MSAs, when compared to that used for the definition of the "initial" dataset.

2.4.2.3 The "*QUERY.enzymes.4189*" Sequence Set

This set of 4189 enzyme sequences that were used as the query sequences in the generation of these examples, will be referred to as the "*QUERY.enzymes.4189*" sequence set throughout this thesis. They show a good distribution of 140 distinct EC classes measured to 4 levels of functional specificity and all 6 general EC classes are represented. Further consideration of the over-representation of certain specific functions is addressed and discussed when required while interpreting particular results at later analysis stages in the thesis.

The bulk of the benchmark analysis, results and conclusions in this thesis are from datasets that have been defined using this particular source set of 4189 query enzyme sequences. In general, these consist of multiple sequence alignments that have been generated through the use of alternative PSI-BLAST sequence database search parameters, allowing comparative analysis between each of the datasets. The procedures used to define these datasets are described in detail below.

2.4.2.4 Methods Used to Define the "Artificial - All1stINCORRECT" Datasets of MSAs

In this section the procedures are described that are used to define some benchmark datasets of MSAs that are repeatedly used throughout the experiments in this thesis. These are defined at this point to avoid unnecessary repetition at later stages. An associated standardized naming convention, used to refer to each of the particular datasets, is also explained. The core methodology used for the PSI-BLAST sequence database search was essentially identical to that previously discussed in this chapter. There were a number of alterations to particular parameters, which are discussed at relevant points, and for clarity the full procedure that was followed is repeated below.

The "*QUERY.enzymes.4189*" sequences were used as the input query protein sequences. A PSI-BLAST database search was then carried out for each of the 4189 target enzyme sequences against the UniProt (Swiss-Prot + TrEMBL) database

(version 4.0), which contained 1,757,967 sequences. Each input sequence was filtered using the SEG low complexity filter (Wootton and Federhen, 1996) and all of the sequences in the search database were filtered using the low complexity, trans-membrane and coiled-coil filter options of the *pfilt* application (Jones and Swindells, 2002). The sequence database search was carried out using 1 iteration of PSI-BLAST version 2.2.10, using an iteration inclusion value (*-h* parameter) of 0.001 and the default BLOSUM62 amino acid substitution matrix, with a gap open penalty of -11 and gap extension penalty of -1. Also, the maximum number of sequences included in the BLAST search output and resultant MSAs, was set at 5000 using the *–v* and *–b* command line parameters. Further, the data content of each of these datasets is a set of MSAs, generated by the 1st iteration of PSI-BLAST, through the use of the *–m 6* command line parameter.

The resulting MSAs were then filtered to remove all sequences not identified as functionally annotated "target" enzyme sequences – *"MSA target enzyme filtering"*. Finally, each of the resulting 4189 BLAST MSAs were processed using the "*All1stINCORRECT*" artificial dataset post-modification procedure, followed by the "*MSA functional diversity criteria*".

A further two parameters were also used in the generation of the BLAST based MSAs. These are: (i) whether composition-based statistics were utilised during the database search, through the setting of the *–t* command line parameter; and (ii) the level of the E-value output threshold parameter, which controls the sequences that are included in the final MSAs through the statistical significance of the sequence similarity between the query and target enzymes. The particular values used for these parameters are defined with each of the specific dataset definitions given below.

With regards to the use of composition-based statistics when generating the MSAs, a discussion related to the reasons for altering this parameter usage is provided in the next chapter.

As for the output E-value threshold parameter, originally the default value of 10 was used. However, due to the nature of the high-specificity function assignment goals of this thesis, it was later decided to use a more stringently filtered dataset of MSAs,

by applying a lower threshold of 0.001. A lower E-value threshold provides alignments that contain sequences with more significant sequence similarity to the query sequence. An outcome of this more stringent alignment filtering is that the MSAs will generally contain fewer sequence homologs and functional false positives. It follows that the number of dataset examples that satisfy the "*MSA functional diversity criteria*", used to ensure a minimum level of functional diversity within the MSAs of the datasets, is also reduced as the E-value output threshold is reduced.

2.4.2.5 Dataset Naming Scheme

To avoid confusion and increase clarity, each of the BLAST generated datasets of MSAs are named using a standardized naming scheme. The elements of this have been chosen to highlight key dataset features and creation parameters that will be discussed at particular experimental stages during this study, namely: "*All1stINCORRECT*" - the MSAs have been modified using the "*All1stINCORRECT*" artificial dataset creation procedure; *"tT"* – composition-based sequence statistics have been used during the sequence database search through setting the *–t* parameter to *T* (true); *"tF"* – composition-based sequence statistics have NOT been used during the sequence database search through setting the *–t* parameter to *F* (false); "*BLOSUM62*" – refers to the particular amino acid substitution matrix used for the database search (in this example the BLOSUM62 matrix); "*masked*" – the residues in the resultant MSAs still contain the sequence masking used to aid the database search; *"unmasked*" – all of the sequences in the MSAs were post-processed to replace all masked "X" amino acid residues with the original amino acid residues from the source, target, Swiss-Prot protein sequences, to generate "unmasked" MSAs; and "*En*" – indicates that the output E-value threshold, which controls the sequence similarity to the query sequence of the MSA sequences, is set as less than or equal to *n*.

2.4.2.6 The "All1stINCORRECT – Using Composition-Based Statistics" Datasets

Two datasets of key interest in this thesis have been created when using composition-based sequence statistics in the BLAST search. These are both the masked and unmasked forms of the dataset that used the default E-value MSA output

threshold of 10. These are referred to as the *"All1stINCORRECT.tT.BLOSUM62.masked.E10"* and *"All1stINCORRECT.tT.BLOSUM62.unmasked.E10"* datasets respectively. After the application of the *"MSA target enzyme filtering"* and "*All1stINCORRECT*" artificial dataset post-modification procedures, followed by the "*MSA functional diversity criteria*", both of these datasets contain the same 4189 MSA examples. The properties of the 4189 query sequences that define these datasets have already been discussed in the earlier "*QUERY.enzymes.4189*" sequence set section.

2.4.2.7 The "All1stINCORRECT – Without Composition-Based Statistics" Datasets

Four additional datasets used in this thesis were created without using composition-based sequence statistics in the BLAST search. These are both the masked and unmasked forms of datasets that used E-value MSA output thresholds of 10 and 0.001.

After the application of the *"MSA target enzyme filtering"* and "*All1stINCORRECT*" artificial dataset post-modification procedures, followed by the "*MSA functional diversity criteria*", the masked and unmasked datasets, which use the E-value<=10 threshold, each contain the same 4054 MSA examples. These are referred to as the *"All1stINCORRECT.tF.BLOSUM62.masked.E10" and "All1stINCORRECT.tF.BLOSUM62.unmasked.E10"* datasets.

When using an E-value threshold <= 0.001, to define which sequences will be part of the generated MSAs, and the application of the "*All1stINCORRECT*" artificial dataset post-modification procedure and the "*MSA functional diversity criteria*", the number of MSA examples in the datasets is reduced to 3527. The masked and unmasked forms of these datasets are referred to as the *"All1stINCORRECT.tF.BLOSUM62.masked.E0.001"* and the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* datasets respectively (see *Appendix I* for more detailed description of these datasets).

2.5 Conclusions

The work in this chapter has aimed to serve two purposes. Firstly, the collection of a large set of enzyme sequences, to allow a study of the functional conservation

accuracy of homology transfer, at high levels of functional specificity through the use of standard sequence homology measures. Secondly, the use of this data to identify datasets that are suitable for the benchmarking of methods intended for improving the prediction of specific enzyme function.

The assessment of the level of enzyme function conservation demonstrates that even close sequence similarity relationships do not suffice to allow confident transfer of specific function in all cases. When placed in comparison to the many previous studies discussed above, some of which draw far more pessimistic conclusions for comparable sequence similarity thresholds, the need can be seen for more powerful methods of discriminating between very similar functional sub-classes. It is the aim of this thesis to investigate some of these methods. Through the use of multiple alignments of homologous sequences it is proposed that sequence features specific to a particular function can be used to separate the different functional types. Evolutionary relationships between groups of homologous sequences, with the same function, can be used to identify amino acid residues that play an essential role in the specific function of the proteins. These are commonly referred to as functional specificity determining residues (fSDRs) and will form a central point of the work in this thesis.

Benchmark datasets have been defined and identified from the analysis carried out above. These are composed of examples where the most significant sequence match from a PSI-BLAST database search is not of the same specific function as the query sequence. Therefore, they fulfill criteria for the assessment of alternative methods that are designed to improve the discrimination of specific functional classes when compared to simple threshold-based sequence similarity methods. An "initial" dataset was first identified for use as a benchmark comparison dataset. However, a larger series of "artificial" datasets were subsequently defined, which supersede the "initial" dataset and are used when assessing the performance of the methods in this thesis. This is because they contain more sequence examples and enzyme functions on which to base the results, lending greater weight to any statistical conclusions drawn from these studies. The larger datasets have also been constructed in a way to provide a guarantee of "sufficient" functional diversity within the aligned sequences

of the examples, with which to aid the analysis of the multiple alignments and the identification of particular inherent evolutionary relationships.

In conclusion, the main goal of this research is to develop and analyse automated techniques for improved high-specificity function prediction, using groups of closely related aligned homologous enzyme sequences. The initial studies carried out in this chapter show why this is an important and timely research problem and also define benchmark datasets to help achieve this goal.

Chapter 3 The Use of Alternative Amino Acid Substitution Matrices for Rescoring the Functional Similarity of Enzyme Sequences

3.1 Introduction

As shown in the previous chapter, it is not always the case that the most significant sequence hit, found through a database search, will have the same specific enzyme function as the query sequence. Neither is a simple sequence similarity threshold sufficient for consistent, high accuracy, functional annotation of protein sequences. The aim of the work in this chapter is the investigation of different scoring metrics, for improved assignment of specific function, when compared to the results from a sequence similarity database search. The hypothesis is that this may provide improved functionally specific ordering of the identified homologous sequences, based on additional sequence features to those used in the statistical homology measures of the original database search.

It has been shown that simply using the most significant "top-hit" from a sequence database search for the prediction of a specific enzyme function can lead to significant levels of incorrect annotation. It is therefore both important and timely, to investigate ways in which groups of sequence homologues identified in a database search, can be scored and re-ranked to improve, both the confidence and the accuracy of the predictions for the specific function of the query sequence.

3.1.1 Overview of Alignment Rescoring Method

A general conceptual overview and the aims behind the alignment re-scoring procedure used in this chapter are discussed in this section. A diagrammatic overview of this procedure is shown in *figure 3.1*. It should be noted that similar, comparable procedures, for the purpose of functionally re-scoring the sequence alignment ordering, are also used to analyse the performance of alternative methods that are investigated in later chapters.

The first three stages depicted in *figure 3.1* are related to the collection and alignment of relevant sequence homologs. This procedure is discussed in detail within *chapter 2* and is also included in this overview diagram to provide context with respect to the functional re-ranking of the identified sequences. An iterative procedure is then carried out to re-score each of the sequences in the multiple sequence alignment (MSA), using a particular scoring method.

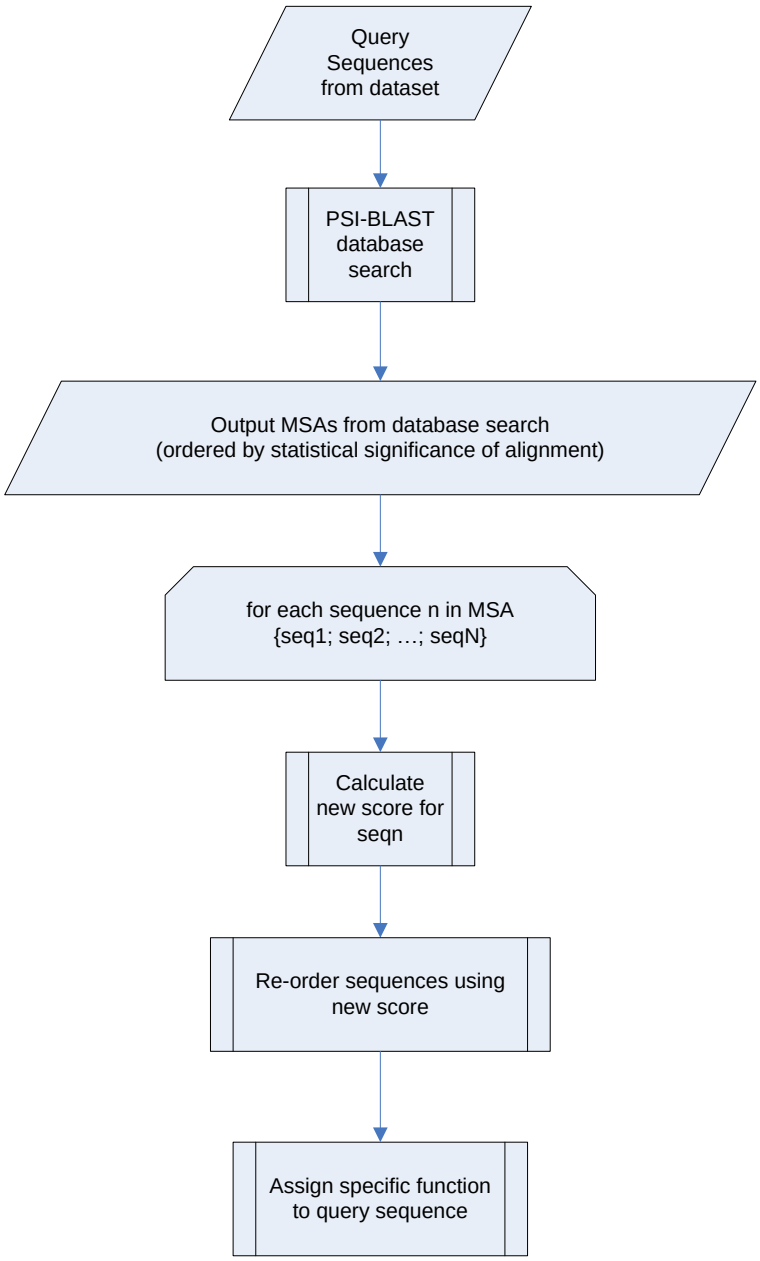

Figure 3.1. Diagrammatic overview of the alignment rescoring procedure.

In this chapter, the re-scoring method comprises pair-wise comparisons between the query sequence and the high scoring "target" sequence homologs, from each of the MSAs in the benchmark datasets. These pair-wise comparisons are carried out using well-established protein sequence alignment metrics. Once all of the sequences within each MSA have been evaluated they can then be re-ordered, using the newly calculated score. Predictions for the specific functional class of the query sequence can then be made based upon this new sequence ordering.

3.1.2 Amino Acid Substitution Matrices

An important consideration when aligning protein sequences and assessing their subsequent level of similarity, is the method used for scoring the similarity between each of the aligned amino acid residues. Evolution determines the structural and functional features of proteins and it is the mutation of amino acid residues that is the main driving force. It therefore follows that, in general, more similar protein sequences are closer in an evolutionary sense and hence show more closely correlated features of specific function.

Analysis of the pattern and rate of change of amino acids during evolutionary divergence was first carried out by Dayhoff (1978). Due to the fact that certain groups of amino acids display similar physical and/or chemical properties (Taylor, 1986), the probability of mutations being accepted through natural selection is greater the more similar the properties are. This becomes clear when considering the need for structural and functional continuity and the likely deleterious effects of a large change in observed amino acid properties during mutation, due to a disruption of function.

Through the alignment of multiple sequences from large numbers of related proteins a probabilistic evolutionary model of the expected mutations from one amino acid to another can be developed. A number of methods and datasets have been used to calculate scoring matrices for particular features and evolutionary distances between proteins (Dayhoff, 1978; Henikoff and Henikoff, 1992; Jones et al., 1992), some of which are discussed below. The simple residue identity type of matrix is first described, followed by two commonly used methods; the percent accepted mutation (PAM) matrices (Dayhoff, 1978) and the BLOSUM series of matrices (Henikoff and Henikoff, 1992).

The form of an amino acid substitution matrix is usually that of a symmetrical matrix of 20 rows by 20 columns, one for each of the 20 common types of amino acid residues. This leads to 210 distinct entries consisting of 190 row and column pairings where the amino acid residues are not the same and 20 further pairings along the matrix diagonal where they are.

3.1.2.1 IDENTITY Matrix

A simple form of substitution matrix is the identity matrix, which consists of a score of one between identical amino acids in an alignment and zero for all other residue comparisons. Although there is no specific evolutionary theory associated with this type of scoring scheme, its simplicity and close association with the commonly used percentage identity measure between sequences means that it is sometimes used for a simple scoring of alignments. The main problem with this matrix is that it rewards and penalises all matched and mismatched residues to the same degree. This is done regardless of the similarities in physico-chemical properties of amino acids or their likelihood of mutation. The following models of amino acid substitution scoring attempt to address these deficiencies.

3.1.2.2 PAM Matrices

The model for generating PAM substitution matrices was developed by Dayhoff (1978) using alignments of closely related groupings of homologous protein sequences with at least 85% sequence identity. Due to the high level of sequence similarity within the groups any observed mutations of the amino acids did not significantly affect the function of the proteins. The next step was to count the number of observed mutations between all pairs of amino acid types, within all the protein groups, allowing an empirical measure of the probability of mutation for each pair of amino acids to be calculated. Finally this data was normalised to remove any bias caused by amino acid composition, mutation rate or sequence length. These calculated amino acid relative mutabilities are those expected within the evolutionary time period defined as 1 PAM unit. For ease of computation these are usually represented in the substitution matrix in their logarithm of odds (log-odds) form, which describes the ratio of the observed frequency of amino acid substitutions divided by the frequency expected by chance. Due to the fact that the model of evolutionary mutation used was a Markov process, it is possible to

calculate larger PAM-N distances through matrix multiplication of the values in the PAM-1 matrix.

3.1.2.3 BLOSUM Matrices

Another commonly used set of substitution matrices for protein sequence alignment and similarity scoring is the BLOSUM series (Henikoff and Henikoff, 1992). The method used to generate these matrices shows a number of important differences to that of the Dayhoff PAM model of amino acid evolution and is based on a larger dataset of protein sequences. Rather than start with very closely related sequences and extrapolate to more divergent ones, the Henikoffs approached the problem by starting with a more divergent set of protein sequences from more than 500 protein families. Using these family alignments, "blocks" of sequence patterns, without gaps, were extracted from the particular families and added to a database. The scoring matrices were then calculated using the log-odds of the types of substitutions found in the conserved pattern of blocks. The different forms of BLOSUM-N matrices (such as BLOSUM62 and BLOSUM50, where N is 62 and 50 respectively) are calculated by first grouping all sequences, within a block, that show an aligned sequence identity above a particular threshold. Each group is then represented by a single sequence with a weighted average of the observed amino acid substitutions within the group. For example, the commonly used BLOSUM62 matrix consists of amino acid substitution data calculated from block patterns that have all sequences, with greater than or equal to 62% identity, clustered into one averaged sequence representative. This reduces the contribution to the matrix from more closely related sequence members of an aligned protein family.

It is important to note that it is not possible to extrapolate from one BLOSUM-N matrix to another as with the PAM matrices, because they are not based on an evolutionary Markovian model. Therefore it is only possible to calculate BLOSUM-N matrices from empirical data of aligned blocks of sequences of the required similarity levels as described above.

It is has been found that the BLOSUM62 matrix generally gives the best overall performance for sequence alignment and sensitive sequence database searching, hence the reason that this matrix is currently used as the default amino acid substitution scoring model in BLAST and PSI-BLAST.

3.2 Methods

3.2.1 Datasets

In this section I will describe the benchmark datasets that are employed to assess the performance of each alignment re-scoring method. The datasets consist of ordered sets of MSAs that are used to determine the specific enzyme prediction accuracy of each re-scoring method. A number of alternative datasets are described, for which three main differences in their method of generation are highlighted. These differences are related to the particular amino acid substitution matrices that are used, in the BLAST database search, to generate the MSAs in each of the datasets. Three different matrices (BLOSUM62, PAM160 and PAM30) are used to allow an investigation into the effect that their use as the database search matrices would have on the functional classification accuracy of the resulting MSAs. In addition, they are used to assess the effects on the functional classification accuracies, of the order in which the particular database search and alignment re-scoring matrices are applied in the functional re-scoring assessment procedure. The reasons for selecting these particular substitution matrices are discussed in detail, in both the relevant method and results and discussion sections below. For the datasets in which the detailed methods are not specified below, the methods used to generate the datasets have been previously defined in detail in *chapter 2*.

3.2.1.1 "Artificial" Dataset Using Composition-based Sequence statistics in BLAST Database Search

Both the masked – *"All1stINCORRECT.tT.BLOSUM62.masked.E10"* – and the unmasked - *"All1stINCORRECT.tT.BLOSUM62.unmasked.E10"* – forms of the 4189 BLAST generated MSAs from these datasets were used in the following analysis.

3.2.1.2 Refinement of the "Artificial" Dataset by Removal of Effect Due to Composition-Based Sequence Statistics

Additional datasets of MSAs were generated, without the use of composition-based sequence statistics during the sequence database search and with an output E-value threshold of 0.001 used to control the sequences included in the output MSAs. Both the masked – *"All1stINCORRECT.tF.BLOSUM62.masked.E0.001"* – and the

unmasked - *"All1stINCORRECT.tF.BLOSUM62.unamsked.E0.001"* – forms of the 3527 BLAST generated MSAs from these datasets were used in the following analysis (see *Appendix I* for more detailed description of these datasets).

3.2.1.3 Generation of a Dataset of MSAs Using a PAM160 Sequence Database Search Matrix

A dataset of MSAs was generated through the use of the PAM160 matrix in a PSI-BLAST protein sequence database search. The steps used in the methodology were as close as possible to those previously described when using the BLOSUM62 substitution matrix. For clarity, the PSI-BLAST search procedure and parameters used is repeated below.

As before, the PSI-BLAST database search was carried out, for each of the 4189 target enzymes in the *"QUERY.enzymes.4189"* sequence set, against the UniProt (Swiss-Prot + TrEMBL) database (version 4.0). Each input sequence was filtered using the SEG low complexity filter (Wootton and Federhen, 1996) and all of the sequences in the search database were filtered using the low complexity, trans-membrane and coiled-coil filter options of the *pfilt* application (Jones and Swindells, 2002). The sequence database search was carried out using 1 iteration of PSI-BLAST (version 2.2.10), an output E-value threshold of 0.001 and the PAM160 substitution matrix. Composition-based sequence statistics were not used during the database search, hence the *–t* parameter was set as *–t F*.

The version of PSI-BLAST used does not implicitly contain support for the PAM160 substitution matrix. Because of this it was necessary to determine the most suitable gap penalty parameters to use in the database search. In comparisons, by Henikoff and Henikoff (1992), between the PAM and BLOSUM series of matrices, the PAM160 matrix is shown to be most closely comparable to the BLOSUM62 matrix. Using this information and that from Reese and Pearson (2002), which suggests similar effective gap penalties for the two matrices, I decided to use the same gap open and gap extension penalties, of -11 and -1 respectively, that were used in the database search with the BLOSUM62 matrix.

An MSA post-processing procedure identical to that used for the BLOSUM62 generated MSAs was then carried out. Firstly, the resulting MSAs were filtered to

remove all sequences not identified as "target" enzyme sequences (see *chapter 2*). Next, each of the MSAs were processed using the "*All1stINCORRECT*" artificial dataset post-modification procedure and finally the "*MSA functional diversity criteria*" was applied (both of these procedures are defined in *chapter 2*).

This resulted in a dataset consisting of 3100 PSI-BLAST generated MSAs, whose query sequences represent coverage of 88 distinct EC classes. During associated analysis and discussion throughout this thesis, the sequence residue masked and unmasked forms of this dataset will be referred to as the *"All1stINCORRECT.tF.PAM160.masked.E0.001"* and *"All1stINCORRECT.tF.PAM160.unmasked.E0.001"* datasets respectively (see *Appendix I* for more detailed description of these datasets).

3.2.1.4 Generation of a Dataset of MSAs Using a PAM30 Sequence Database Search Matrix

One further dataset of MSAs was generated for analysis in this chapter. In this case, a PAM30 substitution matrix was used in the PSI-BLAST sequence database search. Unless specified otherwise, the steps used in the generation of these PAM30 based MSAs are identical to those used in the PAM160 based BLAST MSAs, detailed above.

The important difference in this dataset generation method was that a PAM30 substitution matrix was used in the PSI-BLAST database search. As with the PAM160 case, above, the version of PSI-BLAST used does not implicitly contain support for the PAM30 substitution matrix. Therefore, it was again necessary to determine the most suitable associated gap penalty parameters for use in the database search. The parameters decided upon were: -9 for the gap opening; and -1 for the gap extension penalty. These were selected because two previous studies (Altschul et al., 2001; and Frommlet et al., 2004), which investigate the effects of sequence alignment scoring schemes on statistical alignment parameters, both recommend the use of these gap scoring parameters with the PAM30 substitution matrix.

An MSA post-processing procedure, identical to that used for the PAM160 generated MSAs, was then carried out. This resulted in a dataset consisting of 2110 PSI-BLAST generated MSAs, whose query sequences represent coverage of 82

distinct EC classes. During associated analysis and discussion throughout this thesis, the sequence residue masked and unmasked forms of this dataset will be referred to as the *"All1stINCORRECT.tF.PAM30.masked.E0.001"* and *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* datasets respectively (see *Appendix I* for more detailed description).

3.2.2 Calculation of Alignment Scores Using Non-Standard Amino Acid Substitution Matrices

A method was developed for the parsing of the PSI-BLAST generated multiple alignments. Each of the individual pair-wise alignments, between the query and high scoring "target" sequences from the database search, were then re-scored using a selected set of amino acid substitution matrices. The substitution matrices used in the experimental analysis were:

- *IDENTITY matrix:* This consisted of just two different score entries for all amino acid pairings *s(i, j):*
  - $s(i, j) = 1$ where $i = j$
  - $s(i, j) = 0$ where $i \neq j$
- *PAM matrices:* A number of PAM matrix evolutionary distances were used in this analysis, ranging from: PAM10 to PAM250 in increments of 10 PAM units.
- *BLOSUM matrices:* A variety of BLOSUM matrices were also used in the analysis: (BLOSUM30 to BLOSUM60 in increments of 5; BLOSUM62; BLOSUM70 to BLOSUM90 in increments of 5; and BLOSUM100)

All of the PAM and BLOSUM series of matrices used were downloaded from the following website (ftp://ftp.ncbi.nih.gov/blast/matrices/). The PAM matrices were calculated using *"pam" Version 1.0.6 [28-Jul-93]* and the BLOSUM matrices were calculated from the BLOCKS 5.0 database, at the required sequence cluster percentage level.

3.2.2.1 Alignment Re-Scoring Procedure

The procedure used in these experiments for re-scoring each pair-wise alignment, between query and target sequence in the MSAs, closely follows that shown in

figure 3.1. Each of the individual pair-wise alignments were extracted and all of the aligned residue pairs were then re-scored using the scores defined in each of the distinct substitution matrices described above. It is important to note that it is the local alignments, generated by PSI-BLAST, that are used in this analysis and that no re-alignment of the sequences is carried out. A simplified overview of this process, consisting of only two pair-wise alignments, is shown in *figure 3.2*. This particular example shows two short alignments and the resulting score obtained from using the BLOSUM62 matrix to score each of the aligned residues between the *query* and *sequence_n*. In this case the alignment score of the *query* with *sequence_2* is greater than with *sequence_1*. Therefore, using this scoring scheme, *sequence_2* would be ranked as a closer specific functional match to the query than *sequence_1*.

												score
Query	L	L	A	R	F	Q	V	R	M	G	P	
Sequence_1	I	L	G	Y	M	Q	F	R	K	G	P	
BLOSUM62 score	***2***	***4***	***0***	***-2***	***0***	***5***	***-1***	***5***	***-1***	***6***	***7***	***25***
Query	L	L	A	R	F	Q	V	R	M	G	P	
Sequence_2	L	L	G	L	F	Q	N	R	Y	G	P	
BLOSUM62 score	***4***	***4***	***0***	***-2***	***6***	***5***	***-3***	***5***	***-1***	***6***	***7***	***31***

Figure 3.2. A simplified schematic overview, showing the way that pair-wise sequence alignments are functionally re-scored, using different amino acid substitution matrices (in this particular case BLOSUM62 is used).

3.2.2.2 Treatment of Insertions and Deletions

Insertions and deletions of amino acids play an important role in protein evolution. They give rise to "gapped" sections to provide optimal alignments between sequences. In this analysis two different approaches were taken to the treatment of gaps in the alignments when calculating the re-scored values.

Un-gapped: This method scores all residues aligned to gap positions as 0

Gapped: This method uses the same affine gap penalty model as that used in the BLAST algorithm and is defined below in *equation 3.1*

$$G_n = g_{open} + (n-1) * g_{extend} \qquad \textit{(equation 3.1)}$$

where G_n is the overall gap penalty, g_{open} is the penalty for opening a gap, n is the number of consecutive gaps and g_{extend} is the penalty for extending a gap. In both the

"un-gapped" and *"gapped"* form of analysis the starting and trailing gaps were removed from the ends of all the alignments before carrying out the rescoring calculations.

3.2.3 Assessing Prediction Accuracy

3.2.3.1 Top-hit Method

To assess the improvement in prediction accuracy when re-scoring the MSAs, a simple "top-hit" approach was taken. This is where the specific function of the query sequence is assigned the same specific function as the sequence with the highest score from the pair-wise re-scoring procedure. If the specific functional classes are the same (to a degree of all 4 numbers in the EC hierarchy), then the result is defined as a *"correct"* prediction of specific function, otherwise the result is defined as an *"incorrect"* prediction.

Exceptions to these outcomes are seen when a group of sequences have equal scores, producing a set of tied ranking positions. A group of this kind contains two or more members that may (or not) have the same specific function. If the members all have the same specific function, and it is the same as the query sequence, then this is classed as a correct prediction. Alternatively, if none of them have the same specific function as the query then this is classed as incorrect. A third case is where the sequences in the "tied-rank" group have two or more different functional classes and one of them is the same specific function as the query sequence; in this case it is not possible to differentiate between the correct and incorrect examples and therefore can be classed as "*undecidable"*. For all practical purposes, these types of *"undecidable"* examples should be classified as "incorrect" when considering the functional prediction results, as they cannot be separated from those that are correct using the available information from the defined scoring scheme. In this analysis, these *"undecidable"* examples are indeed treated as "incorrect" predictions.

3.2.3.2 Definition of a Random Sequence Selection Model for Specific Function Prediction

A random model of a simple naive prediction system was defined to provide a baseline comparison with the "top-hit" function prediction results obtained from the different re-scoring methods. This was based upon the concept of randomly

permuting the ranked results of the sequence homologues in each of the MSAs in the dataset. The prediction result was then determined to be correct or incorrect through functional comparison between the specific EC classification of the query sequence and the randomly permuted "top-hit".

A simple, computationally inexpensive way of modelling these random permutations is through the calculation of the probability of randomly selecting a functionally correct sequence (where all 4 levels of the EC hierarchy are equal between the query and randomly selected sequence) from each MSA. The resulting probability calculation, for each MSA, is shown in *equation 3.2*.

$$P_{random\_correct} = \frac{n_{correct}}{n_{all}} \qquad \textit{(equation 3.2)}$$

Where: $P_{random\_correct}$ is the probability of a randomly assigned, correct, functional prediction; $n_{correct}$ is the number of sequences in the MSA with the same (correct) specific function as the query sequence; and n_{all} is the total number of sequences in the particular MSA of interest.

3.2.3.3 Bootstrap Re-sampling Analysis of Results

A computational statistical re-sampling method, known as the "bootstrap" (Efron and Gong, 1983), was used to allow the accurate calculation of statistical properties from data distributions that are not normally distributed. The central limit theorem states that the distribution of a sample of calculated means approximates a normal distribution, when the number of data points is large. Standard statistical calculations can then be made on the resulting, normally distributed, bootstrap re-sampled data.

Sample Mean

The sample arithmetic mean, $\bar{x}$, is calculated using *equation 3.3*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{(equation 3.3)}$$

where n is the number of data points in the sample and x_i is the value of data point i.

Standard Error

The standard error is a metric that is commonly used to approximate the dispersion of a sample statistic, such as the sample mean. The bootstrap sample statistics were used in this calculation, following the method of Good (1999). The standard error (*se*) can be defined as the square-root of the unbiased estimate of the sample variance (see *equation 3.4*).

$$se = \sqrt{\text{var}\,iance(bootstrap\_statistics)} \qquad \textit{(equation 3.4)}$$

Equation 3.5 shows the detailed method of calculation used to compute the standard error (*se*) of a sample containing *B* bootstrap values

$$se = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \overline{\hat{\theta}_b}\right)^2} \qquad \textit{(equation 3.5)}$$

Where $\hat{\theta}_b$ is the bootstrap value, *b*, and $\overline{\hat{\theta}_b}$ is the mean of these bootstrap values.

Outline of the Bootstrap Re-sampling Procedure

The general bootstrap procedures used for the experimental analysis of both the random model and the alignment re-scoring methods are described below, where the number of bootstrap repetitions, *B*, is 10000 in all of the bootstrap calculations.

Using the Random Model Data

- For a dataset of *N* MSA examples, calculate the distribution of the *N* probabilities for "random correct prediction", calculated using *equation 3.2*.

- **Bootstrap:** (repeat steps 1 and 2, *B* times, storing the mean sample estimate from each bootstrap replicate in a vector, ***M***, of length *B*)

    1. Randomly select *n* (where *n=N/2*) data-points, with replacement, from the original sample distribution of $P_{random\_correct}$ values.

    2. Calculate, using *equation 3.3*, the mean of the $P_{random\_correct}$ bootstrap sample values and add to vector ***M***.

- Finally, calculate the standard error (se) of the bootstrap statistics contained in ***M***.

Using the Alignment Re-scoring Top-hit Prediction Data

- For a dataset of *N* MSA examples, apply the re-scoring method, evaluate whether the functional "top-hit" prediction result is "correct" or "incorrect". The dataset will then consist of $n_{correct}$ and $n_{incorrect}$ examples.

- **Bootstrap** (repeat steps 1 and 2, *B* times, storing the calculated sample estimate from each bootstrap replicate in a vector, ***M***, of length *B*)

    1. Randomly select *n* (where *n=N/2*) data-points, with replacement, from the original sample distribution of *N* ($n_{correct}$ and $n_{incorrect}$) examples.

    2. Calculate the fraction of correct examples in the bootstrap sample and add to vector ***M***.

- Finally, calculate the standard error (se) of the bootstrap statistics contained in ***M***.

3.2.4 Calculation of PAM Distance from Sequence Percentage Identity

A PAM 1 mutation matrix is defined to be a specific measure of a unit of evolutionary distance. Therefore, it is possible to define a function that calculates the relationship between PAM evolutionary distances and the changes in amino acid sequence identity. In this chapter these calculations were carried out using the *PerIdentToPam()* function that is available in the Darwin interpreted computer language suite of software tools (Gonnet et al., 2000). This function carries out an iterative procedure using Newton's method for solving equations (see the following section of the Darwin user manual for further details: http://www.inf.ethz.ch/personal/gonnet/DarwinManual/node155.html).

3.2.5 Query Sequence Clustering

The input query sequences that were used as input to the BLAST database search and MSA generation were clustered based on the level of sequence identity through the use of the CD-HIT algorithm (Li and Godzik, 2006). The clustering was done for each of the separate query sequence sets identified by the three dataset generation methods described above. A range of percentage sequence identity levels were used for the clustering (40% - 90% in intervals of 10%) and the recommended default parameters were used for all. The longest sequence in each cluster was used as the

representative. A summary of the cluster properties, at each defined level of sequence identity, is given in the relevant section of results.

3.3 Results and Discussion

3.3.1 Benchmark Prediction Results Using the Artificial Datasets

An initial analysis of the 4189 MSA examples, in the *"All1stINCORRECT.tT.BLOSUM62.masked.E10"* dataset, was carried out to ensure the correct functioning of the alignment re-scoring algorithm. The same amino acid substitution matrix, gap scoring algorithm and gap penalty values, as those employed for the BLAST generation of the alignments, were used for the alignment re-scoring. These were: BLOSUM62; the affine gap penalty scoring method described in *equation 3.1;* and a gap opening (g_{open}) value of -11 and gap extension (g_{extend}) value of -1, respectively.

As has been described previously, the way in which the MSAs in the artificial datasets have been modified ensures that none of them generate a correct "top-hit" functional prediction result, when considering all 4 levels of the EC classification scheme and the sequences have been ranked in ascending E-value order. Therefore, the hypothesis was that by using a score matrix and gap penalty parameters in the re-scoring algorithm, equivalent to those used in the sequence alignment during the BLAST database search, an identical sequence ranking should be observed for each of the MSAs. This was however not the case, as a significant number (2045 out of 4189, or a proportion of 0.49 correct predictions) of examples in the re-scored *"All1stINCORRECT.tT.BLOSUM62.masked.E10"* dataset, showed a correct functional sequence "top-hit" after the functional alignment re-ranking, when using the BLOSUM62 re-scoring matrix and the gapped scoring model.

These results clearly show that the alignment re-scoring algorithm was not producing the expected results during the calibration of the benchmark dataset. This was problematic because it indicated a possible flaw within the re-scoring algorithm, preventing the establishment of a true, reproducible, benchmark comparison between the BLAST generated predictions and those from the re-scored alignments. The reasons for these discrepancies are investigated and discussed further in the following section.

3.3.1.1 Testing and Calibration of Benchmark Datasets Used for Assessing the Prediction Accuracy of the Functional Re-scoring Algorithm

The alignment re-scoring algorithm was carefully tested to ensure that the expected alignment score, for each of the pair-wise alignments, was being calculated. The results from this test showed that the algorithm was generating the expected results when using the specified gap scoring model and amino acid substitution matrix. However, comparisons between these calculated alignment scores, and the "raw" BLAST alignment scores, showed differences that caused the functional ranking discrepancies in the benchmark dataset.

This finding indicated that the differences between the BLAST alignment scores and those calculated with my re-scoring algorithm must be explained by additional parameters in the BLAST alignment score calculations that were not being incorporated into the alignment re-scoring algorithm. Analysis of the parameters used in the BLAST search highlighted the use of sequence composition-based statistics calculations (controlled through the use of the command line *–t* argument), during the generation of the BLAST alignments, as the reason for the observed discrepancies. It was found – using the *"All1stINCORRECT.tF.BLOSUM62.masked.E10"* dataset - that, when compared to no use, composition-based statistics can generate slightly different alignment scores. This can lead to varying statistical significance scores and subsequent differences in the rankings of the sequence homologs identified with BLAST. This was the reason for the observed differences between the "top-hit" function prediction results of the BLAST MSAs, when using composition-based statistics and those from the alignment re-scoring algorithm, when using identical substitution matrices and gap scoring models.

To correct these differences I decided to define an alternative benchmark dataset of MSAs, still generated by BLAST, but without the use of composition-based statistics. This solution was chosen because it allows for an exact reproduction of the aligned sequence ordering, and associated "top-hit" function prediction results, when using the re-scoring algorithm. It also provides a simpler implementation for the alignment re-scoring algorithm because there is no requirement to explicitly calculate the additional effects due to the composition-based sequence statistics.

In summary, when composition-based statistics are not used to generate the alignments, the benchmark re-scoring results are equivalent between both the BLAST-based and alignment re-scoring methods, when using an equivalent substitution matrix, gap scoring model and penalties. Therefore, for the remainder of this chapter, the experimental analysis only uses datasets that contain MSAs that have been generated without the use of composition-based sequence statistics. Also, at this point, a decision was made to concentrate all further analysis on MSAs created through the use of a more stringent output E-value threshold of 0.001, namely the MSAs in the *"All1stINCORRECT.tF.BLOSUM62.masked.E0.001"* and the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* datasets. This was found to not alter the general results and experimental trends observed during the following analysis.

3.3.2 Definition of a Simple Random Sequence Selection Model for Function Prediction

During the analysis in this chapter, comparisons are made between the function prediction results from alternative alignment re-scoring methods and those from associated random sequence selection models. As described in the methods, the random model used for this comparison is based on the probability of randomly selecting a sequence, from a multiple alignment, that has the same specific function as the query sequence. The aim of these comparisons is to assess the difference in prediction performance between the re-scored analysis results and the baseline provided by the random model. Where necessary the random model is defined alongside the associated dataset and analysis under discussion. Also, in *table 3.1*, a summary of the dataset size, bootstrap parameters and calculated mean and standard error (se) statistics is given for the random sequence selection models of each dataset used in this analysis.

3.3.2.1 Probability Distributions and Bootstrapping of the Random Sequence Selection Model

In general, the probabilities for the correct prediction of specific enzyme function, using a model of uniform random sequence selection from each of the MSAs in a dataset, follow a non-normal sample distribution. Due to this, the bootstrap method

can be used (see methods) to calculate associated statistical properties of the distribution, such as the sample mean and standard error.

When calculating the bootstrap statistics for the *"All1stINCORRECT.tF.BLOSUM62.(un)masked.E0.001"* datasets, the number of bootstrap replicates, *B*, used was 10000 and the sample size for each replicate was 1764, which is approximately half of the 3527 MSA examples in the dataset. The resulting statistics, shown in *table 3.1*, for the random sequence selection model for this dataset show a bootstrap mean of 0.502 and a standard error of +/- 0.006.

Dataset	**No. of MSAs (*N*)**	**(bootstrap) Sample Size (*N/2*)**	**(bootstrap) mean +/- se**
All1stINCORRECT.tF.BLOSUM62.E10	4054	2027	0.475 +/- 0.006
All1stINCORRECT.tF.BLOSUM62.E0.001	3527	1764	0.502 +/- 0.006
All1stINCORRECT.tF.PAM160.E0.001	3100	1550	0.522 +/- 0.007
All1stINCORRECT.tF.PAM30.E0.001	2110	1055	0.572 +/- 0.008

Table 3.1. Summary of the dataset size, bootstrap sample size and calculated mean and standard error (se) statistics for the random sequence selection model for each associated dataset of MSAs used in this analysis.

3.3.3 The Effect on the Top-Hit Prediction Performance of Using Alternative Substitution Matrices to Re-score the MSAs

The aim of this section is to analyse the effect, on the performance of the "top-hit" function prediction results, of using alternative amino acid substitution matrices with the alignment re-scoring algorithm. A thorough investigation of the IDENTITY matrix and the BLOSUM and PAM series of amino acid substitution matrices, defined in the methods, is carried out.

Also studied are some of the additional parameters that may affect the alignment re-scoring results, such as sequence residue masking and the gap scoring of the alignments. Alongside these analyses are comparisons to the associated function prediction results, from the baseline random sequence selection model, of the dataset under investigation.

3.3.3.1 Comparison Between the Substitution Matrices When Using MSAs Containing Sequence Masking

For this analysis the *"All1stINCORRECT.tF.BLOSUM62.masked.E0.001"* dataset was used, with sequence residue masking still present in the functionally filtered MSAs. The effect of the amino acid substitution matrices, on the alignment re-scoring results, are compared using the gap scoring model of *equation 3.1*, with the same gap penalties as those used in the original BLAST search: $g_{open} = -11$ and $g_{extend} = -1$ and also with an "un-gapped" scoring model where $g_{open} = 0$ and $g_{extend} = 0$. Both the number, and proportion, of correct function prediction results for a representative set of IDENTITY, BLOSUM-N and PAM-N substitution matrices are shown in *table 3.2*. All four levels of EC functional classification of the top scoring aligned sequences in each re-ranked MSA, are used to predict the specific enzyme function of the query sequences.

	Gapped (-11, -1)		Un-gapped (0, 0)	
Re-score Matrix	**Number (proportion) Correct**	**(bootstrap) mean proportion correct +/- se**	**Number (proportion) Correct**	**(bootstrap) mean proportion correct +/- se**
IDENTITY	1819 (0.52)	0.516 +/- 0.012	1844 (0.52)	0.523 +/- 0.012
BLOSUM30	1507 (0.43)	0.427 +/- 0.012	1909 (0.54)	0.541 +/- 0.012
BLOSUM40	1467 (0.42)	0.416 +/- 0.012	1945 (0.55)	0.552 +/- 0.012
BLOSUM50	1306 (0.37)	0.370 +/- 0.011	1907 (0.54)	0.541 +/- 0.012
BLOSUM60	809 (0.23)	0.229 +/- 0.010	1845 (0.52)	0.523 +/- 0.012
BLOSUM62	0 (0.00)	0.000 +/- 0.000	1850 (0.52)	0.524 +/- 0.012
BLOSUM70	1291 (0.37)	0.366 +/- 0.011	1882 (0.53)	0.533 +/- 0.012
BLOSUM80	1544 (0.44)	0.438 +/- 0.012	1898 (0.54)	0.538 +/- 0.012
BLOSUM90	1589 (0.45)	0.450 +/- 0.012	1882 (0.53)	0.534 +/- 0.012
BLOSUM100	1744 (0.49)	0.494 +/- 0.012	1906 (0.54)	0.540 +/- 0.012
PAM10	2002 (0.57)	0.568 +/- 0.012	2053 (0.58)	0.582 +/- 0.012
PAM20	2018 (0.57)	0.572 +/- 0.012	2124 (0.60)	0.602 +/- 0.011
PAM30	2043 (0.58)	0.579 +/- 0.012	2165 (0.61)	0.614 +/- 0.012
PAM40	2032 (0.57)	0.576 +/- 0.012	2134 (0.61)	0.605 +/- 0.012
PAM50	2049 (0.58)	0.581 +/- 0.012	2086 (0.59)	0.591 +/- 0.012
PAM60	2017 (0.57)	0.572 +/- 0.012	2043 (0.58)	0.579 +/- 0.012
PAM80	1946 (0.55)	0.552 +/- 0.012	1979 (0.56)	0.561 +/- 0.012
PAM100	1828 (0.52)	0.518 +/- 0.012	1985 (0.56)	0.563 +/- 0.012
PAM120	1780 (0.51)	0.505 +/- 0.012	1935 (0.55)	0.549 +/- 0.012
PAM140	1721 (0.49)	0.488 +/- 0.012	1934 (0.55)	0.548 +/- 0.012
PAM160	1712 (0.49)	0.485 +/- 0.012	1928 (0.55)	0.547 +/- 0.012
PAM180	1635 (0.46)	0.464 +/- 0.012	1899 (0.54)	0.538 +/- 0.012
PAM200	1600 (0.45)	0.453 +/- 0.012	1904 (0.54)	0.540 +/- 0.012
PAM220	1660 (0.47)	0.471 +/- 0.012	1911 (0.54)	0.542 +/- 0.012
PAM240	1658 (0.47)	0.470 +/- 0.012	1886 (0.54)	0.535 +/- 0.012

Table 3.2. A comparison between the number, and proportion, of correct functional prediction results for a representative set of substitution matrices used for alignment re-scoring. All results for the number of correct predictions are out of a possible 3527. Also shown are the corresponding mean and standard error (se) results calculated from the bootstrap analysis. Results from both gapped and un-gapped gap re-scoring models are shown, where gap penalties of (g_{open} = -11 *and* g_{extend} = -1) and (g_{open} = 0 *and* g_{extend} = 0) were used respectively.

IDENITITY Matrix

When using the IDENTITY matrix, with the "gapped (-11, -1)" gap scoring model, to re-score the MSAs in the *"All1stINCORRECT.tF.BLOSUM62.masked.E0.001"* dataset, the proportion and number of correct predictions is 0.52 (1819/3527), see *table 3.2*.

BLOSUM-N Matrices

The results, in *table 3.2*, for the "gapped (-11, -1)" re-scoring analysis, clearly show that the expected minimum - of 0 correct predictions - is obtained when the BLOSUM62 matrix is used in the alignment re-scoring algorithm. Also, as the N value of the BLOSUM-N matrices is both increased and decreased, the number of correct predictions increases. This is, perhaps, to be expected, as the definition of the benchmark dataset only allows for the identification of examples that either improve, or do not alter, the accuracy of function prediction. There also appears to be some correlation between an increasing number (or proportion) of correct predictions and the distance of the BLOSUM-N N value from the BLOSUM62 matrix used to calibrate the dataset. For the BLOSUM-N matrices, the maximum fraction of correct predictions, 0.49 (1744/3527), is obtained by re-scoring the alignments using the BLOSUM100 matrix.

PAM-N Matrices

The prediction results, in *table 3.2*, for the "gapped (-11, -1)" re-scoring analysis when using the PAM-N matrices show quite a different trend to those of the BLOSUM-N. Most noticeably, there is no clear minimum for the matrices in the series that is comparable to that of the BLOSUM-N results. This is most striking for the PAM160 matrix, which is the suggested PAM series equivalent to the BLOSUM62 matrix (Henikoff and Henikoff, 1992), because it does not show a comparable prediction performance, of 0 correct predictions, to that of BLOSUM62. The minimum fraction of correct predictions is observed with the PAM200 matrix, whereas, the maximum fraction of correct predictions, 0.58 (2049/3527), is obtained by re-scoring the alignments with the PAM50 matrix.

3.3.3.2 Applying Bootstrap Analysis to the Alignment Re-scoring Results

To obtain a more statistically accurate assessment for the mean fraction of correct prediction results, and the associated standard error, a bootstrap analysis was carried out on the results from the *"All1stINCORRECT.tF.BLOSUM62.masked.E0.001"* dataset. The bootstrap parameters used were the same as those for the associated random model, where the number of replicates, *B*, was 10000 and the sample size of each replicate was 1764 - approximately half the number of MSA examples, 3527, in the dataset. Unless otherwise stated, all remaining analysis comparisons and discussion in this chapter will refer to the bootstrapped form of the function prediction results.

The bootstrap analysis results, with the mean and standard error (se) values for a representative set of the IDENTITY, BLOSUM-N and PAM-N substitution matrices, are shown in *table 3.2*. With regards to the results from the "gapped (-11, -1)" gap scoring model, it can be seen that the mean and standard error for the BLOSUM62 results is 0. This is to be expected as all the examples are defined to be incorrect predictions with this score matrix, which leads to no variation in the sample distribution of predictions used for the bootstrap. Overall, both with and without bootstrapping, the trends of the re-scoring results for all of the substitution matrices are similar.

Maximum predictive performance is still seen when using the PAM50 matrix, with a mean proportion of 0.581 correct predictions. Although there is now significant overlap, of the standard error bars, with the results from PAM10 to PAM40 and PAM60. Each one of these "optimal" matrices shows a large improvement, in the proportion of correct predictions, when compared to the random sequence selection model, which has a mean value of 0.502, shown in *table 3.1*.

3.3.3.3 Comparisons Between the Masked and Unmasked Alignments

As discussed previously, sequence masking was used for the BLAST search and generation of the MSAs in the benchmark dataset. To investigate the effects of sequence masking on the prediction results, the alignments were modified to replace all masked sequence residues, with the amino acid residues present in the associated source protein sequences extracted from the Swiss-Prot database. The key observation to take from these alignment rescoring results, is the consistent

improvement in the proportion of correct functional "top-hit" predictions for all of the substitution matrices investigated, when comparing the respective results from the alignments containing un-masked with those containing masked sequence residues. Overall, the trends in the prediction results are similar to those of the masked sequences, with significant improvement (within 1 standard error difference) shown for all of the matrices, except those results from using the PAM10 matrix. The remaining analyses focus on the results from re-scoring the un-masked versions of the MSAs from each dataset.

The optimal predictive performance, for the "*All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* dataset, is now seen when using the PAM30 matrix to re-score the unmasked sequence alignments, with a bootstrapped mean proportion of 0.606 correct predictions and a standard error of +/- 0.012. This provides a small increase, of 0.025, for the proportion of correct function predictions, when compared to the results from using the PAM50 matrix to re-score the masked alignments. Also shown is an improvement, of 0.104, in the mean proportion of correct predictions, when compared to the random sequence selection model, mean value, of 0.502, shown in *table 3.1*.

3.3.3.4 Comparison Between the "Gapped" and "Un-gapped" Models for Alignment Re-scoring

All of the results shown so far incorporate a "gapped" scoring method into the alignment re-scoring algorithm, which uses an identical scoring model and parameters to that of the default gapped BLAST algorithm with the BLOSUM62 search matrix. In this section, the "un-gapped" method, which scores all residues aligned with gaps as 0, was used to calculate a comparable set of alignment scores (see methods).

The results, shown in *figure 3.3*, provide a comparison between the use of the "gapped" and "un-gapped" models for scoring sequence alignment gaps in the *"All1stINCORRECT.tF.BLOUSM62.unmasked.E0.001"* dataset. It can be seen from these results that a significant increase in the proportion of correct predictions is obtained when the "un-gapped" gap scoring model is used for the alignment rescoring. This is true for the IDENTITY and all of the BLOSUM-N and PAM-N substitution matrices investigated. The clearest example of this is in the difference

between the numbers of correct predictions when using the BLOSUM62 matrix with the "un-gapped" model. When using gap-scores of $g_{open} = -11$ and $g_{extend} = -1$, the masked dataset shows 0 correct predictions, whereas the unmasked dataset has a mean proportion of 0.218 correct predictions. However, with the "ungapped" method (where $g_{open} = 0$ and $g_{extend} = 0$), the mean proportion of correct predictions increases to 0.524 and 0.560, for the masked and unmasked alignments respectively. Further, for the BLOSUM-N matrices, a clear difference can be seen between the trends in prediction results for the gapped and un-gapped scoring models. When using the un-gapped model there is little difference between the proportions of correct predictions for the different BLOSUM-N matrices, especially when taking the overlap of the standard error of the mean into consideration. This is in contrast to the results, described above, for the gapped model of the BLOSUM-N alignment re-scoring. The trends for the PAM-N matrices are similar overall to those seen in the gapped model but show a consistently improved performance.

A further observation is highlighted by the comparison of these un-gapped prediction results to the associated random sequence selection model, where all of the mean values, for the proportion of correct predictions from the un-gapped BLOSUM-N and PAM-N re-scoring results, show a significant improvement when compared to the random model. This is also the case for the results for the IDENTITY matrices and the gapped results from the PAM-N matrices, when N is less than 170.

The optimal prediction result, for all matrices investigated when using the un-gapped model with unmasked sequence alignments, was 0.631, which was observed with the PAM30 matrix and can be seen in *figure 3.3*.

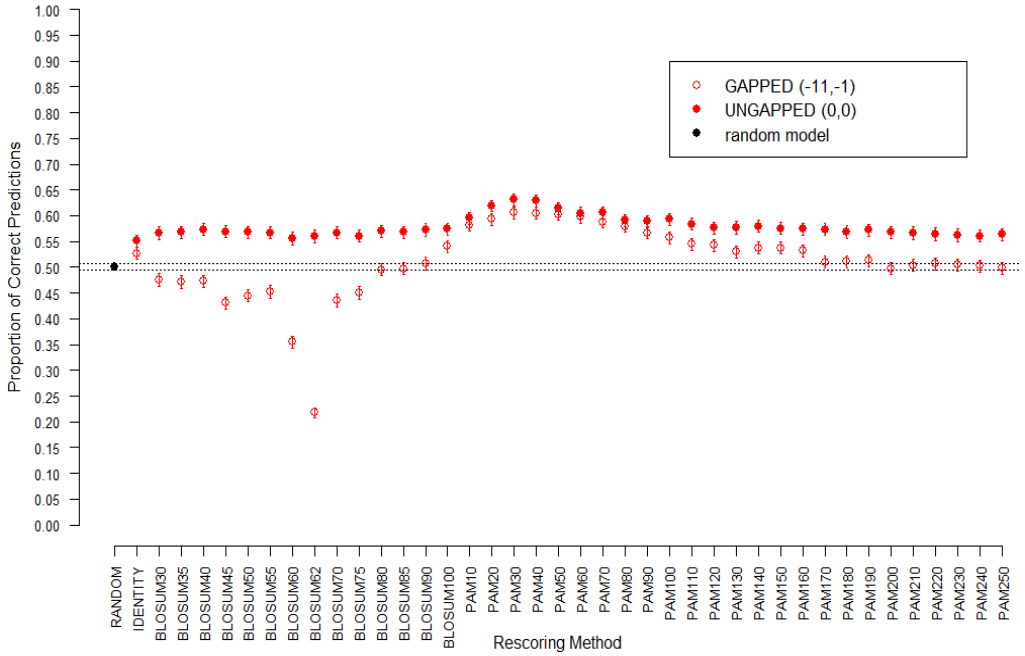

Figure 3.3. A comparison of the proportion of correct predictions obtained for each of the specified substitution matrix re-scoring methods. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. Results are shown for the gapped (-11,-1) and un-gapped (0,0) alignment re-scoring of the Al1stINCORRECT.tF.BLOSUM62.unmasked.E0.001 dataset, when using the IDENTITY, BLOSUM-N and PAM-N substitution matrices. Also shown is the associated random sequence selection model for the dataset, where the dotted lines show 1 standard error deviation from the mean.

3.3.3.5 Comparison Between the Re-Scoring the Alignments from the "Original" and "Artificial" Datasets

To assess whether these observed results were dependent on the nature of the "artificial – All1stINCORRECT" dataset of alignments, a control experiment was carried out. In the previous section, it was shown that the PAM30 matrix was the optimally performing matrix for re-scoring the 3527 unmasked alignments from the "All1stINCORRECT" dataset. However, due to the "artificial" nature of the benchmark dataset used it is not clear whether these results are concealing a potential decrease in performance when re-scoring alignments that already have a correct specific functional hit as the top-ranked sequence. Therefore, the "original" BLAST MSAs (i.e., prior to the generation of the "artificial" dataset, described in *section 2.4.2*) were re-scored using the PAM30 matrix. These were then compared to the

results obtained from re-scoring the "original" unmasked MSAs with the BLOSUM62 matrix.

When using gap-scores of g_{open} = *-11* and g_{extend} = *-1*, the unmasked "original" dataset showed 3459 (out of 3527) and 3465 (out of 3527) correct predictions, when re-scoring with the BLOSUM62 and PAM30 matrices, respectively. In comparison, when using the "un-gapped" scoring model (g_{open} = *0* and g_{extend} = *0*) the unmasked "original" dataset showed 3454 (out of 3527) and 3463 (out of 3527) correct predictions, when re-scoring with the BLOSUM62 and PAM30 matrices, respectively.

These results show that, for both the gapped and un-gapped models, there is a small increase in the proportion of correct predictions when using the PAM30 matrix instead of the BLOSUM62 to re-score the "original" alignments. However, the key observation from these results is that the use of a PAM30 matrix, when compared to the BLOSUM62 matrix used in the BLAST search, does not have a detrimental effect when re-scoring alignments that contain a large proportion of examples that are originally "correct". This result, therefore, provides validation for the use of the "*All1stINCORRECT*" artificial datasets as a benchmark in this thesis.

3.3.4 Investigation into the Effect of the Amino Acid Substitution Matrix Used in the BLAST Search on the Top-Hit Prediction Accuracy

In the alignment re-scoring analysis discussed above, the BLOSUM62 amino acid substitution matrix was used in the BLAST sequence database search that generated the MSAs in each dataset. It was shown that the overall optimum performance, for specific enzyme function prediction, was obtained from re-scoring the MSAs using the PAM30 substitution matrix. To investigate whether this observed prediction improvement was due to the specific ordered combination of BLOSUM62 and PAM30 matrices, this analysis was followed by investigating the use of the PAM equivalent of the BLOSUM62 matrix in the BLAST search procedure.

3.3.4.1 Analysis of the Dataset Obtained from Using the PAM160 Matrix in the Sequence Database Search

The PAM160 matrix is regarded as the closest PAM equivalent to the BLOSUM62 matrix (Henikoff and Henikoff, 1992). The following section analyses the effects of re-scoring the MSAs, from the *"All1stINCORRECT.tF.PAM160.masked.E0.001"* and *"All1stINCORRECT.tF.PAM160.unmasked.E0.001"* datasets, with the same set of non-standard amino acid substitution matrices used in the previous analysis of the "BLOSUM62 generated" datasets.

The main purpose of this analysis is to ascertain whether similar trends of function prediction performance are seen, when using the PAM series equivalent of the BLOSUM62 matrix to generate the source dataset MSAs. Specifically, whether there is a similar peak in performance when the lower N values (such as 30) of the PAM-N series matrices are used in the re-scoring. The hypothesis is that this will test whether the enhanced prediction performance is due to: (i) a particular combined property of the BLOSUM62 and low PAM-N matrices; or (ii) due to a more general case of prediction enhancement that is present regardless of whether a BLOSUM or PAM series matrix is used for the generation of the BLAST-based MSAs.

All of the following analysis was carried out on the "bootstrapped" form of the prediction results. For the derivation of these, the number of bootstrap replicates, *B*, used was 10000 and the sample size for each replicate was 1550 - half the number of 3100 MSA examples in the dataset. The random sequence selection model, for the *"All1stINCORRECT.tF.PAM160.(un)masked.E0.001"* datasets, was calculated using the same bootstrap parameters. The statistical parameters of which, are summarised in *table 3.1*.

3.3.4.2 Comparisons Between the Re-scoring of the "Masked" and "Un-masked" Alignments

As in the previous analysis, a set of "unmasked" MSAs were generated, through the replacement of all masked sequence residues with the amino acid residues present in the associated protein sequences extracted from the Swiss-Prot database. The general trends were recorded between the prediction results for the "masked" and "un-masked" sequence alignments. These were observed to be very similar to the trends seen between the "masked" and "un-masked" datasets generated from using

BLOSUM62 as the BLAST search matrix. Specifically, the prediction results for the un-masked datasets show a similar, consistent improvement, over the results from the masked datasets, when using identical substitution matrices for the alignment re-scoring. For brevity, these comparison results are not shown and the remainder of the analysis in this section will focus on the un-masked dataset of MSAs.

3.3.4.3 Comparisons Between the "Gapped" and "Un-gapped" Models for Alignment Re-scoring

Before discussing the comparisons between the results from the re-scoring of the alignments using the "gapped" and "un-gapped" scoring model, it is of interest to first look at the trends and results from re-scoring using just the "gapped" model. As in the previous analyses the gap-score parameters of $g_{open} = -11$ and $g_{extend} = -1$ are used with the gap scoring model defined in *equation 3.1*. Again, these specific parameters were chosen because they are the same as those used in the BLAST sequence database search used to generate the MSAs.

The "gapped" prediction results, for the *"All1stINCORRECT.tF.PAM160.unmasked.E0.001"* dataset, are shown in *figure 3.4*. It can be seen that the minimum prediction result, with a mean value of 0.202, is a result of using the PAM160 matrix to re-score the alignments. This is expected because, due to the way in which the dataset has been defined when using the PAM160 matrix, all of the top ranking sequences show a different, "incorrect", specific function to the query sequence. This is a similar result to that shown previously, when re-scoring the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* dataset, with the BLOSUM62 matrix that was also used in the BLAST search to generate the MSAs.

The key observation that we can take from these results is the presence of a clear peak in prediction performance, when using the PAM-N matrices of PAM10, PAM20 and PAM30 with the "un-gapped" alignment re-scoring model. This is similar to the trend seen when re-scoring the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* MSAs, using equivalent re-scoring parameters. Thus, indicating that the use of a second, low PAM-N re-scoring substitution matrix, improves the specific function prediction performance of BLAST MSAs generated from both BLOSUM62 and PAM160 matrices.

Comparison of the "gapped" prediction results with those from the "un-gapped" model (where the gap-score parameters $g_{open} = 0$ and $g_{extend} = 0$) is shown in *figure 3.4*. These results show that, for most of the substitution matrices used, a significant increase in the proportion of correct predictions is obtained when the "un-gapped" gap scoring model is used for the alignment rescoring. Interestingly, when using the PAM matrices, ranging from PAM10 to PAM70, there is no significant difference between the corresponding "gapped" and "un-gapped" results.

With regards to the alignment re-scoring results obtained from the IDENTITY matrix, neither the gapped or un-gapped results are particularly large, with the proportion of correct predictions equivalent to and slightly larger than the associated random model values, respectively.

When comparing these results with the random model, it is possible to see, from *figure 3.4*, that all of the prediction results from using the "un-gapped" model are significantly better. Whereas in the case of the "gapped" model only one BLOSUM series matrix, BLOSUM100, and the PAM10 to PAM70 range of matrices show a clear, significant improvement, over the random sequence selection model.

The optimal prediction result shows a mean value, for the proportion of correct predictions, of 0.611, which was obtained by using the PAM30 matrix with the "gapped" form of the alignment re-scoring algorithm. There is, however, no significant difference between both the gapped and un-gapped function prediction results when using either of the PAM10, PAM20, or PAM30 substitution matrices.

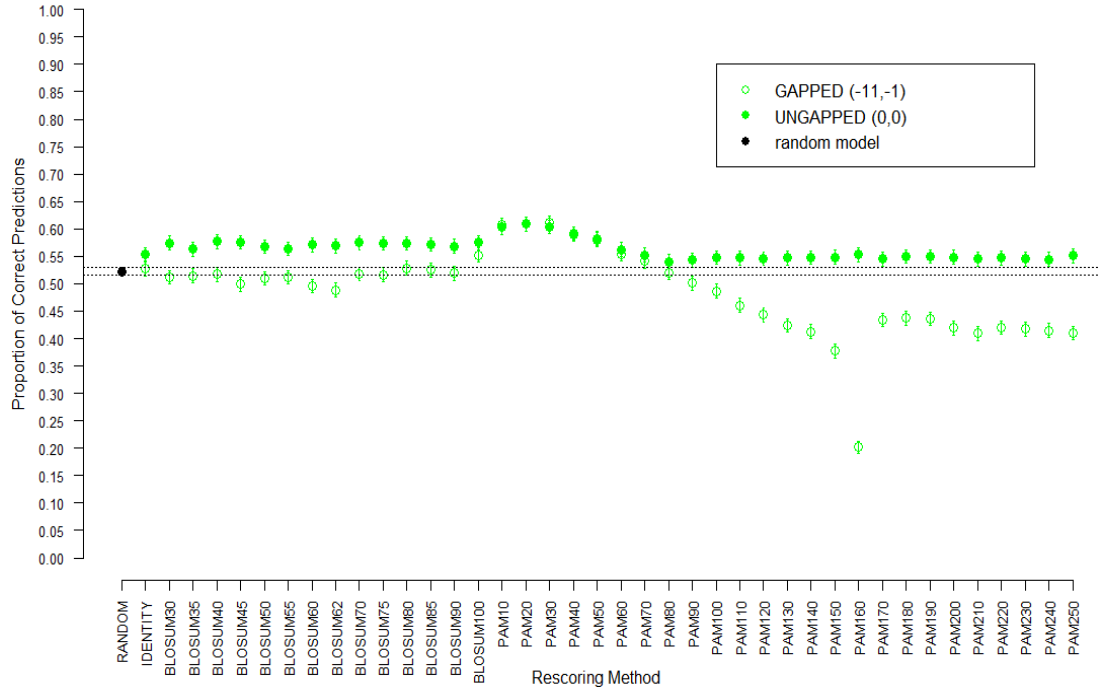

Figure 3.4. A comparison of the proportion of correct predictions obtained for each of the specified substitution matrix re-scoring methods. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. Results are shown for the gapped (-11,-1) and un-gapped (0,0) alignment re-scoring of the Al1stINCORRECT.tF.PAM160.unmasked.E0.001 dataset, when using the IDENTITY, BLOSUM-N and PAM-N substitution matrices. Also shown is the associated random sequence selection model for the dataset, where the dotted lines show 1 standard error deviation from the mean.

3.3.4.4 Comparison Between Results from Re-Scoring the BLOSUM62 and PAM160 BLAST Generated Multiple Alignments

To conclude this part of the analysis, let us compare the prediction results that were obtained from re-scoring the MSAs generated from using both the BLOSUM62 and PAM160 matrices in the PSI-BLAST database search. It has been shown that there are similar peaks in function prediction results, for both the BLOSUM62 and PAM160 generated MSAs, when using the lower PAM-N matrices (where N is in the range between 10 and 50) to re-score the MSAs. Specifically, in both datasets, the PAM30 matrix provides the largest proportion of correct specific enzyme function predictions. In the case of the PAM160 generated alignments the "gapped" model was optimal, whereas for the BLOSUM62 generated alignments the "un-gapped" model was shown to be optimal.

Overall, re-scoring the BLOSUM62 generated MSAs - *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* - with a PAM30 matrix, when compared with the equivalent results from the PAM160 generated alignments - *"All1stINCORRECT.tF.PAM160.unmasked.E0.001"* - that have been re-scored using the PAM30 substitution matrix, shows: (i) a larger mean proportion of correct specific enzyme function predictions, of 0.631, when compared to 0.611; and (ii) a larger improvement over the associated random sequence selection model, of 0.129, when compared to 0.089.

The main conclusion to draw from these results is that an improvement in specific function prediction results is observed, for both the BLOSUM62 and PAM160 BLAST generated alignment datasets, when using an additional PAM30 re-scoring matrix. This indicates that these results are not simply an artefact of the MSAs in the BLOSUM62 generated dataset. Nor are they only due to the specific combination of using a BLOSUM62 matrix to generate the BLAST MSAs followed by a low PAM-N matrix to functionally re-score the alignments. This shows that the use of an additional, carefully selected, substitution matrix can provide a consistent improvement, in the annotation of specific enzyme function.

3.3.4.5 Analysis of the Dataset Obtained from Using the PAM30 Matrix in the Sequence Database Search

Following on from the previous analyses, which looked at the effects of functionally re-scoring BLAST alignments generated with equivalent BLOSUM and PAM amino acid substitution matrices, a set of experiments were carried out to compare the effect of functionally re-scoring BLAST alignments generated with a PAM30 search matrix. The reason for selecting the PAM30 matrix to generate BLAST-based MSAs, was that it has been shown to be the best performing functional re-scoring substitution matrix, when applied to both the BLOSUM62 and PAM160 BLAST generated alignments, and could therefore be used to explore the following outcomes: (i) whether the PAM30 generated MSAs would show a comparable peak in prediction performance when using a BLOSUM62 and/or PAM160 matrix in the alignment re-scoring procedure; (ii) whether the PAM30 generated MSAs would show a comparable peak in prediction performance when using matrices other than the BLOSUM62 or PAM160 matrices in the subsequent alignment re-scoring

procedure; or (iii) whether the PAM30 generated MSAs would show no comparable improvement in specific enzyme function prediction performance when using any of the alternative alignment re-scoring matrices.

The working hypotheses used for this analysis were the following. If outcome (i) was shown to be true, then it would suggest the presence of complementary information between the pair of BLAST creation and alignment re-scoring matrices. Thus resulting in an equivalent enhancement of function prediction performance, independent of the order in which the matrices are applied in the alignment creation and re-scoring procedures. Outcome (ii) would indicate that the alignment re-scoring process had a more unpredictable pattern of behaviour, which is dependent on the specific identity and ordering of the pair of matrices used in the alignment creation and subsequent re-scoring procedures. And outcome (iii) would provide further evidence that MSAs, generated through BLAST database searches using either BLOSUM62 or PAM160 matrices, coupled with subsequent re-scoring with a PAM30 substitution matrix, show the most effective way of observing an improvement in the specific functional annotation of enzyme sequences.

The *All1stINCORRECT.tF.PAM30.unmasked.E0.001* dataset, containing 2110 MSAs, was used for the analysis in these experiments. The bootstrap parameters were: B=10000 for the number of bootstrap replicates; and a bootstrap replicate size, 1055, which is half the number of MSAs in the dataset under analysis. The details of the random sequence selection model associated with this dataset is summarised in *table 3.1*.

Like all previous analysis in this chapter, a series of comparisons were carried out to assess the differences between the alignment re-scoring function prediction results when altering the re-scoring matrices and gap scoring parameters. I will summarise the trends observed and highlight the key findings from these parameter variations that are of relevance to a comparison between these prediction results and those obtained from the BLOSUM62 and PAM160 generated BLAST alignments.

Comparison Between the "Gapped" and "Un-gapped" Models for Alignment Re-scoring

A procedure similar to that used for the gapped and un-gapped re-scoring of the BLOSUM62 and PAM160 PSI-BLAST generated alignments was followed here. Here, the gap-score parameters of $g_{open} = -9$ and $g_{extend} = -1$ are used in the gap scoring model that is defined in *equation 3.1*. Again, these parameters were chosen, for use in the alignment re-scoring with the alternative substitution matrices, because they are the same as those used in the BLAST sequence database search that generated the alignments. The "un-gapped" model again scores both the g_{open} and g_{extend} gap-score parameters equal to 0 during the alignment re-scoring. A comparison of the *Al1stINCORRECT.tF.PAM30.unmasked.E0.001* dataset re-scoring results is provided in *figure 3.5*.

For the "gapped" prediction results, when using the PAM re-score matrices, there is a clear minimum seen, when applying the PAM30 matrix to the alignment re-scoring algorithm, which results in a bootstrap mean value of 0.227 for the proportion of functionally correct predictions. This was expected, due to the way in which the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* dataset was defined. As in the previous analyses of BLOSUM62 and PAM160 BLAST generated datasets, there is a sharp increase in correct predictions when using matrices of both lower and higher *"N"* (BLOSUM-N or PAM-N) values than the particular type of matrix used for the dataset generation. For the PAM10 matrix and the PAM-N matrices, with N values greater than 150, the results approach a level of specific function prediction that is close to that of the random sequence selection model.

With respect to the BLOSUM series of matrices, the results for the "gapped" model show that the proportion of correct predictions, for all of the BLOSUM matrices, are within or below the standard error range of the associated random sequence selection model. Therefore, there is not a minimum of a comparable magnitude to the PAM30 matrix result, or a clear maximum corresponding to the BLOSUM62 re-score results.

Interestingly, for this dataset, the overall maximum proportion of correct predictions, of 0.621, is obtained when re-scoring with the IDENTITY matrix, using the gapped (-9, -1) form of the alignment re-scoring method.

A brief analysis of the results from using the "un-gapped" alignment of re-scoring, shows a broadly flat distribution of mean values for the proportion of correct predictions. This is the case when using both the BLOSUM and PAM series of matrices in the re-scoring algorithm. The results range from a minimum mean prediction value of 0.567, for the BLOSUM75 matrix, to a maximum mean prediction value of 0.577, for the BLOSUM60 matrix, when using the BLOSUM-N matrices. And similar results that range from a minimum mean prediction value of 0.553, for the PAM10 matrix, to a maximum mean prediction value of 0.575, for the PAM140 matrix, when using the PAM-N matrices to re-score the alignments. The corresponding un-gapped re-scoring results for the IDENTITY matrix are found to be less than the results for the random sequence selection model.

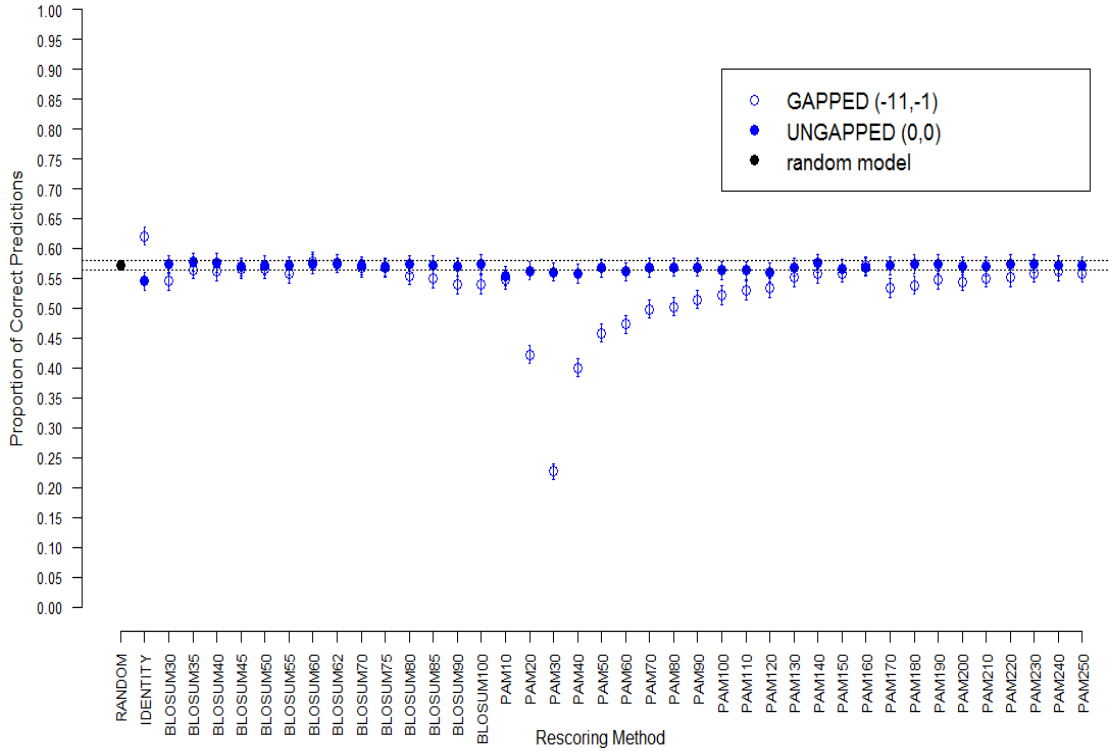

Figure 3.5. A comparison of the proportion of correct predictions obtained for each of the specified substitution matrix re-scoring methods. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. Results are shown for the gapped (-9,-1) and un-gapped (0,0) alignment re-scoring of the Al1stINCORRECT.tF.PAM30.unmasked.E0.001 dataset, when using the IDENTITY, BLOSUM-N and PAM-N substitution matrices. Also shown is the associated random sequence selection model for the dataset, where the dotted lines show 1 standard error deviation from the mean.

In summary, when re-scoring the PAM30 based BLAST MSAs, there are no trends in either the "gapped" or "un-gapped" results, when using the PAM-N or BLOSUM-N re-score matrices, that show a significant prediction peak that is comparable to the results obtained from re-scoring the BLOSUM62 or PAM160 generated BLAST MSAs. There is, however, a clear peak when using the IDENTITY matrix with the gapped form of the alignment re-scoring algorithm, which is a new observation for the *All1stINCORRECT.tF.PAM30.unmasked.E0.001* dataset, when compared to the alignment re-scoring results obtained from the previous datasets.

3.3.4.6 Comparison Between Results from Re-Scoring the BLOSUM62, PAM160 and PAM30 BLAST Generated Datasets

It is now possible to compare and contrast the enzyme function prediction results obtained from re-scoring the MSAs, generated via PSI-BLAST, using the BLOSUM62, PAM160 and PAM30 substitution matrices in the sequence database search. For clarity, I have chosen to only include in this comparison a representative subset of results from each of the datasets analysed. These selected subsets are: (i) the "un-gapped" re-scoring results from the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* dataset; (ii) the "un-gapped" re-scoring results from the *"All1stINCORRECT.tF.PAM160.unmasked.E0.001"* dataset; and (iii) the "gapped" re-scoring results from the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* dataset. These were chosen because they highlight the key alignment re-scoring trends and results from each of the three datasets and alternative substitution matrices investigated.

The proportion of correct predictions of enzyme function obtained from re-scoring the alignments from these three selected subsets, along with the associated random sequence selection models, are shown in *figure 3.6a*, along with an enlarged view of the results when using the IDENTITY and PAM-N matrices, shown in *figure 3.6b.* In both figures, the different re-scoring methods are shown on the horizontal axis and the proportion of correct results for the specific enzyme function prediction shown on the vertical axis.

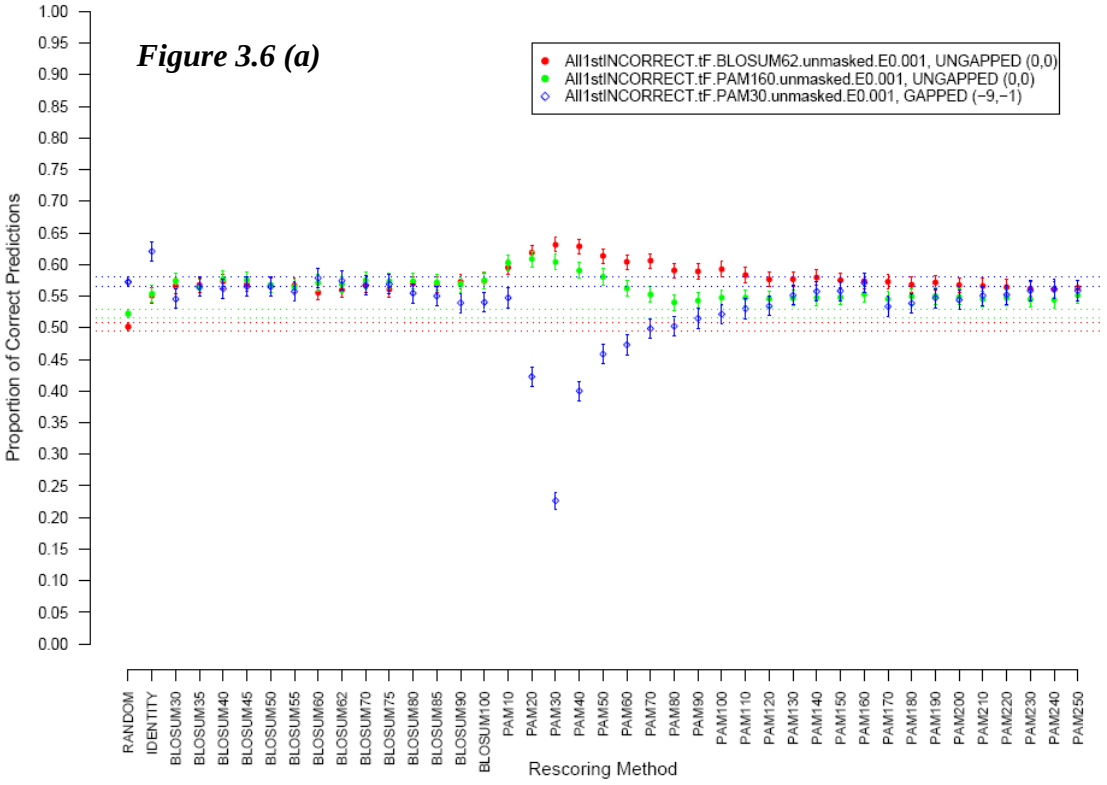
All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001, UNGAPPED (0,0)
All1stINCORRECT.tF.PAM160.unmasked.E0.001, UNGAPPED (0,0)
All1stINCORRECT.tF.PAM30.unmasked.E0.001, GAPPED (−9,−1)
Proportion of Correct Predictions
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
RANDOM
IDENTITY
BLOSUM30
BLOSUM35
BLOSUM40
BLOSUM45
BLOSUM50
BLOSUM55
BLOSUM60
BLOSUM62
BLOSUM70
BLOSUM75
BLOSUM80
BLOSUM85
BLOSUM90
BLOSUM100
PAM10
PAM20
PAM30
PAM40
PAM50
PAM60
PAM70
PAM80
PAM90
PAM100
PAM110
PAM120
PAM130
PAM140
PAM150
PAM160
PAM170
PAM180
PAM190
PAM200
PAM210
PAM220
PAM230
PAM240
PAM250
Rescoring Method

Figure 3.6 (a)

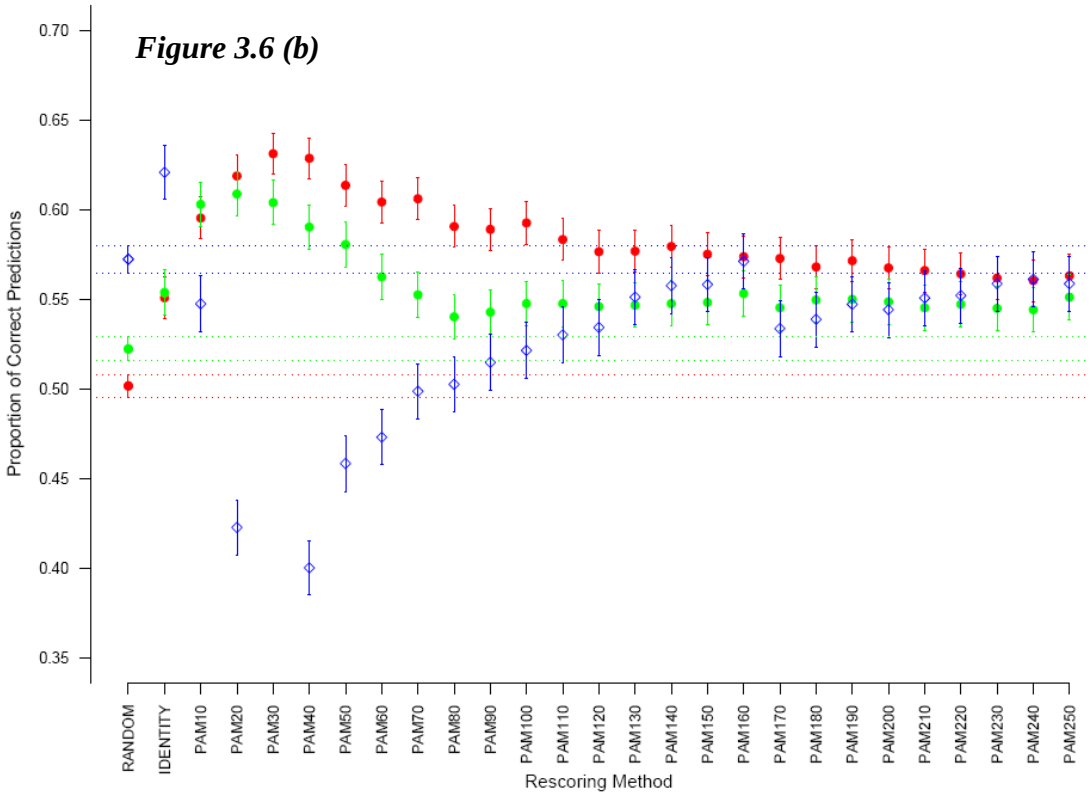

Figure 3.6. A comparison of the proportion of correct enzyme function predictions for each of the specified substitution matrix re-scoring methods. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. Results are shown for the un-gapped (0,0), un-gapped (0,0) and gapped (-9,-1) alignment re-scoring of the All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001, All1stINCORRECT.tF.PAM160.unmasked.E0.001 and All1stINCORRECT.tF.PAM30.unmasked.E0.001 datasets, respectively. Also shown are the associated random sequence selection models for each these datasets, where the dotted lines show 1 standard error deviation from the mean. (a) Shows all the results for the IDENTITY, BLOSUM-N and PAM-N substitution matrices and also the random sequence selection model. (b) Shows an enlarged view of just the IDENTITY and PAM-N matrix re-scoring results and the random sequence selection model. The legend information shown in (a) is also relevant for (b).

This comparison provides an overview of some of the key points that have been discussed so far. We can best see from *figure 3.6(b)* that "un-gapped" re-scoring of the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* alignments, with the PAM30 substitution matrix, produces the largest proportion of 0.631 correct predictions. In addition, this PAM30 re-score result shows a larger difference than any of the other methods, of 0.129, between the mean value of the re-score prediction result and the mean of the associated random sequence selection model. Also, *figure 3.6(a)* shows the difference between the general trends in prediction

results between the three MSA generation methods investigated. In the case of the results from the un-gapped re-scoring of both the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* and *All1stINCORRECT.tF.PAM160.unmasked.E0.001* datasets, the peaks in prediction performance when using the lower PAM-N matrices are clear. These peaks start to become apparent when re-scoring with PAM-N matrices with N values of 70 and below. In contrast, the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* generated results do not show any similar peaks with any of the comparable BLOSUM-N or PAM-N re-score methods used. But, these results do show an improved predictive performance when using the IDENTITY in the alignment re-scoring algorithm, which is almost comparable to that of the un-gapped PAM30 re-scoring results from the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001 dataset*.

3.3.5 Effect from Clustering the Dataset Query Sequences

A series of sequence clusters were defined, using six thresholds of sequence percentage identity (90%, 80%, 70%, 60%, 50%, and 40%), by clustering the query sequences used to create each of the three BLAST-generated sets of MSAs (see *Appendix I* for more detailed description of these datasets). The aim of this was to investigate the effect that any potential bias, due to sequence redundancy within the query sequences used to create the benchmark datasets, may have on the accuracy and trends of the alignment re-scoring prediction results. To some extent, this consideration has already been factored into the previous analysis through the repeated bootstrap sampling of the prediction results. A summary of the sequence identity clustering thresholds and the number of sequence clusters generated is given for each of the datasets, in *table 3.3*, where a 100% identity threshold refers to the dataset compositions prior to any CD-HIT sequence clustering. The number of sequence clusters produced at each threshold, for each distinct dataset, also defines the number of MSAs that constitute the datasets at each of the sequence identity thresholds.

% identity threshold	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Dataset: All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001							
sequence clusters	721	1038	1392	1701	2131	2622	3527
Dataset: All1stINCORRECT.tF.PAM160.unmasked.E0.001							
sequence clusters	608	869	1174	1440	1826	2270	3100
Dataset: All1stINCORRECT.tF.PAM30.unmasked.E0.001							
sequence clusters	403	582	766	925	1191	1503	2110

Table 3.3. A summary of the number of clusters generated for each of the three datasets at each of the specified sequence identity clustering thresholds.

For each set of "clustered" sequence alignments within each dataset, a repeat of the previous alignment re-scoring experimental analysis was carried out, using the same IDENTITY, BLOSUM and PAM substitution matrices. Overall, the prediction results from the alternatively clustered subsets of the three MSA datasets were found to show similar trends to the previously discussed results, obtained without query sequence clustering. A point of note is that the standard error deviation becomes progressively larger as the sequence identity threshold used in the clustering is lowered. This is to be expected because it causes the number of examples in the datasets to decrease, which means that the bootstrap statistics are calculated on progressively smaller sample distributions. An example of this can be seen in *figure 3.7*, which shows how the proportions of correct predictions are altered when using the un-gapped re-scoring model on the alignments from the sequence identity clustered subsets of the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* dataset. For clarity, only the results from the 40%, 60%, 80%, and 100% sequence identity clustered subsets are shown. These results show that the overall trends in the prediction results, through consideration of the mean proportions, are similar for each of the cluster thresholds used. It can been seen, however, that as the clustering threshold is lowered, the best performing re-scoring method on this particular dataset becomes the PAM40, rather than the PAM30 matrix, previously identified when re-scoring the un-clustered sets of MSAs. Also, these results highlight the increasing lengths of the standard error bars as the sequence threshold is lowered, which leads to greater overlap between results from alternative re-scoring methods.

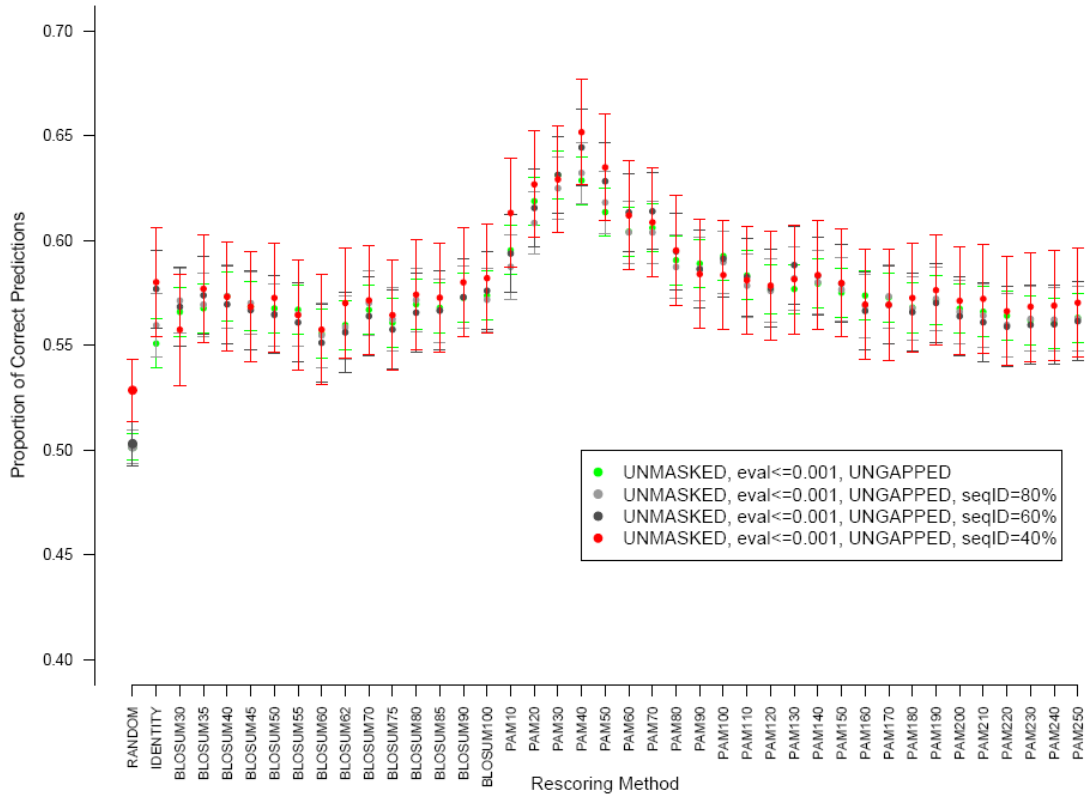

Figure 3.7. Comparison of the proportion of correct predictions from the un-gapped alignment re-scoring results from a selection (40%, 60%, 80%, and 100%) of the sequence clustered subsets of the "All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001" dataset.

A comparison is shown, in *figure 3.8*, between the results from re-scoring the three BLAST generated datasets after a 40% sequence identity threshold has been applied to the constituent query sequences. These results provide an overview of the results obtained from both the gapped and un-gapped scoring models, when using the IDENTITY, BLOSUM and PAM matrices. Also shown are the associated random model statistics for each one of the three clustered datasets.

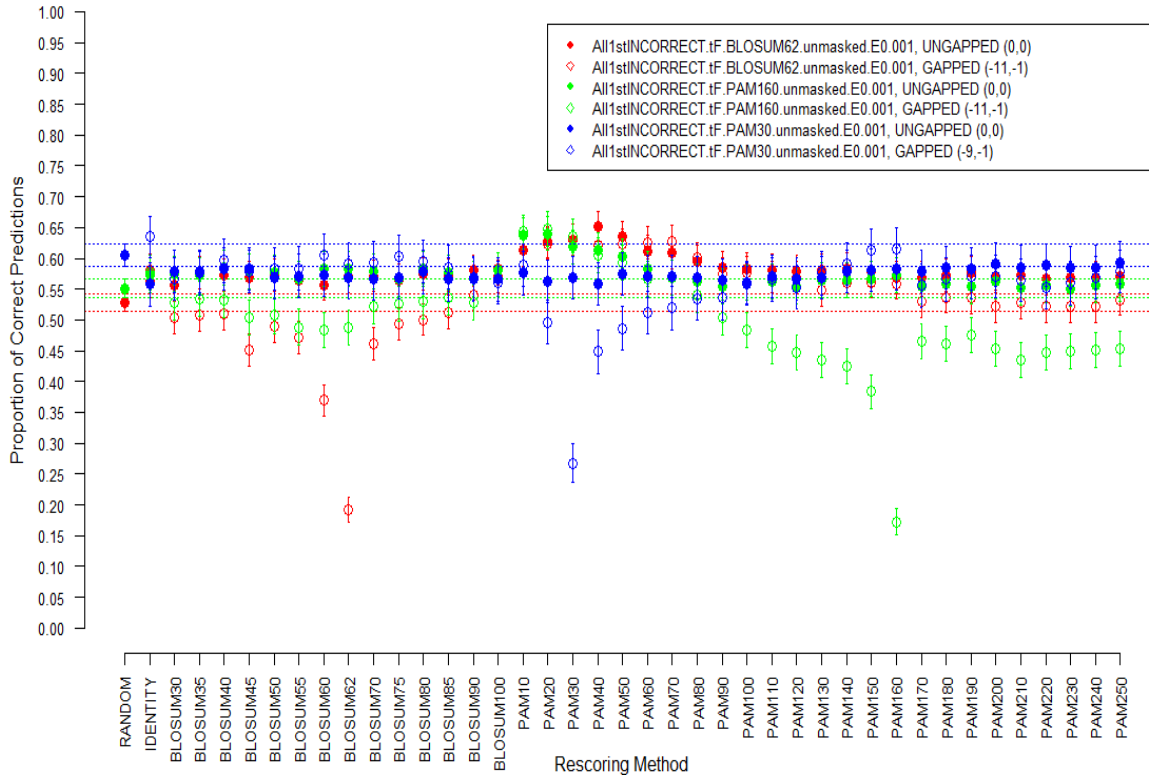

Figure 3.8. A comparison of the proportion of correct predictions obtained for each of the specified substitution matrix re-scoring methods. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. Results are shown for the All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001, All1stINCORRECT.tF.PAM160.unmasked.E0.001 and All1stINCORRECT.tF.PAM30.unmasked.E0.001 datasets, after a 40% sequence identity threshold has been applied to the query sequences. Also shown are the associated random sequence selection models for each of these datasets, where the dotted lines show 1 standard error deviation from the mean.

The results from the 40% sequence identity threshold are shown because they were found to display the largest deviation from the results seen previously when no sequence clustering was used. Although, from the comparisons in this graph it can be seen that the overall trends in the re-scoring results are broadly comparable to those obtained from the datasets where no sequence identity clustering has been applied. A good example of this is seen when analysing the results from re-scoring the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* and *"All1stINCORRECT.tF.PAM160.unmasked.E0.001"* alignments with the PAM-10 to PAM-50 set of matrices, which show comparable improvements in performance.

There are, however, some notable exceptions to this, primarily concerning the results from re-scoring the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* dataset when

a 40% sequence cluster threshold has been applied. These are highlighted in more detail below, with the aid of *figure 3.9*. This graph provides a clearer comparison between the PAM-N and IDENTITY re-scoring matrix results. Comparisons are shown between the re-scoring results from the *"All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001"* and *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"*datasets, with two sequence identity cluster thresholds (100% and 40%), when using the IDENTITY and PAM10-N matrices with both gapped and un-gapped scoring models. With regards to the gapped form of the re-scoring algorithm, it can be seen that the proportion of correct predictions is consistently greater for the MSA subset clustered at 40% query sequence identity than for the un-clustered (100%) dataset. In contrast, the results from the un-gapped re-scoring model are generally more closely correlated when the query sequence clustering is applied.

Of particular interest is the relatively large increase in the proportion of correct predictions seen when applying the PAM150 and PAM160 re-scoring matrices, with the gapped scoring model, to the 40% query sequence clustered subset of the "*All1stINCORRECT.tF.PAM30.unmasked.E0.001*" dataset. This is an interesting observation because it shows the possible start of a peak in prediction performance, when using the PAM-N matrix (PAM160) that is most closely related to the BLOSUM62 matrix used to generate the BLAST MSAs in the comparison dataset.

These results slightly contradict the previous comparisons between the results from the three BLOSUM62, PAM160 and PAM30 BLAST generated MSA datasets, without taking into consideration any query sequence clustering. The lack of a corresponding prediction peak when re-scoring the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* dataset with BLOSUM62 or PAM160 initially indicated that there was no complementary improvement in specific enzyme prediction performance, when reversing the order of application of the BLAST search and re-scoring substitution matrices. The new observations, shown in *figure 3.9*, indicate that there may be some level of complementary information in the PAM30/PAM160 pair of matrices that was previously being masked by the potential query sequence redundancy of the dataset. This is not as clear as the corresponding performance peaks with lower PAM-N re-score matrices.

Also, when re-scoring the 40% sequence identity clustered *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* MSAs, with a BLOSUM62 matrix, an expected peak is not seen. Furthermore, the sharp decrease in correct predictions when using the PAM170 matrix could be an indication that we are simply seeing an artefact of the 40% sequence clustered subset of the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* dataset. It is also possible to see that the IDENTITY matrix continues to produce the largest number of correct predictions regardless of the threshold of query sequence identity applied to the *"All1stINCORRECT.tF.PAM30.unmasked.E0.001"* dataset. These results are shown in more detail in *table 3.4*.

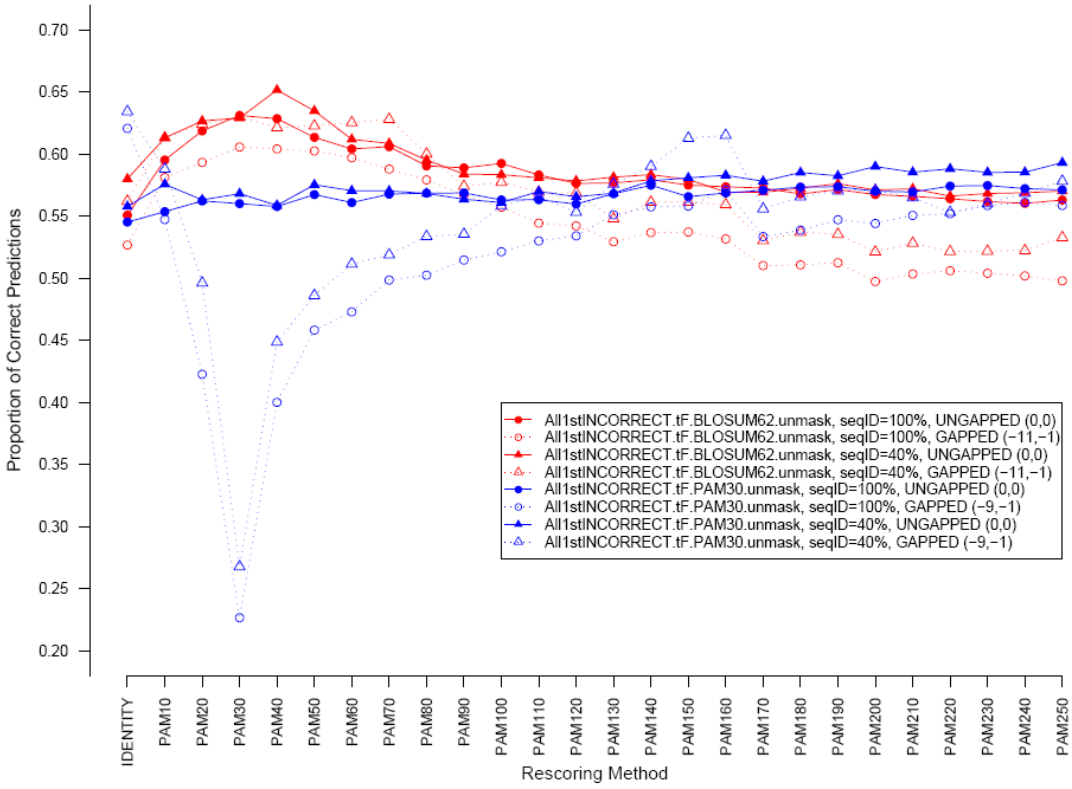

Figure 3.9. A comparison of the proportion of correct predictions obtained for each of the IDENTITY and PAM-N matrix re-scoring methods. The proportions of correct predictions are the bootstrap mean values. Results are shown for the Al1stINCORRECT.tF.BLOSUM62.unmasked.E0.001 and Al1stINCORRECT.tF.PAM30.unmasked.E0.001 datasets, after both 100% and 40% sequence identity thresholds have been applied to the query sequences.

To conclude this analysis, a summary is given in *table 3.4*, which highlights the re-scoring methods that provide the largest number of correct specific enzyme function predictions for each of the datasets and the associated query sequence clustered

subsets (of 100%, 80%, 60% and 40%) investigated. A number of observations can be drawn from these results. Each of the three datasets of MSAs that were investigated, show that similar re-scoring matrices provide the optimal level of specific enzyme function annotation, when applying different sequence identity clustering thresholds to the query sequences. In the case of the results for the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, the optimal results are obtained when using either a PAM30 matrix (with 100% sequence identity clustering) or a PAM40 matrix, with an un-gapped (0,0) gap-scoring method. Similarly, the optimal results for the *All1stINCORRECT.tF.PAM160.unmasked.E0.001* dataset are seen when using either a PAM20 or PAM30 re-scoring matrix, but in this case there is also an additional variation, with the sequence identity clustering threshold, in the gap scoring model that provides these results. Finally, the results for the *All1stINCORRECT.tF.PAM30.unmasked.E0.001* dataset show that the IDENTITY matrix, with a gapped (-9,-1) gap scoring model, is generally the best re-scoring method. Overall, the results from re-scoring the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, consistently show a larger mean proportion of correct enzyme function predictions, with the largest value of 0.652 seen for the subset of MSAs generated when the query sequence cluster threshold is 40% and a PAM40 re-scoring matrix with an ungapped (0,0) scoring model is used.

BLAST search matrix used to generate dataset	**Optimal re-score matrix**	**Gap Penalties (g_{open}, g_{ext})**	**(bootstrap) mean proportion correct +/- se**	**Number Correct (out of)**
Query sequence cluster threshold = 100%				
BLOSUM62	PAM30	(0, 0)	0.631 +/- 0.012	2226 (3527)
PAM160	PAM30	(-11, -1)	0.611 +/- 0.012	1894 (3100)
PAM30	IDENTITY	(-9, -1)	0.621 +/- 0.015	1310 (2110)
Query sequence cluster threshold = 80%				
BLOSUM62	PAM40	(0, 0)	0.632 +/- 0.015	1347 (2131)
PAM160	PAM20	(0, 0)	0.604 +/- 0.016	1103 (1826)
PAM30	IDENTITY	(-9, -1)	0.585 +/- 0.020	697 (1191)
Query sequence cluster threshold = 60%				
BLOSUM62	PAM40	(0, 0)	0.645 +/- 0.018	898 (1392)
PAM160	PAM20	(0, 0)	0.621 +/- 0.020	729 (1174)
PAM30	PAM160	(-9, -1)	0.607 +/- 0.025	465 (766)
Query sequence cluster threshold = 40%				
BLOSUM62	PAM40	(0, 0)	0.652 +/- 0.025	470 (721)
PAM160	PAM20	(-11, -1)	0.648 +/- 0.028	394 (608)
PAM30	IDENTITY	(-9, -1)	0.634 +/- 0.034	256 (403)

Table 3.4. A summary of the re-scoring methods that give the optimal specific enzyme functional predictive performance for each of the MSA datasets and a selected set of associated query sequence clustered subsets. The column - BLAST search matrix used to generate dataset – specifies the amino acid substitution matrices used in the sequence database search to generate the particular dataset of MSAs under consideration. The columns – optimal re-score matrix and gap penalties – show the re-score methods and gap scoring models that give the best predictive performance for the dataset under investigation. Bootstrap values for both the mean proportion, with standard error (se), and number of correct predictions are shown for each identified method.

Overall, the additional results obtained in this section - from clustering the query sequences used to generate the MSAs in the datasets - indicate that the potential sequence redundancy is not distorting the true trends in the alignment re-score prediction results. Some notable exceptions to this have been highlighted, such as the results seen for the *All1stINCORRECT.tF.PAM30.unmasked.E0.001* dataset, when the query sequence identity cluster threshold is 40%, which may be worthy of further study.

3.3.6 Investigation of Potential Correlation Between the Conservation of Enzyme Functional Specificity and the PAM Evolutionary Distance

The aim of this section is to investigate whether there is any correlation between the optimal re-scoring lower PAM-N (such as PAM30) substitution matrices and the conservation of specific enzyme function at the associated PAM evolutionary distances. This was done by using the sequence identities, calculated in the analysis of *chapter 2 – section 2.3.1*, shown in *figure 2.2*, as input to the *PerIdentToPam()* function from the Darwin application. This data was used because it provides a large-scale study of the relationships between pair-wise sequence similarity (and hence PAM evolutionary distance) and enzyme functional class conservation. Therefore providing a logical extension of the function conservation studies presented in *chapter 2*. The outcome of this was *figure 3.10*, which shows the variation of enzyme function conservation, with respect to the PAM evolutionary distance, between pairs of aligned enzyme sequences. Because this thesis is focussed on high functional specificity, the analysis is restricted to the accuracy of conservation at the first three and all four levels of the EC functional classification hierarchy.

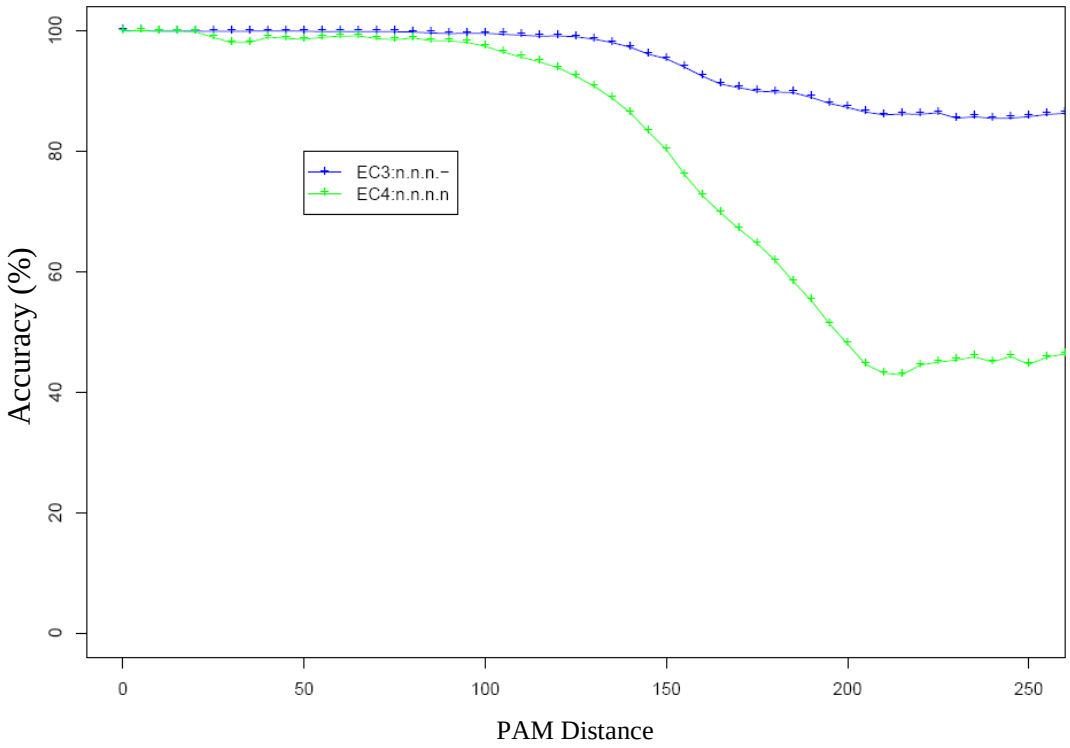

Figure 3.10. Graph showing the functional conservation accuracy, using PAM distances between enzyme sequence pairs from the 1st iteration of the database search results. Where, EC3:n.n.n.- are the results for the first three EC numbers predicted correctly; and EC4:n.n.n.n for all four EC numbers correctly predicted.

When considering the results for the conservation of all 4 levels of EC numbers (EC4: n.n.n.n), a functional conservation accuracy of 95-100% is observed when the PAM distance is less than 100. Between PAM distances of 100 and 200 the accuracy decreases to approximately 40-45%, where it remains for PAM distances greater than 200. These results indicate that there is no clear, unique correlation between the low PAM10-PAM50 evolutionary distances and the accuracy of specific enzyme function conservation, which is what might possibly be expected from the outcome of the PAM matrix re-scoring results. There is however, a clear decrease in accuracy when the PAM distance is 100 and greater. This could be of relevance because this is the PAM distance at which the peaks in function prediction performance begin to become apparent when re-scoring the PAM160 and BLOSUM62 generated MSAs. However, this particular signal is perhaps not strong or convincing enough to provide a reason for the specific function prediction improvements shown for the alignment re-scoring when using the lower N-value (such as PAM10-PAM50), PAM-N matrices.

3.4 Conclusions

In this chapter a number of automated approaches have been presented that investigate the utilisation of alternative amino acid substitution matrices for improving the specific functional classification of enzyme sequences. The aim of this work was mainly two-fold: (1) to assess any improvement in the function prediction accuracy of a PSI-BLAST generated sequence significance ordering, through the use of additional amino acid substitution matrices to functionally re-score the aligned sequences; and (2) to identify any general, significant trends in the analyses that are correlated with the variation of the substitution matrices used.

Three methods for generating datasets of multiple sequence alignments have been investigated. Each dataset was the result of a gapped BLAST sequence database search that used one of either: BLOSUM62; PAM160; or PAM30 as the search amino acid substitution matrix. The constituent MSAs were then modified, to define a series of benchmark datasets, where the enzyme sequence with the most significant sequence similarity to the query protein is classified as functionally incorrect. The purpose of these benchmark sets was to assess the effect of subsequent sequence re-scoring and re-ranking methods on the accuracy of specific enzyme function annotation. An IDENTITY matrix and a wide selection of BLOSUM and PAM amino acid substitution matrices were employed to carry out this analysis. Also investigated were the effects on the functional re-scoring results of: sequence residue masking; gap score penalties; and the generation of MSA subsets using clusters based on the sequence identity of the query sequences.

Initially, the analysis focussed on the sequence alignments obtained from using a BLOSUM62 matrix in the gapped BLAST search – the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset. From these it was shown that the MSAs containing un-masked amino acid residues gave consistently larger proportions of correct function predictions, irrespective of the particular substitution matrix re-scoring method used. Similarly, the "un-gapped" form of the alignment re-scoring algorithm, in which all residues aligned with gaps were scored as zero, consistently outperformed the method that used identical gap penalties to those used in the original BLAST search. Overall, the best performing method for specific functional classification of these MSAs is the one which uses the PAM30

matrix and an "un-gapped" gap scoring model, without sequence residue masking. This resulted in a maximum mean value, for the proportion of correct specific functional classifications, of 0.631 (or 2226/3527 correct classifications). In addition, there is a more general trend towards improved classification results when using PAM-N matrices that have progressively lower N values, culminating in the optimal peak observed with the PAM30 matrix. This trend is seen for both the gapped (-11, -1) and un-gapped (0, 0) alignment re-scoring results.

Further, a control experiment was carried out to assess whether the results were also valid when using the original "non-artificial" dataset of alignments. In this, the PAM30 matrix continued to show an improvement in the number of correct predictions, when compared to the BLOSUM62 matrix. This showed that the PAM30 matrix does not have a detrimental effect when re-scoring alignments that contain examples that are originally "correct" and helps to validate the use of the "All1stINCORRECT" artificial dataset as a benchmark dataset in this thesis.

Following on from these observations, an analysis was conducted to assess whether the above phenomena of improved functional classifications were a unique property of the sequence alignments in the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset. One way in which this was approached was by using the PAM series equivalent of a BLOSUM62 matrix, which is PAM160, when generating the benchmark MSAs. These alignments were then subjected to an identical set of re-scoring analyses, where similar trends were observed. Also, the optimum number of correct specific enzyme function classifications occurred when using the same PAM30 substitution matrix that produced the maximum for the BLOSUM62 based alignments. A difference in the case of the PAM160 alignments was that it was the gapped (-11, -1), rather than un-gapped (0, 0), re-scoring model that gave the maximum proportion of correct classifications, equal to 0.611 (or 1894/3100 correct classifications). However, it was shown that the comparable PAM30 re-score results from the un-gapped model were almost identical and fall within one standard error of deviation of the gapped results. Although the maximum proportion of correct predictions is larger in the BLOSUM62 than the PAM160 generated MSAs, with a small difference between the means, of 0.020, these analyses do indicate that the re-scoring of multiple

sequence alignments with low N value PAM matrices, specifically a PAM30 matrix, results in an increased number of correct specific enzyme sequence classifications, when compared to the other substitution matrices investigated. This suggests that the lower PAM-N matrices do show a general improvement in specific functional classification of enzyme sequences, when either a BLOSUM62 or PAM160 matrix is used to generate the datasets, and the results are not simply due to an artefact of the BLOSUM62 substitution matrix used in the BLAST MSA generation.

To complete this analysis, the same process was again followed, using PAM30 as the substitution matrix in the BLAST sequence database search. The aim of this was to assess whether there would be a comparable improvement, in correct function prediction results, when using a low PAM-N matrix for the initial MSA generation, followed by a BLOSUM62 or PAM160 matrix for alignment re-scoring. The results from this analysis did not show a comparable peak in prediction results when using either the BLOSUM62 or PAM160 matrices to re-score the PAM30 generated MSAs. In-fact, there were no clear peaks of specific function prediction improvement for any of the BLOSUM-N or PAM-N substitution matrices investigated. However, when using the IDENTITY matrix, with the gapped (-9, -1) scoring model, an optimal value of 0.621 for the mean proportion of correct functional predictions was observed. This result is comparable to the two optimal results, described above, from re-scoring the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* and *All1stINCORRECT.tF.PAM160.unmasked.E0.001* datasets.

A further study was then carried out to investigate the effect that any potential sequence redundancy, within the query sequences used to create the benchmark datasets, may have on the accuracy and trends of the alignment re-scoring prediction results. For this, a number of sub-datasets, on which the alignment re-scoring experiments were repeated, were generated using a variety of sequence identity thresholds. The outcome of these additional analyses showed similar results and trends for each of the sequence identity cluster thresholds used. An exception to this was seen when the MSAs were used from the subset of the *All1stINCORRECT.tF.PAM30.unmasked.E0.001* dataset, defined using a 40%

sequence identity threshold, where an increase in specific functional classification accuracy is seen when using a PAM160 re-score matrix.

In general, the results obtained from the alignment re-scoring experiments indicate that the order in which the particular pair of matrices are applied, for MSA generation and subsequent re-scoring, is important for improving the specific enzyme classification. This is shown by the fact that there was mostly no complementary improvement in performance, when reversing the order of application of the BLOSUM62 matrix for BLAST MSA generation and PAM30 for subsequent sequence alignment re-scoring. Although, the exceptions seen for the 40% sequence identity clustered subset of MSAs, indicate that there may be some complementary information in the pair of BLOSUM and PAM matrices used and this phenomenon could be worthy of further study.

A possible explanation for these observations may be found in the intended uses for the particular types of amino acid substitution matrices and therefore the methods used to generate them. The BLOSUM series of matrices are generally used (and found to be optimal) in sequence database searches, such as BLAST, because they tend to generate better quality alignments and provide improved levels of homology detection. This is in contrast to the PAM matrices which are often used to assess the evolutionary origin of sequences and for modelling evolutionary changes across a family of proteins (Mount, 2004). Therefore, the optimal performing PAM matrices may be related to the level of evolutionary distance between the homologous sequences in the specific alignments, thus, providing additional information that improves the specific functional classification of the more closely related sequence homologues. The results from the comparisons between PAM evolutionary distance and the accuracy of specific EC conservation indicate a possible correlation of this type. However, the correlation signal is quite weak and further study would be required before any firm conclusions could be stated regarding these results.

It has also been shown that the results from re-scoring the alignments containing no sequence residue masking are a consistent improvement over those containing the residue masking used in the original database search. A possible reason for this performance improvement could be that the sequence masking, used in the sequence database search, improves the homolog detection, by reducing the false positives

identified from similarities to masked sequence regions of low information content, whereas the subsequent re-scoring of the un-masked locally aligned sequence regions provides additional sequence information that improves the specific functional ordering of the homologous enzyme sequences.

In summary, the results presented in this section highlight some areas of improvement for the accuracy of specific functional assignment, when compared to the sequence similarity based, statistical significance ordering of a BLAST database search. For the BLOSUM62 and PAM160 BLAST generated MSAs, there is a definite trend towards an increase in correct prediction results when using the lower evolutionary distances of the PAM-N matrices, where a maximum is observed for the PAM substitution matrix of 30/40 PAM units. The next chapter aims to improve on these results by implementing a more refined procedure, based on sequence evolution and additional phylogenetic information, for the selection of particular residues to use in the sequence scoring function.

Chapter 4 Identification of Functional Specificity Determining Residues

4.1 Introduction

In the previous chapter, methods based on alternative amino acid substitution matrices, were investigated for re-scoring the functional similarity of aligned homologous enzyme sequences. However, these approaches did not take in to consideration the particular amino acid residues that are most likely to be responsible for the specific functional behaviour of the proteins. In this chapter, the aim is to investigate and benchmark a selection of methods that have been developed to do precisely that and then investigate their use for the improvement of specific enzyme function annotation.

The hypothesis used in these approaches is based on the knowledge that the functional divergence of proteins is determined by selective pressures during molecular evolution. In general, new functions arise in paralogous proteins through the fixation, via natural selection, of a number of key amino acid mutations that are functionally beneficial (Ohno, 1970; Taylor and Raes, 2004; Conant and Wolfe, 2008). This is a particularly important means for the diversification of the substrate binding specificity and the biochemical mechanisms of enzymes. Closely related enzyme sequences, such as those used in this study, are therefore well suited to the identification of amino acid changes that highlight functional differences. This is especially true when considering the (often small number of) mutations responsible for thermodynamically favourable binding of a particular substrate instead of other substrates that are chemically similar.

Considering these observations, regarding the mechanism for the evolution of specific protein functions, it would appear important to develop computational methods to identify these particular residues. An additional driving force is the fact that it is time-consuming and economically expensive to identify each of these residues through experimental methods (Saghatelian and Cravatt, 2005). Most computational approaches to this problem are based on comparisons between multiple sequence alignments (MSAs) containing groups of functionally identical or

similar sequences. Due to the fact that divergent evolution is believed to be much more common than convergent evolution of function (Patthy, 1999) these sequences are generally obtained through homology recognition techniques.

In this chapter, I have implemented and investigated the performance of two methods for the identification of residues that determine functional specificity. Both of these have been previously published and take quite different approaches to solving the problem. The methods chosen and discussed below are: (i) the *"func-MB"* method (Pazos et al., 2006); and (ii) the *"profile-HMM"* based method (Hannenhalli and Russell, 2000). In an earlier study (del sol Mesa et al., 2003), three methods for identifying functionally determining residues were compared. A benchmark was devised that used the distances from predicted residues to bound ligands and hetero atoms to assess the accuracy. It was concluded that there was little difference between the performances of the three methods and furthermore, suggested that a combined approach would be expected to be optimal. It was therefore decided to investigate a modified form of the MB method used in that study, which was later described by Pazos et al. (2006). A non-parametric rank correlation coefficient is used in this method to assess the correlation between specific function and amino acid similarities. This method was chosen over the others because, it was relatively simple to fully automate - which is in contrast to the *SequenceSpace* based method - and also because it contained an implicit representation of the sequence phylogeny.

The *profile-HMM* method was chosen primarily because it has been used previously, with some success, for the identification of residues determining specific function and the subsequent prediction of function using a subset of these residues. This method uses the probability of observing certain residues, within specific functional groupings, to identify the residues most likely to be responsible for the definition of specific functions, meaning that this approach is quite different to the non-parametric rank-order correlation based MB methods.

The main aims for the work in this chapter were primarily three-fold:

1. The implementation and investigation of methods for identifying fSDRs in groups of functionally related enzyme sequences;

2. The provision of a benchmark that compares the ability of selected fSDR subsets, from each of these methods, to improve the functional clustering and specific function prediction accuracy for enzyme sequences; and

3. Demonstrate the performance of the methods when applied to a well-studied example of enzymes that have differing substrate specificities.

These studies have also been designed to enable the definition of a "gold-standard" subset of computationally identified fSDRs. That is, they provide an optimal predictive performance, with regards to their use in the assignment of correct specific enzyme function to the query sequence. This dataset of fSDRs are then used in the experiments of *chapter 5*, to investigate the feasibility of using machine learning techniques to identify fSDRs in MSAs, without prior knowledge of the functional sub-types of the constituent sequences of the alignments.

4.2 Methods

4.2.1 Datasets

As in the previous studies, presented in *chapter 3*, the datasets used for the following experimental analysis consist of multiple alignments of enzyme sequences. Two datasets of MSAs were used for the studies contained within this chapter. To assess the performance of the fSDR based sub-alignment re-scoring methods, the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, which consists of 3527 BLAST generated MSAs, was used in the following analysis. Additionally, a single MSA from the *"initial"* dataset is used to provide a detailed investigation of a specific example, which contains aligned sequences from the lactate and malate dehydrogenase classes of enzymes. The methods used to generate both of these datasets are the same as those used previously in this thesis and are defined in detail in *chapter 2*.

4.2.2 The Functional Mutational Behaviour *"func-MB"* Method

The idea behind this method was originally inspired by the mutational behaviour (MB)-method described by del sol Mesa et al. (2003). In this method, a rank correlation coefficient is used to identify positions, within a multiple protein sequence alignment, that show correlation with the mutational behaviour of the

whole group of homologous sequences. The hypothesis being that a larger correlation coefficient indicates aligned positions that most closely resemble the mutational pattern of the sequence family, hence identifying the positions most important to the specific phylogenetic relationships between the sequences. An extension of this method, which was recently studied by Pazos et al. (2006), investigates the correlation between the specific functional class and the individual residues in an aligned column. The method is referred to as the *Xdet* method in the original paper by Pazos et al. (2006), but will be referred to as the *"func-MB"* method in this thesis, so as to maintain a similar naming scheme with the previously published *"MB-method"*, which has been discussed in other parts of this thesis. The aim of the *func-MB* method is to identify the residues that have a mutational behaviour closely correlated to variations in specific functional properties. The implementation details, which differ slightly to those described by Pazos et al. (2006), are described below.

4.2.2.1 Implementation of the *func-MB* Method

For each pair of sequences in the MSA, a matrix of values, *S*, was constructed to represent the specific "functional similarity" (or distance) between them. Calculation of the functional similarities was done by looking at the number of EC code description levels each of the compared enzyme sequences had in common. For example, if all 4 EC numbers were conserved between a pair of sequences then the associated matrix value would be '4', conservation of the first 3 numbers yields a value of '3', with values of '2' and '1' being used for conservation of 2 and 1 EC number respectively. Finally, a value of '0' was used when the 1st EC number was not common between the sequences. A matrix of these values was calculated once for a particular MSA.

Then, for each of the aligned columns, a corresponding "amino acid similarity matrix", *A*, is calculated, with the same number of elements as the functional similarity matrix, to measure the similarity between each of the residue pairs. An amino acid substitution matrix is used as a measure of "mutational similarity" between each of the amino acid pairs in the columns. Both the BLOSUM62 and PAM30 substitution matrices were used in the work presented here, but any other measure of similarity can be easily integrated into this method.

To calculate the correlation between the functional and residue similarities, the Spearman-rank order correlation coefficient, c_i, (Press et al., 1992) was calculated for each of the aligned columns, *i*, in the MSA, using the following equation:

$$c_i = \frac{\sum_{x,y}(A_{xyi}-\bar{A})(S_{xy}-\bar{S})}{\sqrt{\sum_{x,y}(A_{xyi}-\bar{A})^2}\sqrt{\sum_{x,y}(S_{xy}-\bar{S})^2}} \qquad \textit{(equation 4.1)}$$

where the rank order of amino acid similarity in sequence *x* and sequence *y*, at position *i*, is represented by A_{xyi} ; the rank order of functional similarity in sequence *x* and *y* is represented by S_{xy} ; and the average rank position of these amino acid and functional matrices is given by $\bar{A}$ and $\bar{S}$ respectively.

4.2.3 The Functional *Profile-HMM* Based Method

An alternative method for identifying functionally specific residues has been proposed by Hannenhalli and Russell (2000). The basis of this method is the identification of amino acids that are more likely to be conserved within groups of sequences with the same function, but differ between them. Starting from an alignment of sequences, containing proteins of different molecular functions, a set of alignments are created, each containing only sequences with a single specific function. A hidden Markov model (HMM) profile was created for each of these functional sub-alignments using the *hmmbuild* application provided with the HMMER application (version 2.3.2 – http://hmmer.wustl.edu). The default parameters were used in the creation of all profiles.

The profiles output by *hmmbuild* are in log-odds form. Because the aim of this method is to calculate the probability of a particular type of amino acid occurring in one profile, compared to all others, these scores were converted into probabilities. For each aligned column, *i* (with a match state in the profile HMM), the probability of occurrence of amino acid, *x*, in specific function *s*, was calculated, $P^s_{i,x}$. From the resulting probability profiles, the relative residue conservation between profiles was calculated using the relative entropy (Durbin et al., 1998) of each alignment position, defined as follows:

$$RE_i^s = \sum_x P_{i,x}^s \log \frac{P_{i,x}^s}{P_{i,x}^{\hat{s}}} \qquad \textit{(equation 4.2)}$$

where, the relative entropy for a specific function *s*, at position *i* is defined as RE_i^s and is calculated from the summation of the contribution from all residue types *x* at this position. The union of all specific functional types, except for *s*, is denoted by ŝ, with the probability of occurrence of amino acid *x* at position *i* in this combined alignment, represented by $P_{i,x}^{\hat{s}}$. The relative entropy of an alignment position can be thought of as a measure of the degree of conservation at that position, in a specific function *s*, when compared to all other functions, ŝ.

Two further calculations were required to assess the importance of each alignment position. The first determines the cumulative relative entropy, *CRE*, at each alignment position, *i*:

$$CRE_i = \sum_s RE_i^s \qquad \textit{(equation 4.3)}$$

which aims to assess the discriminatory role of alignment position *i*, when summing the relative entropy contributions over all the specific functional types. Finally, a Z-score is used to assess the overall significance of the cumulative relative entropies, when considered in context to all the aligned positions in the MSA.

$$Z_i = \frac{CRE_i - \mu}{\sigma} \qquad \textit{(equation 4.4)}$$

Where, μ and σ are the mean and standard deviation of the CRE for all positions in the multiple sequence alignment. A larger Z-score indicates greater significance for that aligned column and therefore indicates it is more likely to be a determinant of specific function.

4.2.4 The Sub-Alignment Re-scoring Procedure

This section provides a description of the methods used to select sub-sets of the aligned columns, from each of the MSAs in a particular dataset, that are predicted to determine specific enzyme function (fSDRs). First, however, a description is provided of the procedure that is subsequently used to functionally re-score the

aligned sequences through the incorporation of those selected sub-sets of aligned columns.

4.2.4.1 The fSDR-based Sub-alignment Functional Re-scoring Procedure

The previously described method, of *section 3.2.2.1* (see also the method flowchart in *figure 3.1*), is built upon here to propose an alternative method for sequence re-scoring and re-ranking to determine the functional similarity between a query sequence and related, aligned enzyme sequences. This method is based on the re-scoring of sub-alignments of amino acid residues that have been extracted from the full MSAs in the input datasets. An overview of this procedure is shown below in *figure 4.1*. This shows a simplified overview of the proposed method for the identification of fSDRs and their subsequent use in generating a functionally more informative sub-alignment of amino acids for use in improving classification accuracy.

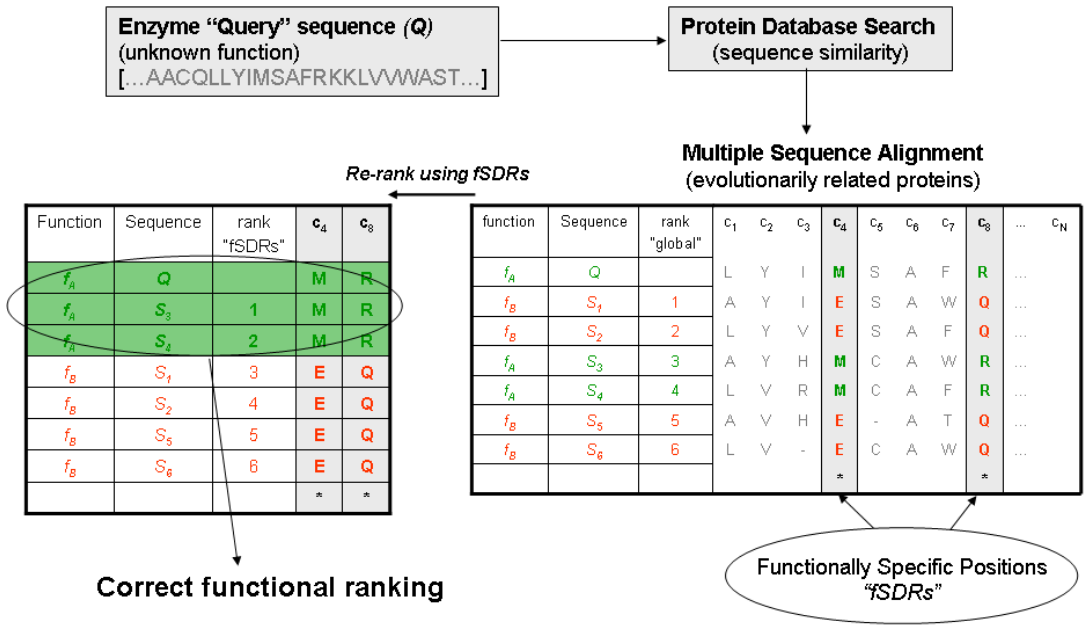

Figure 4.1. Simplified overview diagram of the proposed fSDR-based sub-alignment generation, extraction and functional re-scoring procedure.

In this method an MSA of evolutionarily related sequences is obtained from a sequence database search, the columns (c_4 and c_8 in the example shown in *figure 4.1*) containing potential fSDRs are identified and then extracted to generate a "sub-alignment" of sequences. These consist of the same number of aligned sequences as

the full MSA, but a smaller, selected subset, of the aligned columns. Each of the individual pair-wise sub-set of alignments, between the query and aligned enzyme sequences, are then evaluated using the scores from the amino acid substitution matrix used for the sequence re-scoring. For these studies the BLOSUM62 and PAM30 substitution matrices were used to score the pair-wise residue similarities, however, only the results from using the PAM30 matrix are shown in the following analyses. After the re-scoring of the sub-alignments has been completed the sequences are then re-ordered and their specific functional similarity to the query sequence is assessed. The simplified example, shown in *figure 4.1*, highlights the key concepts behind this approach. It shows a hypothetical situation in which the original functional sequence ordering from the database search generates a ("rank global") sequence ordering where the top-ranked sequence (s_1) has a different function (f_B) to that of the query sequence (Q), which has function (f_A), and therefore results in an incorrect functional classification. However, once the sequences have been re-ranked, using the identified fSDRs ("rank fSDRs"), the top-ranked sequence now shows the same specific function as the query and therefore results in an improved and correct functional classification of the query sequence.

4.2.4.2 Methods for Selecting Aligned Subsets of fSDRs

Three methods were used to select subsets of aligned residue columns from each of the MSAs in the dataset, for use in the subsequent fSDR-based sub-alignment re-scoring procedure: (i) the selection of aligned columns using a cut-off threshold, obtained from the column score – the "column score threshold" method; (ii) the selection of N aligned columns, using the N highest ranking column scores – the "top-N" method; and (iii) the selection of aligned columns using the top X percentage of the highest ranking column scores – the "top-X percent" method. In all three, the column scores are the values obtained from either the Spearman-rank order correlation coefficient or the Z-score, depending on whether the fSDR identification method used was the *func-MB* or *profile-HMM*, respectively.

4.2.5 The Treatment of Gaps in the Sequence Alignments

There are three stages in the sub-alignment re-scoring procedure where the methods used for scoring gaps in the sequence alignments must be considered. Each stage is

defined separately below and where necessary the particular fSDR column identification method of relevance is indicated.

4.2.5.1 The Aligned Column Gap Percentage Threshold of Inclusion

A method for the pre-filtering of aligned columns from the MSAs was used, based on the percentage of gap residues that are contained within a particular column of all types of aligned residues. This method removes all aligned columns from the MSA, prior to the application of the fSDR identification methods, which contain more than a defined percentage of gaps. This is referred to as the *"column gap percentage threshold (colgap_percent)"* and where relevant the specific thresholds used are stated alongside the discussion of the results.

4.2.5.2 Gap Score Penalty Used for Calculating the Amino Acid Similarity Matrix in the *func-MB* Method

When defining the amino acid similarity matrix, *A (see equation 4.1)*, required for calculating the aligned column correlation coefficients for the *func-MB* method, it is necessary to consider aligned residue pairs that may contain gaps. For the following analysis, the method of Pazos et al. (2006) was used, where a gap score of 0 was used for scoring all of the aligned amino acid pairs that contain gaps.

4.2.5.3 Gap Score Penalty Used for the Sub-Alignment Re-scoring

In the following analysis, a single gap penalty of 0 is used for all aligned residue pairs that contains gaps when re-scoring the fSDR-based sub-alignments of sequences.

4.2.6 Methods for Assessing the Accuracy of fSDR-Based Prediction of Specific Enzyme Function

4.2.6.1 Top-hit Method

The "top-hit" assessment method was again used to assess the functional classification accuracy resulting from the functional re-scoring of the enzyme sequences, when using fSDR-based sequence sub-alignments. It is conceptually the same method as that used previously in *chapter 3* (*section 3.2.3.1*). This classifies a prediction as correct if the specific enzyme functional class of the query sequence is

the same as that of the sequence with the highest score, after the sub-alignment based functional sequence re-ranking.

4.2.6.2 Calculation of the Proportion of "Correct" Specific Enzyme Predictions

The same method as that in *chapter 3* was used for calculating the proportion of "correct" specific enzyme predictions (or classifications) obtained from the "top-hit" assessment method. This is defined as $correct_{proportion}$ in *equation 4.5*, where: $n_{correct}$ is the number of "correct" predictions observed from the "top-hit" assessment method and N is the number of MSA examples in the dataset that were used in the analysis.

$$correct_{proportion} = \frac{n_{correct}}{N} \qquad \textit{(equation 4.5)}$$

4.2.6.3 Bootstrap Re-sampling Analysis of Top-hit Results

The same bootstrap statistical re-sampling method (Efron and Gong, 1983), which was previously described in *chapter 3 - section 3.2.3.3*, is again used in this chapter to analyse the statistical significance of the functional classification results obtained from the fSDR-based sub-alignment sequence re-scoring.

4.2.6.4 Definition of a Random Sequence Selection Model for Specific Enzyme Function Assignment

A random sequence selection model was again used to provide a baseline comparison with the "top-hit" function prediction results obtained from the fSDR-based sub-alignment re-scoring result. This was identical to the method described in *chapter 3* (*section 3.2.3.2*), which is based upon the concept of randomly permuting the ranked results of the sequence homologues in each of the MSAs in the dataset. The functional classification result was then determined to be correct or incorrect through functional comparison between the specific EC classification of the query sequence and the randomly permuted "top-hit". For a detailed definition of this selection procedure please refer back to *section 3.2.3.2*, in *chapter 3*.

4.2.6.5 Random Selection of Subsets of Aligned Residue Columns

An additional random model – the *"random column selection method"* - was also used for the assessment of the fSDR-based sub-alignment functional classification results in this chapter. This method implements a randomised procedure for aligned column selection, using a similar logic to the identification of the fSDR-based subsets of aligned residue columns described above. However, for this random selection no regard was given to the actual likelihood of the columns being associated with specific enzyme functional properties (i.e. they are not necessarily high scoring fSDRs). So, unlike with the *profile-HMM* and *func-MB* methods, the columns were (randomly) selected without first ranking them based on the calculated fSDR significance scores. Therefore, from each one of the "complete" MSAs, a subset of *n* aligned amino acid columns was randomly selected (using a uniform distribution to randomly select aligned columns from the MSA), without replacement. The number, *n*, of aligned columns was selected in the same way as for the fSDR-based "top-N" and "top-X percent" column selection methods described above, in *section 4.2.4.2*. Leading to a randomly selected ("random-N" or "random-X percent") sub-alignment of sequences, containing *n* aligned columns of amino acids. This type of model does not naturally lend itself to producing a randomly selected subset of aligned columns that are directly comparable to the "column score threshold" method of sub-alignment generation and therefore one is not provided.

4.2.7 Query Sequence Clustering

An identical procedure to that used in *chapter 3* (*section 3.2.5*) was followed to analyse the effects of query sequence identity clustering on the enzyme functional classification accuracies. The clustering was again done through the use of the CD-HIT algorithm (Li and Godzik, 2006), on the 3527 query sequences that were used to generate the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset of MSAs. A range of percentage sequence identity levels were used for the clustering (from 40% to 90% in intervals of 10%) and the recommended default parameters, for the CD-HIT application, were used for each sequence identity threshold levels. Again the longest sequence was used as the representative from each cluster. A summary of the cluster properties, at each defined level of sequence identity for the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, is provided in *table 3.3*.

4.3 Results and Discussion

4.3.1 Benchmark of Functional Re-scoring Prediction Results Using the fSDR-based Sub-Alignments

This section provides a large-scale investigation into how effective the *func-MB* and *profile-HMM* fSDR identification methods are for improving the classification accuracy of the specific function of enzyme sequences. This builds on the results from the previous analyses, presented in *chapter 3*, which investigated the effects of using all of the aligned sequence information, and alternative amino acid substitution matrices, to functionally re-score the aligned enzyme sequences.

The 3527 MSAs from the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset were used as the benchmark dataset in all of the following analyses. This particular dataset was chosen for two reasons. Firstly, this dataset was one of those used previously in the alternative amino acid re-scoring experiments, discussed in *chapter 3*, allowing a direct comparison between those results and the ones obtained in the following fSDR-based functional re-scoring experiments. Secondly, this dataset was selected over the others investigated in *chapter 3* because it was shown to give the largest overall improvement in specific enzyme function classification accuracy, when using a PAM30 amino acid substitution matrix to re-score the aligned sequences.

For this analysis, both the *func-MB* and *profile-HMM* methods for fSDR aligned column identification were applied to each of the MSAs in the dataset. Selected subsets of these columns were then used to re-score the similarity of the aligned sequences to the query, allowing assessment of the accuracy of this approach for specific enzyme classification. Comparisons were then made between the classification results from using these fSDR-based subsets, with those previously obtained from re-scoring all of the aligned sequence residues with alternative amino acid substitution matrices.

The selection of the particular columns to include in the subsets of aligned residues was controlled by a number of alternative methods. For both the *func-MB* and *profile-HMM* methods, three approaches were used to select the columns for inclusion - based on the significance based ordering of the Spearman-rank order

correlation coefficients and Z-scores, from each of the fSDR identification methods respectively. Each of the selection methods aim to identify slightly different subsets of aligned columns and therefore investigate the best method and associated parameters for improving the enzyme classification accuracy.

One method used was the "top-N" method, which selects a set of aligned columns of fixed size, *N*, based on the ranking of the aligned column scores from the fSDR identification methods. A number of values of *N* were used for each of the MSAs in the dataset and the overall effect on the specific enzyme classification performance was assessed for each. A similar method – the "top-X percent" method - was used to select a subset of aligned columns based on a percentage, X, of all aligned columns in each of the MSAs. Therefore, this method, unlike the "top-N" method, will not generally select the same number of aligned columns for each of the applied subset X percentage selection thresholds. Finally, a method was used that applies a threshold based on the calculated value of the aligned column correlation coefficients, or Z-scores, from the associated fSDR identification methods, to generate the sub-alignments. Again, as in the "top-X percent" method, this "score threshold" selection criteria may generate different numbers of columns in each sub-alignment obtained from the MSAs in the dataset.

The assessment method used for the correct classification of specific enzyme function, when using selected sub-sets of fSDR columns, was the same "top-hit" sequence re-scoring method that was used in *chapter 3*. For both of the fSDR identification methods the bootstrap form of the results were analysed, which allows robust calculation of the mean proportion of correct functional classifications, and the associated standard errors, for each of the functionally re-scored subsets of fSDR sub-alignments.

4.3.1.1 The *func-MB* Method

When using the *func-MB* method to identify potential functional specificity determining residues it was expected that the way in which gaps are treated in the multiple sequence alignments could make an important contribution to the particular columns identified. There are three stages in the *func-MB* based analysis procedure where the gap handling has been considered:

- The selection of which aligned columns should be included when calculating the fSDR significance score – "*the column gap percentage threshold of inclusion"*;

- The way in which gaps were scored during the calculation of the fSDR correlation scores for the *func-MB* method; and

- The way in which gaps were scored during the re-scoring of the enzyme sequences in the fSDR-based sub-alignments.

A number of gap percentage thresholds were used; ranging from no filtering (*colgap_percent* = 100%) to the removal of all columns containing any gaps (*colgap_percent* = 0%), in intervals of 10%. This provides a pre-filtering step for each of the input MSAs.

When constructing the residue correlation matrices for the aligned columns a score of 0 was used for the similarity between any amino acid residues aligned with gaps. This was selected because it was the value used in the study by Pazos et al. (2006).

In the case of the third point, for the following studies it was decided to use a score of 0 for all of the pair-wise sequence re-scoring comparisons between any of the amino acid types and alignment gaps. This value was chosen for the sequence alignment re-scoring stage of the analysis because of the reasons provided earlier in the methods section of this chapter.

The "Top-N" Method for fSDR-based Sub-Alignment Generation

For the "top-N" method of fSDR selection and sub-alignment generation a series of thresholds for the value of N were used. The effects (on the proportion of correct classifications of enzyme function) of gradually increasing the number of aligned columns, selected from each MSA for inclusion in the resulting sequence sub-alignments, are shown in *figure 4.2*. That is, the horizontal axis represents the number of columns, N, of aligned residues that were included in the sub-alignments for functional re-scoring. These were selected through the use of an ordered ranking of the Spearman-rank order correlation coefficients calculated by the *func-MB* method, from which the fSDRs with the highest N ("top-N") correlation coefficients were used to generate the sub-alignments of N aligned columns. These results also show the effects of varying the aligned "column gap percentage threshold of

inclusion", which serves the purpose of removing aligned columns with particular proportions of alignment gaps prior to the functional re-scoring analysis. All of the results show the bootstrap values of the mean proportion of correct functional assignment and the bootstrap calculation of the standard error deviation from the means. To maintain a consistent comparison with the previous bootstrap analyses carried out in this thesis, the parameters for the number of bootstrap repetitions, *B*, was 10000 and the bootstrap sample size of each replicate was 1764 - approximately half the number of MSA examples, 3527, in the dataset. Also highlighted in *figure 4.2* are the bootstrap statistics for the random sequence selection model associated with the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset that is being analysed. This is the same random sequence selection model that was used and defined in *chapter 3* (see *section 3.3.2*).

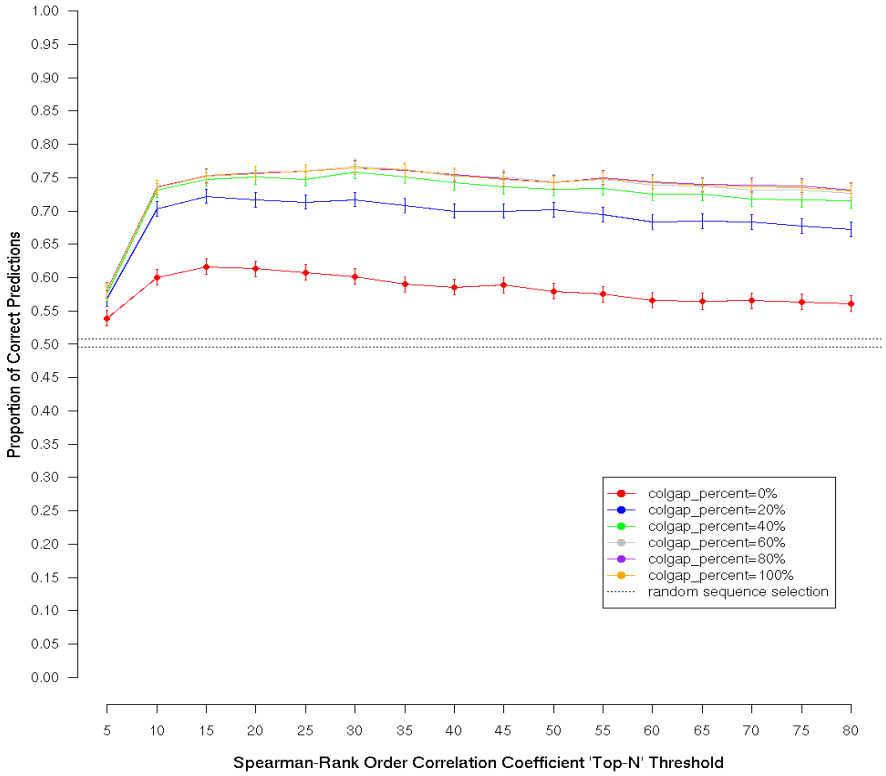

Figure 4.2. A comparison showing the proportion of correct functional predictions obtained as the "top-N" threshold, used to select the subsets of fSDRs used in the functional re-scoring, was varied. The horizontal axis – "Spearman-Rank Order Correlation Coefficient 'top-N' Threshold" – represents the number of aligned columns, with the N highest scoring Spearman-rank order correlation coefficients, that were included in the sequence sub-alignments. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. Enzyme classification results are shown for re-scoring the colgap_percent=0%, colgap_percent=20%, colgap_percent=40%, colgap_percent=60%, colgap_percent=80% and colgap_percent=100% filtered variations of sequence sub-alignments. Also shown is the associated random sequence selection model for the dataset, where the dotted lines show 1 standard error deviation from the mean.

From these results, shown in *figure 4.2*, it can be seen that when using a small sub-set of aligned columns (for example, when N=5) a minimum is observed in the proportion of correct predictions. As the number of columns included in the re-scored sub-alignments is increased, the number of correct enzyme classifications also increases until a maximum is reached, after which point the classification accuracy gradually decreases while the number of included columns in the alignment subset continues to be increased. The actual value of *N* at which the maximum proportion of correct enzyme classifications is obtained is dependent on the value of the *"colgap_percent"* threshold of inclusion. *Figure 4.2* shows that the trends,

with respect to the correlation between sub-alignment size N and the resulting proportion of correct predictions, are very similar for each of the different "column gap percentage threshold of inclusion" thresholds used for the MSA filtering. A further, more detailed, analysis of these results is provided below and summarised in *table 4.1*.

Before this, a number of more general observations related to the results from the top-N sub-alignment re-scoring results, shown in *figure 4.2*, can be explored. It can be seen that the proportions of correct functional predictions, when using the most stringent threshold for pre-filtering aligned columns from the MSAs that contain gaps (*colgap_percent* = 0%), are considerably less than those when using a higher threshold (such as *colgap_percent* = 20% and greater). An explanation for this difference can be provided through a more detailed analysis of the underlying data that was used to calculate the bootstrapped proportions of correct and incorrect enzyme classifications after sequence re-rescoring.

When re-scoring and subsequently re-ranking the sequences contained within each of the sub-alignments, there are a number of possible outcomes when considering a "top-hit" approach to assessing the accuracy of the resulting specific functional classification. These outcomes can be categorised into 2 general states (either: (i) a "correct"; or (ii) an "incorrect" functional classification) but on closer inspection they can also be considered to possess six distinct properties: (i) "top-rank (correct)" – where the top ranked sequence has the same ("correct") specific enzyme class as the query and has a unique score when compared to the other re-scored sequences; (ii) "top-rank (incorrect)" - where the top ranked sequence has a different ("incorrect") specific enzyme class to the query; (iii) "tied-rank same-function (correct)" – where the top ranked sequence shares the same "tied" score (and therefore rank) with one or more other sequences, which all have the same ("correct") enzyme functional class as the query; (iv) "tied-rank different-function (correct and incorrect)" => "undecidable (incorrect)" – where the top ranked sequence has the same "tied" score (and therefore rank) with one or more other sequences, which have both the same ("correct") and different ("incorrect") enzyme functional classes as the query. This in essence means that the sequence re-scoring result is "undecidable" when using the available information and therefore must be

classed overall as an "incorrect" classification result; (v) "tied-rank different-function (incorrect)" – where the top ranked sequence has the same "tied" score (and therefore rank) with one or more other sequences, which all have a different ("incorrect") functional class when compared with the query; and (vi) "empty subset (incorrect)" – where the criteria for fSDR-based aligned column selection does not select any columns for inclusion in the sequence sub-alignment. This therefore means that no sequence re-scoring can be carried out due to the fact that the sub-alignment is "empty" and the classification result is by definition "incorrect".

If these six more detailed classification outcomes are analysed for the top-N results presented in *figure 4.2*, it becomes possible to get an understanding of the reasons for the comparatively poor performance of the *colgap_percent* = 0% classification results. The variation in these properties with each value of *N* used to generate the "top-N" sub-alignments is shown in *figure 4.3*. This clearly shows that the number of "empty subset (incorrect)" examples increases as the "*column gap percentage threshold (colgap_percent*)" parameter is made more stringent (i.e. decreased). This is especially prominent for the results shown when using the most restrictive gap inclusion threshold of *colgap_percent* = 0%. On reflection this is perhaps not a particularly surprising observation, because such a stringent threshold does not allow for a single gap to be present in the aligned columns selected for the sequence sub-alignments. It therefore follows that there will be increasing numbers of MSAs in the analysed dataset that do not contain any aligned columns that satisfy the gap percentage pre-filtering criteria, culminating in the extreme case of no gaps allowed in any of the selected columns. This hypothesis is borne out by the results in *figure 4.3(f)* where a *colgap_percent* threshold of 0% results in 14% (494 out of 3527) of the generated sub-alignments being "empty". In contrast, as the *colgap_percent* threshold is increased to 10% then only 5% (176 out of 3527) of MSAs generate "empty" sub-alignments, and further, once the *colgap_percent* threshold is at 50% and above, hardly any (i.e. approximately 0%) "empty" sub-alignments are being generated.

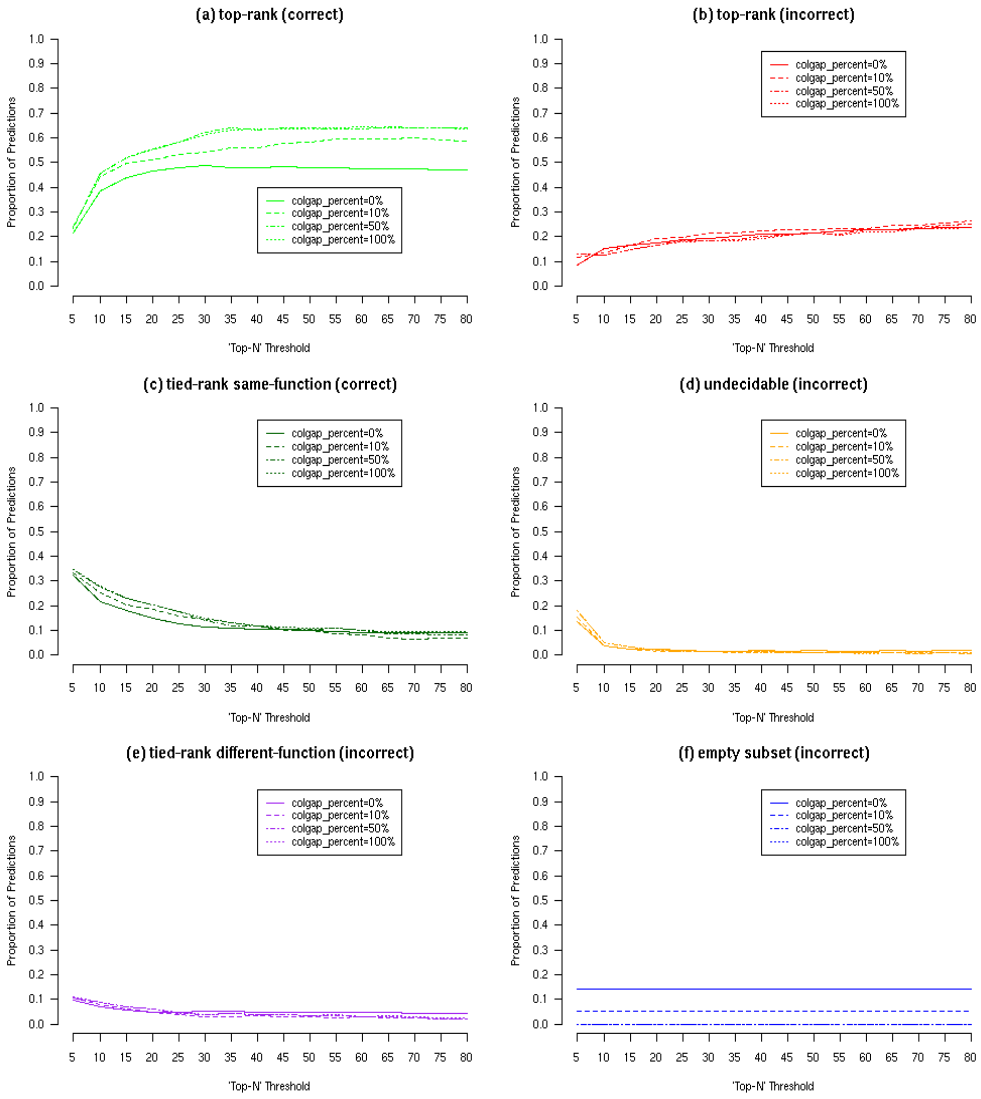

Figure 4.3. A series of graphs showing the variation of the proportions of observed predictions with the specified top-N sub-alignment threshold, for the six distinct prediction outcomes (a) shows the "top-rank (correct)" results; (b) shows the "top-rank (incorrect)" results; (c) shows the "tied-rank same-function (correct)" results; (d) shows the "undecidable (incorrect)"results; (e) shows the "tied-rank different-function (incorrect)" results; and (vi) shows the "empty subset (incorrect)" results. For each of these graphs the results for re-scoring the colgap_percent=0%, colgap_percent=10%, colgap_percent=50% and colgap_percent=100% pre-filtered sequence sub-alignments are shown.

It is possible that this phenomenon could be due to a number of factors, such as the level of evolutionary diversity included within the sequence alignments or potentially misaligned sequences - leading to the incorrect placement of gaps. These possible contributing factors are not explored any further here, but they may be features worthy of further study when considering the selection of particular columns for inclusion in sequence sub-alignments.

A further point to make (with regards to the lower proportions of "correct" enzyme classifications that are observed when the *colgap_percent* threshold is decreased) relates to the method of calculation used for the proportions of correct predictions. The presence of the "empty" sub-alignments (described above) suggests an alternative method for calculating these proportions, using a modified value for *N* in *equation 4.5*. Where, instead of simply using the dataset size, a more refined ("re-normalised") form of calculation could use the number of dataset examples minus the number of "empty" sub-alignment examples for which it is not possible to calculate a re-scored classification result. This modified form of *equation 4.5* is presented in *equation 4.7*

$$correct_{proportion} = \frac{n_{correct}}{(N - n_{empty\_subset})} \qquad \textit{(equation 4.7)}$$

where: $n_{correct}$ and N are the same as in *equation 4.5* and n_{empty_subset} is the number of MSA examples that generate "empty subset (incorrect)" results.

The corresponding proportion of correct classifications obtained from using the method in *equation 4.7* are shown (in parenthesis) in *table 4.1*, alongside those calculated through the use of *equation 4.5*. It can be seen that for these re-normalised results the proportion of correct classifications increases for all of those sub-alignments that have had a more stringent *colgap_percent* threshold applied (i.e. *colgap_percent* <= 40%), due to the presence of a certain number of "empty subset" examples. It should, however, be noted that, by definition, the actual number of correct classifications, at each top-N threshold, was unchanged.

These results show that the difference between the optimal classification accuracies for the sub-alignments, which have been pre-filtered with a more stringent gap filter,

is greatly reduced when applying this alternative accuracy assessment method. In particular, for the *colgap_percent=0%* sub-alignments, the difference between the optimal correct proportions and those of the overall optimal performance (i.e. where *colgap_percent=60%*) is reduced from 0.150 to 0.049. Likewise, for the *colgap_percent=10%* results, the difference is reduced from 0.070 to 0.031. This is still a statistically significant difference, due to the standard error deviation of 0.011 (see *table 4.6*), but it does highlight a potentially informative alternative method for comparing the results of the classifications.

With the aid of the results shown in *figure 4.3*, it is now possible to explore the reasons for the slightly counter-intuitive observations, seen in *figure 4.2*, which show a clear minimum in the proportion of correct enzyme classifications when using the smallest subset (N=5) of aligned columns. This was surprising because it was expected that the subsets consisting of aligned columns, with the strongest correlations between amino acid type and enzyme function, would show the most accurate separation of the specific functional classes in the MSA and therefore the largest proportion of "correct" "top-hit" functional classifications. This was, however, not the case, mostly due to the larger proportion of examples with an "undecidable (incorrect)" result, when using the top-5 rather than the top-10 ranked column correlation coefficients. Where, for all of the *colgap_percent* thresholds investigated there was a sharp reduction in "undecidable (incorrect)" examples and a corresponding increase in "top-rank (correct)" examples when re-scoring sub-alignments generated from the top-5 and top-10 ranked correlation coefficients, respectively.

colgap_percent (%)	**(optimal) "top-N" (N)**	**(bootstrap) mean proportion of correct predictions**	**(bootstrap) mean number of correct predictions**
0	15	0.617 *(0.718)*	2176
10	15	0.697 *(0.736)*	2458
20	15	0.722 *(0.747)*	2546
30	20	0.746 *(0.751)*	2631
40	30	0.759 *(0.764)*	2677
50	30	0.764 *(0.764)*	2695
60	30	0.767 *(0.767)* **(*)**	2705
70	30	0.764 *(0.764)*	2695
80	30	0.765 *(0.765)*	2698
90	30	0.765 *(0.765)*	2698
100	30	0.765 *(0.765)*	2698

Table 4.1. A comparison between the optimal bootstrap results (mean proportion and number of correct "top-hit" specific enzyme predictions) and the top-N subset size that generates them, for each of the colgap_percent thresholds applied. All results for the number of correct predictions are out of a possible dataset size of 3527. (*) indicates the overall maximum predictive performance. The values in parenthesis are the corresponding "re-normalised" proportions (see text) of correct classifications calculated with equation 4.7.

The results, shown in *table 4.1*, provide a summary of the optimal functional re-scoring results for each of the *"colgap_percent"* alignment pre-filter thresholds, along with the number of high scoring aligned columns (fSDRs), *N*, which contribute to the re-scored sequence sub-alignments without re-normalisation. Both the mean proportion and number of correctly classified enzyme functions are shown for comparison, where all results refer to the bootstrap form of the "top-hit" assessed prediction results. It can be seen from the results in *table 4.1* that, overall, the optimal predictive performances of the sub-alignment methods show a minimum (of 0.617 (2176/3527)) when using the *colgap_percent* threshold of 0%, with sub-alignments containing the top-15 scoring columns. And a maximum (of 0.767 (2705/3527)) when using a larger threshold of *colgap_percent* = 60%, with sub-alignments containing the top-30 scoring columns. Further, none of the different sub-alignment re-scoring methods and associated parameters show an improvement in performance when more than 30 of the high scoring fSDR columns are included in the sequence sub-alignments.

The "Top-X Percent" Method for Sub-Alignment Generation

The next method investigated for the automatic selection of which aligned columns should be included in the sequence sub-alignments for functional re-scoring was the "top-X percent" method. This differs from the "top-N" method described above, because the number of columns in each of the resulting sub-alignments is selected based on a specified percentage, *X*, of the columns with the highest scoring Spearman-rank order correlation coefficients. Therefore, unlike the "top-N" method, the "top-X percent" method, in general, selects varying numbers of columns for each sub-alignment, dependent on the particular percentage selection threshold, *X*, used for the inclusion of fSDRs and the query sequence length. The variations in the bootstrapped mean values for the proportions of correct enzyme classifications, when using the "top-X" percentage threshold, are shown in *figure 4.4*. The horizontal axis in this graph represents the percentage, *X*, of aligned columns of residues that were included in the sub-alignments for functional re-scoring. These were selected through the use of an ordered ranking of the Spearman-rank order correlation coefficients, calculated by the *func-MB* method, from which the fSDRs with the highest X% ("top-X percent") of correlation coefficients were used to generate the sequence sub-alignments. Comparisons are also shown between these results when using different *colgap_percent* alignment pre-filtering thresholds, ranging (as in the top-N results) from a value of 0% to 100%.

Overall, the results were similar to those shown for the "top-N" method, in terms of the proportions and numbers of correct classifications resulting from the "top-hit" assessment of the sequences within the re-scored sub-alignments. Initially, when including a small percentage (i.e. when X is less than 5%) of the top-scoring columns in the sub-alignments, the accuracy of classifications was generally low. This observation was mostly due to an increase in the number of "empty subset (incorrect)" examples, when low percentage threshold values were used. This was true for each of the *colgap_percent* thresholds, apart from the exceptional results obtained from using a *colgap_percent* threshold of 0%. For brevity, a presentation of these more detailed results, showing the variation of the six outcomes of the "top-hit" functional assessment method, has not been included here.

An additional point of note is the behaviour of the re-scoring "top-hit" prediction results when all of the columns (i.e. X = 100%) are used to generate the sequence "sub-alignments". As can be seen in *figure 4.4*, the actual proportion of correct predictions varies depending on the *colgap_percent* threshold, however, it shows that as expected the result for the *colgap_percent* = 100% sub-alignment re-scoring is (approximately – due to minor bootstrap variations) the same as the *PAM30 UNGAPPED (0,0)* re-scoring results that were observed for the same dataset, in *chapter 3*.

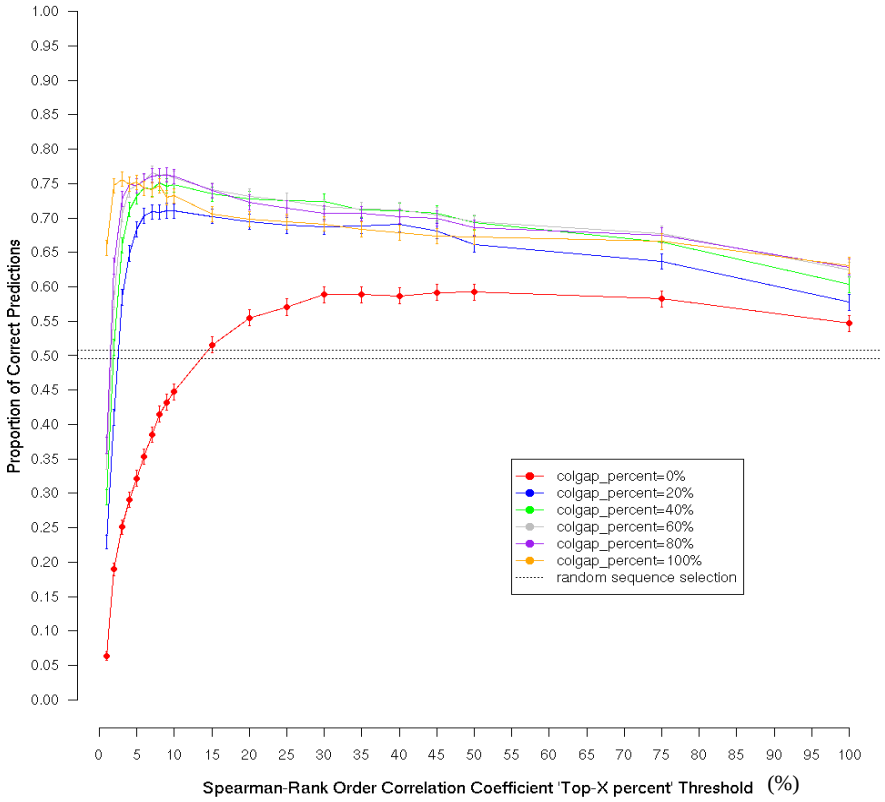

Figure 4.4. A comparison showing the proportion of correct functional predictions obtained as the "top-X percent" threshold, used to select the subsets of fSDRs used in the functional re-scoring, was varied. The horizontal axis – "Spearman-Rank Order Correlation Coefficient 'top-X percent' Threshold" – represents the percentage of aligned columns from each MSA, with the highest scoring Spearman-rank order correlation coefficients, that were included in the sequence sub-alignments. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. The enzyme classification results are shown for re-scoring the colgap_percent=0%, colgap_percent=20%, colgap_percent=40%, colgap_percent=60%, colgap_percent=80% and colgap_percent=100% filtered sequence sub-alignments. Also shown is the associated random sequence selection model for the dataset, where the dotted lines show 1 standard error deviation from the mean.

A summary of the optimal functional re-scoring results for each of the *"colgap_percent"* alignment pre-filter thresholds, along with the percentage of high scoring aligned columns (fSDRs), *X*, that contribute to the re-scored sequence sub-alignments, is shown in *table 4.2*. As for the "top-N" sub-alignment results, both the mean proportion and number of correctly classified enzyme functions, obtained from using the "top-hit" assessment method after fSDR-based re-scoring of the sequence sub-alignments, are shown for comparison. As usual, all results refer to the bootstrap form of the prediction results. It can be seen from the table that, overall, the optimal predictive performances of the sub-alignment methods show a minimum (of 0.592 (2088/3527)) when using the *colgap_percent* threshold of 0%, with sub-alignments containing the top-50% of high scoring columns and a maximum (of 0.769 (2712/3527)) when using a larger threshold of *colgap_percent* = 90%, with sub-alignments generated through the inclusion of the top-8% of aligned columns. There is, however, only a difference of 10 correct predictions in performance between this and the next lowest result of 0.766, when using the top-7% of high scoring aligned columns and a *colgap_percent* threshold of 60%.

***colgap_percent* (%)**	**(optimal) "top-X percent" (X)**	**(bootstrap) mean proportion of correct predictions**	**(bootstrap) mean number of correct predictions**
0	50%	0.592	2088
10	15%	0.691	2437
20	9%	0.711	2508
30	10%	0.739	2606
40	8%	0.751	2649
50	8%	0.761	2684
60	7%	0.766	2702
70	8%	0.763	2691
80	9%	0.763	2691
90	8%	0.769 (*)	2712
100	5%	0.752	2652

Table 4.2. A comparison between the optimal bootstrap results (mean proportion and number of correct "top-hit" specific enzyme predictions) and the "top-X percent" subset size that generates them, for each of the colgap_percent thresholds applied. All results for the number of correct predictions are out of a possible dataset size of 3527. (*) indicates the overall maximum predictive performance.

The *func-MB* "column score threshold" Method for Sub-Alignment Generation

One further method, based on the *func-MB* fSDR calculation method, was used for selecting aligned columns for the inclusion in sequence sub-alignments. This utilised a varying threshold, which was applied to the Spearman-rank order correlation coefficients that were calculated for each of the aligned columns. Therefore, only aligned columns with correlation coefficients greater than or equal to the particular threshold were included in the sequence sub-alignments used for the subsequent functional re-scoring stage. The threshold was varied from a value of 0.0 (essentially a random correlation between the residue similarities and specific enzyme function) to a value of 1.0 (indicating perfect rank correlation between the residue similarities and specific enzyme function). A graph of these results is shown in *figure 4.5*. This graph shows that there is a rapid decrease in the sub-alignment re-scoring accuracy (as measured by the proportions of "correct" predictions) when using a progressively higher threshold for the correlation coefficients associated with each aligned column.

To a certain extent these results were expected because, as the lower limit for correlation coefficient defined inclusion to the sub-alignments is made more stringent, there will be fewer available columns that fulfil the selection criteria. The sharpness, however, of the decline in functional re-scoring accuracy (using the "top-hit" assessment method), when applying a correlation coefficient threshold greater than 0.2, is perhaps surprising. This observation, seen for all *colgap_percent* thresholds, shows that, in general, even though the correlation coefficients are of less significance, the "top-hit" based classification performance increases by including these less correlated columns in the re-scored sub-alignments. Therefore, it shows that, although the nature of the relationship between residue similarity and specific function is generally (i.e. across all 3527 MSAs in the dataset) quite noisy and weak, there is some informative signal present, but it is clearly not as clean and simple a relationship (with regards to the "top-hit" re-scoring accuracy) as might be initially expected and hoped for when using the current dataset.

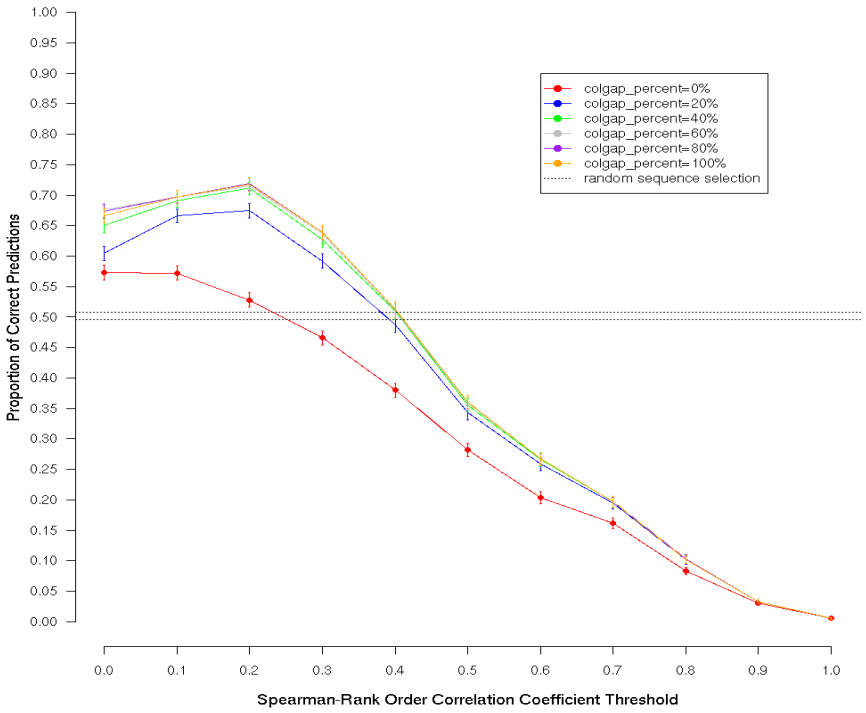

Figure 4.5. A comparison showing the proportion of correct functional predictions obtained as the Spearman-Rank order correlation coefficient threshold, used to select the subsets of fSDRs used in the functional re-scoring, was varied. For this, the aligned columns included in the sequence sub-alignments were those with an associated Spearman-Rank order correlation coefficient greater than or equal to the threshold value shown on the horizontal axis. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. The enzyme classification results are shown for re-scoring the colgap_percent=0%, colgap_percent=20%, colgap_percent=40%, colgap_percent=60%, colgap_percent=80% and colgap_percent=100% filtered sequence sub-alignments. Also shown is the associated random sequence selection model for the dataset, where the dotted lines show 1 standard error deviation from the mean.

A more detailed analysis of these results, shown in *figure 4.6* (using a similar analysis to that provided in *figure 4.3* for the "top-N" results), highlights some of the reasons for this sharp decrease in prediction accuracy as the threshold is increased. It is clear, from *figure 4.6*, that the main cause for the decline in the number of "correct" functional "top-hit" classifications (after sub-alignment re-scoring) is the rapid increase in the number of "empty subset (incorrect)" examples as the aligned column inclusion threshold is increased. There is also an additional contribution from increasing numbers of "undecidable" examples, which occur as the correlation coefficient threshold is increased above 0.2. Therefore, the increase in "incorrect" enzyme classifications is contributed to by both the "empty" sub-alignments (generally after a correlation coefficient threshold of 0.3-0.4) and the "undecidable" examples, whereas the "correct" examples from "tied – same function" do not show a compensatory increase. These results indicate that a Spearman-rank order correlation coefficient threshold, greater than 0.2, does not (in general) include enough columns to informatively discriminate between the "undecidable" examples, when using the "top-hit" method to assess the accuracy of the functionally re-scored sub-alignments of enzymes.

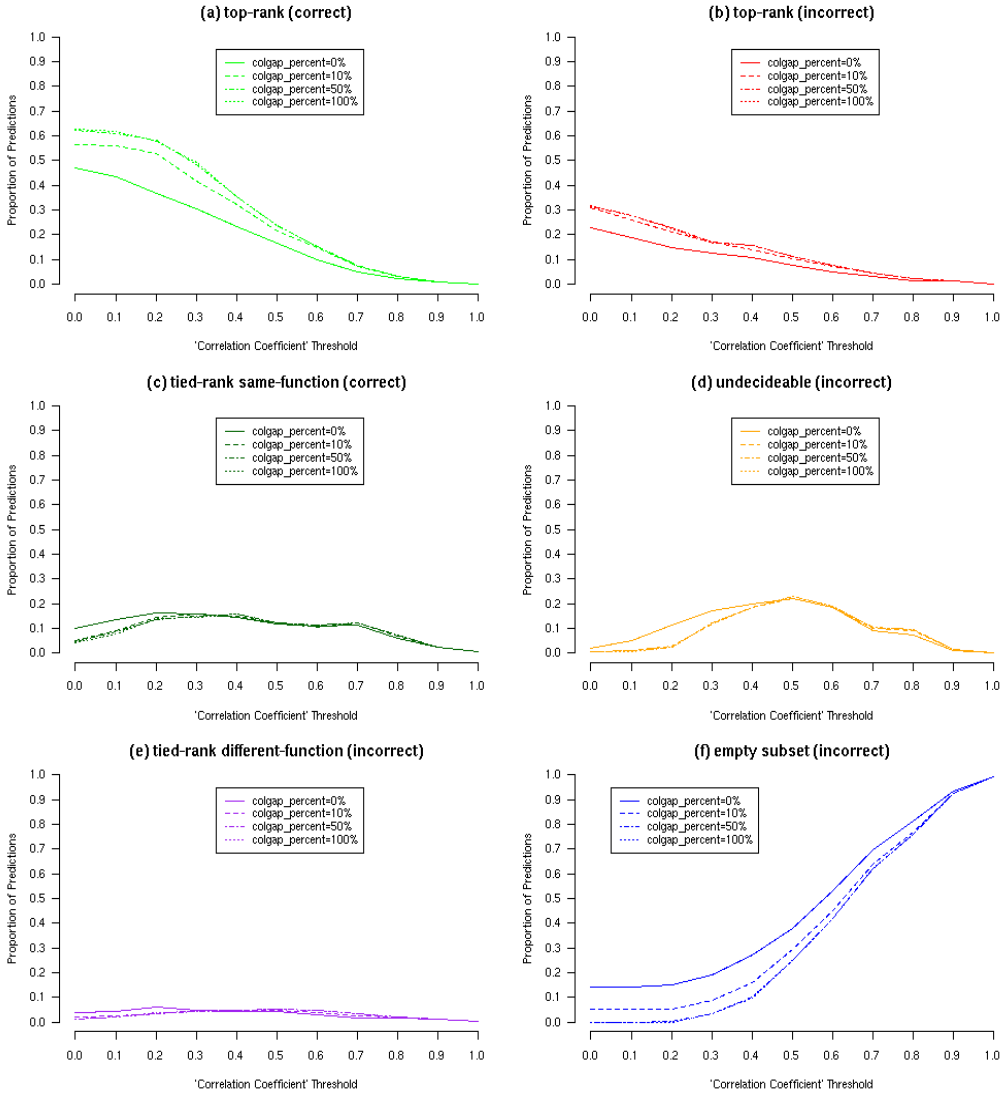

Figure 4.6. A series of graphs showing the variation of the proportions of observed predictions with the specified Spearman-rank order 'correlation coefficient' sub-alignment threshold, for the six distinct prediction outcomes (a) shows the "top-rank (correct)" results; (b) shows the "top-rank (incorrect)" results; (c) shows the "tied-rank same-function (correct)" results; (d) shows the "undecidable (incorrect)"results; (e) shows the "tied-rank different-function (incorrect)" results; and (vi) shows the "empty subset (incorrect)" results. For each of these graphs the results for re-scoring the colgap_percent=0%, colgap_percent=10%, colgap_percent=50% and colgap_percent=100% pre-filtered sequence sub-alignments are shown.

Again, a summary of the optimal results for each of the analysed *colgap_percent* thresholds is provided (see *table 4.3*). These results essentially reinforce the observations above, which show that the optimal "top-hit" based functional re-scoring results are obtained when using sequence sub-alignments containing only fSDRs with correlation coefficients (calculated via the *func-MB* method) that are greater than or equal to 0.2. This was true for all *colgap_percent* thresholds, except those of 0%, with an overall maximum number and proportion of correct predictions, of 0.719 (2536/3527), resulting when *colgap_percent* = 80%. Although, it can be seen that there is little difference between the results once the *colgap_percent* threshold reaches 50%.

colgap_percent (%)	**(optimal) "column score threshold" (Spearman-rank order correlation coefficient)**	**(bootstrap) mean proportion of correct predictions**	**(bootstrap) mean number of correct predictions**
0	0.0	0.573	2021
10	0.2	0.673	2374
20	0.2	0.674	2377
30	0.2	0.691	2437
40	0.2	0.711	2508
50	0.2	0.715	2522
60	0.2	0.716	2525
70	0.2	0.716	2525
80	0.2	0.719 (*)	2536
90	0.2	0.718	2532
100	0.2	0.718	2532

Table 4.3. A comparison between the optimal bootstrap results (mean proportion and number of correct "top-hit" specific enzyme predictions) and the "func-MB column correlation score" threshold used to generate the sequence sub-alignments that generate them, for each of the colgap_percent thresholds applied. All results for the number of correct predictions are out of a possible dataset size of 3527. (*) indicates the overall maximum predictive performance.

4.3.1.2 The Profile-HMM Method

Following on from the methods of selection used above, for the *func-MB* method of fSDR identification, a comparable set of analyses were carried out for the *profile-HMM* method. Again, three alternative methods were used for selecting aligned

columns to be included in the sequence sub-alignments. These were the "top-N", "top-X percent" and the "*profile-HMM* column score threshold" methods. Each of these are based on the same selection procedure as the *func-MB* method, except that in the following analysis the aligned column selection is based on relative ranking of the columns based on the Z-scores (rather than the Spearman-rank order correlation coefficient) calculated by the *profile-HMM* fSDR identification method.

For the *profile-HMM* method, the parameters used in the implementation of Hannenhalli and Russell (2000) were applied in this study, therefore the default settings of *hmmbuild* were used, which meant that all columns with greater than 50% gap residues were not included in the profiles generated for each of the functional sub-classes. It may, however, in future work be informative to investigate changes to the *hmmbuild* gap percentage inclusion threshold when carrying out further analysis. As in the *func-MB* method for sequence sub-alignment generation, it was decided to use a score of 0 for all comparisons between any amino acid types and gaps during the functional sequence re-scoring phase of the analysis.

The *profile-HMM* "top-N", "top-X Percent" and "column score threshold" Methods for Sub-Alignment Generation

Presented in this section are the results - from using the *profile-HMM* "top-N", "top-X Percent" and "column score threshold" methods - for the functional re-scoring of the enzyme sequence sub-alignments, generated by the *profile-HMM* based method for fSDR identification. The "top-hit" assessment method, with bootstrapping, was used to determine the accuracy of the resulting specific enzyme classifications.

Results for the variation in the proportions of correct predictions with varying sub-alignment threshold selection parameters, for the "top-N" and "top-X" percent *profile-HMM* sub-alignment selection methods, are shown in *figure 4.8* and *figure 4.9*, respectively. For brevity, a similar graphical comparison of the results for the enzyme "top-hit" classification accuracy with variation of the Z-score threshold is not shown. It is, however, worth noting that they were observed to follow a pattern similar to that seen when a threshold was applied to the *func-MB* column scores (using the Spearman-rank order correlation coefficients) for sub-alignment generation (see *figure 4.5*). That is, they exhibit a rapid decrease in the number (and proportions) of correct classifications as the (Z-score) fSDR column score threshold

is increased. This decrease occurs after an initial peak, showing a proportion of 0.665 (2345/3527) correct predictions, when the Z-score threshold used for sequence sub-alignment generation was greater than or equal to 0.5. As in the *func-MB* threshold analysis, this behaviour was mainly due to the increasing number of "empty subset (incorrect)" examples in the sequence sub-alignment re-scoring procedure. A summary of these results, along with the best performing "top-N" and "top-X percent" sub-alignment selection methods (for the *profile-HMM* based fSDR selection method) is provided in *table 4.4*.

Sub-alignment Selection Method	**(optimal) Sub-alignment threshold**	**(bootstrap) mean proportion of correct predictions**	**(bootstrap) mean number of correct predictions**
top-N	N = 35	0.673 (*)	2374
top-X percent	X = 30%	0.664	2342
Z-score column score threshold	0.5	0.665	2345

Table 4.4. A summary of the optimal bootstrap results (mean proportion and number of correct "top-hit" specific enzyme predictions) for the profile-HMM based fSDR sub-alignment re-scoring. The thresholds at which these results are obtained are shown for each of the "top-N", "top-X percent" and "Z-score column score threshold" sub-alignment selection methods investigated. All results for the number of correct predictions are out of a possible dataset size of 3527. (*) indicates the overall maximum predictive performance.

It can be seen from these results, in *table 4.4*, and the comparisons of different methods, shown in both *figure 4.8* and *figure 4.9*, that the *profile-HMM* method generally performs worse, when using this particular dataset, than the comparable enzyme classifications obtained from the *func-MB* based sub-alignment re-scoring. It is not immediately clear why there is such a difference in performance between the methods and thus further study into the optimisation of the parameters associated with the *profile-HMM* method as well as a more sophisticated filtering procedure for the input MSA data, prior to the application of the *profile-HMM* fSDR identification method, may be worthwhile.

4.3.1.3 Investigating the Random Selection of Aligned Columns

A method was implemented to calculate the specific enzyme functional classification accuracy from sequence sub-alignments that had been generated through random selection of aligned columns from the MSAs in the dataset. The aim of this was to

provide a comparison with both the *profile-HMM* and *func-MB* based sub-alignment re-scoring classification accuracies and also an assessment of their significance. The reasoning being that if the enzyme classification accuracy from the fSDR-based sub-alignment re-scoring was consistently better than that from the comparable randomly selected sub-alignments, then it would show that the fSDR-based sub-alignment selection procedure was providing additional information for the improvement of functional classification.

As stated in the methods, both the "random-N" and "random-X percent" aligned column selection methods, were used to generate the sequence sub-alignments. However, for this random selection no regard was given to the actual likelihood of the columns being associated with specific enzyme functional properties (i.e. they are not necessarily high scoring fSDRs). So, unlike with the *profile-HMM* and *func-MB* methods, the columns were (randomly) selected without first ranking them based on the calculated fSDR significance scores.

The results for both the "random-N" and "random-X percent" sub-alignment re-scoring are shown in *figure 4.7(a)* and *figure 4.7(b)*, respectively. Both show the effects (on the proportions of correct enzyme predictions) of applying different *colgap_percent* MSA pre-filtering thresholds, before the random column selection was carried out. The "random-N" results show similar behaviour for each of the applied *colgap_percent* thresholds, with overall maximum values of approximately 0.6 seen for the proportions of correct ("top-hit" based) enzyme functional predictions. With regards to the "random-X percent" results, it can be seen that they gradually tend towards the functional classification accuracies for the "X = 100% – all columns selected in the sequence sub-alignment" results, as the percentage of randomly selected columns is increased. This is to be expected, because the random selection of all columns is the same as any other selection method for all aligned columns, when using a gap-scoring function (such as the 0 gap penalty used in this sequence re-scoring study) that does not depend on the sequential ordering of the adjacent, aligned, amino acid residues (unlike that of an affine gap scoring function with non-zero gap penalty parameters). Finally, as with the "top-rank" fSDR-based sub-alignment re-scoring, there were notable exceptions (especially prominent for the "random-X percent" results) seen when using the pre-filter gap threshold of 0%.

figure 4.7 (a)

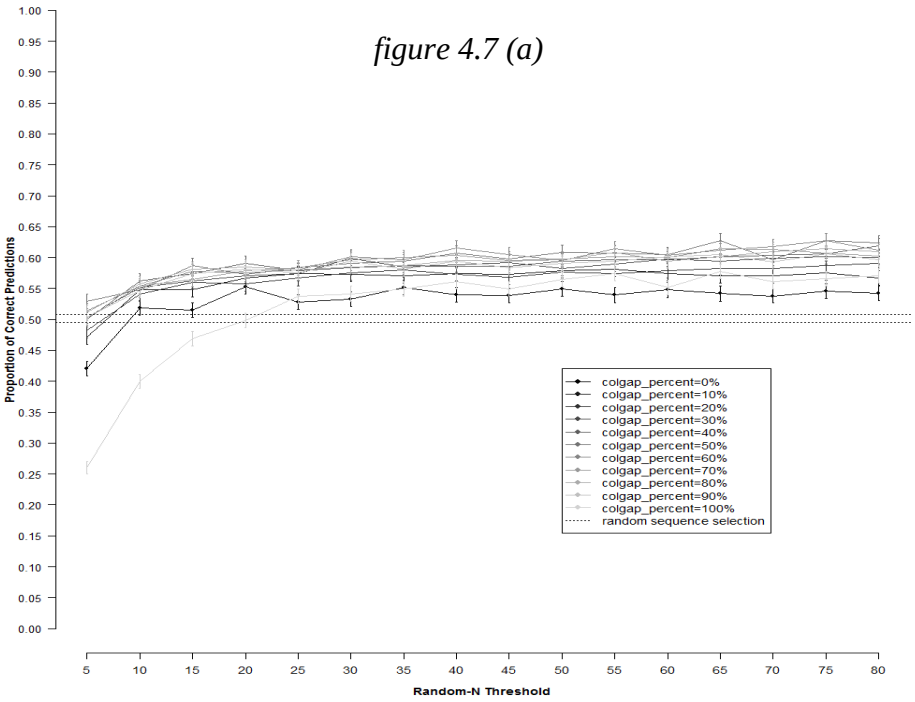

figure 4.7 (b)

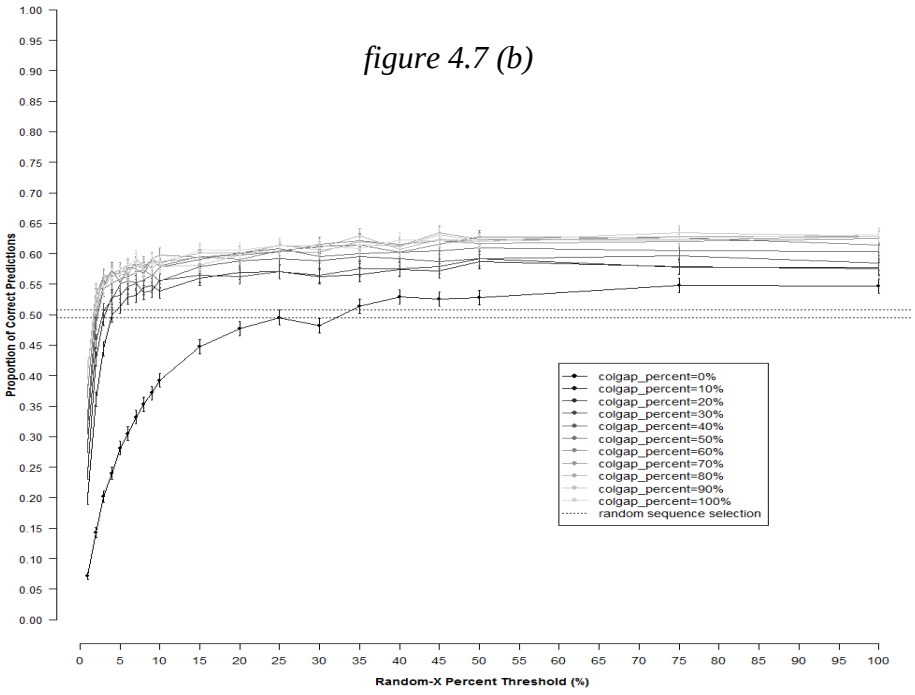

Figure 4.7. A comparison of the proportion of correct predictions obtained at each of the (a) "random-N" and (b) "random-X percent" thresholds used for random sequence sub-alignment generation. For both of these graphs, the proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars. The enzyme classification results are shown for re-scoring the randomly selected aligned columns from the MSAs that have been filtered using colgap_percent thresholds of 0-100% (in intervals of 10%). Also shown are the associated random sequence selection models for the dataset, where the dotted lines show 1 standard error deviation from the mean.

Also shown on both of these graphs is a comparison to the simple random sequence selection model, which was introduced in *chapter 3*. This has the same values - at all sub-alignment fSDR column selection thresholds and for both the "top-N" and "top-X percent" rescoring results - because it is only dependent on the enzyme classifications of the constituent sequences in the MSAs of the associated dataset. From these results, it can be seen that, in general, the functional re-scoring results from random column selection, show a larger number of correct specific enzyme classifications than the associated random sequence selection approach. This was expected to a certain extent because it was shown, in the previous chapter, that the functional re-scoring results were better when using all of the aligned columns (with a PAM30 matrix and gap scoring penalties of 0) rather than random sequence selection. Therefore, although smaller (randomly selected) subsets of these columns are being assessed, in this case the resulting subsets of aligned residues are still functionally more informative than a randomly selected sequence from the MSAs.

4.3.1.4 Comparisons between the Enzyme Sequence Sub-Alignment Functional Re-scoring Methods

To conclude this analysis, comparisons are shown between the different methods that have been investigated so far for the large-scale functional re-scoring and specific classification of enzyme sequences. The results from both the *func-MB* and *profile-HMM* methods, for "top-N" and "top-X percent" fSDR-based sequence sub-alignment selection and functional re-scoring, are compared, see *figure 4.8* and *figure 4.9* respectively. With regards to the *func-MB* calculated results, the particular *colgap_percent* thresholds were selected that gave the best overall classification performance. Therefore, for the "top-N" comparisons the results when using *colgap_percent* = 60% were selected (see the optimal overall enzyme classification accuracy results in *table 4.1*) and for the "top-X percent" comparisons those from using *colgap_percent* = 90% (see the optimal overall enzyme classification accuracy results in *table 4.2*). Also included in both of these comparisons were: the optimal predictive performance from the "column score threshold" studies, where the Spearman-rank order correlation coefficient threshold was 0.2 (see *table 4.3* - where *colgap_percent* = 80%) and the Z-score was 0.5 (see *table 4.4*), for the *func-MB* and *profile-HMM* based methods respectively; the functional re-scoring results from the "random-N" and "random-X percent" sub-alignments; the random sequence

selection model (introduced in the analysis provided in *chapter 3*); and the *PAM30 UNGAPPED (0, 0)* method, which was shown to be the best performing functional re-scoring method from the alternative amino acid substitution studies analysed in *chapter 3*, when using all aligned amino acid residues of the multiple sequence alignments for the functional re-scoring assessment.

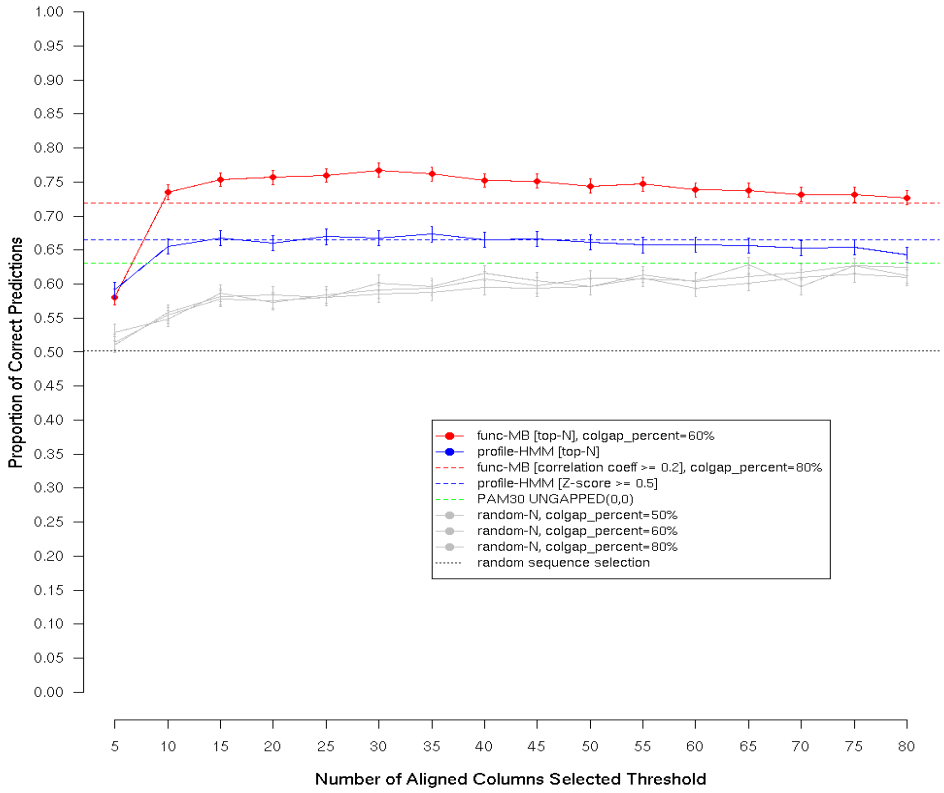

Figure 4.8. A comparison of the proportion of correct predictions obtained for the following selection of "optimal" functional sequence re-scoring methods: (i) the func-MB "top-N" method, using a colgap_percent threshold of 60%; (ii) the profile-HMM "top-N" method; (iii) the optimal func-MB "column score threshold" method, where the spearman-rank order correlation coefficient is >= 0.2, using a colgap_percent threshold of 80%; (iv) the optimal profile-HMM "column- score threshold" method, where the Z-score is >= 0.5; (v) the PAM30 UNGAPPED (0,0) re-scoring method, which was identified as optimal performing in chapter 3; (vi-vii) the "random-N" column selection methods, using colgap_percent thresholds of 50%, 60% and 80%; and (ix) the random sequence selection method. Where shown the error bars refer to 1 standard error deviation from the mean of the bootstrapped results, otherwise, just the mean value of the bootstrapped results are shown to improve clarity.

For the "random-N" and "random-X percent" column selection methods the *colgap_percent* = 50% threshold results are shown, to allow direct comparison with those results from the *profile-HMM* method, and the *colgap_percent* thresholds of

60% and 90% are included for comparison to the optimal *func-MB* based "top-N" and "top-X percent" methods, respectively. All of the results shown are from using the "top-hit" assessment method to calculate the specific functional classification accuracy after re-scoring and re-ordering the aligned enzyme sequences, and as usual they are represented by the bootstrap calculations.

From the functional re-scoring results, of both the "top-N" and "top-X percent" sub-alignment selection methods, a number of interesting observations can be made. It can be seen, in both *figure 4.8* and *figure 4.9*, that the re-scoring results for the optimal *func-MB* fSDR identification methods show an improvement over those of the optimal *profile-HMM* fSDR identification methods. The differences in accuracy between the bootstrapped mean of the proportion (and number) of correct enzyme classifications are: 0.094 (331), for the *func-MB* "top-30" and *profile-HMM* "top-35" results; 0.105 (370), for the *func-MB* "top-8 percent" and *profile-HMM* "top-30 percent" results; and 0.054 (191), for the *func-MB* "Spearman-rank order correlation threshold = 0.2" and *profile-HMM* "Z-score threshold = 0.5" results. Further, it is also possible to see a clear and significant improvement, in correct enzyme "top-hit" classifications, when using the optimal fSDR-based sub-alignments of enzyme sequence (especially in the case of the *func-MB* method), rather than the *PAM30 UNGAPPED (0, 0)* method, which was identified as optimal in *chapter 3* when using all of the aligned sequence residues to assess the functional classification accuracy of the sequence re-scoring procedure. The largest improvement, in proportion (and number) of correct predictions, seen between these methods, is 0.136 (479), when using the *func-MB* "top-8 percent" sub-alignment re-scoring method.

Comparisons between the fSDR-based sub-alignment re-scoring methods and the two alternative random models (i.e., the random sequence selection model that was introduced in *chapter 3*, and the "random-N" and "random-X percent" random column selection methods that were introduced in this chapter) clearly show significant improvements in specific enzyme classification accuracies when using the best performing sub-alignment methods. This is especially true for the *func-MB* method, which has been shown to be a better performing method overall for this benchmark dataset. Furthermore, the consistent improvement seen with the fSDR-based sub-alignment selection re-scoring methods, when compared to the

comparable random columns sub-alignment selection re-scoring methods, indicates that there is a clear, significant and functionally informative advantage to using fSDR-based sequence sub-alignments to re-evaluate the specific enzyme function of an unknown query sequence.

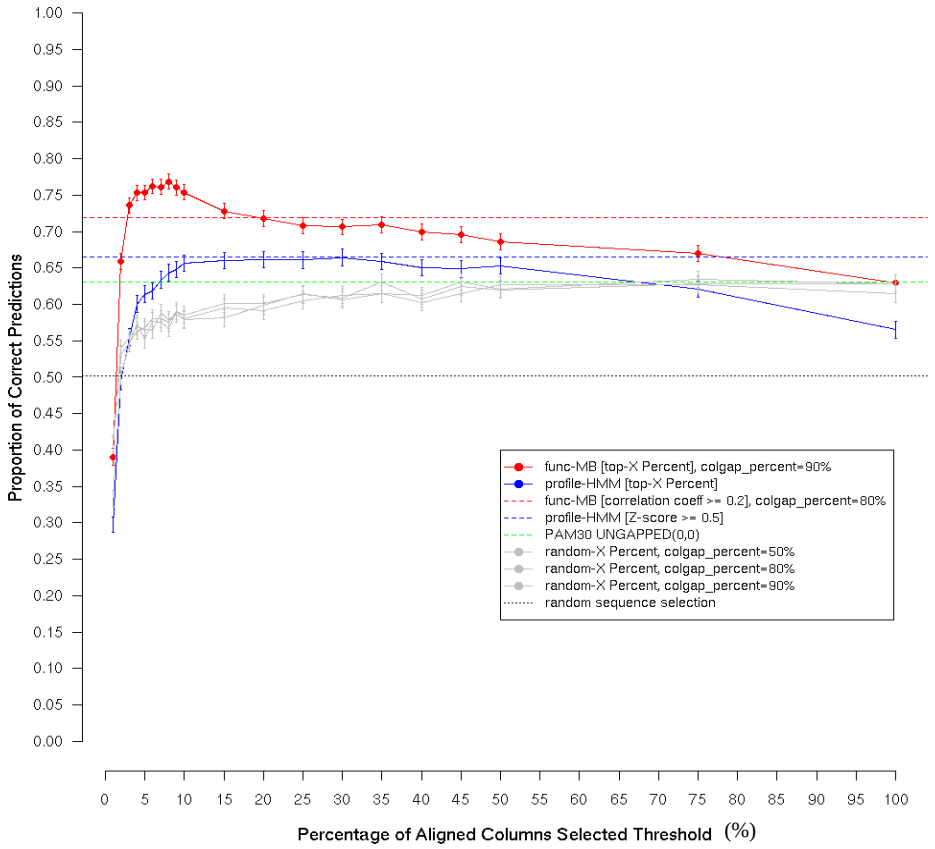

Figure 4.9. A comparison of the proportion of correct predictions obtained for the following selection of "optimal" functional sequence re-scoring methods: (i) the func-MB "top-X percent" method using a colgap_percent threshold of 90%; (ii) the profile-HMM "top-N" method; (iii) the optimal func-MB "column score threshold" method, where the spearman-rank order correlation coefficient is >= 0.2, using a colgap_percent threshold of 80%; (iv) the optimal profile-HMM "column- score threshold" method, where the Z-score is >= 0.5; (v) the PAM30 UNGAPPED (0,0) re-scoring method, which was identified as optimal performing in chapter 3; (vi-vii) the "random-N" column selection methods, using colgap_percent thresholds of 50%, 80% and 90%; and (ix) the random sequence selection method. Where shown the error bars refer to 1 standard error deviation from the mean of the bootstrapped results, otherwise, just the mean value of the bootstrapped results are shown to improve clarity.

Also, the *func-MB* results of *figure 4.9* show that even when including a quite large percentage of aligned columns in the sequence sub-alignments (such as 50% or 75%), there is still some (albeit much smaller) improvement observed in the overall

accuracy of the predictive performance. This is encouraging and to be expected, because any amount of enrichment of the aligned columns, with regards to the correlation between residue similarities and specific function, would be expected to improve the functional information signal in the resulting sequence sub-alignments. This is indeed shown (again in *figure 4.9*) by the gradual improvement in functional classification accuracy as the percentage of lesser correlated aligned columns, included in the re-scored sequence sub-alignments, is decreased, resulting in an optimal performance at the already stated threshold of the top-8%. These results, therefore, show that the specific enzyme functional classification accuracy clearly benefits from the use of a particular, optimally defined, sequence sub-alignment of functionally important residues (especially when using the *func-MB* method for fSDR identification). The most pertinent of these results are summarised for comparison in *table 4.5*.

Functional Re-scoring Method	**(optimal) Sub-alignment threshold**	**(bootstrap) mean proportion of correct predictions**	**(bootstrap) mean number of correct predictions**
func-MB *(colgap_percent=90%)*	top-8%	0.769 (*)	2712
profile-HMM	top-35	0.673	2374
PAM30 UNGAPPED (0,0)	n/a	0.631	2226
random-N			
colgap_percent=50%	N = 65	0.628	2215
colgap_percent=60%	N = 75	0.627	2211
colgap_percent=80%	N = 75	0.615	2169
random-X percent			
colgap_percent=50%	X = 50%	0.627	2211
colgap_percent=80%	X = 35%	0.630	2222
colgap_percent=90%	X = 75%	0.635	2240
random sequence selection	n/a	0.502	1771

Table 4.5. A summary of the optimal bootstrap results from the functional re-scoring assessments analysed in this chapter (mean proportion and number of correct "top-hit" specific enzyme predictions). Where relevant, the sub-alignment selection methods and associated thresholds at which these results were obtained are shown. All results for the number of correct predictions are out of a possible dataset size of 3527. (*) indicates the method with the overall maximum predictive performance.

In conclusion, these comparisons show that when using fSDR-based sequence sub-alignments, improvements are observed for the predictive performance of specific enzyme function classification, when using the "top-hit" assessment method. This is shown especially clearly when contrasting the best results from the *func-MB* method of sub-alignment generation and those from the *PAM30 UNGAPPED (0, 0)* method, which uses all aligned columns of residues to re-score the functional similarity of the aligned sequences. The detailed results for each of these approaches are summarised for comparison in *table 4.5*.

4.3.1.5 The Effects from Clustering the Dataset Query Sequences

In this section a procedure similar to that described in *chapter 3* (*section 3.3.5*) was carried out. This involved the definition of a series of sequence clusters using CD-HIT (Li and Godzik, 2006) - with six thresholds of sequence percentage identity (90%, 80%, 70%, 60%, 50%, and 40%) - by clustering the query sequences used to create the MSAs contained in the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset. As before, the aim was to provide an investigation (additional to the bootstrap re-sampling) into the effect that any potential bias, due to sequence redundancy within the query sequences used to create the benchmark dataset, may have on the accuracy and trends of the enzyme function prediction results from the fSDR-based sub-alignment re-scoring.

A summary of the sequence identity clustering thresholds and the number of sequence clusters generated for this particular dataset was given in *table 3.3*; where a 100% identity threshold refers to the dataset compositions prior to any CD-HIT sequence clustering. The number of sequence clusters produced at each threshold, for each distinct dataset, also defines the number of MSAs that constitute the datasets at each of the sequence identity thresholds.

The above *func-MB*, *profile-HMM* and random column selection analyses were repeated for each set of the clustered sequence alignments, identified in *table 3.3*, for the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset. For the bootstrap re-sampling analysis of the results a sample selection size of half the number of dataset MSA examples was used for each of the associated clustered datasets.

The Effects on the *func-MB* "top-N", "top-X Percent" and "column score threshold" Methods for Sub-Alignment Generation

Repetition of the fSDR-based sub-alignment re-scoring experiments was carried out for each set of MSAs in the clustered sub-sets of data. To allow direct comparison between these results and those shown above, which did not include any query sequence identity clustering, identical column selection methods and *colgap_percent* thresholds for MSA pre-filtering were applied. All three of the *func-MB* based column selection methods (i.e. the top-N; top-X percent; and the Spearman-rank order correlation coefficient threshold methods) were investigated, and the proportion of correct enzyme classifications compared at each cluster percentage threshold. In general, the overall trends and accuracies of correct prediction were similar to the results obtained without any prior clustering of the query enzyme sequence set.

Results from the re-scoring of the "top-X percent" *func-MB* generated sub-alignments are shown in *figure 4.10*. This shows the proportion of correct (specific enzyme) predictions that were observed when re-scoring the sub-alignments created from the MSAs contained within the datasets associated with the 40%, 60%, 80% and 100% query sequence identity clustering process. For clarity, only the results from the functional re-scoring of the sub-alignments, generated after application of the *colgap_percent=90%* threshold for the pre-filtering of aligned columns with a specified number of gaps, are shown. This particular threshold was selected because it gave the best predictive performance in the analysis above, where no sequence clustering was done, and therefore allows a direct comparison between the results. As before, the results show the mean proportion of correct (specific enzyme) predictions and one standard error deviation from the mean, for each of the investigated sub-alignment generation thresholds.

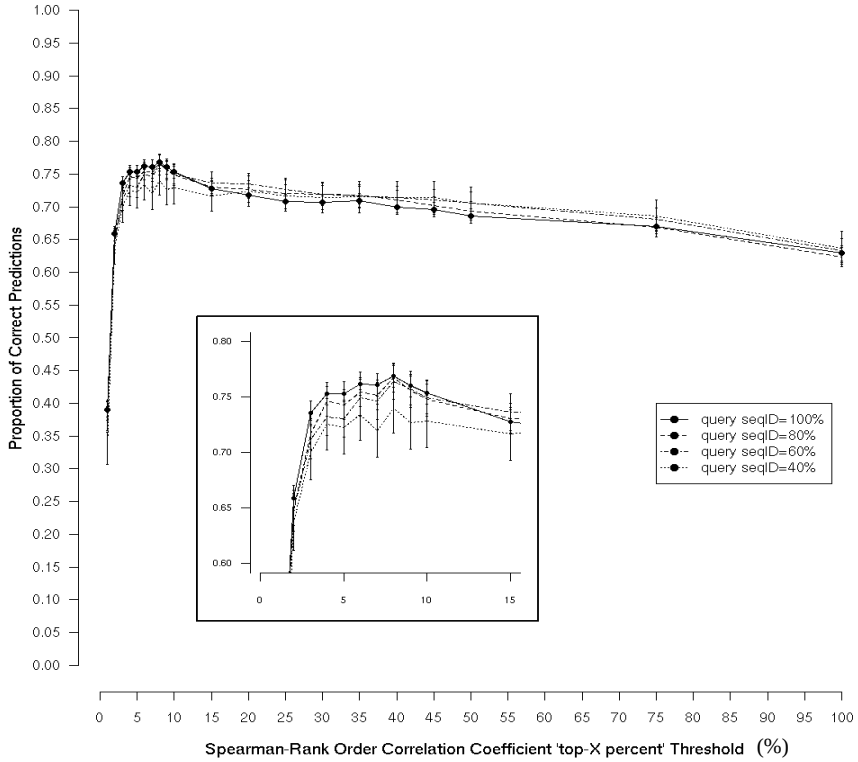

Figure 4.10. A comparison between the proportion of correct predictions obtained at each of the spearman-rank order correlation coefficient "top-X percent" thresholds used for func-MB fSDR-based sequence sub-alignment generation. The enzyme classification results are shown for re-scoring the query sequence identity (seqID) clustered datasets of MSAs (using thresholds of seqID=40%, seqID=60%, seqID=80%, and seqID=100%), which have had a colgap_percent=90% aligned column gap threshold filter applied prior to sequence sub-alignment generation. The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars.

From *figure 4.10*, it can be seen that in general there is a decrease in the observed proportion (and therefore number) of correct enzyme classifications, when re-scoring the fSDR-based sub-alignments from the clustered sub-datasets with aligned column sub-set sizes generated using a selection threshold below X=15%. This is shown in more detail in the inset graph of *figure 4.10*. In contrast, above this percentage threshold, the clustered datasets (in particular those from using a 40% and 60% sequence identity clustering threshold) generally out-perform those using 80% and 100% (i.e. no clustering).

These results indicate that the overall specific classification results may contain a certain degree of bias from particular families of closely related enzyme sequence families that are responding favourably to the specific combination of *func-MB* sub-

alignment selection parameters. Even though the relationship becomes slightly less clear and the predictive signal reduced, there is still significant improvement to be gained from using a sub-alignment of sequences based on the *func-MB* method of functional residue selection. This is shown in *table 4.6*, which shows that the optimal performing *func-MB* fSDR-based re-scoring methods consistently give a larger number of correct enzyme classifications, when compared to the comparable results presented in *chapter 3* that use all available aligned amino acid residues to assess the functional similarities. This is the case for all three of the *func-MB* column selection methods and also each of the query sequence identity clustering thresholds that were investigated.

The Effects on the *profile-HMM* "top-N", "top-X Percent" and "column score threshold" Methods for Sub-Alignment Generation

A similar analysis was also carried out using the *profile-HMM* method for fSDR identification. Some of the resulting effects on the classification accuracy are shown in *figure 4.11* and the optimal results are summarised in *table 4.6*. An interesting observation from these results is that as the query sequence identity clustering threshold is reduced, in general, the proportion of correct classifications increases. This is perhaps best demonstrated with the "Z-score threshold based sub-alignment" results, shown in *figure 4.11*, which shows a clear improvement (when using a Z-score threshold of 1.0 and 1.5) for the 40% and 60% query sequence identity clustered datasets. It is clear, however, from *table 4.6*, that the *func-MB* based re-scoring method continues to correctly classify the specific function of the query sequence in more comparable cases and that unfortunately the *profile-HMM* method continues to not perform as well as was first expected on this data.

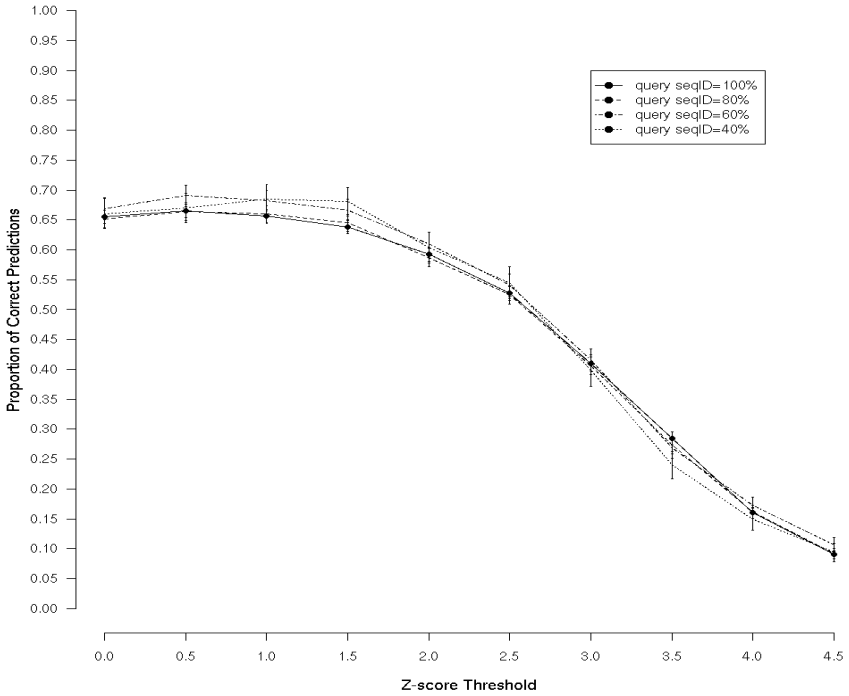

Figure 4.11. A comparison between the proportion of correct predictions obtained at each of the Z-score thresholds used for profile-HMM fSDR-based sequence sub-alignment generation. The enzyme classification results are shown for re-scoring the query sequence identity (seqID) clustered datasets of MSAs (using thresholds of seqID=40%, seqID=60%, seqID=80%, and seqID=100%). The proportions of correct predictions are the bootstrap mean values, shown with the corresponding standard error bars.

The Effects on the *"random-N"* and *"random-X Percent"* Methods for Sub-Alignment Generation

To complete this analysis, the random column selection methods were also applied to these clustered subsets of sequence alignments. The results from functionally re-scoring both the "random-N" and "random-X percent" generated alignments showed no significant difference between the different levels of sequence identity clustering, for all column selection thresholds. For brevity and to avoid repetition these results are not shown here.

	func-MB method			*Profile-HMM method*			**amino acid matrix method**
	top-N	**top-X percent**	**column score**	**top-N**	**top-X percent**	**column score**	
Query sequence cluster threshold = 100%							
(optimal) sub-alignment threshold	top-30 *colgap=60%*	top-8% *colgap=90%*	>= 0.2 *colgap=80%*	top-35	top-30%	>= 0.5	PAM30 (0, 0)
(bootstrap) mean proportion correct +/- se	0.767 +/- 0.011	0.769 +/- 0.010	0.719 +/- 0.011	0.673 +/- 0.011	0.664 +/- 0.011	0.665 +/- 0.011	0.631 +/- 0.012
Number Correct (out of 3527)	2705	2712	2536	2374	2342	2345	2226
Query sequence cluster threshold = 80%							
(optimal) sub-alignment threshold	top-30 *colgap=60%*	top-8% *colgap=90%*	>= 0.2 *colgap=80%*	top-25	top-15%	>= 0.5	PAM40 (0, 0)
(bootstrap) mean proportion correct +/- se	0.763 +/- 0.013	0.767 +/- 0.013	0.723 +/- 0.014	0.676 +/- 0.014	0.664 +/- 0.014	0.664 +/- 0.014	0.632 +/- 0.015
Number Correct (out of 2131)	1626	1634	1541	1441	1415	1415	1347
Query sequence cluster threshold = 60%							
(optimal) sub-alignment threshold	top-30 *colgap=50%*	top-9% *colgap=80%*	>= 0.2 *colgap=80%*	top-30	top-15%	>= 0.5	PAM40 (0, 0)
(bootstrap) mean proportion correct +/- se	0.759 +/- 0.016	0.759 +/- 0.016	0.723 +/- 0.017	0.698 +/- 0.017	0.685 +/- 0.018	0.690 +/- 0.018	0.645 +/- 0.018
Number Correct (out of 1392)	1057	1057	1006	972	954	960	898
Query sequence cluster threshold = 40%							
(optimal) sub-alignment threshold	top-30 *colgap=100%*	top-8% *colgap=90%*	>= 0.1 *colgap=60%*	top-20	top-15%	>= 1.0	PAM40 (0, 0)
(bootstrap) mean proportion correct +/- se	0.748 +/- 0.023	0.740 +/- 0.023	0.713 +/- 0.024	0.695 +/- 0.024	0.681 +/- 0.024	0.685 +/- 0.025	0.652 +/- 0.025
Number Correct (out of 721)	540	534	514	501	491	494	470

Table 4.6. A summary of the re-scoring methods that give the optimal specific enzyme functional predictive performance for each of the sub-alignment selection methods and a selected set of associated query sequence clustered subsets. The column - amino acid matrix method – specifies the optimal amino acid substitution re-scoring matrices and gap penalties previously identified in chapter 3 (see table 3.4). Bootstrap values for both the mean proportion, with standard error (se), and number of correct predictions are shown for each method.

4.3.2 Lactate/Malate Dehydrogenase Example

To complete this part of the study into the possible use of automatically predicted functional specificity determining residues for the improvement of specific enzyme functional assignment, a detailed study of a well characterised example enzyme class was carried out. This provides an opportunity to look at the more detailed aspects of the sub-alignment re-scoring process, in a single specific example, which removes some of the potential complications from using the larger sample dataset size of MSAs in the earlier aggregated study.

For this, an example involving the experimentally and computationally well-studied lactate and malate dehydrogenases (LDH/MDH) was selected. These enzymes are found in a wide range of organisms, often showing quite divergent sequences, although they do, however, share a common substrate binding site and overall catalytic mechanism (Goward and Nicholls, 1994; Wilks et al., 1988). Experimental studies, involving the rational redesign of protein sequences, have shown that the substrate binding specificity can be altered through the mutation of just one key residue in the binding site (Goward and Nicholls, 1994; Wilks et al., 1988), while some of the other substrate binding residues are completely conserved. This is therefore a good example of subtle protein sequence changes causing important differences in specific biochemical enzyme function, which cannot easily be identified by standard "whole" sequence similarity methods. A further attraction for investigating this group of enzymes is that they have been studied previously in both of the studies by Pazos et al. (2006) and Hannenhalli and Russell (2000).

For this particular example a multiple sequence alignment was obtained from the "initial" dataset and was therefore generated through a PSI-BLAST database search using an E-value threshold for sequence inclusion of 10. The input query sequence used was represented by the UniProt database entry with accession code **O08349**. This protein has a length of 294 amino acids and an EC classification of 1.1.1.37 (malate dehydrogenase).

The alignment contained sequences from four specific enzyme functional classes: (i) 158 sequences with (L)-lactate dehydrogenase activity, using NAD(+) as a coenzyme (EC 1.1.1.27); (ii) 128 sequences with malate dehydrogenase activity, using NAD(+) as a coenzyme (EC 1.1.1.37); (iii) 6 sequences with malate

dehydrogenase activity, using NADP(+) as a coenzyme (EC 1.1.1.82); and (4) one sequence representative of the enzyme EC 1.2.1.12, which appears to be functionally quite different to the other three, sharing only the use of a NAD(+) coenzyme and the general dehydrogenase enzyme class. The EC 1.2.1.12 annotated enzyme sequence has an insignificant E-value of 4.1 and a small percentage of alignment overlap with the query sequence of less than 50%. Also, its EC function only shares a general functional class with the other enzymes in the alignment and therefore is not very close in terms of functional specificity. Each of these factors suggests that it is a false positive database hit and should be removed from the analysis to prevent noise from distorting the quality of the scores obtained in the fSDR selection. It is worth noting that this particular sequence would have been removed from the alignment by the use of a more stringent E-value threshold filter, such as that applied to the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset used in some earlier analyses in this thesis.

A further feature of this alignment is slightly more subtle and presents an interesting case for consideration that could also occur in other examples. Of the three remaining functions, EC 1.1.1.27 and EC 1.1.1.37 have distinct substrate binding specificities, but the same coenzyme binding specificity. In contrast, EC 1.1.1.37 and EC 1.1.1.82 have the same substrate binding specificities, but distinct coenzyme binding specificity. These changes in functional detail have been experimentally verified and involve very few amino acid mutations in each case (Goward and Nicholls, 1994). This highlights the potential complications that can affect the automated analysis of functional sequence details. It also shows some of the functional subtleties that can be hidden in schemes of functional classification, such as the EC system. Some of these points are explored further below.

4.3.2.1 Analysis of the Identified Functional Specificity Determining Residues (fSDRs)

The columns identified with the highest scores, calculated using the rank correlation coefficient for the *func-MB* method and the Z-score for the *profile-HMM* method, are shown in *table 4.7*. All columns with greater than 50% gaps were not included in the *func-MB* calculation and all gap comparisons were scored as 0 in the amino acid similarity correlation matrix. Listed in the table are the top 5 fSDR column scores

from each method. These were calculated using all the aligned sequences, which contained the four specific functions described above and therefore no pre-filtering of the alignment was done.

Rank (method)	Score	Alignment position	Residues in malate<->lactate
1 (*profile-HMM*)	3.505	233 (107)	M <-> E
1 (*func-MB*)	0.850	233 (107)	M <-> E
2 (*profile-HMM*)	3.376	215 (102)	R <-> Q
2 (*func-MB*)	0.740	249	(not obvious)
3 (*profile-HMM*)	1.856	322	N <-> N
3 (*func-MB*)	0.690	215 (102)	R <-> Q
4 (*profile-HMM*)	1.781	212 (102)	(not obvious)
4 (*func-MB*)	0.661	1032	P <-> (I, V)
5 (*profile-HMM*)	1.759	324	(not obvious)
5 (*func-MB*)	0.622	995	(A, S) <-> T

Table 4.7. A comparison between the top-5 ranked scores from the func-MB and profile-HMM, fSDR identification methods. "Alignment position" represents the column number in the BLAST alignment and the number in brackets (where comparable) is the cross-reference to the corresponding alignment position used by Hannenhalli and Russell (2000). Where clearly distinguishable, the main residue type in the alignments of the malate and lactate dehydrogenases is shown.

It can be seen, through manual inspection of the MSAs, that both methods identify the main, experimentally verified (Wilks et al., 1988), specificity determining residue switch between Arginine (R) and Glutamine (Q) in the malate and lactate dehydrogenases respectively. This residue occurs at position 215 in the BLAST MSA and corresponds to alignment position 102 in the study by Hannenhalli and Russell (2000) and position 95 in the study by Pazos et al. (2006). When cross-referenced to the query sequence O08349, this position relates to residue number 81 and is annotated in the UniProt FT field as a substrate binding site. Interestingly, both methods report the highest scoring alignment position to be at 233 (107 in Hannenhalli and Russell (2000)), relating to a switch from predominantly Methionine (M) to Glutamic acid (E) residues in the malate and lactate dehydrogenases respectively. No specific experimental evidence seems to be available to quantify the effect on catalysis caused by this switch. However, it does occur in the region of the substrate binding site and would therefore be expected to play some part in the specific substrate recognition.

Hannenhalli and Russell suggest in their study that a Z-score above 3.0 is usually a good indicator of functionally specific alignment positions. Pazos et al. (2006) don't mention a particular threshold value for the correlation coefficient, although they do suggest using 0.6 in their earlier study of the MB-method (del sol Mesa et al., 2003). If these thresholds are applied to the scores in *table 4.7*, we can see that all five of the *func-MB* scores are above 0.6, but only the top two Z-scores are above 3.0.

To investigate this further, the *profile-HMM* method was used to re-calculate the fSDR scores when: (i) removing the sequence with function EC 1.2.1.12; and (ii) only using the aligned sequences with enzyme functions EC 1.1.1.27 and EC 1.1.1.37. The first case aims to investigate effects of unrelated specific functions and sequence profiles of small sequence sample sizes on the Z-scores. The second case explores the difference between Z-scores (especially for the R<->Q change) when the potentially confusing case of two different functional classes (EC 1.1.1.37 and EC 1.1.1.82) with the same substrate binding specificity, but different functional class, is removed. The results in *table 4.*8 compare the findings for these two cases and the unfiltered sequence alignment results from *table 4.*7). The top-5 identified columns and the corresponding Z-scores are shown.

	Unfiltered		**EC 1.2.1.12 removed**		**EC 1.1.1.27 and EC 1.1.1.37 only**	
Column score rank	**Z-Score**	**Alignment position**	**Z-Score**	**Alignment position**	**Z-Score**	**Alignment position**
1	3.505	233 (107)	5.619	1032	6.170	215 (102)
2	3.376	215 (102)	4.613	215 (102)	5.180	233 (107)
3	1.856	322	4.585	233 (107)	5.080	1032
4	1.781	212 (102)	3.129	995	2.676	397
5	1.759	324	2.676	397	2.928	249

Table 4.8. A comparison of the effect of sequence alignment pre-filtering on the top-5 identified fSDR columns and the corresponding Z-scores calculated with the profile-HMM method.

When removing the sequence with function EC 1.2.1.12 it can be seen that the two alignment positions (233 and 215) with highest Z-score (from the unfiltered data) are still identified, but with increased Z-scores. This is also true for the case where only the sequences with functions EC 1.1.1.27 and EC 1.1.1.37 are used. These results would seem to indicate that, as expected, the fSDR score signals improve as

sequences that may cause functional signal noise complications are removed from the analysis.

It would also seem from these results that the profile method is sensitive to functional groups with very few sequence examples and also a low percentage of aligned residues. This appears to be due to the way in which the relative entropy calculations require a consensus match emission state for a particular position in all of the individual functional alignments and the limited amount of enzyme sub-class specific sequence information from which to build the HMM profiles. This highlights a potential limitation of the method and further study targeted towards improving the results from potentially problematic alignments, with poorly aligned sequences containing large gapped regions and enzyme sub-classes with few sequence examples, would be useful. As an example, we can see in *table 4.8* that when the EC 1.2.1.12 example is removed from the alignment, alternative high scoring alignment positions are obtained that previously occurred in a region of the MSA that was not aligned with this sequence. These aligned columns are 1032 and 995 and are also identified by the *func-MB* method (see *table 4.7*). We can therefore see from these results that the identified fSDR results from the *func-MB* method, without any alignment pre-filtering, are comparable to those of the *profile-HMM* method after some filtering. This indicates that the *func-MB* method may be less sensitive than the *profile-HMM* method to some of the potential alignment problems and is a likely contributing factor to the improved overall re-scoring results seen for the *func-MB* method when compared to the *profile-HMM* method in the earlier, larger analysis of the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset.

4.3.2.2 Effects on Functional Rank-Ordering and Grouping

A major aim of this study is the incorporation of information gained from identifying a set of functionally specific amino acids, into a method for improving automatic functional classification of an unknown enzyme sequence. In this section, a detailed analysis is carried out into the effects on the functional rank-ordering of aligned enzyme sequences, when using alternative sub-alignment based re-scoring methods. For this particular example, the LDH/MDH sequence alignment (obtained from query sequence O08349 in the "initial" dataset), with no pre-filtering of the alignments, was again used.

When carrying out a PSI-BLAST database search, using query sequence O08349, the most significant hit (with E-value = $1x10^{-52}$) is to a lactate dehydrogenase instead of a malate dehydrogenase and is therefore functionally incorrect at the 4th EC number, which denotes substrate specificity. The highest ranked sequence with the "correct" function is at rank 35 with an E-value of $1x10^{-43}$. A more detailed analysis of the distribution of the ranking positions, for the sequences with correct (EC 1.1.1.37) and incorrect (EC 1.1.1.27) functions, shows that the ordering is almost the inverse of that required to enable a correct prediction based on homology transfer. *Figure 4.12(a)* highlights this problem with a smoothed density distribution of the rank positions for the sequence homologs with the "correct" and "incorrect" functional classifications from the BLAST output, showing the majority of the correct predictions at the lower ranked positions. A similar, although slightly improved, situation is also observed in the functional rank distributions when using the PAM10 (0, 0) (i.e. using a gap score of 0) substitution matrix (which was found to be optimal for re-scoring the initial dataset – data not shown) to re-rank the sequences. These observations, coupled with knowledge of the substrate binding requirements, show that this is a prime example of a situation where additional information is required to correctly assign the specific enzyme function.

When using sequence sub-alignments, generated from the top-5 highest scoring columns identified by both the *func-MB* and *profile-HMM* fSDR identification methods, a significant improvement in the rank distributions of the correct sequences is observed. This is shown in *figure 4.12(b).* A comparison between these ranking distributions and those from the BLAST and PAM10 (0, 0) results are shown in *figures 4.12(c) and figure 4.12(d).* The distributions of the rankings for the "correct" functional sequences, in *Figure 4.12(c)*, clearly move towards higher ranking positions when only the amino acids from the top-5 scoring aligned columns are used for re-scoring. Conversely, the distribution of the "incorrect" functional sequences, shown in *figure 4.12(d),* show a clear movement towards the lower ranking positions when the fSDR identification methods are used. Therefore, it can be seen from these graphs that both the *func-MB* and *profile-HMM* methods provide significant improvement - in the ordering of the sequences with the same "correct" functional classifications as the query - when compared with the other sequence

similarity re-scoring methods, such as BLAST and the PAM10 substitution matrix, which use all of the aligned sequence residue information.

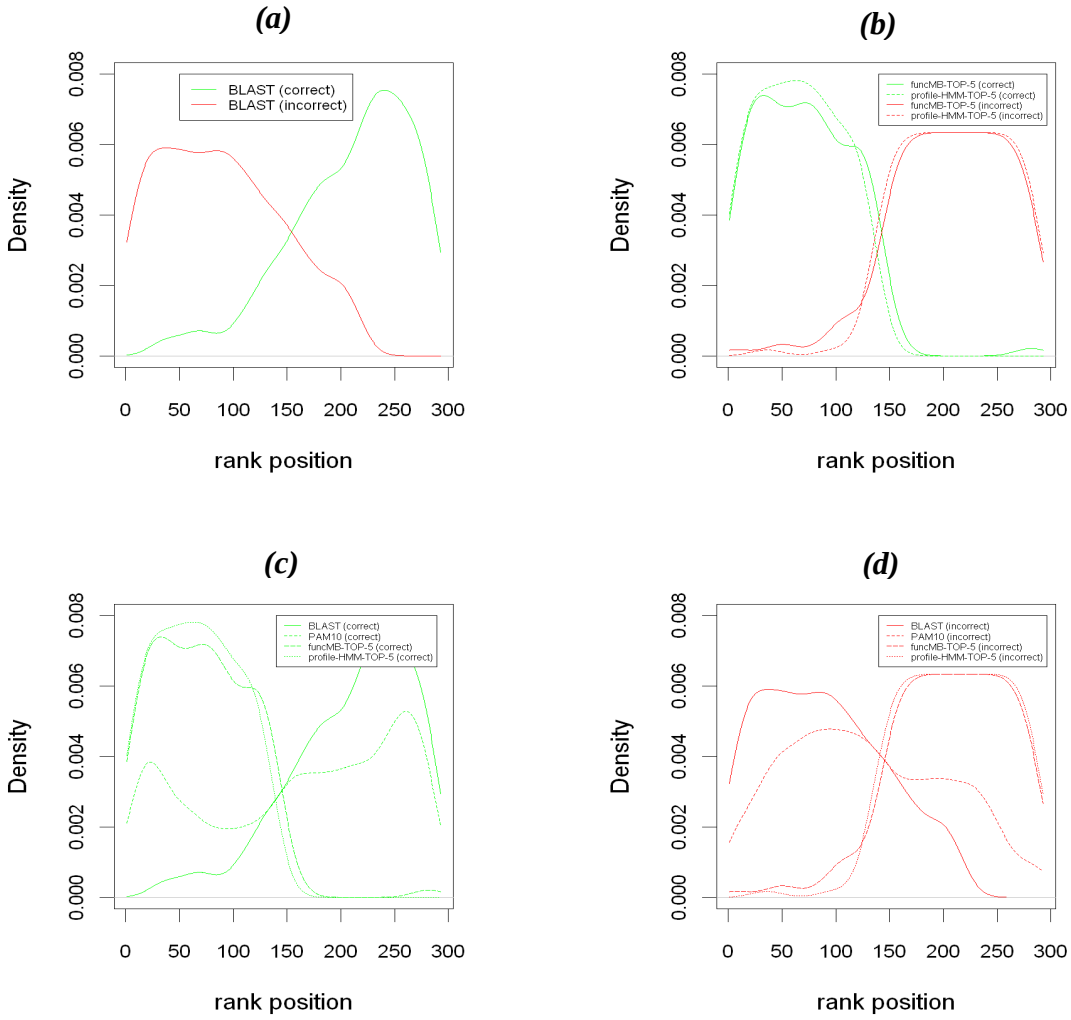

Figure 4.12. Graphs showing the variation of ranking distributions when using different alignment scoring methods: (a) distribution of functionally "correct" and "incorrect" rank positions from the original 1st iteration PSI-BLAST results; (b) distribution of functionally "correct" and "incorrect" rank positions when using the func-MB and profile-HMM "top-5" results; (c) overlay of "correct" ranking distributions from BLAST, PAM10 (0, 0), func-MB top-5 and profile-HMM top-5 results; (d) overlay of "incorrect" ranking distributions from BLAST, PAM10, func-MB top-5 and profile-HMM top-5 results.

Further verification of this improvement in the grouping of sequences with the correct specific function, is shown by the data in *table 4.9.* Here, a more detailed analysis has been carried out, which measures the number of functionally "correct" sequences (the "enrichment") occurring in the top 10, 20, 30, 40 and 50 rank

positions after alignment re-scoring. Again, the data used is from the lactate/malate dehydrogenase alignment data, described above, with no pre-filtering of the alignments. The re-scoring methods compared are: BLAST; global sequence identity (seqID); PAM10 (0, 0) substitution matrix; *func-MB* and *profile-HMM* "top-N" methods (where N=5, 10, 20, and 30).

		Number of sequences (N) with correct functions in the top-N ranked positions after functional re-scoring				
re-scoring method	**Top hit?**	**10**	**20**	**30**	**40**	**50**
BLAST	No	0	0	0	2	2
seqID	No	3	6	7	10	11
PAM10 (0, 0)	No	4	10	17	21	23
***func-MB* top-5**	Yes	9	19	29	39	48
***func-MB* top-10**	Yes	8	9	13	14	15
***func-MB* top-20**	Yes	9	14	18	20	24
***func-MB* top-30**	Yes	1	4	4	4	4
***profile-HMM* top-5**	Yes	10	20	30	39	49
***profile-HMM* top-10**	Yes	10	19	29	38	47
***profile-HMM* top-20**	Yes	7	17	27	34	42
***profile-HMM* top-30**	Yes	5	15	22	28	31

Table 4.9. Comparison between the level of "enrichment" of correct prediction results in the top 10, 20, 30, 40 and 50 rank positions, after re-scoring the aligned sequences using the methods listed. The "Top hit?" column indicates whether the top ranked sequence position shows a correct specific functional hit to the query. The alignment data used is from the lactate/malate dehydrogenase alignment from the "initial" dataset with no sequence pre-filtering.

The main outcome of this comparison is the large improvement in the number of correct predictions in the top ranking positions, when using the fSDR-based amino acid subset to re-rank the aligned sequences, compared to the "whole" alignment sequence similarity methods (such as BLAST, seqID and PAM10). Both the *func-MB* and *profile-HMM* methods show that almost all the top 50 rank positions are populated by correct functional predictions, when the top-5 scoring fSDR columns are used for re-scoring. This is a very promising result, because it shows the potential of the fSDR identification methods for improving the functional re-ordering of the sequences using this particular example. Therefore, it provides further evidence for the use of the proposed fSDR-based sequence sub-alignment re-scoring method for the improvement of specific enzyme functional assignment.

A final point is that, as the number of aligned fSDR positions included in the subset increases, the corresponding amount of top ranking enrichment tends to decrease when using both methods. This is to be generally expected because, as the fSDR score value decreases, the ability of the residues in the aligned sequence subset, to separate the specific functional classes, will be reduced. This must, however, be qualified with some caveats. Firstly, *table 4.9* shows that each methods is quite sensitive to the particular number of columns included in each of the sub-alignments (for example, the *func-MB* top-20 method appears to be performing slightly better than the top-10, whereas the top-30 is significantly worse than both). Also, the optimal top-N values (of N=30 and N=35 for the *func-MB* and *profile-HMM* methods respectively) obtained from the earlier large-scale study of the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, are not always likely to be optimal for a single specific example. This is especially important when taking into consideration a more detailed assessment scheme, such as top-rank enrichment, in comparison with the "top-hit" method and also highlights potential limitations in the use of single, averaged thresholds obtained from all the MSA examples in a large-scale analysis.

4.4 Conclusions

In this chapter, two different methods, for the automatic identification and scoring of aligned amino acids that are expected to play a role in determining functionally specific protein properties, have been described and compared. These potential fSDRs, identified by both the *func-MB* and *profile-HMM* methods, were then used to generate sub-alignments of enzyme sequences, of varying sizes, via a number of aligned column selection methods. The sub-alignments were then functionally re-scored, using the PAM30 amino acid substitution matrix and the resulting functional classification accuracy was assessed using the "top-hit" method. In addition to this, a comparable method for the random selection of aligned columns was developed and the functional classification performance of the resulting sub-alignments was assessed and compared to the fSDR-based methods. Finally, a detailed analysis of fSDR identification and their subsequent use in functional re-scoring was carried out for a multiple sequence alignment of the well-studied lactate and malate dehydrogenases.

The optimal functional re-scoring results, from the *func-MB* method, show a significant improvement in the level of enzyme classification accuracy, when compared to all of the other methods investigated. This is the case for all three of the sequence sub-alignment selection methods, in particular those using small subsets of residues predicted to be correlated with specific function. Overall the best results are obtained through use of the "top-N" (where N=30) and "top-X percent" (where X=8%) methods of residue selection, where the proportion (and number) of correct predictions is 0.767 (2705/3527) and 0.769 (2712/3527), respectively. These are obtained through use of an alignment pre-filtering method that removes all aligned columns with a percentage of gaps greater than 60% and 90% respectively. Although, in general, there is no significant difference between the optimal results for any of the pre-filtered alignments, once the *colgap_percent* threshold is greater than or equal to 50%. These results represent a significant improvement over those obtained from using the *PAM30 UNGAPPED (0,0)* method of re-scoring, which uses all of the aligned amino acids in the functional assignment procedure and was found to be optimal in the analysis presented in the previous chapter.

With regards to the other *colgap_percent* thresholds that were investigated, there was a clear minimum, in classification performance, when setting the threshold to preclude all columns from the sub-alignments that contain any gaps. This result was expected to a certain degree, because the stringent alignment pre-filtering process is likely to cause the inclusion of a larger number of well conserved columns in the top-ranked set used for the functional re-scoring. The effect of this was the observed decrease in classification accuracy due to fewer columns that are sufficiently diverse and strongly correlated with the functional specificities of the aligned enzyme sequences. Also, the use of a strict filtering method, such as this, for the presence of aligned gap residues, means that alignments which have a greater degree of sequence diversity will in general have fewer columns of amino acids without any gaps in the alignment. This results in the observed sharp increase in the number of "empty subset (incorrect)" examples when the *colgap_percent* threshold is lowered towards 0%.

It has been shown that when using the "top-hit" assessment method for the large-scale analysis of the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset,

the *profile-HMM* method of sub-alignment generation does not perform as well as the *func-MB* method. This is the case for the functionally specific assignment accuracies obtained from each of the three thresholding methods used for sub-alignment generation. The optimal *profile-HMM* method – from the "top-N" (where N=35) columns selection method - shows an improvement (of 4.2% or 148 correct classifications) over that of the "all aligned amino acids" *PAM30 UNGAPPED (0,0)* re-scoring method. Although this is a statistically significant improvement (when considering a deviation of one standard error from the means of the bootstrap results) it is relatively small and is especially disappointing when viewed in comparison to the 13.8% (or 486 correct classifications) improvement in correct assignments when using the best *func-MB* based method.

It is not clear why the *profile-HMM* method for fSDR-based identification did not provide a larger improvement in performance, when re-scoring the generated sequence sub-alignments. Further study is required into the limitations of this method and the ways in which to obtain a better general improvement in specific enzyme classification accuracy through the use of this method.

When comparing the classification results obtained from functionally re-scoring the two fSDR-based sub-alignment generation methods, with those from the equivalent random column selection methods, a general improvement is seen. This is especially prominent for the fSDR selection thresholds used to generate the sub-alignments that produced the largest number of correct classifications. It was also shown that the random column selection model introduced in this chapter, generally results in a greater number of correct enzyme classifications than the previously described model of random sequence selection.

As in the previous chapter the effect of any potential sequence redundancy, within the query sequences used to create the benchmark datasets, was again investigated. A comparable sequence clustering method was used to assess the accuracy of the sub-alignment based functional classifications. As before, a number of sub-datasets (on which the alignment re-scoring experiments were repeated) were generated by applying a variety of sequence identity clustering thresholds to the query sequences. There was an overall reduction in the optimal number of correct classifications observed when applying the *func-MB* re-scoring method to the MSAs obtained from

the more stringently clustered sequence identity thresholds, such as 40%. In contrast, the *profile-HMM* method showed small improvements when comparing the results from using the MSAs obtained from progressively more stringent sequence identity clustering thresholds. These results, however, do not alter the conclusions that have already been drawn – that the classification results from using the *func-MB* re-scoring methods consistently outperform those from the comparable *profile-HMM* methods and, in general, similar results and trends were observed for each of the sequence identity cluster thresholds used.

The detailed study of the alignments of lactate and malate dehydrogenase sequences demonstrated, through a relevant experimentally well-studied example, how fSDR-based aligned subsets of residues can be used to improve the ranking enrichment of specific functions, when compared to the ranks generated by BLAST and other sequence similarity measures. These results show that there is significant improvement in the functionally specific sequence grouping and ordering, when using only the amino acids in the identified high-scoring fSDRs for functional scoring.

This particular example was chosen because the substrate binding specificity and other functional details of these enzymes has been well studied experimentally, which provides valuable experimental verification for some of the identified residues. Also, this example showed a good example of a case where the BLAST generated functional ranking results were the opposite of that required to make a correct specific functional assignment. This was clearly shown in *figure 4.12*, where the rank distributions of the functionally correct enzyme sequences are essentially inverted when only a subset of high scoring specificity determining residues were used to functionally score and order the aligned sequences, rather than the functional ordering generated by the original BLAST sequence similarity search. Although this is only one example, it clearly shows the potential of the approach for improving problematic specific functional classifications. It also highlighted a number of interesting factors regarding the operation of the methods and the sort of considerations necessary for further study and alternative methods for assessing the functional classification quality of the re-ranked sequences. In particular, whether any pre-filtering of the sequences in the multiple alignments should be carried out

before running fSDR score calculations and also, whether the relative ranking of sequences, other than the "top-hit", should be considered when assessing the functional classification accuracy.

In summary, the results in this chapter show a proof-of-concept for the use of a subset of amino acids – that are predicted to be indicative of specific protein functions - for improving the functional classification accuracy for enzyme sequences, when compared to standard sequence similarity measures. The large-scale benchmark study has shown that this approach does, in general, improve the classification accuracy when using functionally informative sub-alignments instead of functional scoring measures that use all of the aligned residues in BLAST generated sequence alignments. The following chapter aims to investigate ways to improve the methods used to assess the functional classification and enable the definition of a dataset of automatically identified fSDRs. The aim of which is their use in the training and validation of a machine learning approach for the automatic identification of functionally specific residues in sequence alignments and their subsequent use in the assignment of specific enzyme function.

Chapter 5 Towards the Automatic Identification of Functional Specificity Determining Residues Using Support Vector Machines

5.1 Introduction

In the previous chapter it was shown that functional specificity determining residues (fSDRs) could be used to significantly improve the classification of specific enzyme functional classes. These automatically identified fSDRs were used to define and re-score sub-alignments of enzyme sequences and the resulting classification accuracies were favourably compared to the functional re-scoring when using all aligned residues. There are, however, limitations to these methods of functional classification, for which possible solutions are presented and analysed in this chapter.

The main disadvantage of the two previously studied *func-MB* and *profile-HMM* methods for automatic identification of fSDRs, is the need for prior knowledge of the specific functional classes of the aligned sequences. This is problematic because it limits their use to alignments of functionally well-characterised sequences, thus preventing a more general approach to the classification problem and limiting the possible uses to a much reduced sample space of functionally annotated sequences. In an attempt to circumvent this requirement, it is proposed that machine learning methods could be used for the automatic identification of fSDRs in multiple sequence alignments (MSAs). The analysis in this chapter investigates the feasibility of this approach through the use of support vector machines (SVMs) (Vapnik, 1995) to discriminate between aligned columns of amino acids that are important for the determination of a specific function (fSDRs), or not ("non-fSDRs"). Support vector machines are classifiers that provide a means for distinguishing between different classes data. Features representing the input data are transformed into a multi-dimensional feature space through the use of kernel functions, which can be either linear or non-linear in form. It is then possible to identify a hyper-plane that provides an optimal separation of the two classes of

"positive" and "negative" data examples, which results in an associated set of unique kernel parameters found during the SVM training. In this chapter, the information contained within the aligned residues, associated with the two classes of positive (fSDR) and negative (non-fSDR) data, was encoded into input feature vectors, without reference to the functional classes of the associated protein sequences.

Unfortunately, there is not a well-established, large-scale and experimentally verified dataset of function specific residues that is suitable for this purpose. An enzyme specific database of catalytic residues, called the catalytic site atlas (CSA) (Porter et al., 2004), has been developed but it is not designed to catalogue the residues responsible for determining the substrate specificity of enzymes. Rather, this database concentrates on the identification and detailed classification of enzyme residues that are thought to be directly involved in the reactions catalysed by the associated enzymes. Because of this, these residues tend to be quite conserved across sequence homologs and are generally more indicative of the general enzyme functional class, rather than the specific functional sub-classes that are of interest in this thesis.

Therefore, it was necessary to first define a suitable benchmark dataset for these studies. For this, a method was developed for automatically characterising the aligned columns of amino acids in each MSA as either important for the determination of a specific function (fSDRs) or not (non-fSDRs). A modified form of the fSDR-based, sub-alignment, re-scoring method (which was introduced in the previous chapter) was used for this. For this, the previously studied "top-hit" functional assessment method was extended to include a measure of the "functional enrichment" of the top ranked enzyme sequences after re-scoring. Thus, allowing a suitable set of aligned fSDR columns to be identified for this particular problem.

To my knowledge, no previous studies have addressed this important problem by first using automated methods - to define a benchmark dataset of function specificity determining residues - and then using SVMs for their identification within multiple sequence alignments. There are, however, a number of previous studies (Gutteridge et al., 2003; Petrova and Wu, 2006; Tang et al., 2008) that have used machine learning approaches, such as artificial neural networks (NNs) and SVMs, for the classification of residues contained within the CSA database. These studies

demonstrated the feasibility of using machine learning methods for CSA identification from protein sequence and structural information. Although the identification of CSA residues is not the same problem as identifying specificity determining residues; these studies do provide inspiration for appropriate methods to use for pre-processing the data, prior to SVM training, as well as techniques for assessing the quality of the prediction results.

The fully automated identification of functional residues within proteins continues to be a challenging area of study. An equally important and related problem is the improved classification accuracy of the specific functional properties of enzymes, and other proteins, in a fully automated way. The following analysis aims to provide some novel ideas for building datasets and the use of SVMs towards solving these problems.

5.2 Materials and Methods

5.2.1 Datasets of Multiple Sequence Alignments

As in the previous studies, presented in *chapters 3* and *4*, datasets of multiple sequence alignments (MSAs) were used as the basis of the studies contained within this chapter. Two datasets of MSAs were primarily used.

5.2.1.1 The *"targets_only"* Dataset of MSAs

The *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, which was used in the previous chapter and consists of 3527 BLAST generated MSAs, was used in the following analysis for the assessment of the performance of the fSDR-based sub-alignment re-scoring methods. These alignments can be thought of as the *"targets_only"* dataset of MSAs because they have been filtered to only include sequences that have an annotated "target" enzyme functional classification (EC number) in the Swiss-Prot database (see *section 2.4.2* for details).

5.2.1.2 The "BLAST - raw output" Dataset of MSAs

Additionally, a second set of MSAs were used for the SVM analyses in this chapter. These differ from the *"targets_only"* dataset in that they have had no sequence filtering applied to the MSAs obtained from the BLAST sequence database searches. In particular the MSAs were not subject to the *"MSA target enzyme filtering"*

process or the "*All1stINCORRECT*" artificial dataset post-modification procedure, which were previously defined in *chapter 2*. Therefore, these alignments are referred to, throughout this chapter, as the *"BLAST - raw output"* dataset of MSAs.

5.2.2 The "Rank Enrichment" Method for Assessing the Accuracy of fSDR-Based Classification of Specific Enzyme Function

An additional method for assessing the performance and accuracy of specific enzyme functional classification, using fSDR-based alignment re-scoring, was investigated. This aims to build upon the limitations of the "top-hit" method, used previously (see *section 3.2.3.1* and *section 4.2.6.1*), by incorporating a measure of the ability of the functional re-scoring methods to group enzyme sequences that have the same functional specificity as the query sequence, in the top ranking positions after re-ranking using the alignment re-scoring procedure.

The proposed "rank enrichment" method aims to provide a score that calculates a measure of the overall change in the rank-ordering of sequences with the same specific enzyme functional classification as the query sequence. In order to achieve this, a method was implemented to calculate the number of "correct" (i.e. sequences with the same specific enzyme EC classification as the query sequence) sequences present in the top ranking, *N*, positions, after the application of a particular functional alignment re-scoring method. This is formally represented in *equation 5.1*.

$$E_{score} = \frac{N_{correct}}{N_{rank\_positions}} \qquad \textit{(equation 5.1)}$$

Where: E_{score} is the "functional enrichment score", which measures the enrichment level of the number of sequences with "correct" functional classifications - represented by $N_{correct}$ - that occur in the top ranking positions of interest - represented by $N_{rank\_positions}$. The E_{score} is bounded between a minimum of *0.0*, which is obtained when none of the sequences in the $N_{rank\_positions}$ show a correct functional match to the query, and a maximum of *1.0*, which is observed when all of the sequences in the $N_{rank\_positions}$ show a correct functional match to the query (i.e.

when $N_{rank\_positions} = N_{correct}$). Two examples that demonstrate the calculation of these functional enrichment scores are shown in *figure 5.1*.

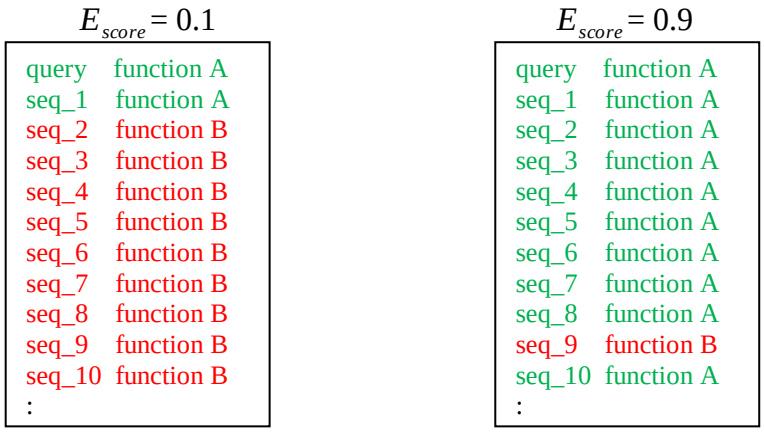

Figure 5.1. Two examples showing the way in which functional enrichment scores are calculated. On the left, only one sequence occurs with the same function as the query in the top-10 (i.e. $N_{rank\_positions} = 10$) ranking positions, leading to a score of E_{score} = 0.1. On the right, nine sequences are now found in the top-10 ranking positions, leading to an improved score of E_{score} = 0.9.

In the following studies, two values were used for the $N_{rank\_positions}$ parameter, these were: $N_{rank\_positions} = 10$; and $N_{rank\_positions} = N_{correct\_sequences\_in\_MSA}$. The value of *10* was used because it provides a calculation of the proportion of correct sequences occurring within the top-10 ranked positions after alignment re-scoring. Due to the way in which the functional class composition of the sequences within the MSAs of the "*targets_only*" dataset was defined (see *chapter 2, section 2.4.2*), this was an appropriate number of rank positions to consider. Each MSA in this dataset was defined to contain at least *10* sequences with an annotated EC classification identical to the query sequence. Therefore, the use of the top-10 ranked positions in the enrichment score, E_{score}, calculation ensures that it is theoretically possible for every MSA to return the maximum possible score of *1.0*, where all 10 top ranked positions are populated with functionally "correct" sequences. The alternative value that was used for $N_{rank\_positions}$, was $N_{correct\_sequences\_in\_MSA}$, which equals the number of sequences, in each MSA, with the same specific EC classification (i.e. "correct") as the query sequence. This means that this value will be variable between the different MSAs that constitute the analysed dataset, dependent upon the number of "correct"

functional sequences contained within each alignment. This form of the functional enrichment score aims to provide a description of how the alignment re-scoring methods affect the rank ordering and functional grouping of the "correct" enzymes that occur in the lower ranking positions (i.e. the sequences with "correct" functional classifications that are ranked outside the top-10 places). Also, this additional way of calculating E_{score} provides a means to differentiate between functional re-scoring methods that result in the same enrichment score, for a particular MSA example, when using only the top-10 ranked positions.

Again, this method creates a bounded value, between 0.0 and 1.0, for the enrichment score. Where the minimum is obtained when none of the sequences in the $N_{rank\_positions}$ show a correct functional match to the query, and a maximum of *1.0*, when all of the sequences in the $N_{rank\_positions}$ show a correct functional match to the query (i.e. when $N_{rank\_positions} = N_{correct}$). So, in this instance, a value of 1.0 can only be observed when all of the sequences that have the same functional classification as the query (i.e. $N_{correct\_sequences\_in\_MSA}$) are ranked above all the other sequences in the alignment that have "incorrect" classifications.

5.2.3 The Functional Alignment Re-scoring Procedures

The analyses within this chapter use the alignment re-scoring procedures previously described in both *chapters 3* and *4* (see *sections 3.2.2* and *4.2.4* respectively). The results from using these methods to functionally re-score the *"targets_only"* set of alignments were then analysed and their functional enrichment scores compared.

5.2.4 Definition of a Benchmark Dataset of Functional Specificity Determining Residues (*fSDRs)* within Enzymes

For these studies it was necessary to obtain a benchmark dataset of fSDRs that could be used for the training and validation of the optimal SVM parameters. There is no pre-existing data source that could be used as a "gold-standard" for a large-scale investigation of this type - which involves the automatic identification of amino acid residues that determine function specific sequence properties. Therefore, for these studies, it was necessary to develop a method for the selection and definition of a benchmark dataset of fSDRs for SVM training and validation. Due to the limitations

of available experimentally verified data, it was decided to use an automated method for this selection procedure and therefore this dataset can be thought of more as a "silver-standard" benchmark dataset of automatically selected fSDRs. The methods used in this selection procedure are outlined in detail below.

5.2.4.1 The Aligned Column Gap Percentage Threshold of Inclusion

For the analysis carried out in this chapter, a single percentage threshold was used for the removal of aligned columns from the sequence alignments, prior to the functional re-scoring of the sequences. This was described previously, in *chapter 4 – section 4.2.5.1*, as the *"column gap percentage threshold (colgap_percent)"* method. As before, the application of the *colgap_percent* threshold was only relevant to the *func-MB* and the *"random column selection method"* (see *chapter 4 – section 4.2.6.5*) methods of alignment re-scoring. For the following analyses a *colgap_percent* threshold of 10% was used. This particular threshold was selected because it was the same threshold as that used to pre-filter the MSAs in the study by Pazos et al. (2006), which first discussed the *func-MB* method of fSDR identification. A discussion of the possible limitations to this single threshold approach is provided later, in *section 5.3.4*.

5.2.4.2 Identification of the Optimally Performing Sub-sets of fSDR Columns in Each MSA

A method was used to identify fSDRs that generate the optimal functional enrichment scores, E_{score}, in each multiple alignment. For this, *equation 5.2* was used to calculate the difference in enrichment scores between the enzyme sequence ordering obtained from the original BLAST generated MSAs, in the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, and those from the re-scored sequence ordering.

$$diff_{enrichment\_score} = E_{score}^{OPTIMAL\_fSDR\_method} - E_{score}^{BLAST\_BLOSUM62} \qquad \textit{(equation 5.2)}$$

Where, for each MSA in the dataset: $diff_{enrichment\_score}$ represents the difference between the enrichment score from the optimal fSDR-based re-scoring method(s), represented by $E_{score}^{OPTIMAL\_fSDR\_method}$; and that from the original BLAST generated

sequence ordering, which uses the BLOSUM62 substitution matrix and gapped alignment scoring), represented by $E_{score}^{BLAST\_BLOSUM62}$. Because the aim was to identify the best performing fSDR subsets in each of the MSAs, the enrichment score differences were calculated for all of the *func-MB* and *profile-HMM* based sub-alignment generation and re-scoring methods that were studied in the previous chapter.

The enrichment score differences were calculated using the difference in functional composition of the top-10 ranking positions (i.e. where $N_{rank\_positions} = 10$ in *equation 5.1*). The distribution of these optimal score differences, for all of the 3527 MSAs, are shown in the black bars of the histogram in *figure 5.2*. This figure shows that there are a majority of examples with a positive $diff_{enrichment\_score}$ value. From *equation 5.2*, it can be seen that these are the examples that show an improvement in the number of "correct" enzyme functional sequences in the top-10 ranking positions of the re-scored alignments, when using the identified optimal fSDR-based re-scoring methods. The examples which have a zero $diff_{enrichment\_score}$ value represent those examples where there is no change in the enrichment score when using the "optimal" fSDR-based re-scoring methods. Also, the small numbers of examples with negative $diff_{enrichment\_score}$ values are those that show a reduction in the number of "correct" enzyme sequences in the top-10 ranking positions, with respect to the original BLAST MSAs, when the "optimal" fSDR-based re-scoring method is used.

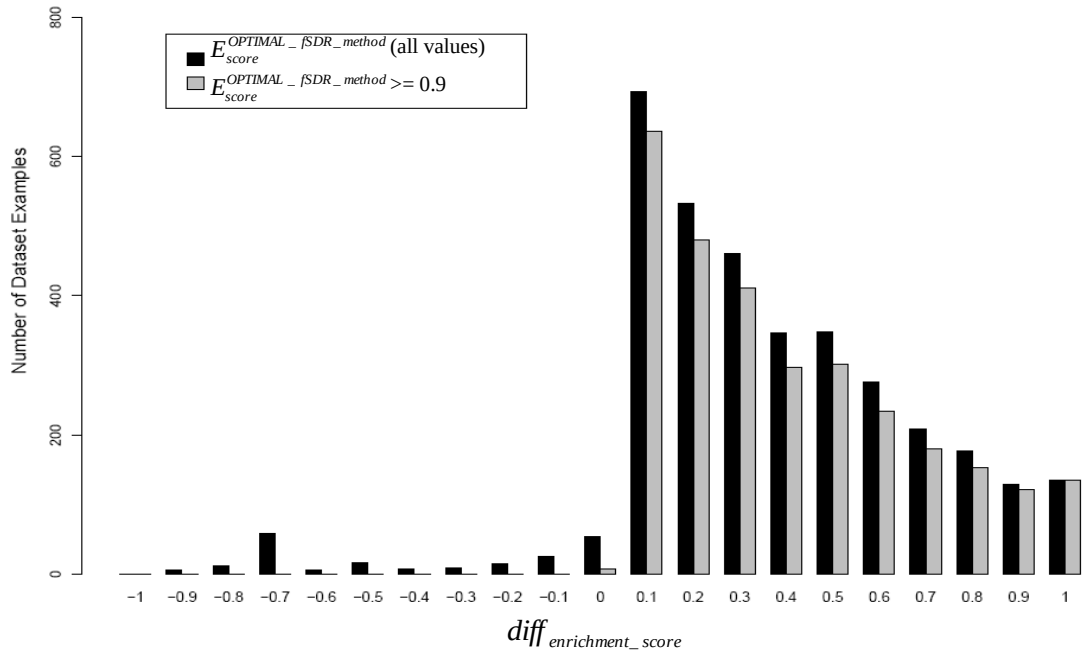

Figure 5.2. Histogram showing the differences, $diff_{enrichment\_score}$, calculated with equation 5.2, between the optimal top-10 functional enrichment scores and those from the original BLAST MSAs, in the All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001 dataset. Where the black bars represent the number of dataset examples with the specified $diff_{enrichment\_score}$ values, regardless of the $E_{score}^{OPTIMAL\_fSDR\_method}$ score (i.e. $E_{score}^{OPTIMAL\_fSDR\_method}$ (all values)); and the specified $diff_{enrichment\_score}$ values, when the optimal $E_{score}^{OPTIMAL\_fSDR\_method}$ score is greater than or equal to 0.9 (i.e. $E_{score}^{OPTIMAL\_fSDR\_method} >= 0.9$).

It can be seen, from *figure 5.2*, that there are a clear majority of dataset examples with an improved (i.e. positive) $diff_{enrichment\_score}$ value, when using the optimal fSDR-based re-scoring methods. However, this does not ensure any particular level of correct functional sequence enrichment in the top-10 ranking positions. Because the main aim of this analysis was the identification of a benchmark dataset of fSDRs that provide a clear differentiation between specific enzyme functions, only the aligned columns from the MSA examples that had an enrichment score of $E_{score}^{OPTIMAL\_fSDR\_method}$ >= 0.9, after sub-alignment re-scoring and ranking with the optimal fSDR-based methods, were considered for inclusion. Although this was a somewhat arbitrary threshold, it was decided that this was a suitable threshold of enrichment as it ensures that at least 9 of the top 10 ranking sequences are of the same "correct" specific enzyme function as the query sequence and therefore provides a 90% chance of a correct specific enzyme functional classification via

annotation transfer from the top-10 ranking sequence homologs after re-scoring. These examples are shown for comparison in *figure 5.2* and are represented by the grey bars.

Applying this enrichment score threshold meant that only a subset of the MSAs were included in the dataset from which the positive and negative classes, of fSDR and non-fSDR aligned columns respectively, were selected. From this data, a subset of 2959 MSAs satisfy both the enrichment score based selection criteria, of $E_{score}^{OPTIMAL\_fSDR\_method}$ >= 0.9, and also the criteria which ensures that the $diff_{enrichment\_score}$ value, relative to the original BLAST alignment sequence ordering, is greater than zero and therefore an improvement.

Once these 2959 MSAs had been identified, an additional stage was incorporated into the selection process for identifying the optimal subset of fSDRs. The purpose of this second selection step was the provision of a method for distinguishing between sub-alignment re-scoring methods that had equal values of the $E_{score}^{OPTIMAL\_fSDR\_method}$ enrichment score, when considering only the top-10 re-scored sequence ranking positions. An additional functional enrichment score was used for this, which took into consideration the change (and hence improvement) in the ranking of the enzyme sequences with the "correct" functional classifications, outside of the top-10. As described earlier, this was done by using $N_{rank\_positions} = N_{correct}$, in *equation 5.1*, when calculating the functional enrichment scores for each re-scored alignment.

After the application of this additional MSA selection procedure, examples were identified that continued to generate the same optimal, functional enrichment scores, $E_{score}^{OPTIMAL\_fSDR\_method}$, when using more than one different sub-alignment re-scoring method. For each of these examples, the re-scoring method which utilised the largest number of fSDR columns was identified and the associated fSDR columns were also identified for inclusion in the positive fSDR dataset. It was decided to include the largest possible subset of aligned columns in the positive (fSDR) dataset because it would maximise the amount of information available to the positive dataset for the training and validation of the SVMs. As a consequence the expected

(and indeed observed) disparity in the numbers of positive to negative class examples in the training and validation datasets was reduced.

Also, it was difficult to justify the use of any other method of positive class (fSDR) selection. This was primarily because a method that uses fewer fSDR columns, to achieve the same level of functional enrichment creates a situation where the negative set of non-fSDR columns contains examples that are positive (fSDR) examples if a different column selection method, with an identical enrichment score assessment criteria, is used. This could result in sub-optimal SVM learning through the inclusion of ambiguously classified positive data examples in the negative (non-fSDR) dataset. A possible solution to this could be the creation of a third set of aligned columns – the "unclassifiable" set - that are not considered as either part of the positive or negative set and therefore not used for the SVM training. It was decided that this would add an extra level of complexity to the selection of fSDRs and therefore, to keep the fSDR selection approach as simple as possible, was not used in this analysis.

These selection criteria were applied to the dataset of MSAs, and the associated methods for selecting the optimal subset of aligned columns (the fSDRs) were identified. The resulting fSDR and non-fSDR columns from each MSA example were separated into the positive and negative datasets, respectively, ready for encoding and use in the SVM analysis.

5.2.5 Removal of "Non-specific Serine/Threonine Protein Kinase" Query Sequence MSA examples

Analysis of the EC functional classes represented by the query sequences used to generate the 3527 MSAs in the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, revealed that a number of classes were represented a relatively large number of times. In particular, the EC 2.7.1.37 class, which specifically denotes enzyme sequences as belonging to the "non-specific serine/threonine protein kinase" functional class, was the functional annotation for 841 of the 3527 query sequences used to generate the dataset. The MSA examples generated by these query sequences were removed from the dataset of MSAs from which the SVM training and validation datasets of fSDRs were extracted. After this, a dataset of 2686 MSA examples remained.

The reason for this removal was that the sequences within these examples were only showing a general kinase functional relationship, and not the detailed differences in specificity, when being functionally re-scored to predict the EC class. Therefore, these MSAs were possibly more suitable examples for correctly predicting the 1st or 2nd EC class numbers, rather than all 4 levels required in a high-specificity enzyme functional annotation task.

5.2.6 Creation of SVM Cross-Validation Training Datasets

When carrying out training and validation for machine learning model parameter optimisation, it is important to have sufficiently non-redundant datasets to prevent over-training on many similar data examples. Also, it is important to define separate groups of non-redundant datasets for both the training and the independent validation stages. This is commonly achieved through the use of n-fold cross-validation. In this section the steps are described for the removal of sequence redundancy and the definition of cross-validation datasets for the training and validating the SVMs.

5.2.6.1 Query Sequence Clustering

Sequence identity based clustering was used to reduce the potential sequence-level redundancy of the query sequences used to generate the 2686 MSAs identified previously. The same query sequence clustering procedure as that used in the previous two chapters (see *sections 3.2.5* and *4.2.7*) was again followed. The CD-HIT algorithm (Li and Godzik, 2006) was applied to the query sequences, using a range of percentage sequence identity clustering thresholds, with the recommended default parameters. See *table 3.3* for a summary of the cluster properties, at each defined level of sequence identity, for the query sequences associated with the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset.

With regards to the SVM training and validation datasets, the main purpose of this sequence clustering was the overall reduction in sequence redundancy of the query sequence datasets. This has the associated effect of limiting any potential redundancy within the datasets of aligned residues (both fSDR and non-fSDR) extracted from the MSAs and therefore limits the potential for the SVM to over-learn an over-represented subset of data. Because of this, only the most stringent sequence identity threshold, of 40%, was used to reduce the redundancy of the dataset for the

subsequent SVM analysis. This resulted in a dataset containing 357 query enzyme sequences that represent 84 specific EC classes (see *Appendix I* for further description of this dataset).

5.2.6.2 Additional *BLASTCLUST* Query Sequence Clustering to Define the Non-Redundant Cross-Validation Datasets

An additional step to reduce the sequence redundancy was then applied to the 357 query enzyme sequences identified above. For this, the *BLASTCLUST* (Altschul et al., 1990) sequence clustering application was used. A stringent E-value based threshold, of 0.01, was used to cluster the query sequences by setting the *BLASTCLUST –e* parameter to be 0.01. This was done to remove any significant level of sequence homology that may have remained between the sequences assigned to different cluster groupings after the initial CD-HIT 40% sequence identity clustering. The outcome of the *BLASTCLUST* clustering process was 58 sequence clusters, with the smallest and largest clusters containing 1 and 47 query sequences, respectively.

A commonly used method for assessing the performance of machine learning classification methods (such as SVM) is that of n-fold cross-validation. For this, *n* equally sized datasets are defined and then used for training and validation purposes. A 5-fold cross-validation procedure was used to assess the SVM analysis carried out in this chapter. To do this, the 58 *BLASTCLUST* generated sequence clusters were randomly partitioned into five (approximately) equal sized groupings of sequence clusters, therefore ensuring no significant level of sequence homology between any sequences in the distinct groupings. Due to the fact that it was not possible to get exactly equal numbers of sequences into 5 groupings, from a dataset of 357, three of the groups contained 71 query sequences and the two remaining groups contained 72. These five dataset groupings are referred to as *"GROUP_1"* to *"GROUP_5"*. Finally, an all against all BLAST comparison, with an E-value threshold of 0.01, was carried out for each of the sequences in one group against those in the other four. This was done to ensure that the *BLASTCLUST* method had not missed any significant sequence homology between the five groups. The result of this analysis showed that there was indeed no significant sequence similarity overlap between the sequences in the five datasets.

5.2.6.3 Creation of the SVM Training and Testing Datasets

The purpose of the 5-fold cross-validation SVM training and testing procedure is the provision of a series of five separate training and testing datasets. These provide a method of optimising the machine learning parameters – using the training datasets - and a separate means of evaluating the performance of the learned parameters using additional data - the test datasets - that has not been used in the training. A commonly used method of partitioning the data into five pairs of (training and testing) subsets was followed here, which involves successively partitioning 4/5 of the data into the training data, with the remaining 1/5 of the data held back for use in testing.

Therefore, for the SVM analysis in this chapter, the five groups of MSAs (*GROUP_1 to GROUP_5*, defined above) were combined to form five distinct pairs of training (*TRAIN_1_2_3_4 to TRAIN_2_3_4_5*) and testing (*TEST_1 to TEST_5*) datasets. A detailed breakdown of the positive (fSDR) and negative (non-fSDR) SVM class compositions identified in these pairs of datasets is provided in *table 5.1*.

5.2.6.4 Random Balancing of the Positive and Negative SVM Classes

It can be seen from *table 5.1* that there is a disparity in the number of positive (fSDR) and negative (non-fSDR) examples in the SVM datasets. The non-fSDR columns occur at a greater frequency than the fSDR examples in all five of the testing and training datasets. When considering the data in *table 5.1* (obtained from using an E-value sequence inclusion threshold of 10^{-3}) the average ratio of negative to positive classes was 15.4 and 12.9 in the testing and training datasets, respectively. Datasets of this type are referred to as class "un-balanced" in the following discussion.

To improve the computational efficiency of the SVM training process, an additional set of training datasets was defined, which contained approximately equal numbers of fSDR (positive) and non-fSDR (negative) SVM class examples. Similar approaches have been used in a number of previous SVM and NN based studies (Gutteridge, et al., 2003; Petrova and Wu, 2006; Tang et al., 2008) on un-balanced datasets. To create these "balanced" datasets, each of the constituent MSAs was subjected to a procedure, which randomly selected from the negative class examples, a subset equal (where possible) to the number of positive (fSDR) examples within

the MSA. There were a small number of instances where complete equality was not possible, due to the fact that there were originally more positive than negative columns identified in a particular MSA and therefore it was not possible to select enough non-fSDR columns to compensate for this. This accounts for the ratios of positive to negative examples, in the randomly "balanced" datasets shown in *table 5.1*, being, in general, slightly less than 1. To improve the representative sampling of the negative class examples selected by the random balancing procedure, it was repeated five times for each of the MSA examples. Thus generating five randomly balanced sets of data associated with each training dataset.

With regards to training and testing the SVMs, the randomly balanced datasets are used in the training stage and the un-balanced versions of the associated testing datasets are used to assess the performance of the generated SVM models.

5.2.6.5 The Composition of the fSDR and non-fSDR Classes in the 5-fold Cross-Validation Datasets of Multiple Sequence Alignments

It is now possible to provide a final breakdown of the number of positive (fSDR) and negative (non-fSDR) classes that were found within the SVM training and testing datasets, defined above. This is best done through analysis of the data summary provided in *table 5.1*, which provides a comparison between the data composition of the ten individual sets of training and testing datasets. The table includes the following information for each dataset: (i) the "number of MSA examples" that constitute each of them; (ii) the "total number of aligned columns" contained within; (iii) the number of positive (fSDR) and negative (non-fSDR) aligned columns identified in each; and (iv) a comparison of the ratio between the number of negative and positive class examples within each. Also shown is a comparison between the number of non-fSDR columns (and subsequent ratios) in the datasets, both before ("un-balanced") and after ("randomly balanced") the process of randomly balancing the number of positive and negative classes within each dataset. Further, the composition of the datasets is shown for two alternative E-value thresholds used for controlling sequence inclusion within the MSAs. The details and relevance of these different thresholds is provided later in this chapter, in the results and discussion of the SVM analysis.

It can be seen that there is quite a wide variation in the ratio of negative (non-fSDR) to positive (fSDR) class examples in the individual groups of testing datasets (i.e. *GROUP_1* to *GROUP_5*). Due to the way in which these groupings of MSAs were constructed, through the *BLASTCLUST* based non-redundant clustering and subsequent random partitioning of these clusters, it was not possible to avoid this outcome. It can, however, be seen that in the larger, combined sets of training datasets the ratio of un-balanced class differences is less wide spread and therefore more comparable between the different datasets.

Dataset	Number of MSA Examples	Total Number of Aligned Columns	POSITIVE (fSDR) Columns		NEGATIVE (non-fSDR) Columns				Ratio of classes (NEGATIVE/POSITIVE)			
					Un-balanced		Randomly Balanced		Un-balanced		Randomly Balanced	
			10^{-3}	10^{-15}	10^{-3}	10^{-15}	10^{-3}	10^{-15}	10^{-3}	10^{-15}	10^{-3}	10^{-15}
***GROUP_1* (*TEST_1*)**	71	50213	1682	1605	48531	48608	1682	1605	28.9	30.3	1.00	1.00
***GROUP_2* (*TEST_2*)**	72	37909	4515	3810	33394	34099	4349	3644	7.4	8.9	0.96	0.96
***GROUP_3* (*TEST_3*)**	71	35834	2558	2408	33276	33426	2402	2264	13.0	13.9	0.94	0.94
***GROUP_4* (*TEST_4*)**	72	32807	1598	1509	31209	31298	1547	1462	19.5	20.7	0.97	0.97
***GROUP_5* (*TEST_5*)**	71	28944	3127	2780	25817	26164	2982	2659	8.3	9.4	0.95	0.96
TRAIN_1_2_3_4	286	156763	10353	9332	146410	147431	9980	8975	14.1	15.8	0.96	0.96
TRAIN_1_2_3_5	285	152900	11882	10603	141018	142297	11415	10172	11.9	13.4	0.96	0.96
TRAIN_1_2_4_5	286	149873	10922	9704	138951	140169	10560	9370	12.7	14.4	0.97	0.97
TRAIN_1_3_4_5	285	147798	8965	8302	138833	139496	8613	7990	15.5	16.8	0.96	0.96
TRAIN_2_3_4_5	286	135494	11798	10507	123696	124987	11280	10029	10.5	11.9	0.95	0.95

Table 5.1. A breakdown of the number of positive (fSDRs) and negative (non-fSDRs) aligned columns that contribute to each of the SVM training and testing datasets. This table includes the following information for each dataset: (i) the "number of MSA examples" that constitute each of them; (ii) the "total number of aligned columns" contained within; (iii) the number of positive (fSDR) and negative (non-fSDR) aligned columns identified in each; and (iv) the ratio of the number of negative and positive class examples within each. Also shown is a comparison between the number of non-fSDR columns (and subsequent ratios) in the datasets before ("un-balanced") and after ("randomly balanced") the process of randomly balancing the number of positive and negative classes. Also shown are comparisons between the dataset contents when using two E-value thresholds (10^{-3} and 10^{-15}) to control sequence inclusion in the MSAs. These are indicated by the column headings of 10^{-3} and 10^{-15} respectively.

5.2.7 SVM software, kernels and learning parameters used

For the SVM analysis, the *SVM*^{light} (Joachims, 1999) and *SVM*^{perf} (Joachims, 2006) software applications were used. Two different learning kernels were used; with the radial basis function (RBF) kernel used in *SVM*^{light} and the linear kernel used in *SVM*^{perf}. The *SVM*^{perf} application was used for the linear kernel based learning because it provides significant improvements in computational efficiency when compared to the linear kernel learning capabilities of *SVM*^{light}. However, these same improvements are not available for RBF kernel based learning and therefore *SVM*^{light} was used for this purpose. A grid search optimisation of the *C* and *gamma* learning parameters was carried out for the relevant SVM kernels, using the guidelines suggested by Hsu et al., (2008), to optimise the learning models generated during the SVM training.

5.2.8 SVM Feature Vector Encoding

Below are descriptions of the methods that have been used to encode the aligned column information into the necessary SVM feature vector format for input into the *SVM*^{light} and *SVM*^{perf} applications. The following feature encoding was carried out in the same way for both the positive (fSDR) and negative (non-fSDR) subsets of data, which consist of aligned columns of amino acids taken from the MSAs that constitute the training and testing datasets.

5.2.8.1 The Amino Acid Composition

A feature vector was calculated to represent the amino acid composition of each aligned column of residues. This was represented by a vector of length 22, which represented the fractional occurrence of each of the 20 standard amino acid types, as well as two additional fractional occurrences for the number of gaps and also the number of unidentified, masked "X" residues within each of the aligned columns being encoded. All of these values were calculated as fractional frequencies of occurrence and therefore they all lie within the range from 0.0 to 1.0, inclusive. This feature vector is referred to as the "amino acid composition (*AA_composition*)" where relevant during this analysis.

5.2.8.2 The Number of Amino Acid Types

Another feature used to represent the aligned residues, was "the number of amino acid types (*NumberOfAATypes*)", which describes the number of distinct amino acid types within a particular aligned column. For the encoding of this feature only the occurrence of the 20 standard amino acid types within the aligned column was considered.

Initially, a simple count of the number of distinct amino acid types found within an aligned column of interest was considered, therefore generating a feature vector with a discrete and bounded value, ranging from 0 to 20, inclusive. Where: 0 signifies that there are no standard amino acid types occurring in the column (and therefore could contain either all "X" residues, or a mixture of gaps and "X" residues), and 20 signifies that all of the standard amino acid types occur, at least once, within the aligned column. The *NumberOfAATypes* feature was then modified, to incorporate a threshold, based on the percentage frequency of occurrence of the amino acid types found within the aligned column of interest. This will be referred to as the "*NumberOfAATypes_threshold_X%"* feature, where *X* represents the applied percentage threshold.

Formally, the percentage frequency of occurrence of each distinct amino acid type, f_{AA}, with respect to all the residues in the aligned column, was calculated. Then, for each of the distinct amino acid types within the aligned column, if the percentage frequency of occurrence, f_{AA}, was greater than or equal to the applied percentage threshold of occurrence, *X*, the amino acid type was classed as occurring within the aligned column and therefore was included in the *NumberOfAATypes_threshold* feature value for the column. For example, consider a column containing 100 aligned residues and 4 distinct residue types, with frequencies of occurrence of 65, 30, 3 and 2. If a percentage threshold of *X=5%* was then applied, the resulting *NumberOfAATypes_threshold* feature would have a value of 2, because only 2 of the amino acid types occur with a percentage frequency greater than or equal to the specified threshold of 5%.

This modification to the *NumberOfAATypes* feature (i.e. use of a threshold) aims to reduce noise in the feature from relatively small instances of an amino acid type

within an aligned column. A more detailed discussion of the threshold selection is provided below, in *section 5.3.2.4*.

Finally, for the purposes of SVM optimisation, it is generally recommended to have all feature values of comparable numerical magnitude and range. Therefore, to be comparable to the *AA_composition* feature values, the *NumberOfAATypes* feature was subsequently re-scaled to a value within the range of 0 to 1.

5.2.9 Assessment of the SVM Model Classification Performance

Three measurements were used to assess the predictive performance of the SVM classifiers. These were: the true positive rate (TPR); the false positive rate (FPR); and the Matthews correlation coefficient (MCC), and are defined below in *equations 5.3*, *5.4* and *5.5*, respectively.

$$TPR = \frac{TP}{TP + FN} \qquad \textit{(equation 5.3)}$$

$$FPR = \frac{FP}{FP + TN} \qquad \textit{(equation 5.4)}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}} \qquad \textit{(equation 5.5)}$$

Where TP, FP, TN, and FN represent the number of examples that are correctly classified as belonging to the positive class (i.e. true positives), the number that are incorrectly classified as belonging to the positive class (i.e. false positives), the number correctly classified as belonging to the negative class (i.e. true negatives), and the number incorrectly classified as belonging to the negative class (i.e. false negatives), respectively. As in the studies of Petrova and Wu (2006) and Gutteridge et al. (2003), because of the un-balanced nature of the testing datasets the MCC is used to assess the performances of the SVM classifications.

Also, the chi-squared test, with one degree of freedom, is used to assess the statistical significance of the MCC values, using *equation 5.6* (which is defined in Baldi and Brunak (2001)).

$$\chi^2 = N \times MCC^2 \qquad \textit{(equation 5.6)}$$

Where: N is the total number of predictions made by the classifier; and the chi-square statistic measures whether the prediction is significantly better than random (i.e., an MCC value of 0).

5.3 Results and Discussion

Before presenting the results obtained from using SVM classifiers to automatically identify fSDRs, an analysis is shown that introduces the use of the "functional rank enrichment" method to assess the level of correct specific functional classification. This method, described earlier, incorporates a measure of the functional re-scoring method's ability to group enzyme sequences - with the same functional specificity as the query sequence - in the top ranking positions. The results presented in the following section are an extension of the "top-hit" based functional re-scoring results that were obtained from the experiments in the previous two chapters. They also serve to show the reasoning behind the methods used to define the dataset of positive (fSDR) and negative (non-fSDR) examples that were used in the subsequent analysis of the SVM classifiers in this chapter.

5.3.1 Using the "Functional Rank Enrichment" Method for Assessing Specific Enzyme Functional Classification

The following analysis is used to investigate an alternative to that of the "top-hit" functional assessment method that has been used previously in this thesis (see *sections 3.2.3.1* and *4.2.6.1*). The methods for assessing the functional classification and differentiation between specific enzyme sub-classes used in this part of the analysis were developed to include a measurement of the changes in specific functional grouping after re-scoring of the sequence alignments.

A large-scale study was carried out into the functional enrichment scores obtained from re-scoring selected sub-alignments of the MSAs contained within the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset. Both the *func-MB* and *profile-HMM* methods for computational fSDR identification were used, along with the three column selection methods used in the previous chapter (see *section 4.2.4.2*) for the associated fSDR-based sub-alignment generation. This analysis allowed a comparison between the enzyme enrichment scores resulting from the best performing methods from each of the alignment re-scoring methods.

5.3.1.1 Comparisons Between Functional Enrichment Scores

To maintain consistency of approach with the work carried out in the previous chapter, in which the "top-hit" method of functional assessment was used, the functional enrichment scores were calculated for the re-scored sequence alignments obtained from these previous analyses. Both the *func-MB* and *profile-HMM* fSDR identification methods were compared and each of the "top-N", "top-X percent" and "column score threshold" methods of aligned column selection, and subsequent sequence sub-alignment generation, were studied. The PAM30 amino acid substitution matrix was again used to calculate and re-order the resulting pair-wise amino acid comparisons, with all residue-residue pair comparisons involving gap characters scored as zero.

The functional enrichment scores for the 3527 MSAs in the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset were first calculated. *Figure 5.3(a)* shows a comparison between re-scoring the sub-alignments, when using the "top-X percent" method of sub-alignment generation. The figure provides a breakdown of the functional enrichment scores, E_{score}, calculated from the top-10 ranking positions (i.e. from using a value of $N_{rank\_positions} = 10$ in *equation 5.1*) after sub-alignment re-scoring. Ten distinct score ranges are shown that represent the E_{score} values between 0 and 1.0. An identical range of "top-X percent" column selection thresholds were applied to both the *func-MB* and *profile-HMM* fSDR scoring methods, allowing a comparison between the number of dataset examples that result in specific enrichment scores from each of these methods.

A key observation from *figure 5.3(a)* relates to the correlation between the "top-X percent" threshold and the fractions of examples with a top-10 functional enrichment score that is greater than or equal to 0.9. The *func-MB* and *profile-HMM* methods both show larger fractions of examples in this highest enrichment score range as the re-scored sub-alignments are generated with progressively lower percentage selection thresholds. This is because they contain smaller numbers of aligned columns that are more highly correlated with the specific functions of the aligned enzyme sequences. Therefore, resulting in an enhanced number of functionally

"correct" enzyme sequences (nine or more in this case) within the top-10 ranking positions, after alignment re-scoring.

Regarding the *func-MB* method, the optimal performance within this enrichment score range was obtained by re-scoring the sub-alignments generated from the highest scoring 5% (top-5%; X=5%) of the fSDRs (as calculated from their Spearman rank-order correlation coefficients). Whereas, for the *profile-HMM* method, the top-7% (X=7%) of the fSDRs, as calculated from the associated Z-scores, produced the optimal performance. These two optimal re-scoring methods resulted in 45% (1601 out of 3527) and 43% (1534 out of 3527) of the dataset examples with a functional enrichment score greater than or equal to 0.9, when considering the *func-MB* and the *profile-HMM* based methods, respectively. This shows that the *func-MB* method shows slightly more examples than the *profile-HMM* method in this functional enrichment score range. However, the overall results are closely comparable between the two sub-alignment generation methods for all of the enrichment score ranges.

Analysis of the enrichment score ranges between 0.1 and 0.9 ($0.1 \leq E_{score} < 0.9$), generally shows a different trend to the results seen in the $0.9 \leq E_{score} \leq 1.0$ score range. That is, in all of these individual score ranges there are consistently more examples observed when re-scoring the sub-alignments generated from larger percentages of aligned columns (e.g. when the "top-X percent" threshold is X=100%, rather than a lower value of X=10%). This is to be expected and is mainly due to the way in which the functional enrichment score data has been partitioned and presented in *figure 5.3(a)*. This presentation means that each individual method of threshold selection that is shown has to have a total number of examples equal to the total number of examples in the dataset under investigation (or, the summed fraction of dataset examples, for each re-scoring method must equal 1.0).

When analyzing the results from the poorest performing functional enrichment score range, where $E_{score} < 0.1$, exceptions to the previously observed trends occur and the data interpretation becomes less clear. There are a relatively large number of examples within this score range, when compared to the other score ranges below 0.9. This can be partly explained by the fact that it is the enrichment score range in

which any “empty subset (incorrect)” examples will appear, which are defined as having an E_{score} of 0.

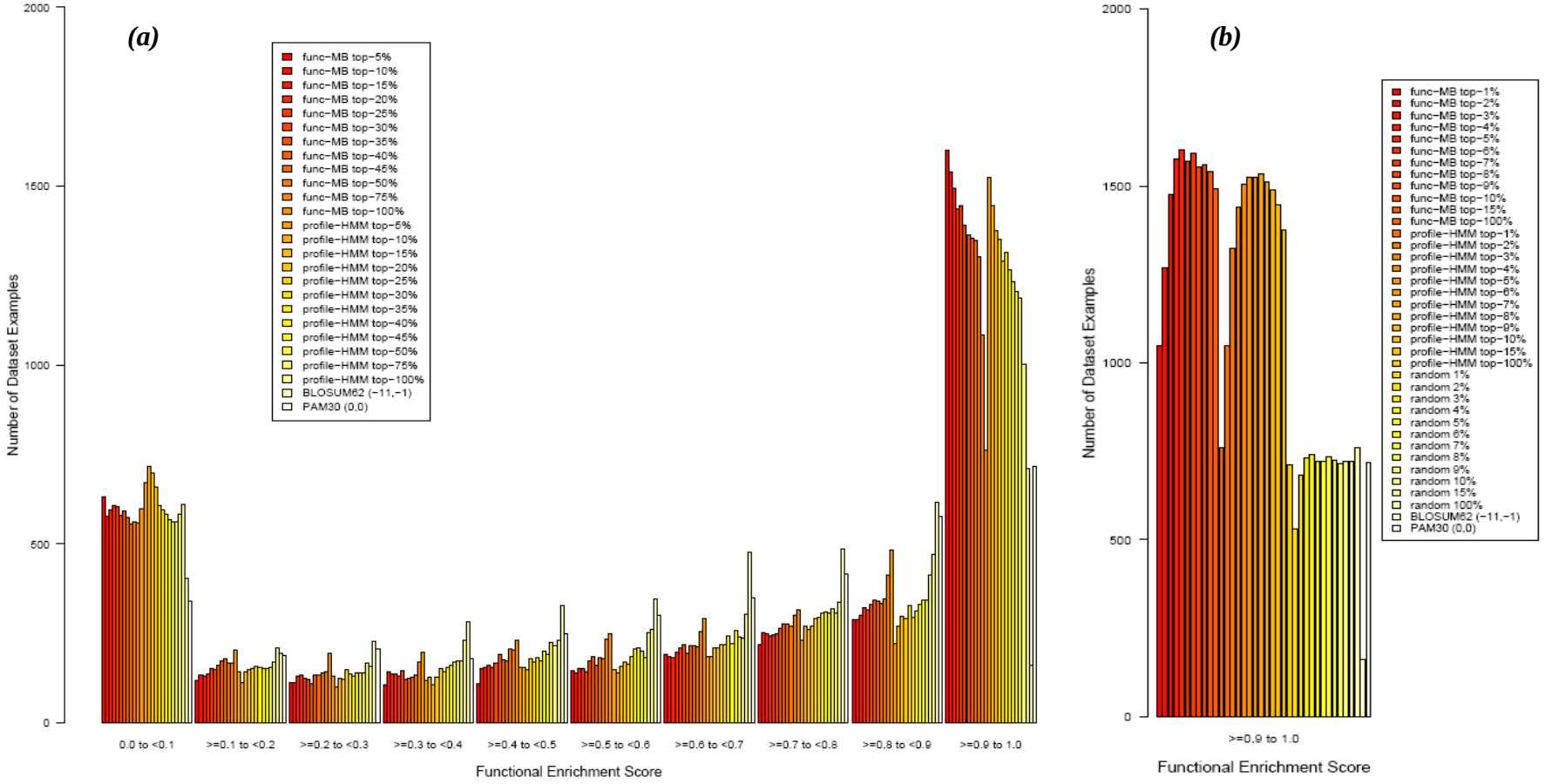

Figure 5.3. A comparison of the number of dataset examples obtained in each of the defined ranges of functional enrichment scores, when using the "top-X percent" method of sub-alignment re-scoring, from func-MB and profile-HMM identified fSDRs. Also shown are the results from the BLOSUM62 (-11,-1) and PAM30 (0,0) methods of alignment re-scoring. (b) Highlights the results for which the functional enrichment score was >=0.9. Also shown are the results for re-scoring with the comparable "random-X percent" sub-alignment generation methods.

This is, however, a somewhat disappointing result; especially considering the fact that the number of examples resulting in these poor functional enrichment scores does not appear to be reduced by using smaller sub-alignments of more highly correlated fSDRs. In fact, for both the *func-MB* and *profile-HMM* methods, a slight tendency to the opposite trend is shown. The presence of these under-performing cases; coupled with the observation that even the best performing single sub-alignment re-scoring method only results in 45% of the dataset examples, when $E_{score} \geq 0.9$, suggested that an alternative approach should be followed for the definition of an optimal dataset of fSDRs, for use in the SVM analysis. This approach was explained in detail, in *section 5.2.4.2*, and provides a more exhaustive search space for finding the optimal subset of fSDRs than a single threshold for fSDR selection could. Therefore, this approach provides a benchmark set of fSDRs, for the SVM classification experiments, that has the best possible functional enrichment scores and also maximizes the number of examples from which the SVM data is comprised.

Also shown in *figure 5.3(b)* are the functional enrichment results from re-scoring sequence sub-alignments that have been generated through the "*random column selection method*", previously described in *section 4.2.6.5*. To provide a comparison with the results from the *func-MB* and *profile-HMM* generated sub-alignments, the same "top-X percent" thresholds (of: *X=1-10% (by 1%); X=15%;* and *X=100%*) are shown for all methods. Although this only compares a selection of top-X percent thresholds, it successfully illustrates the generally observed behaviour for the *random column selection method* in the enrichment score range of $0.9 \leq E_{score} \leq 1.0$. That is, much larger numbers of examples with high functional enrichment scores are observed when using the high scoring fSDRs, identified with the *profile-HMM* and *func-MB* methods, rather than the randomly generated equivalents. Therefore, these results show that there is no clear correlation between the sub-alignment size and improved ranking performance, for "correct" enzyme sequences, when re-scoring randomly selected columns from the sequence alignments.

For brevity, a detailed analysis of the functional enrichment scores obtained from the "top-N" and the "column score threshold" re-scoring methods (described in the previous chapter) is not provided. The optimal performing results for these methods

are, however, shown in *table 5.2*. In summary, the results from these other two methods show similar trends to those seen with the "top-X percent" sub-alignment generation methods, and also the comparable "top-hit" assessment results seen in the previous chapter. That is, for the "top-N" column selection method, a generally improved performance was observed when using smaller subsets of fSDRs that are most highly correlated with the specific functional classes. And the "column score threshold" based results show a peak in peak performance, before a rapid deterioration due to the increasing numbers of "empty set (incorrect)" examples at high column scores.

5.3.1.2 Comparison Between the Functional Enrichment Score Results for the Optimal Alignment Re-Scoring Methods

To complete this analysis of the functional enrichment score improvements, comparisons between the optimal methods are given in *table 5.2*. This table summarises the fraction (and number) of dataset examples that result in an enrichment score greater than or equal to 0.9. The optimal methods shown were taken from each of the "top-N", "top-X percent" and the "column score threshold" methods, used with the *func-MB* and *profile-HMM* methods for scoring potential fSDRs. Also shown, in *table 5.2*, are the results from using the *PAM30 (0,0)* method of "un-gapped" sequence alignment re-scoring. This was shown to be the best performing functional re-scoring method from the alternative amino acid substitution studies, analysed in *chapter 3*. Finally, a contrast is provided to the sequence ordering of the functionally "correct" enzymes in the original "artificial" dataset of BLAST generated MSAs (i.e. from the *All1stINCORRECT.tF.BLOSUM62.masked.E0.001* dataset). This is denoted in the table as *"BLOSUM62 (-11,-1)"*, referring to the fact that the BLOSUM62 matrix was used in the sequence database search and that the gap penalties used were -11 and -1.

These results show that the number of examples with an enrichment score greater than or equal to 0.9 was similar for five of the optimally performing sub-alignment re-scoring methods. These were: the *func-MB* "top-5 percent" and *profile-HMM* "top-7 percent" methods, with dataset fractions of occurrence of 0.45 (1601/3527) and 0.43 (1534/3527), respectively; the *func-MB* "top-10" and *profile-HMM* "top-

15" methods, with dataset fractions of occurrence of 0.45 (1594/3527) and 0.43 (1531/3527), respectively; and the *profile-HMM* "Z-score threshold >= 2.0" method, with a dataset fraction of occurrence of 0.43 (1533/3527). The remaining sub-alignment re-scoring result, using the *func-MB* "Spearman-rank order correlation threshold >= 0.2" method, shows a slightly lower optimal value for the dataset fraction of occurrence, of 0.37 (1304 out of 3527).

			fraction (number) of dataset examples (when $0.9 \leq E_{score} \leq 1.0$)
func-MB	top-N	top-10	0.45 (1594)
	top-X percent	top-5%	0.45 (1601)
	column score	>=0.2	0.37 (1304)
Profile-HMM	top-N	top-15	0.43 (1531)
	top-X percent	top-7%	0.43 (1534)
	column score	>=2.0	0.43 (1533)
amino acid matrix	PAM30 (0, 0)		0.20 (718)
	BLOSUM62 (-11, -1) (*)		0.05 (162)

Table 5.2. A summary of the sub-alignment re-scoring methods that generate the largest number of examples with a functional enrichment score greater than or equal to 0.9 (when $N_{rank\_positions}$=10). The fraction (and number) of dataset examples within this enrichment score range are shown for each method. The "amino acid matrix" re-scoring methods show comparable results for the optimal amino acid substitution re-scoring matrices and gap penalties previously identified in chapter 3 (see table 3.4). (*) indicates the sequence ranking results for the original BLAST All1stINCORRECT.tF.BLOSUM62.masked.E0.001 MSAs (i.e. generated through a residue masked sequence database search with the BLOSUM62 matrix and gap penalties of -11 and -1).

All six of these sub-alignment methods show a clear improvement when compared to the *PAM30 (0,0)* and *BLOSUM62 (-11,-1)* methods, which result in dataset fractions, within the $0.9 \leq E_{score} \leq 1.0$ score range, of 0.20 (718/3527) and 0.05 (162/3527), respectively. Therefore, these results demonstrate a consistent improvement in the number of examples with high functional enrichment scores, when re-scoring sub-alignments of residues that are predicted to be functionally important. With the largest overall improvements, of 883 dataset examples (or 25% of the total dataset) and 1439 dataset examples (or 40% of the total dataset), seen when comparing the

functional re-scoring results of the *func-MB* "top-5 percent" method to those of the *PAM30 (0,0)* and *BLOSUM62 (-11,-1)* methods, respectively.

In conclusion, the functional enrichment score improvements shown are closely comparable to those seen in the previous chapter, when using the "top-hit" method of assessment for specific functional classification. That is they show that enrichment of the "correct" functional sequences is much improved when using sub-alignments of functionally important residues to re-score and re-rank the aligned sequences; rather than using all aligned residues or randomly selected sub-alignments of residues. However, they also demonstrate that a single fSDR selection method is not optimal for all of the MSA examples. This highlighted the need for a more flexible approach when defining an optimally performing dataset of fSDRs for use in SVM experiments. Hence the development of the more exhaustive process for identifying optimally performing sub-sets of fSDR columns, which was previously described in *section 5.2.4*.

5.3.2 Analysis of the SVM Classification Performance

To conclude the enzyme functional classification studies presented in this thesis, a set of analyses were carried out to investigate the use of support vector machines (SVMs) for the identification of functionally specific residues in aligned sequence homologs. The previous work in this chapter has outlined the collection and definition of the training and testing datasets of fSDRs for this task.

These initial SVM experiments concentrate on a small number of simple protein sequence features that were expected to be of importance for the determination of specific enzyme functional properties. Therefore, the following results and conclusions are intended as initial studies into the feasibility of using the defined datasets, with a selection of commonly used SVM kernels and learning parameters, for the identification of fSDRs from multiple alignments of sequences without prior knowledge of their specific functional classifications.

5.3.2.1 Datasets

The feature vector encoding for the training and testing of the SVM classifiers was carried out using the MSAs from the *"BLAST - raw output"* dataset. This data was used because it was expected that it would provide additional evolutionary sequence

information when compared to the MSAs of the *"targets_only"* dataset and therefore enhance the machine learning performance. This is because the *"targets_only"* MSAs only contain sequences from a carefully selected set of fully annotated enzyme sequences. Whereas the *"BLAST - raw output"* alignments generally contain a larger sample of sequence homologs. Also, the overall aim of the SVM analysis is the development of an automated classification system that can identify functionally informative regions within a multiple sequence alignment without any prior knowledge of the functional classification of the constituent protein sequences. Therefore, it was deemed sensible that data of this form, with a minimum amount of alignment pre-processing, was used in these studies towards the development of the classifiers. To aid the SVM learning, and improve computational efficiency, all of the following experiments use the "randomly balanced" form of the datasets for training the SVMs. The un-balanced form of the encoded MSA data is used to assess the classification performance of the learned models on each of the test datasets. This is to ensure that the classification performance is assessed on the type of alignment data that would be expected in a novel fSDR classification problem.

5.3.2.2 Optimisation of the SVM Learning Parameters

The performances of both the linear and radial basis function (RBF) learning kernels were investigated, using the SVM^{perf} and SVM^{light} software applications, respectively. In the case of the linear kernel based learning, the *C* parameter (i.e. the *–c* command line parameter in SVM^{perf}) was progressively altered between a minimum value of $1x10^{-5}$ and a maximum value of $1x10^{7}$, to provide a thorough analysis of the SVM performance. For each *C* parameter used, the classification performance of the resulting SVM models was assessed for each of the 5-fold cross-validation datasets. When using the RBF kernel, both the *C* and *gamma* parameters (i.e. the –c and –g command line parameters respectively, in SVM^{light}) were systematically varied to carry out a thorough analysis of the pairs of learning parameters. For this, a grid search method was used; where exponentially growing sequences of *C* and gamma parameters were used, with $C = 2^{-5}, 2^{-3}, ..., 2^{15}$ and $gamma = 2^{-15}, 2^{-13}, ..., 2^{5}$ (Hsu et al., 2008).

To assess the fSDR classification performance of the SVM training parameters the Matthews Correlation Coefficient (MCC) was calculated for each of the SVM

models. The classification performance of the resulting SVM models was assessed for each of the five cross-validation datasets. For each set of training parameters the MCCs obtained from the classification performance on the associated test sets were averaged. The cross-validation was then repeated for each of the five training sets of randomly balanced fSDR and non-fSDR aligned columns of residues. It was then possible to identify the learning parameters (for both linear and RBF kernels) that provided the optimal classification performance. This was done by identifying the SVM parameters that provided the largest MCC value when averaged across the five cross-validation datasets and the five randomly balanced training sets.

5.3.2.3 The Effect of the Amino Acid Composition Features on the SVM Classification Performance

SVM training was carried out using the 22 "amino acid composition (*AA_composition*)" features as the input feature vectors, which were calculated from each of the *"BLAST - raw output"* alignments in the 5 groups of training and testing datasets. The classification statistics for each of the models were averaged (using the mean) across the randomly balanced training and test set groupings. The optimal results are shown in the *AA_composition (E-value <= 0.001)* section of *table 5.3*.

Comparisons between the RBF and linear kernel classification results showed that better MCC scores are generally observed when using the RBF kernel when training the SVMs. The parameters found to give the best overall fSDR classification performance with the RBF kernel, were *C=0.125* and *gamma=8.0*. Analysis of these results shows that a large number of false positive predictions were being made by the fSDR classifiers, resulting in a large average false positive rate of 0.33. This was the case for all five of the cross-validation datasets, with some variation depending on the particular test set used for the assessment, with a minimum FPR of 0.22 and a maximum of 0.40.

With regards to the MCC results, it can be seen that although the MCC values are generally quite low, with an average of 0.30, the results do show that the SVM classification is performing better than a random predictor, due to the MCC values being greater than 0. This is true even for the lowest MCC value, of 0.18, seen with the *TEST_4* dataset. Nevertheless, the relatively low MCC results and the high rate

of false positives are disappointing. Because of this, further analysis of the input data and the investigation of other possible features for encoding the functionally specific information were investigated to try to aid the classification process.

5.3.2.4 Analysis of the Sequence Alignment Data and Additional Features

The poor quality of the SVM classification results – using the *AA_composition* features calculated from the *"BLAST - raw output"* MSAs - led to the consideration of using additional features for encoding the sequence alignment information and also the analysis of the signal quality of the input sequence alignment data. For this, an investigation of the benefits resulting from the use and incorporation of the *NumberOfAATypes* feature, into the SVM based classification of fSDRs, was carried out. Also investigated was the effect of altering the E-value threshold used for controlling the inclusion of sequences in the MSAs.

Analysis of the *NumberOfAATypes*

The *NumberOfAATypes* SVM feature vector was generated by calculating the number of distinct types of the 20 standard amino acid residues occurring in the fSDR or non-fSDR aligned columns. A comparative analysis of the distributions of the number of amino acid types in both the fSDR (positive SVM class) and non-fSDR (negative SVM class) columns was then carried out. When considering the data from the *"BLAST - raw output"* MSAs, the analysis showed that both the fSDR and non-fSDR columns contained a relatively high frequency of occurrence of examples with large numbers (i.e. greater than 15) of distinct amino acid types. This was a surprising observation because it was expected that the non-fSDR columns would show an increased tendency to contain larger numbers of amino acid types than the "positive" fSDR columns. This is because the fSDR columns were identified as those that provide an optimal level of rank-order improvement of aligned sequences with "correct" specific enzyme functional classes and were therefore expected to display a smaller number of amino acid types, in general, than the non-fSDR columns which have no correlation to the functional specificity.

This observation led to a hypothesis that the input sequence alignments (and therefore the encoded data from the fSDR and non-fSDR aligned columns) may contain a relatively large amount of noise, which could be causing SVM

classification problems. To investigate this, two methods were used. Firstly, a more stringent (i.e. lower) E-value was used to control the sequences included within the encoded MSAs; and secondly, a percentage threshold was applied to the calculation of the number of distinct amino acid types within each aligned column.

The Application of an Additional E-value Sequence Inclusion Threshold before Encoding the Multiple Sequence Alignments

The sequence alignments used so far in the SVM analysis were taken from the *"BLAST - raw output"* dataset. In contrast, the majority of the analysis that has been previously carried out in this thesis was derived from the more carefully defined MSAs of the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset, which contain only well annotated "target" enzyme sequences. Therefore, it is possible that the use of the *"BLAST - raw output"* alignments, containing less well-defined sequences from the full UniProt database, may lead to an increased source of alignment errors and unwanted noise. Possible reasons for this are: the inclusion of non-enzyme sequence homologs; an increased number of false positive homologs; and poorer alignments due to sequence fragments and sequencing errors. Also, in general there will be a larger number of sequences within each MSA, leading to an increased probability of more distinct amino acid types occurring within each aligned column, regardless of the expected correlation to enzyme functional specificity.

In an attempt to reduce the impact of these sources of potential alignment problems, it was decided to investigate the effect - on the *NumberOfAATypes* SVM feature and the SVM classification results - of using a more stringent (i.e. lower) E-value threshold to control the sequences included within the alignments. For this a number of progressively lower thresholds, from 10^{-5} to 10^{-60}, were considered and a value of 10^{-15} was selected. The main reason for selecting this threshold was that it was previously shown, in *figure 2.1*, that the use of an E-value threshold of less than or equal to 10^{-15}, resulted in the presence of enzyme sequences sharing three levels of the EC classification hierarchy (with the query sequences) with an accuracy of greater than 90%. Therefore, increasing the likelihood of more closely related functional homologs occurring within each MSA, while also ensuring that some functional diversity was present within the remaining aligned enzyme sequences.

The *"BLAST - raw output"* alignments were therefore filtered to include only protein sequences that had been identified in the sequence database search, with an associated BLAST E-value score of less than or equal to 10^{-15}. These filtered MSAs are referred to as the *"BLAST - raw output (10^{-15})"* dataset. An analysis of the number of distinct amino acid types, occurring within the identified fSDR and non-fSDR aligned columns, was then carried out using the MSAs from this dataset. This allowed the effect of the E-value filter, on the distributions of the number of amino acid types in the data used for the positive and negative SVM classes, to be examined.

This comparison, between alignments from using the 10^{-3} and the 10^{-15} thresholds, showed some small variations in the number of amino acid types counted within the fSDR and non-fSDR aligned residue columns. In particular, the frequency of occurrence of examples with more than 15 distinct types was reduced when the more stringently filtered *"BLAST - raw output (10^{-15})"* dataset was analysed. This was expected, due to the expected reduction in false positive homologs and the number of sequences in each MSA, and was observed in both the fSDR and non-fSDR examples. Also, the *"BLAST - raw output (10^{-15})"* dataset showed a shift towards more examples with very few (e.g. only one) amino acid types present. Again, this observation was expected to a certain degree in both fSDR and non-fSDR data examples. This is because the number of aligned sequences was generally reduced, causing an associated reduction in the sequence and functional diversity of the more closely related sequences remaining in the alignments.

Updating the Positive (fSDR) SVM Class Examples

A number of minor modifications to the positive and negative SVM class partitions (defined for the E-value<=10^{-3} filtered dataset) were necessary. This is because some of the fSDR columns were found to be "fully conserved" and therefore not providing any informative value for determining the functional specificity. Therefore it was decided to redefine them as negative (non-fSDR) examples when encoding the SVM feature vectors using the *"BLAST - raw output (10^{-15})"* dataset.

A further consideration for the more stringently filtered dataset was the level of functional diversity in the aligned sequences. To assess this, the number of EC classes represented by the enzyme sequences in each of the MSAs was calculated.

Using this, those that only contained sequences with the same specific EC class as the query were considered to display no "functional diversity" and therefore all aligned columns that were previously identified as fSDRs were subsequently altered to be non-fSDRs. Comparisons between the breakdown of fSDR and non-fSDR aligned columns, identified in each of the *"BLAST - raw output"* and the modified *"BLAST - raw output (10^{-15})"* datasets are shown in *table 5.1*.

Investigating the Use of a Percentage Occurrence Threshold Version of the *NumberOfAATypes* Feature

A method was implemented to try and improve the ability of the *NumberOfAATypes* feature to differentiate between the fSDRs and non-fSDRs. This involved applying a series of percentage thresholds to the number of distinct amino acid types occurring in each of the aligned columns. It was decided to explore the application of a threshold to this information because the "un-thresholded" form of the amino acid occurrence frequencies showed a poor level of differentiation between the fSDR and non-fSDR columns. This was primarily due to the occurrence of examples with large numbers of distinct residue types in both the fSDR and non-fSDR data.

A Specific Example Highlighting the Use a Percentage Threshold

To highlight the reasoning behind the use of a percentage occurrence threshold, it is useful to focus on a specific example. For this, an alignment of lactate and malate dehydrogenases (LDH/MDH) was again investigated. The MSAs analysed were taken from the *"BLAST - raw output (10^{-15})"* dataset and the *"targets_only"* dataset with an E-value sequence threshold of 10^{-15} applied. The input query sequence used to generate these MSAs was represented by the UniProt database accession code of **O08349**, which has an EC classification of 1.1.1.37 (malate dehydrogenase).

In the MSAs from the *"targets_only"* and *"BLAST - raw output (10^{-15})"* datasets, there were 210 and 420 aligned sequences, respectively; and they represented only 2 fully annotated EC classes (EC 1.1.1.27 and EC 1.1.1.37). For these particular MSAs there were 5 columns identified as fSDRs. However, for this analysis it is sufficient to concentrate on one of these, the experimentally well studied arginine (R) and glutamine (Q) residues that contribute towards determining substrate binding specificity in malate and lactate dehydrogenase enzymes, respectively. These

residues are generally found aligned to the arginine (R) occurring at residue number 81 in the query, MDH, sequence.

When looking at the number of distinct amino acid types within this column of aligned residues, it was found that the *"targets_only"* and *"BLAST - raw output (10^{-15})"* data contained 9 and 16 different types of amino acids, respectively. This was a surprisingly high number, especially for the *"targets_only"* MSA, because it only contained sequences from two specific enzyme functional classes and therefore was expected to contain only (or close to) the two R and Q residue types. Indeed, these were the predominant residue types found in the aligned column, accounting for over 90% of all the aligned residues, with 27% (56/210) arginine and 65% (137/210) glutamine residues. There were, however, a remaining 7 amino acid types that had low frequencies of occurrence. Additionally, when considering the equivalent column of residues in the *"BLAST - raw output (10^{-15})"* MSA, there were 14 different types of amino acid occurring at a low frequency.

Both of these observations highlight the difficulties associated with potentially misleading information and signal noise, which could arise from using a simple feature (such as a count of the number of amino acids in an aligned column) in the SVM classification. For this reason, it is suggested that a percentage threshold, applied to the number of aligned amino acid types, may reduce the potential signal noise that arises from residues with a relatively low frequency of occurrence, which appears to be inherent within this type of sequence alignment data.

ROC Analysis to Determine the Optimal Percentage Threshold

A series of different percentage thresholds, between 1% and 40%, were applied to the number of distinct amino acids in the fSDRs and non-fSDRs, identified in the MSAs from the *"BLAST - raw output (10^{-15})"* dataset. To assess which of these thresholds would best differentiate between the two classes of aligned residues a receiver operator characteristic curve (ROC) based analysis was carried out. In terms of maximising the area under the ROC curves (AUC), and therefore the predictive differentiation between the two classes of columns, these analysis results showed that a threshold of 12% was optimally performing, with a calculated AUC of 0.73. In comparison, the AUCs when using thresholds of 0% (i.e. no threshold), 1%,

5%, 10%, and 20% were 0.62, 0.65, 0.70, 0.72 and 0.71, respectively. Selected ROC curves associated with these results are shown in *figure 5.4*.

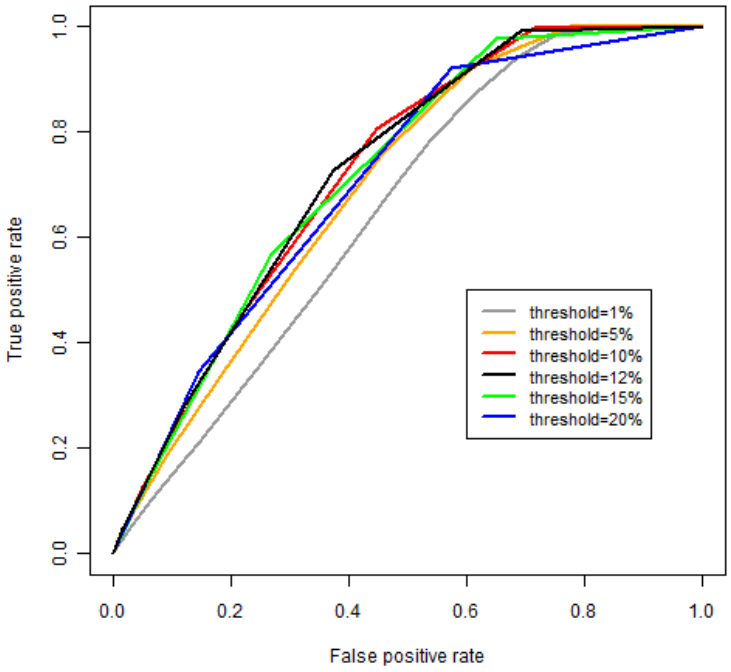

Figure 5.4. ROC curves for the number of amino acid type thresholds of: 1%; 5%; 10%; 12%; 15%; and 20%.

It is also worth noting the results of applying a percentage threshold of 12% to the residues aligned with query sequence (O08349) residue 81 in the LDH/MDH examples discussed above. For the MSAs, from both the *"targets_only"* and *"BLAST - raw output (10^{-15})"* datasets, only 2 amino acid types (R and Q) occur with a frequency of occurrence larger than the 12% threshold. This is an encouraging result as these are the two residue types that dominate substrate specificity within this particular group of enzymes.

5.3.2.5 The Effect of the E-value Based Sequence Filter and the *NumberOfAATypes_threshold_12%* Feature on the SVM Classification Performance

To conclude this analysis, the effects on the SVM classification performance are investigated when using the more stringently filtered *"BLAST - raw output (10^{-15})"* dataset of sequence alignments and also the additional *NumberOfAATypes_threshold_12%* feature. Two separate SVM training runs were carried out and the results are shown in *table 5.3*. Firstly, to provide a comparison between the previous SVM prediction results from the *"BLAST - raw output"* dataset (i.e. with an E-value sequence inclusion filter of 0.001), the same 22 *AA_composition* features were encoded from the *"BLAST - raw output (10^{-15})"* MSAs. An identical 5-fold cross-validation procedure was followed for the SVM training and testing and the results shown in the *"AA_composition (E-value <= 10^{-15})"* section of *table 5.3*.

Dataset	TPs	FPs	TPR	FPR	MCC
AA_composition (E-value <= 0.001)					
TEST 1	1470	10591	0.87	0.22	0.28
TEST 2	4183	11678	0.93	0.35	0.38
TEST 3	2356	9857	0.92	0.30	0.34
TEST 4	1265	11815	0.79	0.38	0.18
TEST 5	2924	10447	0.94	0.40	0.33
cross-validation average	**RBF** *(C=0.125 , gamma=8.0)*		**0.89**	**0.33**	**0.30**
AA_composition (E-value <= 10^{-15})					
TEST 1	1287	9159	0.80	0.19	0.27
TEST 2	3131	12475	0.82	0.37	0.28
TEST 3	1980	10367	0.82	0.31	0.27
TEST 4	1063	9986	0.70	0.32	0.17
TEST 5	2443	11026	0.88	0.42	0.27
cross-validation average	**RBF** *(C=2.0, gamma=8.0)*		**0.81**	**0.32**	**0.25**
AA_composition + NumberOfAATypes_threshold_12% (E-value <= 10^{-15})					
TEST 1	1375	11159	0.86	0.23	0.26
TEST 2	3558	15875	0.93	0.47	0.28
TEST 3	2110	11463	0.88	0.34	0.28
TEST 4	994	10576	0.66	0.34	0.14
TEST 5	2673	13657	0.96	0.52	0.26
cross-validation average	**RBF** *(C=128.0, gamma=0.000195)*		**0.86**	**0.38**	**0.24**

Table 5.3. A comparison of the SVM classification results (TPs, FPs, TPR, FPR, and MCC) for the three sets of input feature vectors used. For each of the three SVM training runs the averaged results from the five datasets of "randomly balanced" data are shown. Results are shown for the five individual TEST sets of the 5-fold cross-validation training sets. Also shown are the averaged FPR, TPR and MCC results for the 5-fold cross-validation, along with the SVM learning kernel parameters that produced the classification results.

These results were somewhat surprising because they show a decrease in the quality of the fSDR classification from the optimised SVM models. As before, the optimally performing SVM models were selected via the highest observed MCC values. The RBF kernel was again found to be optimal. Both the MCC and true positive rates were found to be worse than when using the *"AA_composition (E-value <= 0.001)"* input features; whereas there was a slight improvement in false positive rate. In particular: the average MCC decreased from 0.30 to 0.25; the average TPR decreased from 0.89 to 0.81; and the average FPR decreased from 0.33 to 0.32. It is not clear why this should be. One possibility may be that the use of a more stringent E-value based sequence inclusion threshold, while aiming to minimise signal noise from alignment problems of more distant sequence homologs, inadvertently reduced the sequence and functional diversity of the information contained within the encoded alignments to the detriment of the functionally specific information.

Finally, the *AA_composition* and *NumberOfAATypes_threshold_12%* features, encoded from the *"BLAST - raw output (10^{-15})"* alignments, were combined to form an SVM input feature vector of length 23 for each of the aligned columns of residues. The aim of this was to assess the effect that this additional input feature would have on the classification performance of the SVMs. Again, the 5-fold SVM cross-validation procedure was followed and the results are shown in the *"AA_composition + NumberOfAATypes_threshold_12% (E-value <= 10^{-15})"* section of *table 5.3*. These results show that there was an improvement in the TPR (from 0.81 to 0.86), but a corresponding deterioration in FPR (from 0.32 to 0.38), when compared to the *"AA_composition (E-value <= 10^{-15})"* results. Also, a slight decrease is observed in the average MCC, from 0.25 to 0.24.

This result appears to demonstrate that the *NumberOfAATypes_threshold_12%* feature does not provide any additional information for the purposes of SVM based differentiation between these fSDR and non-fSDR columns. In-fact, it appears to have a negative effect on the SVM classifiers when considering the change in MCCs and FPRs. Due to this result, further investigations of the *NumberOfAATypes_threshold_X%* feature – such as the combination with the *"AA_composition (E-value <= 0.001)"* input feature vectors - were not carried out.

In summary, these studies have investigated the feasibility of using SVMs towards the classification of functionally specific residues in protein sequence alignments. They have used simple measures of the amino acid composition, and the number of amino acid types, within each aligned column of residues and compared the use of alternatively filtered MSAs. In general, the relatively low MCCs and the large number of false positives that were observed for each SVM experiment in this chapter were disappointing. In previous studies for identifying catalytic residues (in the CSA database) through the use of machine learning techniques (such as NNs and SVMs), similar results were reported. For example, Petrova and Wu (2006) report an MCC of 0.23, a TPR of 0.90 and a FPR of 0.13, when using SVMs to identify CSA residues. Also, an earlier study by Gutteridge et al. (2003) that uses neural networks for the same problem, reports an MCC of 0.28 and a high number of detected false positives, where 56% of catalytic residues were identified correctly, but only 1 in 7 of the positive classifications are correct. Both these MCC values are lower than the 0.30 observed for the *"AA_composition (E-value <= 0.001)"* input feature vectors in this chapter. Although these are not directly comparable to the studies of functionally specific residues contained in this chapter, they do possibly demonstrate some of the inherent difficulties in accurately differentiating between functionally (catalytic or specificity determining) and non-functionally important residues.

It must be concluded, however, that these poor classification results present problems for the incorporation into an accurate, automated, method for improving the functionally specific ordering and classification of homologous enzyme sequences. Further work is clearly needed in this area and a more detailed discussion of other possible avenues of study for this SVM based classification are provided in the further work section of this thesis.

5.3.3 Additional Investigation of the Performance of the SVM Classifier

In this section an analysis of the performance of the SVM classifier is carried out to investigate the functionally predictive performance on sequences taken from three well-studied classes of enzymes. These are: the lactate/malate dehydrogenases (LDH/MDH), which have been used in previous examples in this thesis; the nucleotidyl cyclases (cyclases); and the serine proteases. The functional rescoring results from both the "top-hit" and the functional enrichment, of the top-10 enzyme sequences, are considered in this analysis.

5.3.3.1 Generation of the SVM Classifier

The following method was used to generate the SVM model used for the classification of the fSDRs in the three enzyme examples presented below. Using the 5-fold cross-validation results, shown in *table 5.3*, it was decided to use the *"AA_composition (E-value <= 0.001)"* feature vectors as input for the SVM model generation. This particular dataset and feature vectors were selected for the SVM model generation because they showed the largest average MCC value (0.30) of the three alternative feature encoding methods that were investigated (see *table 5.3*).

To generate the SVM classification model to be used for the fSDR classifications, a combined input dataset of the randomly balanced data was used. It was necessary to use the randomly balanced sets of fSDR and non-fSDR data, because the cross-validation (and hence the optimisation of the SVM learning parameters) was done with this. Although five sets of randomly balanced datasets were used in the cross-validation procedure, only one was used for the SVM model generation. The randomly balanced dataset with the largest MCC value was selected for this purpose. However, the small difference between the set with the largest MCC value, of 0.302, and that with the lowest MCC value, of 0.301, would suggest that there would be no significant difference in the predictive performance of the resulting SVM classifiers generated with any of these five datasets.

Finally, the randomly balanced data from the 357 MSAs, defined as belonging to *GROUP_1* to *GROUP_5*, were encoded using the *"AA_composition (E-value <= 0.001)"* method of calculating feature vectors. This combined dataset was then used

to generate an SVM classifier, using *SVM*^{light} with an RBF kernel and learning parameters of *C=0.125* and *gamma=8.0*. These particular learning parameters were used because they were shown, in *table 5.3*, to produce the optimal predictive performance, when considering the MCC values, in the cross-validation of the SVM parameters.

5.3.3.2 Analysis of the Performance of the SVM Classifier on Three Enzyme Examples

To conclude this section of the thesis, the SVM classifier was used to automatically identify potential fSDRs within three previously well-studied families of enzymes coupled with an analysis of their use in the classification of specific enzyme function. These three types of enzymes were selected because their mechanisms and substrate binding properties have been previously investigated using both experimental (Fersht, 1999) and computational methods (Hannenhalli and Russell, 2000; Pazos et al., 2006).

For this analysis, the fSDRs predicted by the SVM classifier were first compared to those experimentally identified as being important for determining specificity. The predicted fSDRs were then used to generate sequence sub-alignments, which were subsequently re-scored to allow an assessment of their use in assigning specific enzyme function classifications. The resulting re-scoring results were then compared to a number of alternative functional re-scoring methods that have been previously discussed within this thesis.

The methods that were compared were: (i) "*BLAST*" – which uses the significance ordering of a BLAST database search; (ii) "*PAM30 (0,0)*" – which uses a PAM30 amino acid substitution matrix with both gap opening and gap extension set to 0 (see *chapter 3*); "*func-MB top-10*" - which uses the aligned residues with the top-10 ranking Spearman-rank order correlation coefficients as calculated by the *func-MB* method - with a *colgap_percent* threshold of 10% (see *chapter 4*); "*func-MB top-30*" - which uses the aligned residues with the top-10 ranking Spearman-rank order correlation coefficients as calculated by the *func-MB* method - with a *colgap_percent* threshold of 50% (see *chapter 4*); "*optimal*" – which uses the sequence sub-alignment that was found to give the optimal functional enrichment score

performance after re-scoring (see *section 5.2.4.2*); and "*SVM predicted*" – which uses the sub-alignments generated from the residues predicted as fSDRs by the SVM classifier. Two further "randomised" methods were also compared: the "*random column selection*", which used (1000) repeated random selections of *n* aligned columns of residues to generate the sequence sub-alignments for re-scoring; and "*random sequence selection*", which compares the probability of randomly selecting a functionally correct sequence from the MSA.

The reasons for selecting these particular methods for comparison were as follows. The *"BLAST"* method provides a baseline comparison to a gapped BLAST sequence database search and the *"PAM30 (0,0)"* method shows the functional classification performance after re-scoring the BLAST generated MSAs with a PAM30 matrix. As in previous sections of this thesis, the *"PAM30 (0,0)"* was used because it was shown to be the optimally performing method in the studies presented in *chapter 3*.

The two methods based on the *func-MB* method of sub-alignment selection were selected because they have been previously identified as providing optimal functional re-scoring performance, when using sequence sub-alignments that have been identified via automatic fSDR identification methods. The "*func-MB top-30*" method was chosen because it was shown to be the best performing of the automatic fSDR identification methods (see *table 4.6*) – when using the "top-hit" method of assessing functional re-scoring success. A *colgap_percent* threshold of 50% was applied to the MSAs in these comparisons because it was found that there was no significant difference in the functional classification accuracy when using thresholds greater than this. An additional *func-MB* based method (*"func-MB top-10"*) was also used in the comparison because it was shown to perform well when using the "functional enrichment" method of assessing the success of functionally specific alignment re-scoring. A *colgap_percent* threshold of 10% was applied to the MSAs in these comparisons. This was also used in the earlier benchmark analysis of the functional enrichment scores obtained from applying the *func-MB* method of sub-alignment identification. Although other methods of sub-alignment selection give comparable functional re-scoring results, they were not found to be significantly different in performance and were therefore not investigated further in this comparison.

The *"optimal"* method provides a comparison to the aligned residues and sub-alignment re-scoring results obtained from using the optimally performing sub-set of residues. Finally, the *"random column selection"* and *"random sequence selection"* methods were used to provide comparisons equivalent to the "random" methods used in previous chapters in this thesis.

For each of these methods (except for the "*random sequence selection*") the success of functional classification was assessed in terms of both the "top-hit" assessment method and the level of functional enrichment of the top 10 ranking sequences (see *section 5.2.2*). For each of the examples, the SVM classifications were made using the data encoded from the *"BLAST - raw output"* form of the MSA and the assessment of the functional re-scoring was carried out on the MSAs from the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset.

5.3.3.3 Lactate/Malate Dehydrogenases

The first example that was investigated was taken from the lactate and malate dehydrogenase (LDH/MDH) families of enzymes. They are a divergent set of enzymes that are generally difficult to separate into the two specific functional sub-types using simple measures of sequence similarity. Both LDH and MDH enzymes have a well defined substrate binding site and an experimentally determined substrate specificity switch, from MDH to LDH, where arginine is replaced with glutamine in the equivalent Arg-102 residue position - described by Fersht (1999). The example shown, used the *UniProt* sequence *O08349 [MDH_ARCFU]* as the query sequence to generate the MSA. This is an example of a malate dehydrogenase (EC 1.1.1.37), which has a sequence length of 294 residues and an associated crystal structure in PDB, with identifier, *2x0i*.

SVM Classification and Functional Re-scoring Results

The SVM classifier identified 146, out of 294, residues to be relevant for determining the functional specificity of the aligned LDH and MDH sequences. This prediction consisted of 5 true positives (TPs), 141 false positives (FPs), 148 true negatives (TNs) and 0 false negatives (FNs), when compared to the five "optimal" positive fSDR residues that were identified for this sequence. The resulting MCC value was 0.132, which is low, but better than expected from a random classifier,

which would be expected to have an MCC of 0. Using *equation 5.6*, a chi-squared statistic of 5.123 and a p-value of 0.0236 is observed for this MCC value.

A subset of the multiple sequence alignment, generated by BLAST with *O08349 [MDH_ARCFU]* as the query sequence, is shown in *figure 5.5(a)* – using the *jalview* software (Waterhouse et al., 2009). The actual number of sequences in the MSA was much larger than shown in the figure, for clarity only a subset of the sequences is shown here. The query sequence is shown in the centre, with a number of aligned MDHs (EC 1.1.1.37) above and LDHs (EC 1.1.1.27) below. The sub-set of aligned residues that are shown, define: the columns of residues that were aligned to the five residues that were defined as fSDRs by the "optimal" method. These are highlighted using the "Taylor" colour scheme, defined in the *jalview* software used to illustrate the MSA; with the corresponding residue indices, in the *2x0i* crystal structure, shown above; the section of the MSA that is associated with the active site loop, described by Fersht (1999), which is contained in the orange box and corresponds to 13 residues (98-110 in *2x0i* and 77-89 in *O08349*); the residue positions that were positively classified as fSDRs by the SVM, are indicated by an *(*)* above the aligned residues. Also shown, in *figure 5.5(b)*, is the crystal structure of the enzyme *2x0i* – generated using *PyMol* (DeLano, 2008) - with a number of residues highlighted. This also highlights the five "optimal" fSDR residues as defined by the *func-MB* method. These are labelled as ARG-102, MET-107, LEU-110 and, in blue, as ALA-237, PRO-250. The active site loop is shown highlighted in orange and the residues (in addition to the 5 TPs) that were classified as fSDRs by the SVM are highlighted in pink. The enzyme NADH cofactor is shown in grey.

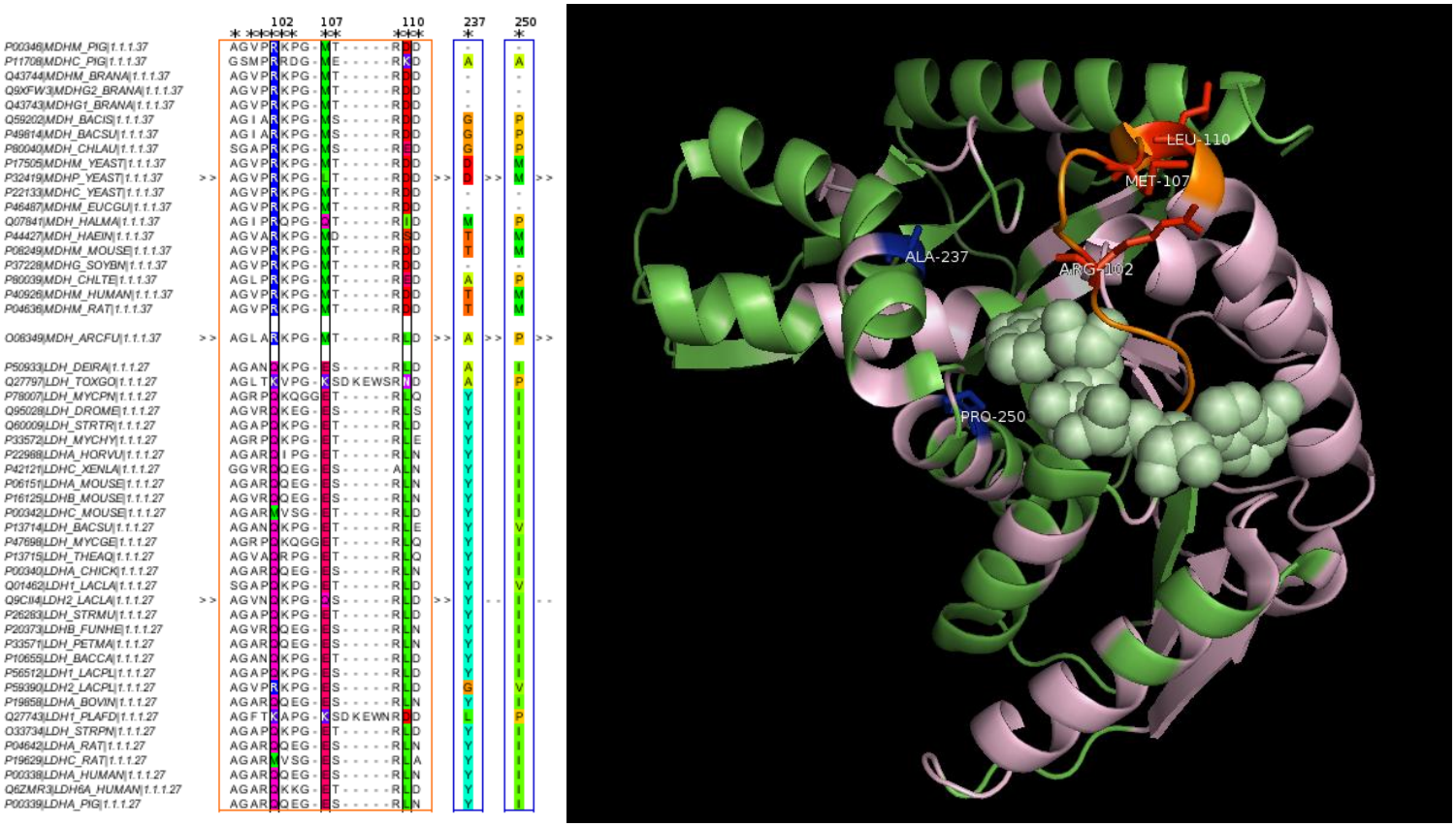

Figure 5.5. (a) Selected columns and sequences from a jalview generated MSA subset of lactate/malate dehydrogenases. The five aligned residues highlighted in the Taylor colour scheme, and marked with the residue index of the structure shown in (b), are those defined as "optimal". Residues marked with an (*) indicate a positive SVM prediction. The orange box indicates the "Fersht active site loop region" (see text). All sequences show associated UniProt identifiers and EC classifications and ">>" denotes deleted sections of the MSA. (b) PyMol crystal structure of enzyme 2x0i, showing: (i) SVM TPs (red and blue); (ii) SVM FPs (pink); (iii) TNs (green); (iv) "Fersht active site loop region" (orange and red); and (v) the five optimal fSDRs (labelled in white). The NADH cofactor is shown in grey space-fill representation.

The ARG-102 and MET-107 residues, which relate to residues ARG-81 and ARG-85 in the O08349 sequence in UniProt, were also both identified as specificity determining sites in the studies by Pazos et al. (2006) and Hannenhalli and Russell (2000). These positions clearly show a preference for: arginine (R) in MDHs and glutamine (Q) in LDHs, at position 102; and methionine (M) in MDHs and glutamic acid (E) in LDHs, at position 107. A further residue (LEU-110) was highlighted, by the *func-MB* method, as an fSDR in the active site loop region. Interestingly this was not highlighted as one of the high-scoring specificity determining residues in the studies by Pazos et al. (2006) and Hannenhalli and Russell (2000); it can, however, be seen that there is a clear preference for aspartic acid (D) in the MDHs and leucine (L) in the LDHs. Although the MDH query sequence shows a leucine in this position, the associated Spearman-rank order correlation coefficient of all the aligned residues, of 0.81, was the second highest in this MSA and therefore identified as a high-scoring fSDR.

The other two "optimal" fSDR residues (ALA-237 and PRO-250) are outside the active site loop region and were not highlighted by the previous studies. They do not show quite as clear a distinction between the two enzyme sub-groups but there does appear to be a tendency for tyrosine (Y) and isoleucine (I) in positions 237 and 250, respectively, of the LDHs. Their relatively close proximity to the NADH cofactor suggests that they may be involved in its binding.

The false positive residue predictions (except for those in the active site loop, which are highlighted in orange) made by the SVM are shown in pink, in *figure 5.5(b).* These and the 5 true positive residue positions were then used to generate a sequence sub-alignment, which was subsequently re-scored using a PAM30 substitution matrix. The results from this and the other re-scoring methods are shown in *table 5.5*. These results show that using the residues predicted by the SVM gives a functional enrichment score of 0.9 (i.e., 9 out of the top-10 ranked sequences, after re-scoring, had an MDH EC classification, of 1.1.1.37) and a functionally correct "top-hit" classification of the query sequence. This shows an improvement over all the other methods except for the "optimal" method. In particular it improves on both the BLAST and PAM30 (0,0) methods, which did not show any functionally correct MDH classified sequences in the top-10. Further, it can be seen that the SVM

method performs better than the random column selection methods, in terms of both "top-hit" classification and the level of enrichment of the functionally correct sequences after re-scoring. In particular, when 146 columns were repeatedly randomly selected from the query sequence, an average functional enrichment score of 0.14 was observed and a functionally correct "top-hit" based classification was made in only 13.7% of the randomly selected sub-alignments.

5.3.3.4 Nucleotidyl Cyclases

The second example was taken from the nucleotidyl cyclase group of enzymes, which consist of the adenylyl cyclases (ACs) and the guanylyl cyclases (GCs). They are a well-studied family of membrane associated enzymes that, in the case of the adenylyl cylases, synthesise cyclic-AMP from an ATP substrate; and in the case of the guanylyl cyclases, synthesise cyclic-GMP from a GTP substrate. Two enzyme classification numbers are used to describe these groups of enzymes: (i) EC 4.6.1.1 denotes an adenylyl cyclase enzyme; and (ii) EC 4.6.1.2 denotes an enzyme of type guanylyl cyclase. Mutagenesis studies, carried out by Tucker et al., (1998), showed that it was possible to alter the substrate specificity, from an adenylyl to a guanylyl cyclase, by the mutation of only two residues. Both of these residue changes were required to confer the change in specificity and they are discussed in more detail in the example below.

For this example the *UniProt* sequence *P0A4Y0 [CYA1_MYCTU]* was used as the query sequence to generate the MSA. This sequence is annotated as an adenylyl cyclase (EC 4.6.1.1), has a sequence length of 443 residues and an associated PDB crystal structure, with identifier *1yk9*. Only the catalytic domain of the sequence (residues 245-428) is present in the crystal structure, as the first half of the sequence is part of a trans-membrane region. The MSA used to assess the functional re-scoring was taken from the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset and contained 77 sequences; of which 34 were classified as EC 4.6.1.1 and 43 classified as EC 4.6.1.2. The *"BLAST - raw output"* MSA, used for the SVM vector encoding and classification, contained 586 sequences.

SVM Classification and Functional Re-scoring Results

From this sequence alignment the SVM classifier identified 154 of the 443 query sequence residues as being associated with determining the functional specificity of the aligned cyclases. This consisted of: 6 TP; 148 FP; 289 TN; and 0 FN classifications, when compared to the six "optimal" positive fSDR residues that were identified for this sequence. In this example the residues for "optimal" re-scoring performance were identified as those selected by the *profile-HMM* method, with a Z-score threshold of greater than or equal to 2.5. As in the LDH/MDH example above, the resulting MCC value of 0.161 was low, but was again shown to be performing

better than a random classifier. In-fact, using *equation 5.6*, a chi-squared statistic of 11.483 and a statistically significant p-value of 0.0007 was observed for this MCC value. Again, all of the "optimal" specificity determining residues were correctly identified by the SVM.

A subset of the MSA is shown in *figure 5.6(a)*. The query sequence (*P0A4Y0 [CYA1_MYCTU]*) is shown in the centre, with the guanylyl cyclases above and the adenylyl cyclases below. For clarity, a reduced, representative set of the sequences within the alignment are shown. The sub-set of aligned residues that are shown, define: (i) the six columns of residues that were defined as fSDRs by the "optimal" method. Again, these are highlighted using the *jalview* "Taylor" colour scheme, with residue indices, corresponding to the *1yk9* crystal structure, shown above; (ii) the residues that were sequentially mutated in the study by Tucker et al., (1998) are shown in the orange box, which correspond to 5 residues, from 365 to 359 in *1yk9* and *P0A4Y0*).

Figure 5.6(b) shows the structure of *1yk9*, with a number of residues highlighted. Those highlighted in red are the six "optimal" fSDR residues defined by the *profile-HMM* method; they are labelled, as GLU-293, ILE-295, GLU-296, ARG-361, CYS-365 and TRY-367. The residues mutated in the Tucker et al., (1998) study are shown highlighted in orange, with the residues (in addition to the 6 TPs) that were classified as fSDRs by the SVM highlighted in pink. It should be noted that there are some discrepancies between the residue types in the UniProt sequence and the PDB sequence, in particular: at position 296 there is a Glu (E) instead of a K (Lys) in the structure; and at position 365 there is a Cys (C) instead of an Asp (D) in the structure.

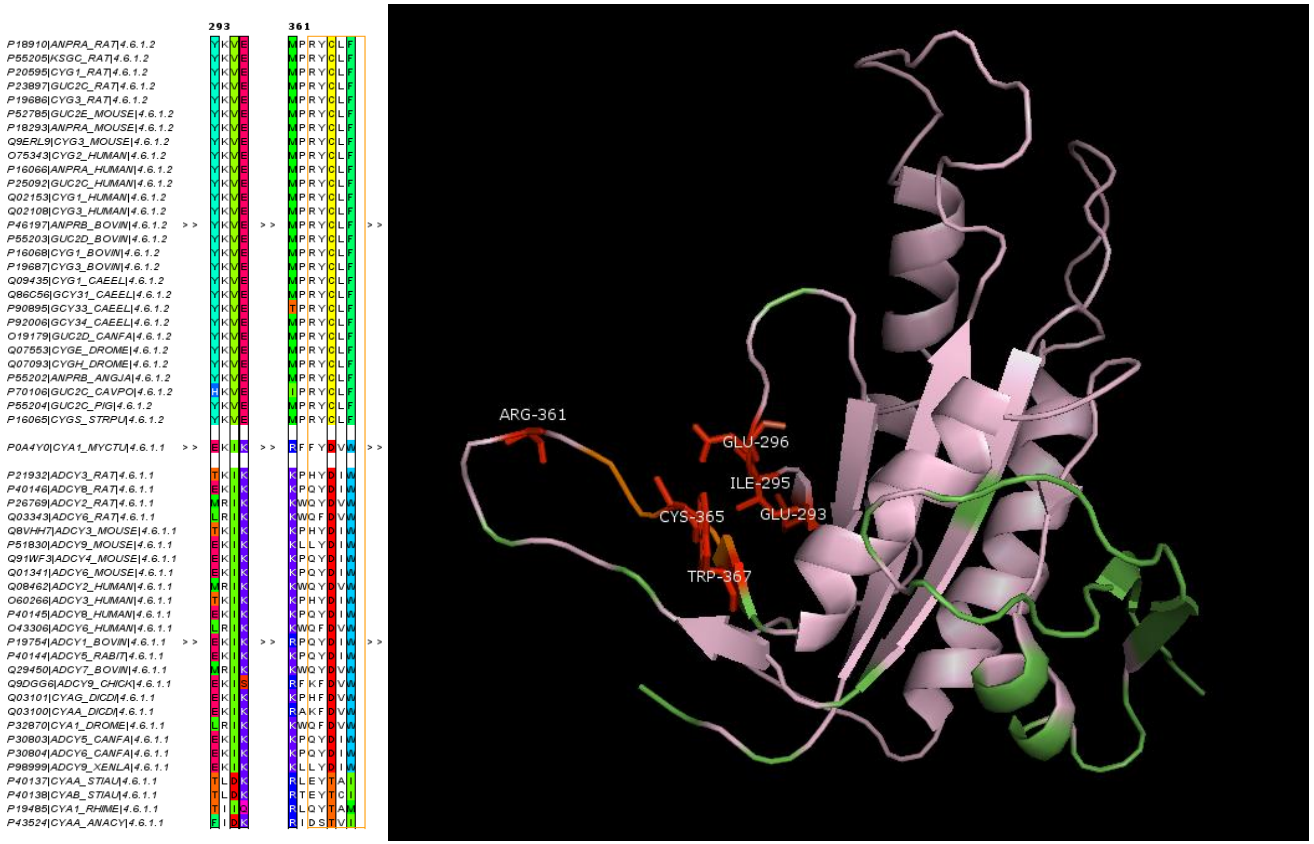

Figure 5.6. (a) Selected columns and sequences from a jalview generated MSA subset of nucleotidyl cyclases. The six aligned residues highlighted in the Taylor colour scheme are those defined as "optimal" fSDRs. Residue indices shown above the MSA correspond to the sequence shown in (b). The orange box indicates the five residues mutated in the study by Tucker. All sequences show associated the UniProt identifiers and EC classifications, with ">>" denoting deleted sections of the MSA. (b) PyMol crystal structure of enzyme lyk9, showing: (i) SVM TPs (red); (ii) SVM FPs (orange and pink); (iii) TNs (green); (iv) "Tucker mutated residues region" (orange and red – residues: CYS-365 and TRP-367); (v) and the six optimal fSDRs (labelled in white).

All of the "optimal" residue positions, except for GLU-293, were also identified as specificity determining in the study carried out by Hannenhalli and Russell (2000). It can be seen, in *figure 5.6(a)*, that the six highlighted columns show a preference for a different residue type in both the ACs and GCs; thereby demonstrating their importance for determining the specific functional sub-type of the enzymes and showing that residues of biological significance are being correctly identified. That is, the two residue positions (296 and 365) identified, by Tucker, as being of key importance for determining functional specificity were also found to have the largest Z-scores.

The 148 false positive classifications made by the SVM are highlighted in pink, on the structure shown in *figure 5.6(b)*. The results from functionally re-scoring the aligned sequences, using a sub-alignment of the residues aligned to the 154 (6 TPs and the 148 FPs) positive SVM residue classifications and a PAM30 matrix, are provided in *table 5.5*. Disappointingly, these show that the sub-alignments generated by the SVM classified residue positions are only performing at a level comparable to the random column selection method. That is, a functional enrichment score of 0.6 (6 correct sequences in the top-10) was seen, when using both the SVM predicted residues, and an average of randomly selecting 154 residues (using 1000 iterations) from the *P0A4Y0* sequence. Also, this random selection method gives a correct "top-hit" prediction of specific function, in 68.8% of the 1000 iterations, whereas the sub-alignment from the SVM predicted residues gives an incorrect "top-hit" assignment.

Although this example demonstrates a lower level of success than the previous LDH/MDH example (and the serine protease example discussed next), it does show a marked improvement over the functional ranking of the sequences generated by the original BLAST search method. This resulted in a functional enrichment of just 0.1 and a functionally incorrect "top-hit". It should be noted that for this example (and the LDH/MDH example) the results from the BLAST method are not affected by the "artificial" dataset creation method, described in *section 2.4.2*. That is, the "top-hit" and functional enrichment scores reported in *table 5.5* were observed in the original gapped BLAST database search. Therefore, this result shows that although the SVM classification method is not performing as well as the other re-scoring methods, for

this cyclase example, it is showing clear improvement when compared to the BLAST method, which is an encouraging result.

5.3.3.5 Serine Proteases

The final example that was investigated was from the trypsin-like serine protease superfamily of enzymes. The trypsin-like serine proteases are a well-studied superfamily of proteins, with an enzyme classification of EC 3.4.21.-, which encompass a number of more specific functional sub-types; three of which are: trypsin (EC 3.4.21.4); chymotrypsin (EC 3.4.21.1); and elastase, which is represented by two EC classes (EC 3.4.21.36 and EC 3.4.21.37). These sub-types are generally found to have similar kinetic properties and catalytic mechanism (Fersht, 1999), which involves the hydrolysis of a specific peptide bond in the protein substrate. The difference between the substrate specificity of the three sub-types described above is related to the structure of the binding pocket. The trypsins generally contain a charged aspartate residue, which allows the binding and subsequent cleaving (via the serine in the catalytic triad) of peptide bonds next to lysine or arginine residues. In the case of chymotrypsins the aspartate is generally mutated to a serine to allow the binding of the large hydrophobic residues. Finally, the elastases provide specific binding of smaller hydrophobic residues, such as alanine, due to steric hindrance provided by a valine residue at the entrance to the binding pocket.

For this example, a trypsin sequence (with the *UniProt* identifier *P00775 [TRYP_STRGR]* and an associated PDB structure of *1os8*), with a specific enzyme classification of EC 3.4.21.4, was used as the query sequence. The resulting BLAST MSA contained 2171 sequences, 229 of which had complete Swiss-Prot designated enzyme annotations. Due to the evolutionarily diverse nature of this serine protease superfamily, this alignment contained sequences with 38 specific functional sub-classes. All of these were of the EC classification type: EC 3.4.21.-. Most of these were present in only small numbers of sequences; four classes had more than 10 sequence representatives in the MSA, these were: trypsin (EC 3.4.21.4); chymotrypsin (EC 3.4.21.1); kallikrein (EC 3.4.21.35); and tryptase (EC 3.4.21.59), which had 54, 13, 27 and 12 sequences representatives respectively. The functional

class composition of this MSA is in contrast to the LDH/MDH and cyclase examples discussed earlier, which both contained only two specific sub-types.

SVM Classification and Functional Re-scoring Results

When applying the SVM classifier to the encoded sequence information, an MCC value of 0.454 was observed, from a total of 178 SVM predicted residues. This constituted a total of 84 TPs, 94 FPs, 80 TNs and 1 FN, when compared to the 85 residues (from a total sequence length of 259) that were identified as "optimal" fSDRs, via the *func-MB* method, with a Spearman-rank order correlation threshold of greater than or equal to 0.1. It can be seen that the predictive performance of the SVM, as measured by the MCC value, was better than the two previous examples. Although a significant number of false positive residues were identified by the SVM as functionally specific, the resulting functional enrichment score from re-scoring the alignment with these identified residues, was 0.9 (i.e., 9 out of the top-10 ranked sequences after sub-alignment re-scoring showed the same specific functional class as the query). When assessing the statistical significance of this MCC value, using *equation 5.6*, a chi-squared statistic of 53.38 and a very statistically significant p-value of much less than 0.0001 was observed. This was a promising result and is compared to the other methods, in *table 5.5*, and discussed in more detail below.

As for the previous two examples, an MSA and crystal structure are shown, in *figure 5.7(a)* and *(b)*, respectively. The MSA shows a sub-alignment of the 85 "optimal" residues, coloured using the "Taylor" colour scheme in the *jalview* software. To improve clarity, where the aligned residues are not sequential, no delimiters (i.e., the ">>" notation in *figures 5.4(a)* and *5.5(a)*) are shown. The sequences shown in this alignment are grouped into five specific functional sub-classes, they are: (i) chymotrypsins (EC 3.4.21.1); (ii) trypsins (EC 3.4.21.4); (iii) elastases (EC 3.4.21.36/37); (iv) kallikreins (EC 3.4.21.35); and (v) tryptases (EC 3.4.21.59). These display the diversity of the aligned sequences and the difficulty in defining a sub-set of aligned residues that can determine the functional specificity differences within such a diverse protein family.

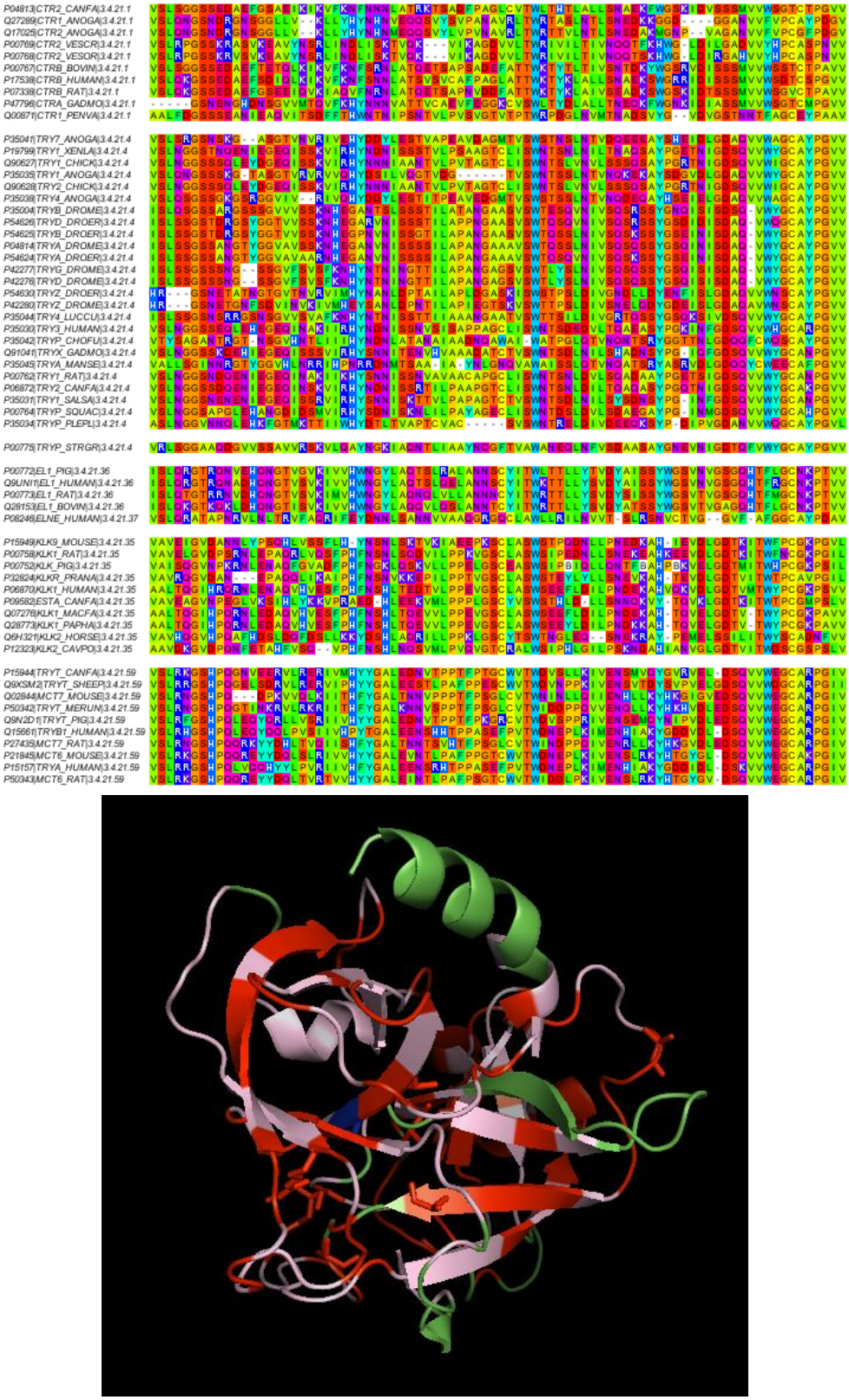

Figure 5.7. (a) A jalview generated sub-alignment showing the 85 "optimal" fSDRs (highlighted in the Taylor colour scheme) and selected sequences from five functional sub-classes of serine proteases. All sequences show associated UniProt identifiers and EC classifications. (b) PyMol crystal structure of enzyme 1os8, showing: (i) SVM TPs (red); (ii) SVM FPs (pink); (iii) SVM TNs (green); and (iv) SVM FNs (blue).

It can be seen, in *figure 5.7(b)*, that the "optimal" fSDRs (highlighted in red) and the SVM predicted residues (TPs, FPs and FN residues are highlighted in red, pink and blue, respectively) are not located in a specific area of the protein structure. This is in contrast to the LDH/MDH and nucleotidyl cyclase examples, which showed a tendency for the small number of "optimal" fSDRs to be found in the enzyme binding sites. One reason for this difference may be the increased diversity of the serine proteases and the much larger number of specific functional sub-classes contained within the alignment, meaning that larger numbers of residues were required to describe the differences between the specific functional sub-types.

To conclude this analysis of the serine proteases a more detailed view of the functional re-scoring results (obtained from a selection of the re-scoring methods shown in *table 5.5*), is shown in *table 5.4*. This table shows a comparison of the top 10 ranked sequences, after five (*BLAST*, *PAM30 (0,0)*, *func-MB top-10*, *optimal*, *SVM predicted*) of the functional scoring methods have been applied. For each of the methods, the sequence identifiers and the 4th terms from the serine protease EC classification of EC 3.4.21.-, are shown. The proteins with the same specific function as the query (*TRYP_STRGR*) are shown in green, with a **4** to indicate that the EC classification was EC 3.4.21.**4**. The proteins with a different EC classification to the query are shown in red.

BLAST		PAM30 (0,0)		*func-MB* (top-10)		Optimal		SVM predicted	
KAL_MOUSE	34	TRY1_ANOGA	4	TRYZ_DROER	4	TRYB_DROER	4	TRYB_DROME	4
KAL_RAT	34	TRYB_DROME	4	UROK_BOVIN	73	TRYA_DROME	4	TRY1_ANOGA	4
TRY7_ANOGA	4	TRYT_SHEEP	59	TRYZ_DROME	4	TRY1_ANOGA	4	TRY1_CHICK	4
KAL_HUMAN	34	TRY1_CHICK	4	UROK_PIG	73	TRYB_DROME	4	TRY2_CHICK	4
FA11_MOUSE	27	TRY2_CHICK	4	TRY4_ANOGA	4	TRYA_DROER	4	TRYA_DROER	4
TRY1_XENLA	4	MCT7_MOUSE	59	TRY7_ANOGA	4	TRYG_DROME	4	TRYD_DROME	4
TRY1_CHICK	4	CTR2_CANFA	1	TRY4_LUCCU	4	TRYD_DROME	4	TRYG_DROME	4
TRY1_ANOGA	4	TRY1_XENLA	4	TRY3_HUMAN	4	TRY1_XENLA	4	TRYT_SHEEP	59
TRY2_CHICK	4	TRYD_DROER	4	TRYP_CHOFU	4	TRYX_GADMO	4	TRYD_DROER	4
TRY4_ANOGA	4	TRYT_MERUN	59	TRYX_GADMO	4	TRYD_DROER	4	TRYB_DROER	4

Table 5.4. Table showing the top-10 ranked sequences, for the methods: BLAST; PAM30; func-MB top-10; optimal; and SVM (see text for method descriptions). The sequences are shown in rank order, from 1 to 10, descending the page, with green indicating a functionally specific correct match to the query (TRYP_STRYGER – EC 3.4.21.4) and red an incorrect match. The UniProt identifier is shown for each sequence, alongside the number (X) of the more general, EC 3.4.21.X, serine protease functional classification.

These comparisons show that the quality of specific functional classification (with regards to the functional enrichment score and "top-hit" assessment methods) increases in the order of: BLAST; PAM30 (0,0); *func-MB* (top-10); SVM predicted; and "optimal". To some extent this was expected as the methods which use a functionally informative sub-set of residues, such as *func-MB* (with a small subset of highly correlated residues) and the "optimal" fSDRs, have been shown to provide better functional classification and re-scored sequence rankings than the BLAST and PAM30 (0,0) "whole alignment" methods. However, a promising outcome from this result was the observation that the use of residues, predicted by the SVM method, was more effective than all of the others (except for the "optimal" subset of fSDRs) when considering the number of functionally correct sequences in the top-10, after re-scoring. Further, it can also be seen, in *table 5.5*, that the SVM method, with a functional enrichment score of 0.9, is also performing considerably better than the random column selection method (when n=178), which showed an average functional enrichment score of 0.65.

Method		LDH/MDH O08349 [MDH_ARCFU]		Nucleotidyl Cyclases P0A4Y0 [CYA1_MYCTU]		Serine Proteases P00775 [TRYP_STRGR]	
		"top-hit"	enrichment	"top-hit"	enrichment	"top-hit"	enrichment
BLAST		NO	0	NO	0.1	NO	0.6
PAM30 (0,0)		NO	0	YES	0.8	YES	0.6
func-MB top-30		YES	0.2	YES	1.0	NO	0.7
func-MB top-10		YES	0.9	YES	1.0	YES	0.8
Optimal		YES	1.0	YES	1.0	YES	1.0
SVM predicted		YES	0.9	NO	0.6	YES	0.9
Random column selection (n)	***10***	40.7%	0.37 (avg)	50.8%	0.50 (avg)	25.9%	0.30 (avg)
	30	35.2%	0.29 (avg)	57.1%	0.53 (avg)	38.6%	0.40 (avg)
	SVM	13.7%	0.14 (avg)	68.8%	0.60 (avg)	75.5%	0.65 (avg)
Random sequence selection		0.42	*NA*	0.44	*NA*	0.25	*NA*

Table 5.5. The performance of the functional re-scoring methods (described in the text) for each of the three enzyme examples. For each of these the "top-hit" classification result and functional "enrichment" score, from the top-10 ranked sequences, is shown. A "top-hit" result of "YES" or "NO" indicates a correct or incorrect specific functional match, respectively. For the "Random column selection (n)" method the "top-hit" result refers to the percentage of correct "top-hits" observed from 1000 repeated iterations and the "enrichment" is the average functional enrichment score obtained in 1000 repeated iterations, where n is the number of aligned residues randomly selected in the MSA subsets. "Random sequence selection" refers to the probability of randomly selecting a functionally correct sequence from the MSAs. For the BLAST method the All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001 dataset sequence ordering was used.

The results from the three example enzymes investigated in this section show that the SVM method of automatically predicting functional specificity determining residues is, in general, identifying functionally informative subsets of residues from the MSAs. This is seen, in particular, for the examples from the LDH/MDH and serine protease classes of enzymes. Both of these show that the functional re-scoring results, from using the SVM predicted subset of aligned residues, are better than (or equal to) all of the other methods (except for "optimal") for identifying functional specificity determining subsets of amino acids. This is especially pronounced when they are compared to the results from the "whole sequence" re-scoring methods, of BLAST and PAM30 (0,0). Those methods, like the SVM method, do not use the functional class of the aligned sequences as a pre-requisite for performing the alignment re-scoring and are therefore most closely comparable. In addition, the observation that the SVM method is performing at least as well as the *func-MB* based methods is very encouraging and shows that the SVM based method is able to perform comparably to methods that use additional functional information to aid the fSDR predictions.

The example from the nucleotidyl cyclases was less successful, with the residues identified by the SVM method only resulting in a functional enrichment score of 0.6 and an incorrect "top-hit" sequence. This example did, however, still show an improvement in the level of functional sequence enrichment, over the original functional sequence ordering from the BLAST database search. This can still be regarded as a positive result and shows that the SVM method is not reducing the level of functional specificity determining information and providing a worse result, when compared to the BLAST and random column selection results.

In summary, two of the three examples showed a marked improvement over the BLAST and PAM30 (0,0) methods, as well as equivalent or improved results when compared to the two *func-MB* methods shown in *table 5.5*. These are promising results that provide evidence for the advantages of using an SVM classifier to identify functionally specific residues, coupled with their subsequent use to improve the automatic assignment of enzyme function using only protein sequence based information.

5.3.4 Observations Regarding the *colgap_percent* Threshold Used in this Analysis

To complete this initial analysis of methods for the automatic identification of fSDRs using SVMs, it is important to address potential limitations of the single *colgap_percent* threshold, of 10%, that was used for the analysis of the *func-MB* and the *"random column selection method"* methods of aligned column selection and functional re-scoring. In retrospect, it may be that a value of 10% was not the best choice of the *colgap_percent* threshold for this purpose. This is highlighted by the results obtained in *chapter 4* (see, for instance *table 4.5* and *table 4.6*), which investigated the (*func-MB* based) functional re-scoring and sub-alignment selection methods that provided the best specific enzyme functional predictive performance. These showed that the optimal results, when using the "top-hit" functional assessment criteria, were achieved with varying *colgap_percent* thresholds that were dependant on the method of aligned column selection used to generate the sequence sub-alignments.

An alternative approach to the use of a single *colgap_percent* threshold, such as the 10% used in this analysis, could be through the implementation of a more sophisticated selection procedure for identifying the optimally performing set of fSDRs. In such a method, the identification of the fSDRs for inclusion in the SVM datasets would be carried out through an analysis of the functional enrichment score improvements when applying all of the analysed *colgap_percent* thresholds, rather than a single, fixed threshold that was used here. Alternatively, a higher single *colgap_percent* threshold of 50% may have been more appropriate. Firstly, because it is comparable to that used in the *profile-HMM* sub-alignment re-scoring method and also because thresholds above 50% were shown, in *chapter 4*, to have a minimal effect on the "top-hit" functional classification accuracy, when using the different *func-MB* sub-alignment selection methods.

It may be beneficial, in further studies, to extend this analysis to include a more detailed and thorough examination of the data in the ways suggested above. Including an analysis of both the functional enrichment scores and the resulting SVM classification performance, when using the "optimal" subset of fSDRs

obtained from altering the *colgap_percent* threshold prior to identifying the fSDR datasets.

Although there are possible ways in which the selection of the *colgap_percent* threshold could have been improved, it does not make the results obtained in the current analysis, from the use of a single threshold value of 10%, invalid. This is because the methods that have been used to identify the optimally performing sub-sets of fSDR (and non-fSDR) columns from each of the MSAs were designed to be optimal, regardless of the *colgap_percent* thresholds applied. Therefore, the subsequent analysis was optimal and valid for the particular datasets of fSDRs identified. The use of alternative, higher, *colgap_percent* thresholds may however, have identified some functional enrichment score improvements for certain MSA examples, thereby providing possible improvements to the SVM classification performance.

5.4 Conclusions

The aims of the analysis within this chapter were mainly twofold. To identify an optimally performing dataset of functional specificity determining residues (fSDRs) and to then use these to investigate methods for their automatic identification, from multiple alignments of homologous sequences, without prior knowledge of their functional classes.

To aid the identification of a benchmark dataset of fSDRs, a method for assessing the change in the rank-ordering of functionally "correct" sequences, from a series of sub-alignment based sequence re-scoring experiments, was implemented. This "functional enrichment score" was then used in conjunction with the *func-MB* and *profile-HMM* methods to identify potential functional specificity determining residues from sequence alignments where the specific functional sub-classes are known. It was shown that the use of a single threshold, for the selection of the fSDRs, was not optimal and therefore a more exhaustive method was implemented to search for those columns of residues (i.e., fSDRs) that provided an optimal rank-ordering of functionally "correct" enzyme sequences. This was based primarily on the functional enrichment of the top-10 ranking positions, but was also supplemented by more detailed selection criteria where necessary.

It has been acknowledged that some aspects of this fSDR selection process could be improved, for example by a more thorough search of the alignment re-scoring results obtained from using different levels of the *colgap_percent* threshold, rather than the 10% used in this study. However, until a large-scale, well defined and experimentally verified dataset of residues that determine enzyme functional specificity is available, it is suggested that a computational optimisation and selection process such as that used in this analysis is a valid alternative.

After identifying the fSDRs in each of the relevant sequence alignments, a non-redundant set of 357 alignments was partitioned into five cross-validation datasets. Because the number of positively identified fSDRs was considerably smaller than the non-fSDRs, it was decided to create five "randomly balanced" datasets containing equal numbers of fSDR and non-fSDR examples. This approach was inspired by previous SVM studies, involving the identification of catalytic residues within datasets of disparate numbers of positive and negative class examples. And was designed to dramatically improve the computational time and the SVM optimisation during training.

Initially, the aligned columns of fSDRs and non-fSDRs were encoded using the composition of the aligned residues. A 5-fold cross-validation training procedure was then carried out to assess the SVM parameters that generated the best models for the automatic differentiation between the two classes. Due to the un-balanced nature of the number of fSDRs and non-fSDRs in the testing datasets, averaged MCC values were used to determine the optimally performing SVM parameters. The resulting SVMs were shown to be generating poor predictions of the fSDR class of aligned residues, with an MCC, of 0.30, observed for the 5-fold cross-validation using the five randomly generated balanced sets of training data. Although this MCC indicates that the classifier was performing better than random, it was apparent that the large number of false positives being incorrectly classified as fSDRs was having a detrimental impact on the overall predictive performance of the classifiers.

In attempts to counteract these high rates of false positive detection, two approaches were investigated. Firstly, a more stringent E-value threshold, of 10^{-15}, was used to control which sequences would be included within the MSAs that were used as the source of the encoded SVM features. The aim of this was to improve the quality of

the input alignment data by reducing any potential noise contributed via false positive homologous sequences and sequence alignment errors. Secondly, the use of an additional input feature to the SVM was investigated. This was based upon an analysis of the number of distinct types of amino acids contained within the fSDRs and non-fSDRs. During the analysis of this feature it was shown, through the analysis of the well-studied substrate specificity determining residues of lactate and malate dehydrogenases, that a percentage threshold may improve this feature by reducing signal noise from residues that occur with relatively low frequencies.

Disappointingly, neither the more stringent E-value threshold nor the additional input feature was able to improve upon the earlier fSDR classification results. In fact they were both found to have a detrimental impact on the overall MCC values. In future work, further exploration of these results should be carried out, along with the investigation of additional input features. The large number of false positives that were identified by the SVMs is a problem that would benefit from further investigation, especially with regards to their use in a fully automated system for the recognition of functional specificity determining residues. The results, from using these quite simple input features to the SVM, suggest that there may be a high degree of ambiguity between the fSDRs and non-fSDRs that is causing the high levels of incorrect false positive classifications. This could be due to a number of factors, such as: the methods used to define the "optimal" fSDRs in each alignment; the lack of informative content in the input features used; and the quality of the sequence alignments used in the feature encoding. These are all important areas that would benefit from further study and are addressed in more detail in *chapter 7* (further work).

To complete this study, the performance of the SVM classifier on the functional scoring of three well-studied enzyme classes was compared to a number of other methods. Two of the these examples showed a marked improvement over the BLAST, PAM30 (0,0), and random selection methods; as well as equivalent or improved results when compared to the two *func-MB* methods, which have been shown to generally out-perform the other methods throughout this thesis. Although, as expected, the SVM method did show a tendency to predict many potentially false positive residues, these results did show that the sub-alignments identified by the

SVM classifier can be used to improve the functionally specific ranking of aligned enzyme sequences, in a fully automated system.

In conclusion, this chapter has described an automatic method for defining a dataset of functional specificity determining residues within alignments of protein sequences. This was necessary due to the current lack of a large-scale, experimentally verified benchmark dataset containing this information. Initial experiments were carried out that looked towards using support vector machines for the automatic classification of these functionally determining residues, from multiple alignments of homologous sequences. The overall classification results were found to be quite disappointing, especially with regards to the high rate of observed false positive predictions. The MCC values, although low, do however, indicate the SVMs are performing better than a random classifier and therefore show that further analysis of the SVM features should be able to improve the predictive performance.

Promising results were also obtained from the detailed investigation of three enzyme classes. These showed the advantages of using an SVM classifier to identify functional specificity determining residues, which can then be subsequently used to improve the automatic assignment of enzyme function, using only protein sequence based information, without any prior knowledge of the specific functional properties of the aligned sequences. These were all key goals of this thesis and show the further success that could be obtained from this method of functionally specific assignment.

Chapter 6 Summary and Conclusions

Due to the continuing growth of available protein sequence data from high-throughput genome sequencing projects and structural information from structural genomics initiatives; there is an important requirement, in bioinformatics and the biological sciences in general, for accurate, reliable and fully automated methods for the prediction of protein function. The time-consuming and expensive nature of experimentally assigning protein function continues to exacerbate this situation. These observations have driven the main aims of this thesis, which were the development and investigation of methods for improving the computational prediction of high specificity protein function. This effort has concentrated on approaches that make use of sequence information, from evolutionary related enzymes, to automatically assign high specificity details of the molecular function of enzymes. As noted previously, the difference between available protein sequence and structural data is significant. This disparity led to the decision to limit the functional classification task in this thesis to the use of protein sequence information because of its much wider availability with respect to that of experimentally determined protein structures.

In pursuit of these goals, four key areas of study were identified. First, a benchmark set of enzyme sequences, with a well-defined, specific, functional classification, which consist of multiple sequence alignments (MSAs) of evolutionarily related proteins were identified and defined. Following this, an investigation into the effect of alternative models of amino acid substitution – on the specific functional ordering of aligned enzyme sequences – was carried out. The third area of study looked into the use of methods, designed to score and identify residues that are closely related to the functional specificity of proteins (fSDRs), to improve the functional re-scoring and re-ordering of these enzymes. Finally, the possible use of these fSDRs to implement an automated method based on machine learning techniques, for the identification of functionally informative sections of aligned sequence homologs, was investigated.

This chapter provides a summary and explores the conclusions that can be drawn from the experimental outcomes of each of these studies. The salient points of

interest from the studies in each of the thesis chapters are described here within separate sections, which are structured according to the thesis chapter from which the experimental conclusions are drawn. In each section, a discussion of the main conclusions is presented, as well as a section that describes the practical incorporation of the findings into a fully automated, computational system for high specificity assignment of protein functions, using only sequence information.

6.1 *Chapter 2* – Investigation into the Functional Conservation of Enzyme Sequences and Dataset Definitions

This thesis started with the collection of a large set of functionally well annotated enzyme sequences from the Swiss-Prot protein sequence database. These sequences were then used to provide a study of the accuracy of the functional conservation of enzyme classes when using standard sequence similarity based measures of homology, to assign functional annotations to closely related sequences. It was shown during this assessment that even close relationships of sequence similarity, such as a sequence identity level of greater than 40%, do not, in general, provide confident transfer of specific enzyme molecular function. It was noted that conclusions of this nature had previously been reached from other studies carried out in this area. Although there is debate in the literature as to the exact level of accuracy that can be attributed to sequence similarity based methods of functional annotation, it is clear that there is certainly room for improvement. This is especially the case when considering the high specificity aspects of functional classification. The results presented in *chapter 2* demonstrate this, showing that as sequence similarity between proteins is reduced, the level of functional conservation, and therefore the accuracy of functional annotation via measures of sequence homology, is also reduced.

It was these outcomes, from assessing the performance at higher levels of functional specificity, using comparisons between the variation in the 3rd and 4th levels of the enzyme commission (EC) scheme of the sequence's enzyme functional descriptions, that highlighted the need for more powerful methods of discrimination between similar functional sub-classes. Towards this goal, a benchmark set of enzyme sequences were defined to enable an assessment of the effect that each of the

alternative methods investigated within this thesis had on improving the specific functional similarity between enzyme sequences.

The form of the benchmark dataset was that of a set of multiple sequence alignments of closely related protein sequences, which were generated through the use of a PSI-BLAST sequence database search. A major criterion for this benchmark dataset was the ability to provide a comparison between methods of improved functional classification, when using the "top-hit" method of assigning functional specificity. That is, the specific function of the sequence deemed to be of greatest functional similarity to the "unknown" query sequence would be directly transferred to the query sequence. For this purpose, it was decided that the baseline method of comparison should be that of the ability of the PSI-BLAST sequence similarity based rankings, to assign enzyme function, with a specificity of all four levels of the EC classification scheme. Therefore, examples were selected that showed the occurrence, in all cases, of an enzyme at the highest level of sequence similarity, but with a specific function different to the query.

However, due to a limitation in the number of available examples of this type, a larger set of "artificial" sequence alignments were identified through the use of a method that progressively removed the most significantly similar sequences with an identical function to that of the query sequence. This expanded dataset provided a much larger number of sequence examples, and associated enzyme functions, on which to base the results and associated statistical inferences.

In conclusion, the initial studies that were carried out in this part of the thesis were able to illustrate the importance and timely nature of the research problem, showing the need for improved methods to reliably assign specific properties of molecular function, in an accurate and automated system.

6.2 *Chapter 3* – The Use of Alternative Amino Acid Substitution Matrices for Rescoring the Functional Similarity of Enzyme Sequences

The observation that a simple sequence similarity threshold was not sufficient for consistent, accurate, functional annotation of enzyme sequences, led to the aim of developing methods that could provide improved power when discriminating

between sub-types of specific function. The studies, presented within *chapter 3*, looked at using non-standard amino acid substitution matrices towards the goal of rescoring the functional similarity of enzyme sequences. This approach investigated the use of a range of BLOSUM and PAM matrices, as well as a residue IDENTITY matrix, to calculate a modified sequence similarity score between the query sequence and associated enzyme homologs within each of the MSAs in the benchmark datasets. This recalculated similarity score was then used to re-rank the aligned sequences and provide an assessment of the effect that each of the alternative amino acid scoring matrices had on the level of correct functional classification. The "top-hit" method of assessing the functional assignment accuracy was used throughout this analysis.

Three alternative methods for generating datasets of multiple sequence alignments were used in these studies. In each of these, a residue-masked, gapped BLAST sequence database search was used to generate multiple alignments of homologous sequences; for which, one of either: BLOSUM62; PAM160; or PAM30 was used as the search amino acid substitution matrix. Suitable gap-scores were researched and used for each matrix. The resulting MSAs were subsequently modified to form three "artificial" datasets, for which the sequence with the most significant level of sequence similarity to the query (used in the database search) was functionally "incorrect" (i.e., annotated with a different specific function to the query enzyme sequence).

Initially, the dataset formed from using a BLOSUM62 matrix was analysed, using a gap opening and extension penalty of -11 and -1, respectively, for the gapped BLAST alignments. The outcome of these studies showed that, in general, MSAs containing un-masked amino acid residues resulted in consistently larger proportions of correct, specific, functional assignments, when they were used in the alignment re-scoring. This result was observed regardless of the parameters used for the alignment re-scoring, or the parameters used for the generation of the three analysed datasets. It is thought that the reason for this observed improvement, in assigning functional similarity when not including masked residues, may be explained by the intended use of sequence masking. In general, it is recommended that sequence regions of low information content should be masked when carrying out sequence

database searches. This aims to reduce the number of false positive homologous sequences that are identified, whereas it is thought that the un-masked sequence residues provide additional functional information, resulting in improved specific functional ordering of the aligned enzymes.

Studies using the un-masked form of this dataset subsequently showed that the PAM30 substitution matrix, with an "un-gapped" gap scoring model, provided the largest proportion of correct functional classifications. This resulted in an observed bootstrap mean value of 0.631, that is, 2226 out of 3527 correct classifications of specific enzyme function. When compared to the optimally performing BLOSUM series matrix, which was BLOSUM-100, this showed a percentage increase of 14.1%, or 482 correct classifications. Furthermore, there was a general trend towards improved levels of correct functionally specific classifications when re-scoring the alignments with PAM-N matrices with progressively lower N values. With an optimal performance observed with the PAM30 matrix.

To assess whether these results were uniquely related to the particular dataset of MSAs used (i.e., the *All1stINCORRECT.tF.BLOSUM62.unmasked.E0.001* dataset), two alternative sets of sequence alignments were generated and investigated. In particular, this was done to investigate whether the observed improvement in functional classification was due to the specific ordered combination of the matrix used in the gapped BLAST database search and that used for the subsequent functional alignment re-scoring.

First, a PAM160 matrix, with gap opening and extension penalties of -11 and -1, respectively, was used to generate the sequence alignments from the database search. This matrix (and gap penalties) was used because it was identified as the closest PAM equivalent to the BLOSUM62 matrix, which was used to generate the previous dataset. An identical set of non-standard amino acid substitution matrices were used to functionally re-score these alignments. It was found that a similar functional re-scoring outcome was obtained for the alignments generated with a PAM160 matrix as was obtained from the BLOSUM62 based alignments. Specifically, there was a very similar peak in correct functional "top-hit" performance when the PAM-N series matrices, with low *N* values (such as 30), were used for the re-scoring. This resulted in a bootstrapped mean proportion of correct predictions equal to 0.611,

which corresponded to 1894/3100 correct classifications of specific enzyme function. This result suggested a conclusion of a general, and comparable, improvement in the specific functional classification of enzyme sequences when using "low" PAM-N substitution matrices (such as PAM30) to functionally re-score BLAST alignments that were generated with either BLOSUM62 or PAM160 matrices.

Following this, a PAM30 matrix, with gap opening and extension penalties of -9 and -1, was used to create a further comparison dataset of BLAST-based MSAs. These sequences were again re-scored using the same substitution matrices. The aim of this third set of analyses was an investigation into the effect that the order, in which the PAM30 and BLOSUM62/PAM160 matrices were applied in the BLAST-based alignment generation (and subsequent functional re-scoring), had on the resulting level of correct functional classifications. The results showed that there were no peaks in functional classification performance, when using either the BLOSUM62 or PAM160 matrices to re-score the PAM30 BLAST-generated MSAs. This led to the conclusion that the improvements in functional classification were due to the particular ordered combination, of the BLOSUM62 or PAM160 substitution matrices used for the alignment generation, and the PAM30 (and other lower PAM-N matrices) used for the subsequent alignment re-scoring.

It was postulated that the observed results were possibly due to the intended usage of the different types of amino acid substitution matrices. For instance, it was noted that a common use of the BLOSUM62 matrices is in sequence database search applications, such as PSI-BLAST. This is primarily because this particular substitution matrix has been shown to provide optimal performance for sequence homology detection, when searching over a diverse sequence space, while also providing high quality alignments. Also, it was suggested that the PAM matrices that performed best when functionally re-scoring the alignments, may be associated with the evolutionary distance between the homologous sequences. Therefore, the lower PAM-N matrices could, in general, be providing additional evolutionary information, which improved the separation of the functionally specific properties of the closely related enzyme homologs that were identified in the database search.

These results were supplemented through a reduced set of enzyme query sequences that were clustered, at differing levels of sequence identity, with the aim of investigating the effect that potential sequence redundancy (within the benchmark datasets) may have had on the accuracy and trends seen in the functional re-scoring results. From this investigation it was shown that, in general, similar results and trends in the classification results were obtained when using comparable re-scoring matrices with each of the clustered sets of query sequences. An approach to automatically assigning a specific molecular function to an unknown enzyme sequence, using these findings, is shown in the "PAM30" branch of *figure 6.1*, which is discussed in more detail, in *section 6.5*.

In conclusion, the experimental outcomes, from functionally re-scoring multiple alignments of homologous enzyme sequences, show that there is significant evidence for using additional PAM matrices to improve the assignment of functionally specific enzyme properties. In particular, the PAM-N matrices, with lower evolutionary distances, showed a general tendency towards increasing levels of correct functional classification. This was clearly seen when functionally re-scoring alignments (which were generated via gapped BLAST with a BLOSUM62 (or PAM160) substitution matrix) using a PAM30 matrix. Although, there was significant improvement in the accuracy of specific functional classification when adopting this approach; there was still room for improvement in the overall proportion of correct classifications arising from any one method.

6.3 *Chapter 4* – Identification of Functional Specificity Determining Residues

In an attempt to improve on the levels of correct functional classification achieved through the use of additional substitution matrices, a more refined process of functionally re-scoring sets of aligned sequence homologs was assessed in *chapter 4* of the thesis. For this, automatic methods for the scoring and identification of functionally important amino acids were implemented. It was decided that two previously studied methods for this purpose would be investigated; they were the *func-MB* method and the *profile-HMM* method. A key hypothesis behind this approach was that, through the identification of subsets of aligned residues that were closely correlated to the functionally specific properties of the enzyme sequences, it

would be possible to generate more accurate automatic assignments of specific function. In contrast, the methods described in *chapter 3*, used all of the amino acids that were aligned in the multiple sequence alignments, returned from the gapped BLAST database search, rather than a functionally correlated subset.

Both the *func-MB* and *profile-HMM* methods of fSDR identification were used to score the columns of amino acids that were aligned to the query enzyme sequences, in each MSA, from the benchmark dataset. For the *func-MB* method, the outcome was a set of aligned column scores represented by the Spearman-rank order correlation coefficient. This was calculated through a comparison between the correlation of the aligned residue similarities and the specific enzyme functional class of the aligned sequences. Whereas, the *profile-HMM* method used a scoring system that takes into account the variation in relative entropy, of each of the aligned residues, when taking into consideration the specific functional sub-types. The basis of this was the scoring of amino acid positions, using a Z-score, which calculated the relative degree of residue conservation within enzyme groupings with the same specific functions. In turn, this enabled a ranking of residues that were most likely to be conserved within groups of aligned sequences with the same function, but differ between them.

These scores, associated with the potential fSDRs, were then ranked and a number of approaches, with varying thresholds, were used to generate sub-alignments from each of the MSAs in the benchmark dataset. It was then possible to re-score the aligned sequences using only those aligned amino acids within the selected sub-alignments. For this, a PAM30 amino acid substitution matrix was used and the resulting accuracy of specific enzyme classification was determined via the "top-hit" method of assessment. An additional method, designed to randomly generate sub-alignments of aligned residues, was also compared to the results from each of the *func-MB* and *profile-HMM* based methods, as well as those from the amino acid re-scoring outcomes from the analyses of *chapter 3*.

When investigating the functional classification accuracy resulting from the *func-MB* method, there was significant improvement observed when compared to the other methods of functional re-scoring looked at in this chapter. In particular, the best results were obtained when using the "top-N" and "top-X percent" methods of

selecting sub-sets of aligned residues were used, for which N=30 and X=8%, respectively. Both of these thresholds, for selecting sequence sub-alignments, provided very similar levels of correct enzyme predictions after functional sequence-based re-scoring was carried out and then assessed using the "top-hit" method. The overall optimal result was shown when re-scoring the sequence sub-alignments constituting the top-8% of aligned residues, as calculated from the Spearman-rank order correlation coefficients, which showed a (bootstrap mean) proportion (and number) of correct functional assignments of 0.769 (2712/3527). However, these were not significantly different to those of the top-30 method. It was also concluded that, in general, there was no significant difference in the functionally specific re-scoring results when using any pre-filtered alignments that have an applied *colgap_percent* threshold of 50%, or more.

It was shown, in comparable studies, that when the *profile-HMM* method was used as the basis for scoring the potential fSDRs - and therefore defining the sequence sub-alignments to be used in the functional re-scoring – the accuracy of "top-hit" based enzyme classification was less than that of the *func-MB* method. It was concluded that further study is required to fully understand the reasons behind the relatively disappointing results, when using the sub-alignments generated through use of the *profile-HMM* scored fSDRs, rather than those from the *func-MB* method.

Although the optimal results seen from re-scoring the sequence alignments with the *profile-HMM* method were not as impressive as those of the *func-MB* method, they were still a significant improvement upon the best "non-fSDR" based methods of sequence alignment re-scoring. This was seen when comparing with the optimal re-scoring results, from *chapter 3*, which used all of the amino acids in the multiple alignments to functionally re-score each of the enzyme sequences. For instance, the optimal profile-HMM method, which used the columns with the top-35 Z-scores to form the sub-alignments from each of the dataset MSAs, showed an improvement of 4.2% (i.e., an additional 148 correct classifications) when compared with the *PAM30 UNGAPPED (0,0)* method of alignment re-scoring, described in *chapter 3*.

This limited, yet still important observation, added weight to the overall conclusion drawn from this area of research; which was that the use of a functionally correlated subset of amino acids generally provides an improved method for the assignment of

specific enzyme function. This was most clearly seen with the large improvement - of 486 correct functional classifications, which corresponds to a difference of 13.8% - when using the highest scoring 8% (or top-30) of the aligned residues calculated from the *func-MB* method, rather than the *PAM30 (0,0)* method of alignment re-scoring, described in *chapter 3*.

As in the analyses of *chapter 3*, a series of sequence identity clustering thresholds were applied to the query sequences that provided the source of the benchmark dataset of MSAs. This resulted in a number of sub-datasets of MSAs, which provided a means to investigate the potential effects that sequence redundancy, within those query sequences, may have on the measured levels of functional classification accuracy. A repeat of the sequence sub-alignment based functional re-scoring, using each of these sub-datasets of MSAs, showed slight variations in results from using both the *func-MB* and *profile-HMM* based methods. Namely, as more stringent sequence identity clustering thresholds were used on the query sequences; the observed specific functional classification accuracies were shown to tend towards increases for the *profile-HMM* based re-scoring method, and decreases for the *func-MB* based method. These slightly contrasting results did not, however, detract from the main conclusions drawn from this area of study: that the *func-MB* based method of sub-alignment generation, and subsequent re-scoring, consistently outperforms the comparable profile-HMM based method; and that there is a general improvement in the accuracy of specific enzyme functional classification when relatively small sub-sets of functionally correlated amino acids are used for the purpose of functionally re-scoring and re-ranking alignments of homologous sequences.

To conclude the studies in this chapter, an in-depth analysis of the experimentally and computationally well-studied family of lactate and malate dehydrogenases (LDH/MDH) was presented. These enzymes provided a good example of the variations, in substrate binding specificity, which can result from evolutionary divergence in small areas of the protein sequence, while also providing experimental verification for some of the predicted fSDRs. This example clearly showed the benefits of using sub-alignments of sequence residues - that were highlighted to be well correlated with the substrate specificity of these enzymes – to provide marked

improvements in the functional ranking (and grouping) of sequences with the same specific function as the query enzyme. This was shown through the use of both the *func-MB* and *profile-HMM* based methods of fSDR identification; with clear improvements over the original BLAST sequence ordering seen for both of the methods.

These observations suggested an alternative, improved, approach for automated functional assignment, when compared to both the BLAST and the PAM30 (0,0) functional re-scoring methods. In particular, the *func-MB* method showed the largest improvement in correctly assigning specific functional classifications after re-scoring. The use of this automated approach is highlighted in *figure 6.1*, in the "*func-MB*" branch of the flowchart, and is discussed in more detail below, in *section 6.5*.

6.4 *Chapter 5* – Towards the Identification of Functional Specificity Determining Residues Using Support Vector Machines

It was concluded that it is beneficial to use sub-alignments of predicted fSDRs to improve the functional assignment accuracy of enzyme sequences. However, implicit within the *func-MB* and *profile-HMM* based automated approaches is a prior knowledge of the specific functional classes of the aligned protein sequences. Therefore, it was proposed that alternative methods should be explored, which had the aim of identifying functionally important sub-sets of amino acids, which could be subsequently used to improve the functional assignment of specific functional information – without prior knowledge of the functional classes of the aligned sequence homologs. For this purpose it was suggested that a machine learning based approach, such as support vector machines (SVMs), could be appropriate.

It was first necessary to gather appropriate data for the training and validation of an SVM based method that could identify functionally specific residues, within MSAs. For this, a method was described that identified a, mutually exclusive, set of fSDRs and non-fSDRs from each of the multiple sequence alignments. A "functional enrichment score" was used to identify the aligned residues that gave an optimal rank-ordering of the enzyme sequences, with the same "functionally correct" specific

EC classification as the query sequence. Both the *func-MB* and *profile-HMM* methods, for scoring and identifying possible functionally informative residues, were used as the basis for generating a number of sequence sub-alignments from each of the MSAs. These sub-alignments were then re-scored and their "functional enrichment scores" were compared.

Through this analysis it was shown that the use of a single threshold, for selecting the predicted fSDRs that form these sub-alignments, was not optimal. This led to the implementation of a method that systematically compared the "functional enrichment scores" that were obtained from applying the: "top-N"; "top-X percent"; and "column score threshold", sub-alignment selection criteria (described in *chapter 4*) to the scores calculated with the *func-MB* and *profile-HMM* methods. The functional enrichment score calculations were based primarily on the number of sequences in the top-10 ranking positions, after alignment re-scoring, with the same "correct" enzyme class as the query. The result of this was a thorough comparison between the functional re-scoring performances of each of the alternative sequence sub-alignments. This allowed the definition of an optimal set of fSDRs (and therefore also the non-fSDRs) associated with each of the MSAs in a benchmark dataset of function specificity determining residues.

A stringent sequence clustering procedure was used to remove any potential sequence redundancy from the multiple sequence alignments. This was important because the sets of fSDRs and non-fSDRs that were extracted from these alignments were to be used as the training and validation datasets for the parameter optimisation of the SVM. The resulting 357 multiple sequence alignments were then partitioned into five, non-redundant, cross-validation datasets. It was observed that the balance between the number of fSDRs (positive SVM classes) and non-fSDRs (negative SVM classes) was unequal, with a much larger number of negative than positive class examples. This was not unexpected, as it was previously shown that a general performance increase, in functional classification accuracy, was seen when using relatively small subsets of fSDRs to re-score and assign specificity. In order to improve SVM optimisation and improve computational training times, a similar approach taken by previous studies in a related area, which had used machine learning techniques to identify catalytic residues, was followed. This involved

randomly selecting an equal number of fSDR and non-fSDR examples from each of the MSAs in the 5-fold cross-validation training datasets.

Features that were expected to describe the functionally informative relationships, between the aligned sequence homologs, were used to explore the feasibility of the proposed SVM-based approach to fSDR identification. To assess the performance level of the classifiers, a 5-fold cross-validation procedure was used to identify the optimal SVM learning kernels and associated parameters. Although the SVM training was carried out using the "randomly balanced" datasets, it was noted that it was important to assess the classifier performance on the full "un-balanced" test datasets of MSAs. This was because this is the form of the data that would be used as input to the SVM in a real biological example, which required the classification of functionally determining residues. The un-balanced nature of the positive (fSDR) and negative (non-fSDR) examples in the test datasets meant that a Matthews Correlation Coefficient (MCC) was used as the measure of classification performance.

The composition of the types of aligned amino acids (*AA_composition*), as well as measures of the number of distinct amino acids types (*NumberOfAATypes* and *NumberOfAATypes_threshold_X%*) occurring within each of the aligned columns of fSDRs and non-fSDRs, were used to encode the feature vectors for input to the SVMs. Two alternative forms of the multiple sequence alignments were used as the input for encoding these feature vectors. They were those generated from using an E-value threshold of either 10^{-3} or the more stringent threshold of 10^{-15} during their generation via a BLAST database search. The lower threshold was selected, in conjunction with an analysis of the distributions obtained from analysing the properties of the *NumberOfAATypes* feature vector, with the aim of reducing the number of more distantly related sequence homologs present in the alignments. The reasoning behind this was that it may reduce the number of false positive detections of fSDRs, made by the SVM classifiers (when using the 10^{-3} threshold); by reducing the level of potential signal noise contained within the sequence alignments.

Interestingly, it was shown that the best MCC classification results were obtained from the SVM classifiers when using the *AA_composition* feature vectors, encoded from the alignments generated using the less stringent E-value threshold of 10^{-3}.

Although the SVM classification results were shown to feature a relatively large number of false positive fSDR predictions; an overall MCC of 0.30 was observed. This was considerably better than a random classifier and comparable to other studies that have used machine learning methods to identify catalytically important residues from multiple sequence alignments. It can, therefore, be concluded that an SVM-based approach - to the problem of automatically identifying residues that determine functional specificity - provides a novel and successful approach to this important area of study.

An additional assessment of the SVM-based classification method was carried out through a detailed investigation of its application to three well-studied classes of enzymes. These were: the lactate/malate dehydrogenases; the nucleotidyl cyclases; and the serine proteases. For each of these, an SVM was used to identify a set of potential function specificity determining residues. These were then used to form sequence sub-alignments, which were functionally re-scored using a PAM30 amino acid substitution matrix. The performance of this sequence re-scoring was assessed through comparison with a number of methods that have been studied throughout this thesis. It was shown that, in general, the SVM-based method was performing well when compared to these alternatives. In particular, the serine protease and lactate/malate dehydrogenase examples showed a clear improvement in comparison to the both the "whole" alignment (i.e., *BLAST* and *PAM30 (0,0)*) and the random selection methods. In addition, the results for these two examples were shown to be equal to, or better than, the two *func-MB* based methods that were used in the comparison. This was a particularly encouraging result and provides important evidence for the use and continued study of the SVM-based method for the automatic assignment of functional specificity to enzyme sequences.

6.5 Summary of Methods

To conclude this thesis, a summary of the key methods that have been investigated and their potential application to the task of annotating the specific molecular function of an enzyme sequence, is given. An overview of the key methods and their practical application is shown in the flowchart of *figure 6.1*. The figure shows three alternative routes that could be followed to help determine the specific functional class of an unknown sequence. The initial step for each of these three methods is the

generation of a multiple sequence alignment, via a gapped BLAST database search, using the unknown protein. Before the sequence database search is carried out a residue masked form of the database should be generated, using the *pfilt* application. Also, *figure 6.1* highlights the BLAST search parameters that were used to generate the MSAs that have been analysed throughout the thesis and from which the results and conclusions have been drawn. They were, the use of: a single search iteration; a BLOSUM62 amino acid substitution matrix; gap scoring penalties of -11 (gap opening) and -1 (gap extension); an E-value threshold of 0.001 to determine the sequences that should be included in the MSA; and the application of the SEG low complexity residue filter to the query sequence. Once the MSA has been generated, the flowchart shows three alternative methods (the "select method" junction) that can then be followed to assess the specific functional class of the query sequence.

The *PAM30 (0,0)* Method

One method, denoted by the *"PAM30 (0,0)"* branch of the diagram, shows the application of the PAM30 amino acid substitution matrix to functionally re-score the aligned sequences. This method refers to the optimal alignment re-scoring method that was identified in the analyses of *chapter 3*. After generation of the MSA, any masked sequence residues should be replaced with the residues present in the source sequences. The MSA can then be re-scored (using a PAM30 substitution matrix, with both gap opening and extension penalties of zero) and the sequences re-ordered accordingly. Functional classification of the unknown sequence can then be carried out, via annotation transfer, from the sequence with the "top-hit" (after re-scoring) to the unknown query.

The *func-MB* Method

The *"func-MB"* branch of *figure 6.1* provides an alternative method for the automatic functional classification of the unknown query sequence. This method refers to the optimal *func-MB* sub-alignment re-scoring method, which was identified in the analyses of *chapter 4*. After generation of the MSA, a series of steps are carried out to identify the aligned residues that are predicted to be most closely correlated with the specific molecular function of the aligned sequences. Any residue masking is first replaced with the amino acids present in the source sequences. Following this, because the *func-MB* method requires prior knowledge of

the specific functional sub-classification of all the aligned sequences, it is necessary to remove those sequences which do not have complete, specific, functional annotations (i.e., complete annotation at all four levels of the EC classification scheme) from the MSA. A further filter should then be applied to the MSA to remove all aligned columns with a proportion of gaps greater than 50%. For each of the remaining columns of residues, aligned to the query sequence, a Spearman-rank order correlation coefficient is calculated for the correlation between the residue similarity and the specific functional similarity of each of the aligned sequences. From this, the residue positions with the highest correlation coefficients can be identified and subsequently used to extract a sequence sub-alignment that is enriched with high ranking function specificity determining residues. It was shown, in the results of *chapter 4*, that the columns with the top-30 (or top-8%) largest correlation coefficients should be used to form these sub-alignments.

Once the sub-alignment of fSDRs has been defined, the same sequence re-scoring procedure as used in the *PAM30 (0,0)* method, should be followed. That is, a PAM30 amino acid substitution matrix, with gap penalties of zero, should be used to score and re-rank the functional similarity of the aligned sequences to the query. Functional classification of the unknown query sequence can then be carried out via annotation transfer from the highest ranking ("top-hit") sequence, after re-scoring.

In *chapter 5*, an alternative "functional enrichment score" method of assessing the level of correct functional classification was also investigated. This was primarily used to identify a benchmark set of fSDRs; however, it can also be used to add an additional level of validation to the functional assignment of an unknown query sequence. It was shown that only the residues with the ten largest correlation coefficients are required to form sub-alignments that generate the largest proportion of functionally correct sequences in the highest ranking positions, after alignment re-scoring. Therefore, in the practical application of this method, if the majority of sequences in the first 10 ranking positions, after alignment re-scoring, have the same specific function as the "top-hit" from re-scoring the sub-alignment of the top-30 fSDRs, further confidence will be added to the likelihood of a correct functional classification of the query.

The *SVM* Method

The *"SVM"* branch, in *figure 6.1*, highlights the stages required to apply the SVM classification method to the problem of automatically determining the specific molecular function of an unknown protein sequence. First, the SVM input is encoded directly from all of the sequences within the BLAST generated MSA, using the *AA_composition* feature vector. The best performing SVM classifier, defined in *chapter 5*, is then used to identify the residues in the query sequence that are predicted to be function specificity determining (fSDRs). A sequence sub-alignment of these "positive" SVM-predicted fSDRs should then be extracted and any residue masking should be replaced with the amino acids present in the source sequences. Finally, a PAM30 amino acid substitution matrix, with gap penalties of zero, should be used to re-score and re-rank the functional similarity of the sequences, in the sub-alignment, to the query. Functional classification of the unknown query sequence can then be carried out.

As with the *"func-MB"* method, the confidence of a correct functional classification will be increased if a majority (ideally, 90% or greater) of sequences, with the same specific molecular function, are observed with the closest similarity to the query sequence. Also, if the "top-hit" sequence has the same specific functional class as this majority group, further confidence can be attached to the transfer of this specific molecular to the function of the unknown query sequence.

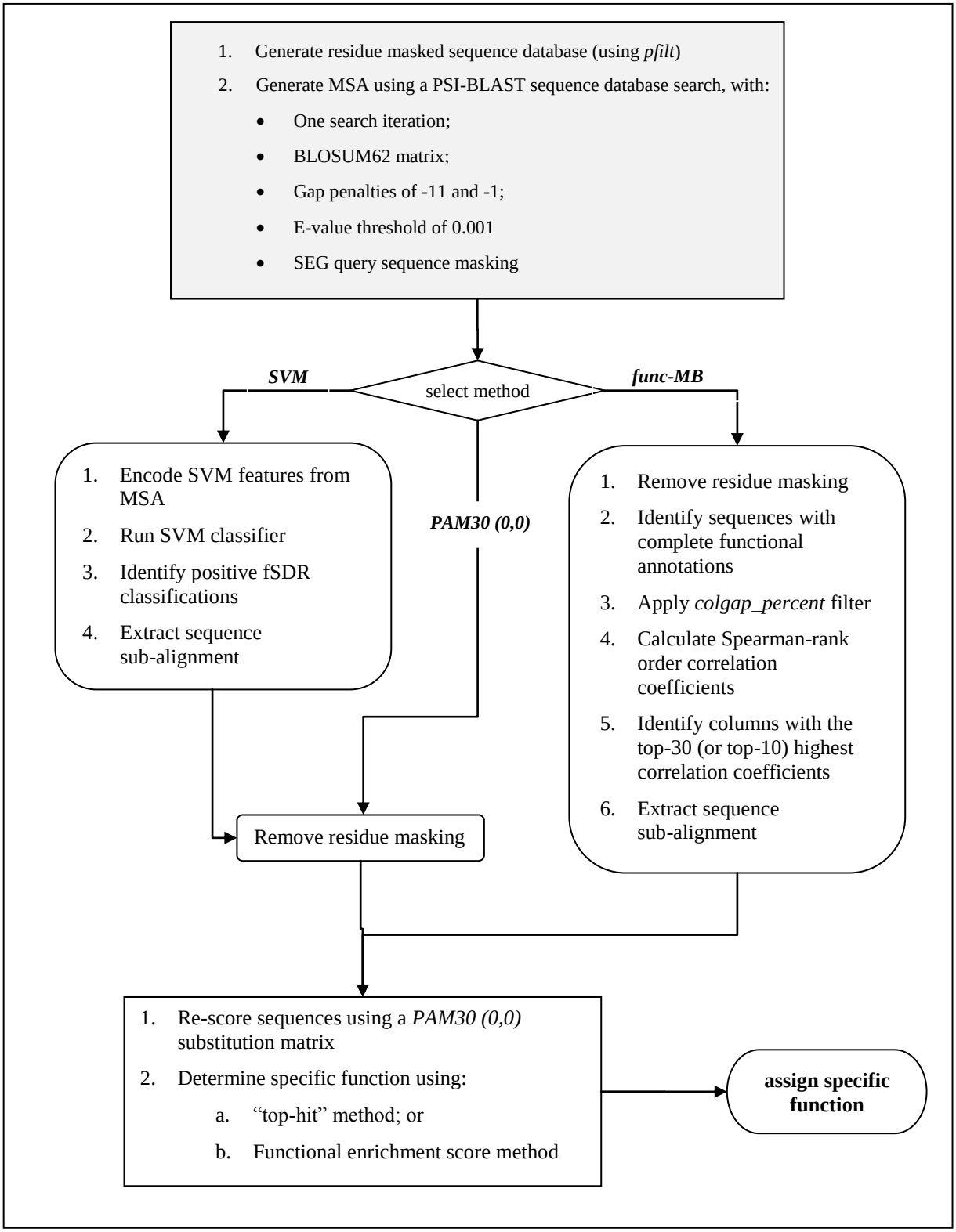

Figure 6.1. A flowchart showing an overview of the alternative functional re-scoring methods that have been researched in this thesis. It shows a summary of the practical stages required to assign a specific molecular function to an enzyme sequence of unknown function.

6.6 Towards Implementation of a Production System

To provide the wider biological research community with access to the prediction methods investigated in this thesis, it is proposed that a public domain web-based system should be developed. This would provide a user-friendly interface allowing submission of an unknown sequence, in FASTA format, for specific functional analysis and subsequent annotation. Depending on the processing times, which would require benchmarking, the results would be returned directly on the web-page or via an e-mailed link to the results.

The key stages that would need to be implemented for this automated function prediction system are described in the flowchart shown in *figure 6.1*. After sequence submission, a single iteration PSI-BLAST search (of a *pfilt* filtered version of the *UniProt* sequence database) would be carried out, using a BLOSUM62 matrix, gap-penalties of -11 and -1, and an E-value threshold of 0.001. The input sequence would also be filtered for low complexity, using SEG, prior to the BLAST search. The resulting MSA would then be automatically processed in accordance with the methods described in the relevant branches of the flowchart. In particular, the methods of most importance for a production system would be those of: (i) the optimally performing *func-MB* method; (ii) the novel SVM prediction method; and (iii) to provide a comparison, the original BLAST search results. Until further benchmarking has been carried out on the SVM method, it is the *func-MB* alignment re-scoring results that provide the most informative and reliable predictive results for specific enzyme function. Detailed descriptions of the steps necessary for generating a specific enzyme prediction for both of these methods are provided in *section 6.5* and *figure 6.1*.

An interactive analysis of the prediction results would be provided to the user, with links to the *UniProt* sequence database and the key parameters associated with each of the functional re-scoring methods highlighted. This would include the BLAST search results, the generated MSA and the fSDRs used for the sub-alignment re-scoring in each of the *func-MB* and SVM methods (along with the associated Spearman-rank order coefficients and SVM classifier results). In order for the system to make a high quality, reliable, specific function prediction a majority (ideally, 90% or greater) of sequences, with the same specific enzyme function,

should be observed in the top 10 ranking positions. As shown in *chapter 4* (*figure 4.9* and *table 4.6*), when using the optimal *func-MB* re-scoring methods the benchmark top-hit accuracy of a correct specific enzyme function assignment is between 76.7% and 76.9%. Also, if the "top-hit" sequence has the same specific functional class as this majority group, the system would attach greater confidence to the prediction. A further test of the quality of the prediction should be provided by comparing the results from the *func-MB* and the SVM methods. If they agree the user would be able to take more confidence from the prediction. In addition to the prediction of function for individual sequences, this system could also be tailored towards a high-throughput analysis and genome-wide annotation of specific enzyme functional class.

6.7 Overall Conclusions

The aim of this thesis was the investigation of computational methods for improving the prediction of the specific molecular function of protein sequences. Due to the increasing disparity between the numbers of proteins deposited in sequence databases and those with high quality functional annotations, it was important that these methods were fully automated. In approaching this, a benchmark set of functionally well-annotated enzyme sequences were defined and a series of associated BLAST generated multiple sequence alignments generated. Three main areas of research were undertaken in this work, which each investigated alternative, but related, approaches to the problem.

The first approach used non-standard amino acid substitution matrices to functionally re-score the sequences within the BLAST alignments. This showed that it is possible to obtain improved levels of specific functional classification by using the PAM30 substitution matrix to re-calculate the similarity between the sequences identified by BLAST and the query sequence. Following on from this simple approach, it was shown that the use of methods, designed to automatically identify subsets of function specificity determining residues, are able to provide a further improvement to the level of correct functional classification.

The disadvantage of these "fSDR-based" methods was their inherent reliance on the prior knowledge of the specific functional classes of the aligned sequences. To counter this, a novel method, which used an SVM classifier to predict function

specificity determining residues in multiple sequence alignments, was developed. Using thorough cross-validation, it was shown that this predictor was performing much better than random. Further, the detailed analysis of three well-studied enzyme examples showed that, in two thirds of the examples, the SVM method gave specific functional classification results comparable to the earlier fSDR methods. This was a particularly important result, because it showed that the SVM method, which does not require prior knowledge of the functional classes of the aligned sequences, can be favourably compared to those methods which do.

This was particularly advantageous for a fully automated method of this type, where the number and diversity of the pre-existing, specific functional annotations may be limited. Further work is still required to obtain a thorough benchmark of the predictive performance of the SVM classifier, when compared to the other method, however these results were very encouraging and they successfully fulfilled the key aim of this thesis.

Chapter 7 Further Work

This chapter details some additional avenues of study that could be explored to improve on and provide comparative analysis to the methods of enzyme functional analysis, presented in this thesis. Where applicable, the proposed suggestions for further work are described within separate sections that are structured according to the thesis chapter in which the additional analysis is most closely related. This is not an exhaustive list of further work and it is not known to what extent the proposed work would impact on the current results. It is, however, expected that the following analyses would improve upon the work presented so far and would be complementary to the work presented in this thesis.

7.1 *Chapter 3* – The Use of Alternative Amino Acid Substitution Matrices for Rescoring the Functional Similarity of Enzyme Sequences

The work presented in *chapter 3* looked at using non-standard amino acid substitution matrices for rescoring the functional similarity of enzyme sequences. Two main sets of amino acid substitution matrices were investigated in this analysis, the BLOSUM and the PAM series of matrices. Also, these were compared to the use of a residue identity based matrix – the IDENTITY matrix. Below are some further suggestions as to which additional types of substitution matrices could be added to this analysis, as well as ways in which to carry out a more detailed analysis of the alignment re-scoring parameters used.

7.1.1 Investigation of Additional Substitution Matrices

7.1.1.1 JTT-PAM-N Matrices

The alignment re-scoring analysis in this chapter could be extended by comparing the results obtained from using the PAM-N series of matrices with the updated JTT-PAM-N matrices (Jones et al., 1992). These matrices use a similar evolutionary model to the "original" Dayhoff form of PAM matrices for matrix generation, but use a larger and updated dataset. The main reason for including these types of matrices in the analysis is that they should provide a good comparison between the

two types of PAM matrices to see if there are comparable trends in the prediction accuracies between the two.

7.1.1.2 Enzyme Specific Matrices

A further approach could be the use of an "enzyme specific" substitution matrix for functionally re-scoring the aligned sequences. This would be done by applying the PAM model for matrix generation to the enzyme sequences gathered from Swiss-Prot. During the matrix calculation, amino acid mutations should only be considered between protein sequences with the same specific function. This would then provide a measure of the mutational probabilities for each of the amino acid types when enzyme functional constraints are in place.

7.1.2 Further Analysis of Gap Scoring Parameters

A further area of study, related to the analysis carried out in *chapter 3*, is the optimisation of the parameters used for scoring gaps when re-scoring the aligned sequences. The analysis presented in this chapter currently uses two methods for scoring the gap opening (g_{open}) and gap extension (g_{extend}) parameters. They are: (i) the "un-gapped" gap scoring model, which scores all residue alignments to gaps as 0; and (ii) the "gapped" scoring model, which uses the same gap scoring penalties as those used in the BLAST sequence database search to generate the MSAs for each of the datasets analysed.

It is known that when aligning sequences, the optimal gap penalties can be sensitive to the particular amino acid substitution matrices that are used for the alignment scoring (Frommlet *et al.*, 2004; Reese and Pearson, 2002). This is of particular importance for the pair-wise sequence alignments that are carried during a database search (such as BLAST) for sequence similarity. Although the analysis of the sequence alignment re-scoring, carried out in *chapter 3*, does not involve the re-alignment of any of the sequences it may still benefit from using gap parameters that are optimised to the particular matrix used in the alignment re-scoring. To do this, it is proposed that a grid-based, optimisation procedure could be carried out to provide an assessment of the effects, on the alignment re-scoring results, of a series of alternative pairs of gap opening and gap extension parameters. Therefore, this approach would build upon the results shown in *chapter 3*, by allowing the

identification of an optimal pair of gap scoring parameters for each amino acid substitution matrix, as measured by their ability to correctly assign the specific enzyme function to the query sequence.

7.2 *Chapter 4* – Identification of Function Specificity Determining Residues

This chapter of the thesis looked at two methods (*func-MB* and *profile-HMM*) for the automatic identification and scoring of function specificity determining residues (fSDRs) in enzymes. These residues were then ranked and subsets of them were used to functionally re-score the aligned enzyme sequences. The functional re-scoring performance of these methods was also compared to selected alignment re-scoring results from the alternative amino acid substitution matrix studies, carried out in *chapter 3*, as well as to random models for sequence and aligned residue subset selection. Below are suggestions for further studies that could be used to enhance the identification of functionally determining residues and also improve the subsequent functional alignment re-scoring of the aligned sequences.

7.2.1 Analysis of Additional Methods for the Identification of fSDRs

There are a number of published methods for the identification of functional specificity determining residues in protein sequences, a number of which were discussed in detail in the literature review provided in *chapter 1*. Although it would be prohibitively time consuming to carry out a thorough comparison between all of these methods and of the resulting functional alignment re-scoring results obtained from the identified functional residues, it may be beneficial to expand the benchmark analysis carried out in this thesis. An additional method for this purpose is proposed below, for the identification of aligned columns containing specificity determining residues.

7.2.1.1 "Unsupervised" Method for fSDR Identification

To complement the "supervised" fSDR identification methods (i.e., those that require functional information as input) described in *chapter 4*, an "unsupervised" method could also be studied. One example of this is the *MB-method*, described by del sol Mesa *et al.* (2003), which only uses overall sequence, rather than functional, similarity measures to calculate correlation scores. It is not expected that this type of

method will perform as well as the supervised ones, but it may provide useful additional information. For example, Pazos *et al.* (2006) suggest using the "supervised" methods when the sequence phylogeny does not agree with the functional classification. Therefore, it may be that for examples where the columns identified by the supervised and unsupervised methods are the same, the sequence phylogeny will closely follow the functional grouping, and vice versa for examples where the identified columns do not agree.

Additionally, the *MB-method* for identifying functional residues would provide an additional comparison to the later results obtained in this thesis from using SVM classification. This is because neither of these methods requires any functional information associated with the aligned sequences, prior to the identification of the specificity determining residues. The appropriate benchmark for comparing the performance of these two methods would be through an analysis of their functional re-scoring performances. This could be assessed with both the "top-hit" and functional enrichment score based assessment methods.

7.2.2 Re-alignment of the Protein Sequences

In the analysis carried out in this thesis all of the multiple sequence alignments are obtained directly from a PSI-BLAST sequence database search, which are generated from a concatenation of pair-wise sequence alignments to the query sequence. Because of this, they are not "true" multiple sequence alignments and therefore may not be the optimal alignment when considering all of the sequences. Taking this into consideration, it may be beneficial to use an additional multiple sequence alignment application (such as: CLUSTAL-W (Larkin *et al.*, 2007); MAFFT (Katoh *et al.*, 2002); or MUSCLE (Edgar, 2004)) to optimally re-align the BLAST identified sequence homologs. A disadvantage to this process would be the additional computation time required for the multiple alignments.

A major aim of this thesis was the investigation of the performance of computationally efficient methods for improving the functionally specific classification of enzyme sequences. Therefore, it is suggested that PSI-BLAST generated sequence alignments should remain as an integral part of the system, due to their computationally efficient generation. Rather, the sequence re-alignment procedure would be used for producing more refined input to the methods for

scoring and identifying the benchmark set of fSDRs. It would then be possible to identify the residues in the query sequence that were deemed to be functionally determining, which could be used to extract the corresponding aligned columns of residues from the PSI-BLAST generated alignments. Finally, these aligned columns of functionally specific residues would then be used to train the SVM classifier. This is important, because it ensures that the classifier is trained and validated using the same information (from PSI-BLAST multiple sequence alignments) that would also be used as the input to the SVMs in the final functional classification system.

7.2.3 Further Analysis of Gap Scoring Parameters

7.2.3.1 Gap Scoring in the Sequence Alignments

The experimental analysis provided in *chapter 4* only investigates the effects of using a single value, of 0, when scoring the alignments of amino acids to gaps. Additional analysis, similar to that suggested above for the alignment re-scoring experiments that use alternative amino acid substitution matrices, in *chapter 3* of the thesis, could also be carried out for the gap scoring parameters used for the sub-alignment re-scoring. However, due to the general nature of these sub-alignments, it would not be possible to extend this analysis to one which investigates the change in the gap opening (g_{open}) and gap extension (g_{extend}) parameters. This is because the aligned residues in the sub-alignments are not necessarily from consecutive residues in the aligned sequences; meaning that the use of a gap scoring method, such as *equation 3.1*, which incorporates the length of consecutive gap residues in a sequence alignment, would not be valid for non-consecutive sub-alignments of residues. Therefore, a range of single values could be used to score the pair-wise alignment of gap residues to amino acids, in the extracted sequence sub-alignments.

7.2.3.2 Gap Scoring in the Amino Acid Similarity Matrices of the *func-MB* Method

For the *func-MB* method of fSDR identification, it is possible to carry out an additional set of analyses into the use of different scores to assess the effects of gap residues in the sequence alignments. The description of the *func-MB* method stated that the calculation of the Spearman rank-order correlation coefficient is based upon the comparison between two matrices of similarity: the residue similarity and

functional similarity matrices. It is the amino acid residue similarity matrix that is of interest for this proposed additional gap scoring analysis. In this thesis, when scoring the similarity between amino acids and gaps in the alignments a single score of 0 has been used in the amino acid similarity matrices. It may, however, be of interest to systematically assess the effects of changing this "gap-residue" similarity measure on the subsequent alignment re-scoring performance.

7.2.4 The Use of Alternative Amino Acid Substitution Matrices

Following on from the above section, it is also possible to systematically vary the amino acid substitution matrices used in both the sub-alignment re-scoring and the calculation of the amino acid similarity matrices (in the *func-MB* method). Careful consideration would need to be given to the optimisation of the alternative amino acid substitution matrices and the different gap scoring parameters; because it is expected that optimal parameter of one would be closely dependent on the other.

7.2.5 Pre-filtering of the Multiple Sequence Alignments

Prior to applying the fSDR scoring and identification methods, it may be beneficial to apply a number of filtering steps to improve the quality of the aligned sequence data. These could take on either a "sequence based" or a "functional class based" form, each of which would lead to differences in their application.

An example of a "sequence based" form of alignment pre-filtering is the removal of sequences that have a pair-wise alignment overlap, with the query sequence, that is less than a defined percentage threshold. This would be considered "sequence based" because its application would not require any additional information other than that present in the aligned sequences. The application of this particular sequence filter would be the removal of sequences with potentially poor alignments and also any potential false positive homologous sequences. This would be applied to the MSAs used for both the identification of the fSDRs extracted in the definition of the benchmark SVM training datasets, as well as to the input MSAs to the SVM-based functional classification system. A potential problem with this "percentage overlap" method of alignment filtering is that it may discard important alignments to a functionally relevant domain, which is part of a longer aligned homologous protein.

Another method of sequence alignment filtering, which is based upon the specific functional classes of the aligned protein sequences, could be used to improve the quality of the input alignment data for the fSDR identification methods used to define the benchmark datasets for SVM training. The proposed method for this "functional class based" sequence filtering would look at the effects of removing sequences that do not share particular levels of the EC classification hierarchy. For example, it may be possible to improve the identification of specific functional residues through the use of sequence alignments that only contain sequences that share the same first three levels of the EC functional classification hierarchy, with the query sequence. The reasoning behind this is that the resulting sequence alignments will only contain information that is associated with the most specific functional differences in the sequences and may therefore improve the scoring of the fSDR identification methods.

A disadvantage of this method may be the reduction in both functional and sequence diversity in the resulting alignments used as input for the subsequent fSDR scoring and identification methods. Also, it is not intended that this method of sequence alignment filtering would be used to pre-filter the alignments of the final SVM-based classification of functionally unknown query sequences. This is because the unknown function of the query sequence makes the functional class based filtering step impossible.

7.3 *Chapter 5* – Towards the Identification of Functional Specificity Determining Residues Using Support Vector Machines

This section discusses suggested improvements and alternative methods of investigation for the SVM studies presented in *chapter 5*. The key areas identified for potential improvement and additional study, are: the definition of the benchmark dataset of fSDRs; the parameters used in the machine learning classification method; and a benchmark assessment of the use of the SVM predicted residues to functionally re-score the aligned sequence.

7.3.1 Optimisation of the Functional Enrichment Score and the Definition of the Benchmark Dataset of fSDRs

7.3.1.1 Investigation of Additional *colgap_percent* Thresholds

There are a number of alternative ways in which the benchmark dataset of fSDRs, used to train and validate the SVM classifiers, could be defined. One modification to the process of dataset definition has already been discussed in detail at the end of *chapter 5*, in *section 5.3.4*. Briefly, the method of dataset definition currently used in *chapter 5* uses only a single *colgap_percent* threshold, to pre-filter aligned columns of amino acids with particular levels of gaps in the alignments, prior to applying the *func-MB* method of fSDR identification. As discussed, in *section 5.3.4*, it would be of interest to extend this analysis to include a more thorough analysis of the functional enrichment scores obtained from using the *func-MB* method with a series of different *colgap_percent* thresholds. This in turn would lead to an alternative benchmark dataset of fSDRs for training and validating the SVM classifiers.

7.3.1.2 Incorporation of the "top-hit" Assessment Method

The benchmark dataset of fSDRs that are used for training and validating the SVM classifiers may be improved by the additional use of a "top-hit" based assessment method, in conjunction with the functional enrichment score based method, for their generation. At present, the benchmark dataset of fSDRs, which was defined in *chapter 5*, uses an approach that selects the subset of fSDRs to include in the benchmark fSDR dataset via the optimisation of the functional enrichment scores obtained after re-scoring each of the MSAs. This method currently uses a number of criteria to ensure that the there is a high level of "functionally correct" enzymes in the optimally re-scored alignments (such as: demanding that at least 9 out of the top 10 ranking sequences after functional alignment re-scoring are the same "correct" specific function as the query sequence). However, this procedure of optimising the functional enrichment score, to determine selection of the benchmark set of fSDRs, may be improved further by incorporating an additional selection criterion, which also requires that the "top-hit", after sequence sub-alignment re-scoring, is functionally correct (i.e., the same as the query sequence).

The main benefit to this improved fSDR selection procedure would be that the SVM would be trained and validated on fSDR data that guarantees both a particular level of functional enrichment and a functionally correct "top-hit" sequence, after alignment re-scoring. Therefore, it would add an additional level of confidence, to any specific functional classifications that are based upon the use of the SVM-based system of automatic function classification.

7.3.2 SVM Analysis

The final part of this thesis concentrated on the initial investigations that were made into the use of an SVM for the automatic identification of fSDRs. There are a number of ways in which this analysis could be extended, such as: through the repeated analysis and comparison using the alternative methods for fSDR dataset definition, described above; and through the investigation of additional SVM learning parameters and features for data encoding.

7.3.2.1 The Investigation of Additional Input Feature Vectors

The results presented in *chapter 5* are based upon the use of a limited number of input feature vectors to the SVMs. To improve the way in which the functional specificity determining information is encoded from the aligned sequences it is expected that the use of additional input features would improve the classification performance of the SVM classifiers. Without carrying out the SVM training and validation experiments it is not possible to determine which particular features will provide an improved predictive performance. However, some of the features, mentioned below, have been used previously in machine learning approaches, involving data in the form of multiple sequence alignments, and may be of interest. Also discussed are suggested extensions to the input features already presented in the SVM classification studies of *chapter 5*.

Additional Percentage Thresholds for the *NumberOfAATypes* Feature

One extension to the features already used in the SVM analyses, of *chapter 5*, could be the use of alternative percentage thresholds of amino acid occurrence with the *NumberOfAATypes* feature. A single threshold of 12% was used for this feature, however, it may be advantageous to analyse the effects of alternative thresholds on the performance of the SVM classifiers. Also, the use of multiple percentage

thresholds, and therefore multiple additional SVM input features, may improve the SVM classification performance. It is expected that this approach would provide additional information as to how the number of distinct amino acid types, in each aligned column of residues, varies as the threshold is altered and therefore provide a measure of the level of "signal noise" in the *NumberOfAATypes* features.

Alternative Methods for Calculating the Amino Acid Composition

The *AA_composition* feature, described in *chapter 5*, used a relatively simple method for calculating the fractional occurrence of each amino acid type contained within the aligned columns of the two classes of fSDR and non-fSDR residues. A number of alternative methods could be investigated for encoding the amino acid composition information for input into the SVMs. One such method is that of the PSIPRED application (Jones, 1999), which uses the position specific scoring matrices (PSSMs) generated from a PSI-BLAST database search. PSSMs are generated as part of the iterative sequence search procedure of PSI-BLAST and they incorporate weighted calculations of the frequency of occurrence of certain amino acid types within the multiple alignments. This weighting uses pseudo-counts and prior knowledge of common amino acid occurrences and mutations, which in turn incorporates evolutionary information into the sequence profiles of the aligned families of similar protein sequences. Therefore, a PSSM based method may improve the encoding of the amino acid composition, in each of the aligned columns, through the incorporation of weighted, evolutionary based calculations into the SVM input features.

Residue Conservation Score

Previous studies that have used machine learning methods for the automated identification of CSA residues, from MSAs, have found the use of a sequence conservation score to be beneficial. This aims to calculate the level of residue conservation within each aligned column and encode it within a score for use as an input feature into the SVMs. There are a number of well-studied methods for calculating the positional conservation in protein sequence alignments, such as: the *scorecons* method (Valdar, 2002), used by Petrova and Wu (2006), Gutteridge et al. (2003) and Tang et al. (2008) for the classification of CSA residues; and also the *AL2CO* method, described by Pei and Grishin (2001). The CSA residues are

generally found to be well conserved; therefore the level of residue conservation is expected to be a more informative feature of these residues that describe general catalytic properties, than those of the specificity determining residues of interest in this thesis. However, the use of a conservation score may be found to be advantageous to the SVM classification of fSDRs and should be investigated.

Residue Window Based Encoding of the SVM Input Features

A further way in which the SVM classification could be improved is through the use of a "residue window" of input features. This approach is proposed to take into account the properties of all the aligned residues within a window, *w*, of the particular aligned residue of interest. This then leads to a total residue window size of *2*w+1*, formed from the residues that are w positions "up" and "down" the query sequence from the "central" aligned residue of interest. Two possible alternative approaches that could be used to encode this residue window based information, for SVM input, are: (i) the "feature averaging" method, used by Youn *et al.* (2008); and (ii) the "multiple feature vectors" method, used by the PSIPRED method (Jones, 1999). The "feature averaging" method uses the same number of input feature vectors, as when not using a residue window, but the information for each feature vector is calculated using an average of the properties for all of the aligned columns of amino acids contained within the sequence window. In contrast, the "multiple feature vectors" method uses the same method for calculating the feature vectors as when not using a residue window, however, the total number of input feature vectors would be a multiple of the window size (i.e., if the number of input features for each column is f, then the number of SVM input feature vectors would be *f*(2*w+1)*).

Both of these residue window methods may improve the ability of the SVM classifiers to differentiate between fSDR and non-fSDR columns of residues, by incorporating additional information from neighbouring sequence residues into the feature vectors.

7.3.2.2 Additional SVM Optimisation and Assessment Strategies

In conjunction with the alternative methods of SVM feature vector encoding, described above, additional methods for optimising the SVM learning parameters and assessing the performance of the classifiers could be used.

An improved method for optimising the parameters of the SVM learning kernels (i.e., the *C* parameter and the *C* and *gamma* parameters for the linear and RBF kernels, respectively) could use an additional, more fine-grained, approach to varying these parameters. That is, once the optimal set of parameters have been identified using the parameters and cross-validation procedure described in *chapter 5*; an additional grid-search of the kernel parameter space could then be undertaken to see if any smaller incremental variations, close to these parameters, could generate any further improvements in the classification.

It was shown in the SVM classification results, in *chapter 5*, that there were a large number of false positive fSDR predictions being made, resulting in high false positive rates. One possible cause of this is the inherent lack of a clear distinction between a significant proportion of the fSDR and non-fSDR classes. This may manifest in the generation of a number of SVM classifications that are clearly scoring as positive or negative classes (i.e., the TPs and the FPs), but also a large number of examples with very similar SVM classification scores that do not allow a clear method of class differentiation. For this type of outcome it may be beneficial to convert the scores from the SVM classification into a measure of the probability (i.e., a p-value) that the particular example is to be found in the positive or negative class. This can be thought of as a way of incorporating a measure of the distance of each example from the optimally separating hyper-plane of the SVM. A method for calculating probabilities of this form has been described by Platt (1999) and may, through the application of a probability threshold, provide a way to reduce the number of generated false positives and provide a confidence level for the SVM predictions.

7.3.2.3 Assessment of the Functional Re-scoring Performance when using the fSDRs Identified with the SVM Classifiers

One further stage of analysis that is important for assessing the performance of the SVM classifiers, when identifying fSDRs, is the subsequent ability to re-score the aligned sequences. That is, using the potential fSDRs that have been identified by the SVM classifiers a sub-alignment of residues should be extracted and used to functionally re-score and re-rank the aligned sequences. This was addressed to some extent, in *chapter 5*, when looking at the effect of functionally re-scoring the three

enzyme examples, using the SVM predicted residues. The aim of this further analysis would be to provide a larger performance benchmark of the functional annotation accuracy obtained when using the SVM predicted fSDRs to re-score and classify specific enzyme function. This method of assessing the performance of the SVM classifiers may result in the observation that a certain level of false positive classifications can be tolerated, without unduly affecting the functional re-scoring and subsequent functional assignment performance.

Appendix I – Dataset Statistics

This appendix presents a more detailed description and summary statistics related to the EC class and sequence composition of selected datasets that have been investigated in the thesis.

EC Class Distributions of the Dataset Query Sequences

An analysis of the EC classes represented by the query sequences used to generate the datasets of MSAs analysed in the thesis is presented below. An overview of the frequency (*figure A-1 (a)*) and percentage (*figure A-1 (b)*) of occurrence of the six top-level EC classes, for four of the datasets (see *section 2.4.2.7*, *section 3.2.1*, *table 3.1* and *section 5.2.6*), is shown. The four datasets are: (i) the *All1stINCORRECT.tF.BLOSUM62.E0.001* dataset (denoted as "BLOSUM62"); (ii) the *All1stINCORRECT.tF.PAM160.E0.001* dataset (denoted as "PAM160"); (iii) the *All1stINCORRECT.tF.PAM30.E0.001* dataset (denoted as "PAM30"); and (iv) the cross-validation dataset of MSAs used in the SVM analysis of *chapter 5* (denoted as "SVM"). *Table A-1* lists the associated frequency and percentage of occurrence of the 6 general EC classes associated with the query sequences from these four datasets. This analysis shows that the oxidoreductases (EC 1.-.-.-) and the transferases (EC 2.-.-.-) are the dominant top-level EC classes in all 4 of the datasets; associated with greater than 75% of query sequence representatives in the BLOSUM62, PAM160, and PAM30 BLAST generated MSAs; and 61.4% of those in the SVM dataset. Except for the isomerases (EC 5.-.-.-), which only occurs once in the SVM and twice in the BLOSUM62 datasets, respectively. Each of the other general EC classes are well represented in the datasets, therefore, showing a good overall coverage of general EC classes in these four datasets.

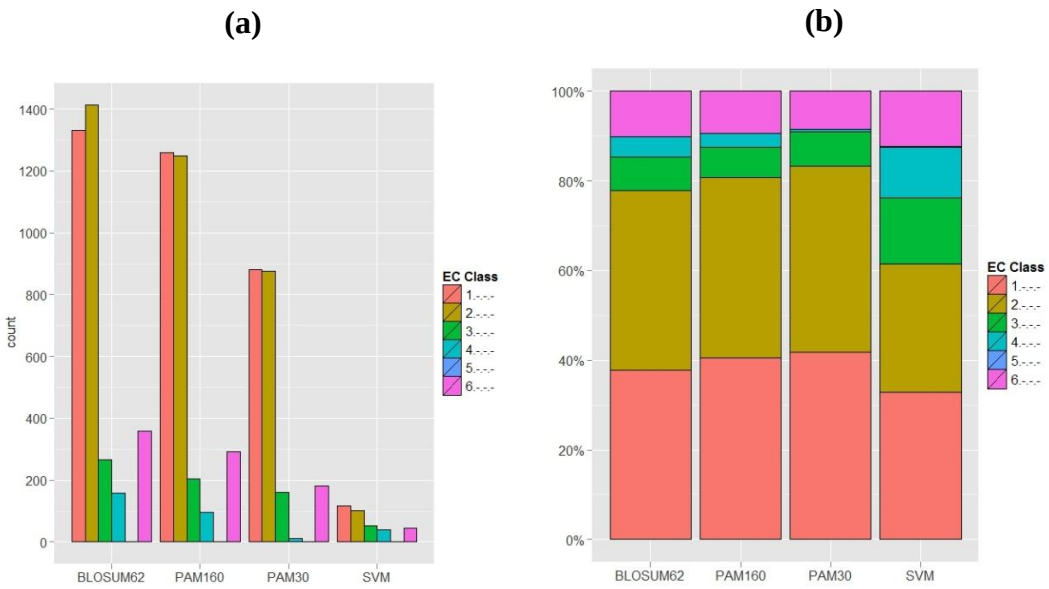

Figure A-1. Overview of (a) the frequency and (b) the percentage of occurrence of top-level EC classes associated with the query sequences of the following four MSA datasets: (i) "BLOSUM62" - the All1stINCORRECT.tF.BLOSUM62.E0.001 dataset; (ii) "PAM160" - the All1stINCORRECT.tF.PAM160.E0.001 dataset; (iii) "PAM30" - the All1stINCORRECT.tF.PAM30.E0.001 dataset; and (iv) "SVM" - the cross-validation dataset of MSAs used in the SVM analysis of chapter 5.

	BLOSUM62		PAM160		PAM30		SVM	
EC class	**count**	**%**	**count**	**%**	**count**	**%**	**count**	**%**
EC 1.-.-.-	1330	37.7	1258	40.6	881	41.8	117	32.8
EC 2.-.-.-	1413	40.1	1248	40.3	876	41.5	102	28.6
EC 3.-.-.-	265	7.5	205	6.6	161	7.6	53	14.8
EC 4.-.-.-	159	4.5	97	3.1	12	0.5	40	11.2
EC 5.-.-.-	2	0.06	0	0	0	0	1	0.28
EC 6.-.-.-	358	10.2	292	9.4	180	8.5	44	12.3
Total	**3527**		**3100**		**2110**		**357**	

Table A-1. The frequency of occurrence ("count") and the percentage (%) of top-level EC classes associated with the query sequences of the four MSA datasets shown in figure A-1.

A similar analysis was also carried out on the top-level EC classes of the query sequences representatives used to generate the 80%, 60% and 40% CD-HIT clustered subsets of MSAs, defined in *section 3.3.5*. The frequency and percentage of EC class occurrence of these three "seqID" subsets is compared (in *figure A-2* and *table A-2*) to the un-clustered (*seqID=100%*) *All1stINCORRECT.tF.BLOSUM62.E0.001* dataset.

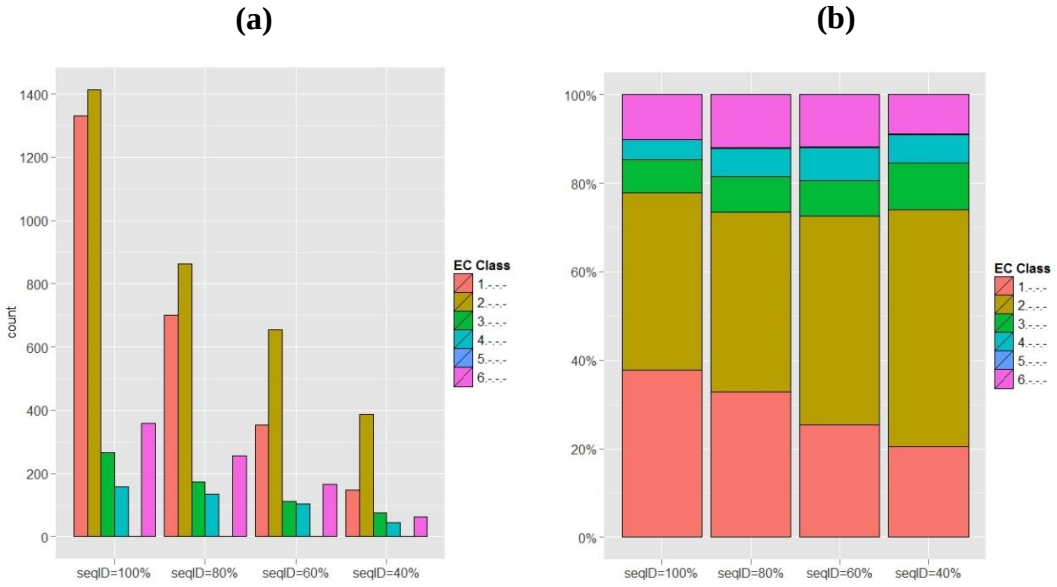

Figure A-2. Overview of (a) the frequency and (b) the percentage of occurrence of top-level EC classes associated with the CD-HIT query sequence representatives of four MSA subsets from the All1stINCORRECT.tF.BLOSUM62.E0.001 dataset: (i) "seqID=100%" – no query sequence clustering; (ii) "seqID=80%" – 80% sequence clustering; (iii) "seqID=60%" – 60% sequence clustering; and (iv) "seqID=40%" – 40% sequence clustering.

	seqID=100%		seqID=80%		seqID=60%		seqID=40%	
EC class	**count**	**%**	**count**	**%**	**count**	**%**	**count**	**%**
EC 1.-.-.-	1330	37.7	700	32.8	354	25.4	147	20.4
EC 2.-.-.-	1413	40.1	864	40.5	655	47.1	387	53.7
EC 3.-.-.-	265	7.5	173	8.1	112	8.0	75	10.4
EC 4.-.-.-	159	4.5	136	6.4	104	7.5	46	6.4
EC 5.-.-.-	2	0.06	2	0.09	2	0.14	2	0.28
EC 6.-.-.-	358	10.2	256	12.0	165	11.9	64	8.9
Total	**3527**		**2131**		**1392**		**721**	

Table A-2. The frequency of occurrence ("count") and the percentage (%) of top-level EC classes associated with the query sequences of the MSA datasets shown in figure A-2.

Again, this analysis shows that the oxidoreductases and transferases are the dominant top-level EC classes in each of the datasets. As the clustering becomes more stringent the percentage of EC 1.-.-.- examples decreases, from 37.7% in the *seqID=100%* subset, to 20.4% in the *seqID=40%* subset. Whereas, the number of EC 2.-.-.- examples increases from 40.1% to 53.7% in the same subsets. The other 4 classes have a similar proportion of representation in each of the clustered subsets.

A more detailed breakdown of the specific EC class distributions, at all four levels of EC classifications, is shown in *figure A-3* and *table A-3*. The concentric pie-chart, in *figure A-3,* shows the proportion of each of the four EC levels of classification represented by the query sequences of the *All1stINCORRECT.tF.BLOSUM62.E0.001* dataset (*seqID=100%*). The six classes of the first EC level are in the centre, followed by the second and third levels, represented by the second and third outer rings, respectively. The 107 fourth level EC class annotations associated with the query sequences in this dataset are represented and labelled on the outer ring of *figure A-3*. To supplement the pie-chart, *table A-3* lists the associated frequency and percentage of occurrence of these 107 EC classes. Also listed in the table are the frequency and percentage of occurrences of these EC classes in the clustered datasets (*seqID=80%*, *seqID=60%*, and *seqID=40%*) and the 84 EC classes in the SVM dataset. This analysis shows that, except for a small number of specific classes (i.e., EC 2.7.1.37 (non-specific serine/threonine protein kinase), EC 1.6.5.3 (NADH dehydrogenase), EC1.14.14.1 (unspecific monooxygenase), EC 2.7.1.112 (protein-tyrosine kinase), EC 4.1.3.27 (anthranilate synthase)), the majority of the EC classes represent much less than 5% of the total and therefore result in a fairly even distribution of specific EC classes within the datasets.

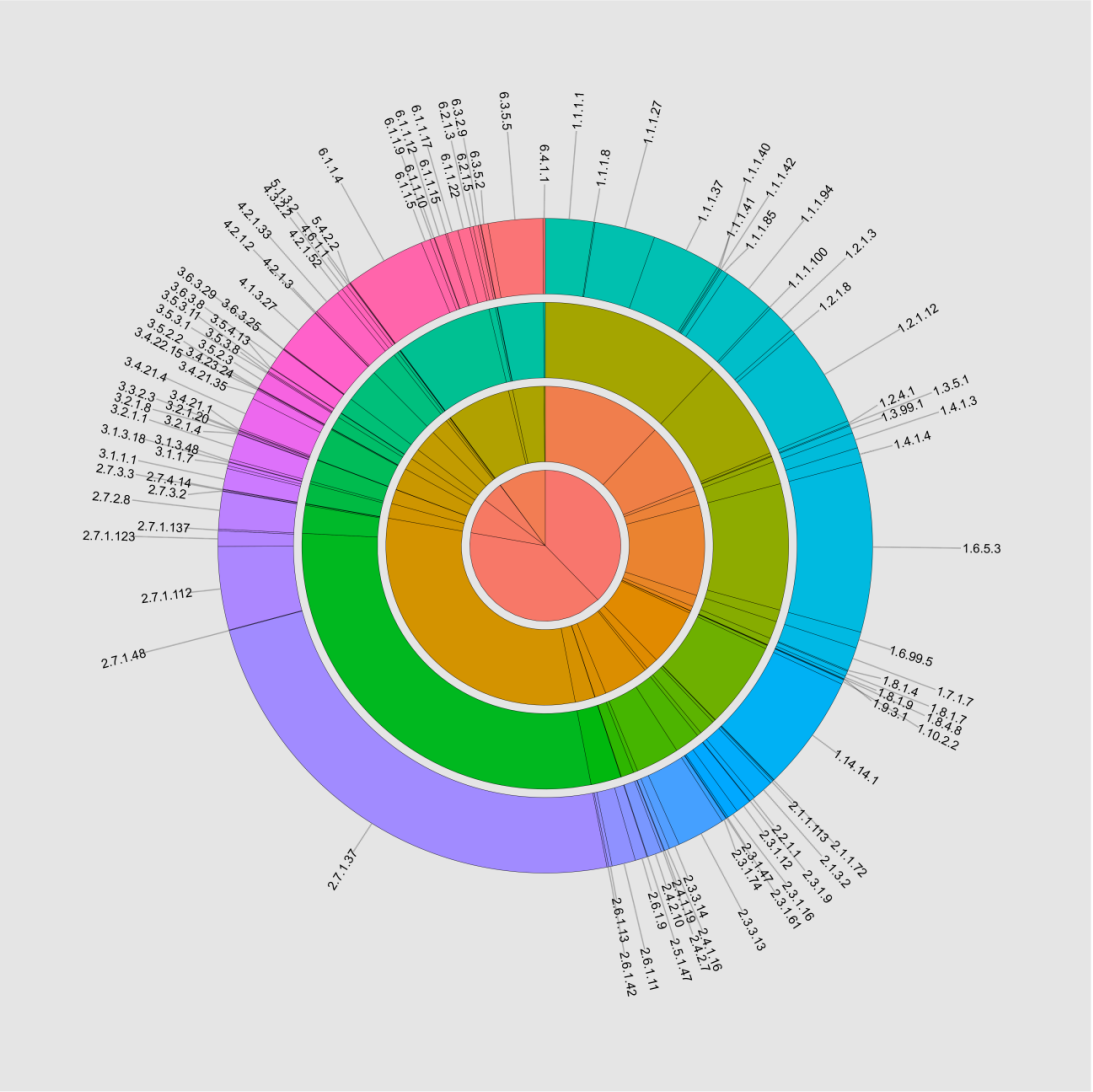

Figure A-3. Concentric pie-chart showing the proportion of each EC level of classification represented by the query sequences of the All1stINCORRECT.tF.BLOSUM62.E0.001 dataset (seqID=100%). The first EC level is in the centre. The second and third levels are represented by the second and third outer rings, respectively. The 107 fourth level EC class annotations associated with the query sequences in this dataset are represented and labelled on the outer ring.

EC	seqID=100		seqID=80		seqID=60		seqID=40		SVM	
	count	%	count	%	count	%	count	%	count	%
1.1.1.1	84	2.38	37	1.74	16	1.15	7	0.97	6	1.68
1.1.1.8	2	0.06	1	0.05	1	0.07	1	0.14	0	0.00
1.1.1.27	105	2.98	39	1.83	25	1.80	8	1.11	7	1.96
1.1.1.37	120	3.40	65	3.05	41	2.95	10	1.39	11	3.08
1.1.1.40	4	0.11	3	0.14	2	0.14	1	0.14	0	0.00
1.1.1.41	2	0.06	1	0.05	1	0.07	0	0.00	0	0.00
1.1.1.42	3	0.09	2	0.09	2	0.14	2	0.28	1	0.28
1.1.1.85	12	0.34	9	0.42	5	0.36	3	0.42	2	0.56
1.1.1.94	90	2.55	62	2.91	43	3.09	16	2.22	16	4.48
1.1.1.100	4	0.11	2	0.09	2	0.14	2	0.28	1	0.28
1.2.1.3	52	1.47	29	1.36	15	1.08	8	1.11	4	1.12
1.2.1.8	8	0.23	4	0.19	2	0.14	1	0.14	1	0.28
1.2.1.12	176	4.99	79	3.71	20	1.44	4	0.55	4	1.12
1.2.4.1	7	0.20	5	0.23	3	0.22	2	0.28	2	0.56
1.3.5.1	2	0.06	2	0.09	1	0.07	0	0.00	0	0.00
1.3.99.1	13	0.37	10	0.47	6	0.43	4	0.55	1	0.28
1.4.1.3	27	0.77	9	0.42	7	0.50	2	0.28	2	0.56
1.4.1.4	26	0.74	16	0.75	8	0.57	2	0.28	1	0.28
1.6.5.3	295	8.36	168	7.88	85	6.11	33	4.58	30	8.40
1.6.99.5	28	0.79	24	1.13	19	1.36	14	1.94	4	1.12
1.7.1.7	41	1.16	12	0.56	3	0.22	2	0.28	2	0.56
1.8.1.4	2	0.06	2	0.09	1	0.07	1	0.14	1	0.28
1.8.1.7	9	0.26	6	0.28	4	0.29	1	0.14	1	0.28
1.8.1.9	5	0.14	3	0.14	3	0.22	3	0.42	2	0.56
1.8.4.8	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
1.9.3.1	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
1.10.2.2	8	0.23	3	0.14	3	0.22	3	0.42	1	0.28
1.14.14.1	202	5.73	104	4.88	33	2.37	14	1.94	14	3.92
1.14.18.1	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
2.1.1.72	2	0.06	2	0.09	2	0.14	2	0.28	2	0.56
2.1.1.113	5	0.14	5	0.23	5	0.36	5	0.69	4	1.12
2.1.3.2	37	1.05	26	1.22	16	1.15	6	0.83	6	1.68
2.2.1.1	10	0.28	7	0.33	3	0.22	2	0.28	2	0.56
2.3.1.9	1	0.03	1	0.05	1	0.07	0	0.00	0	0.00
2.3.1.12	27	0.77	23	1.08	18	1.29	10	1.39	8	2.24
2.3.1.16	19	0.54	8	0.38	7	0.50	3	0.42	3	0.84
2.3.1.47	3	0.09	3	0.14	3	0.22	2	0.28	2	0.56
2.3.1.61	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
2.3.1.74	8	0.23	1	0.05	1	0.07	1	0.14	0	0.00
2.3.3.13	84	2.38	57	2.67	35	2.51	6	0.83	5	1.40
2.3.3.14	18	0.51	15	0.70	12	0.86	6	0.83	3	0.84
2.4.1.16	9	0.26	9	0.42	7	0.50	5	0.69	5	1.40
2.4.1.19	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
2.4.2.7	5	0.14	4	0.19	4	0.29	1	0.14	1	0.28
2.4.2.10	25	0.71	20	0.94	16	1.15	5	0.69	5	1.40
2.5.1.47	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
2.6.1.9	21	0.60	18	0.84	17	1.22	8	1.11	8	2.24
2.6.1.11	41	1.16	31	1.45	28	2.01	8	1.11	7	1.96
2.6.1.13	6	0.17	5	0.23	2	0.14	1	0.14	1	0.28
2.6.1.42	3	0.09	2	0.09	2	0.14	1	0.14	1	0.28
2.7.1.37	841	23.84	484	22.71	371	26.65	263	36.48	0	0.00
2.7.1.48	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
2.7.1.112	145	4.11	70	3.28	52	3.74	30	4.16	26	7.28
2.7.1.123	27	0.77	15	0.70	11	0.79	6	0.83	4	1.12
2.7.1.137	3	0.09	3	0.14	3	0.22	3	0.42	0	0.00
2.7.2.8	65	1.84	47	2.21	31	2.23	6	0.83	5	1.40
2.7.3.2	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
2.7.3.3	2	0.06	2	0.09	2	0.14	1	0.14	1	0.28
2.7.4.14	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
3.1.1.1	36	1.02	20	0.94	9	0.65	7	0.97	6	1.68
3.1.1.3	5	0.14	3	0.14	2	0.14	2	0.28	2	0.56
3.1.1.7	6	0.17	5	0.23	4	0.29	3	0.42	3	0.84
3.1.3.18	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
3.1.3.48	5	0.14	3	0.14	3	0.22	2	0.28	0	0.00
3.2.1.1	45	1.28	32	1.50	20	1.44	14	1.94	9	2.52
3.2.1.4	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
3.2.1.8	3	0.09	3	0.14	3	0.22	3	0.42	0	0.00
3.2.1.20	2	0.06	2	0.09	2	0.14	2	0.28	0	0.00

EC	seqID=100		seqID=80		seqID=60		seqID=40		SVM	
	count	%	count	%	count	%	count	%	count	%
3.3.2.3	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
3.4.21.1	3	0.09	2	0.09	2	0.14	2	0.28	0	0.00
3.4.21.4	56	1.59	34	1.60	17	1.22	7	0.97	6	1.68
3.4.21.35	17	0.48	9	0.42	4	0.29	1	0.14	1	0.28
3.4.22.15	2	0.06	2	0.09	2	0.14	2	0.28	0	0.00
3.4.23.24	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
3.5.2.2	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
3.5.2.3	29	0.82	24	1.13	18	1.29	9	1.25	9	2.52
3.5.3.1	2	0.06	2	0.09	1	0.07	1	0.14	1	0.28
3.5.3.8	2	0.06	2	0.09	2	0.14	1	0.14	1	0.28
3.5.3.11	7	0.20	5	0.23	3	0.22	3	0.42	3	0.84
3.5.4.13	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
3.6.3.8	37	1.05	17	0.80	12	0.86	8	1.11	8	2.24
3.6.3.25	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
3.6.3.29	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
4.1.3.27	81	2.30	64	3.00	55	3.95	23	3.19	21	5.88
4.2.1.2	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
4.2.1.3	3	0.09	2	0.09	1	0.07	1	0.14	1	0.28
4.2.1.33	48	1.36	44	2.06	23	1.65	7	0.97	5	1.40
4.2.1.52	10	0.28	10	0.47	10	0.72	6	0.83	6	1.68
4.3.2.2	10	0.28	9	0.42	8	0.57	2	0.28	1	0.28
4.6.1.1	6	0.17	6	0.28	6	0.43	6	0.83	5	1.40
5.1.3.2	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
5.4.2.2	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
6.1.1.4	140	3.97	95	4.46	60	4.31	22	3.05	13	3.64
6.1.1.5	16	0.45	14	0.66	11	0.79	6	0.83	2	0.56
6.1.1.9	7	0.20	6	0.28	6	0.43	4	0.55	0	0.00
6.1.1.10	1	0.03	1	0.05	1	0.07	1	0.14	0	0.00
6.1.1.12	20	0.57	14	0.66	10	0.72	4	0.55	3	0.84
6.1.1.15	3	0.09	3	0.14	3	0.22	2	0.28	2	0.56
6.1.1.17	19	0.54	15	0.70	13	0.93	5	0.69	5	1.40
6.1.1.22	21	0.60	13	0.61	7	0.50	2	0.28	2	0.56
6.2.1.3	6	0.17	4	0.19	2	0.14	2	0.28	1	0.28
6.2.1.5	9	0.26	5	0.23	1	0.07	1	0.14	1	0.28
6.3.1.5	4	0.11	4	0.19	3	0.22	2	0.28	2	0.56
6.3.2.9	1	0.03	1	0.05	1	0.07	1	0.14	1	0.28
6.3.5.2	12	0.34	9	0.42	7	0.50	2	0.28	2	0.56
6.3.5.5	95	2.69	68	3.19	38	2.73	8	1.11	9	2.52
6.4.1.1	4	0.11	4	0.19	2	0.14	2	0.28	1	0.28

Table A-3. The frequency of occurrence (“count”) and the percentage (%) of specific EC classes associated with the query sequences of the seqID=100%, seqID=80%, seqID=60% and seqID=40% MSA subsets of the All1stINCORRECT.tF.BLOSUM62.E0.001 dataset and the “SVM” cross-validation dataset of MSAs.

Analysis of the Sequence and Functional Class Composition of the MSA Datasets

In the above section the analysis was based upon the EC class distributions of the enzyme query sequences that were used to generate the MSA datasets. Here, an analysis is presented of some properties of the sequences that are contained within these MSAs. The 3527 MSAs of the *All1stINCORRECT.tF.BLOSUM62.E0.001* dataset were used for this analysis. This dataset was chosen because it is investigated in most detail throughout the thesis.

To obtain an overview of the distribution of the numbers of sequences present in each of the constituent MSAs the histogram, shown in *figure A-4*, was constructed. The bin size used in this histogram was 5 (starting at 1 to 5, inclusive) and the minor grid-lines on the horizontal and vertical axes are 20 and 10, respectively. This distribution shows a wide and reasonably well distributed number of sequences in each MSA (i.e., between 20 and 760), with the larger number of examples containing between 25 and 70 sequences and a maximum of 61-65 sequences.

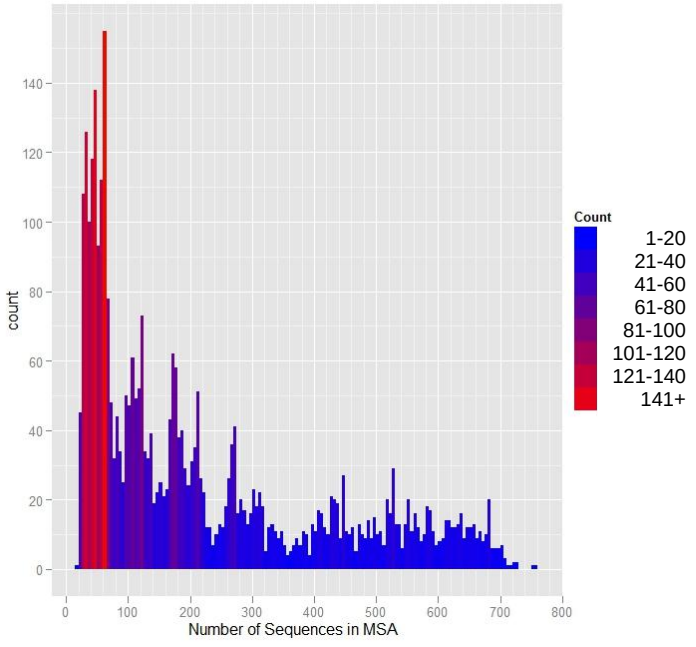

Figure A-4. Histogram of the number of sequences in MSAs of the All1stINCORRECT.tF.BLOSUM62.E0.001 dataset. The bin size is 5.

An analysis of the EC class distribution of the sequences in these 3527 MSAs is shown in *figure A-5*. In particular, *figure A-5 (a)* shows a scatter plot of the number

of “correct” (i.e., the number of sequences matching the query sequence to all 4 levels of EC specificity – “EC4”) and “incorrect” functional assignments in each MSA. This data is also shown, in *figure A-5 (b)*, as a histogram of the percentage of “correct” sequences in each of the MSAs. These plots show the correlation and wide spread in the distribution of “correct” to “incorrect” functional sequence assignments within the MSAs of the dataset. A comparison to this is provided in *figure A-5 (c)*, which shows a comparable histogram when assessing the percentage of functionally “correct” sequences in each MSA using the less specific measure of the first 3 levels of EC classification (i.e., “EC3”). This comparison clearly shows that when the level of specific functional measure is reduced from “EC4” to “EC3” the number of MSAs containing 100% “correct” sequences increases dramatically. This trend is continued when decreasing the level functional specificity further to the first two and just the first level of EC classification (results not shown).

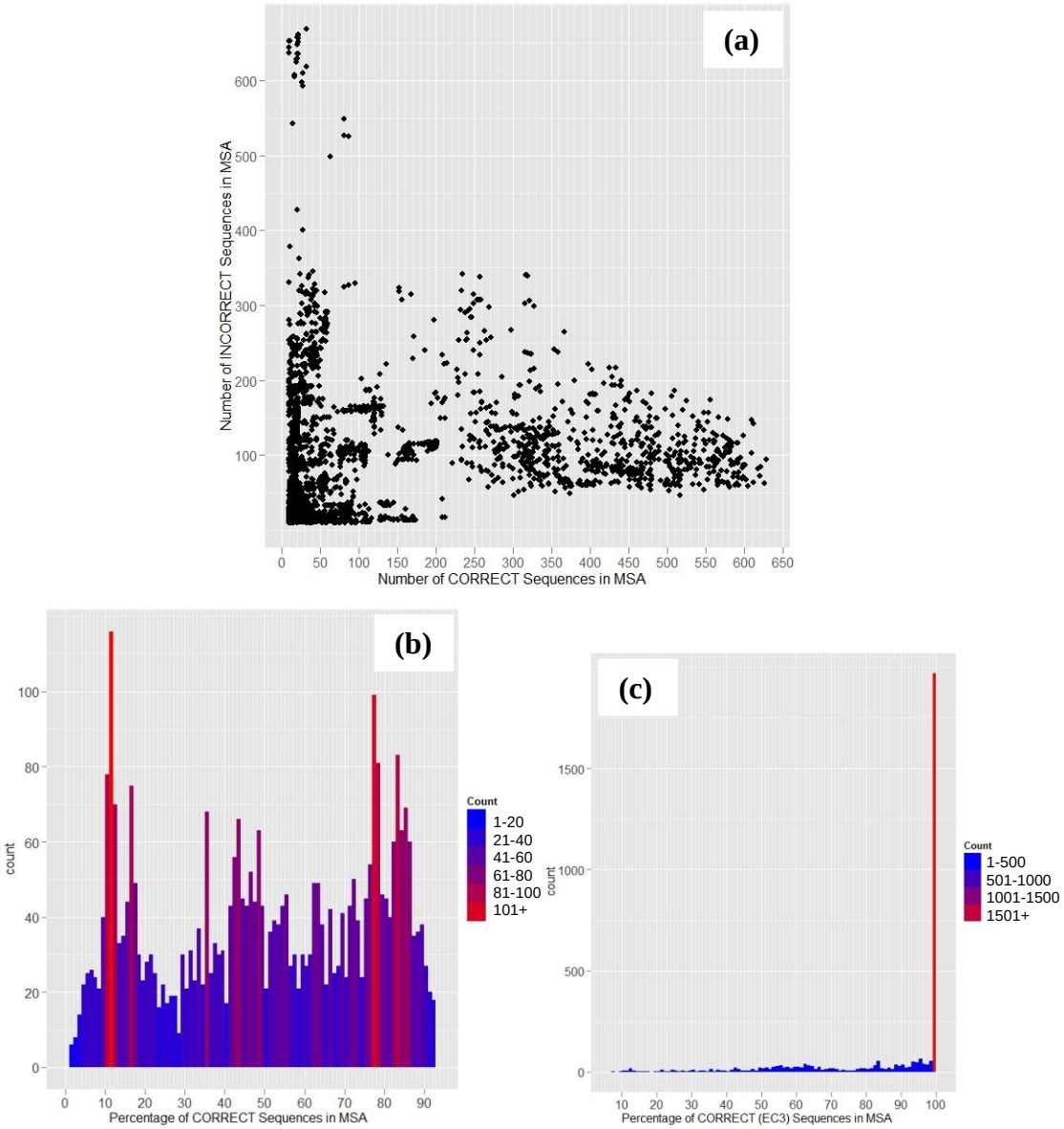

Figure A-5. EC class distributions of the sequences in the MSAs of the All1stINCORRECT.tF.BLOSUM62.E0.001 dataset. (a) Scatter plot of the number of "correct" (i.e., number of sequences matching the query sequence at all 4 EC levels) and "incorrect" sequences in each MSA. (b) and (c) are histograms of the percentage of "correct" sequences in each of the MSAs at "EC4" and "EC3" levels of functional specificity, respectively. The bin size is 1% in both.

Bibliography

Aloy P, Querol E, Aviles FX, and Sternberg MJE. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of Molecular Biology* **311**, 395-408 (2001).

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410 (1990).

Altschul SF, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).

Altschul SF, Bundschuh R, Olsen R, and Hwa T. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research* **29**, 2, 351-361 (2001).

Andrade MA, Casari G, Sander C, and Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biological Cybernetics* **76**, 441-450 (1997).

Andrade MA, *et al.* Automated genome sequence analysis and annotation. *Bioinformatics* **15**, 391-412 (1999).

Apweiler R, *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **32**, D115-D119 (2004).

Armon A, Graur D, and Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology* **307**, 447-463 (2001).

Ashburner M, *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).

Attwood TK, *et al.* PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research* **31(1)**, 400-402 (2003).

Babbitt PC. Definitions of enzyme function for the structural genomics era. *Current Opinion in Chemical Biology* **7**, 230-237 (2003).

Bairoch A. The Enzyme Data-Bank. *Nucleic Acids Research* **21**, 3155-3156 (1993).

Bairoch A. The ENZYME database in 2000. *Nucleic Acids Research* **28**, 304-305 (2000).

Baldi P and Brunak S. Bioinformatics: The Machine Learning Approach. Massachusetts Institute of Technology, 2nd Edition (2001).

Bateman A, *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**, D138-D141 (2004).

Berezin C, *et al.* ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **20**, 1322-1324 (2004).

Berman HM, *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

Brenner SE. Errors in genome annotation. *Trends in Genetics* **15**, 132-133 (1999).

Camon E, *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research* **32**, D262-D266 (2004).

Casari G, Sander C, and Valencia A. A Method to Predict Functional Residues in Proteins. *Nature Structural Biology* **2**, 171-178 (1995).

Chelliah V, Chen L, Blundell TL, and Lovell SC. Distinguishing Structural and Functional Restraints in Evolution in Order to Identify Interaction Sites. *Journal of Molecular Biology* **342**, 1487-1504 (2004).

Conant GC and Wolfe KH. Turning a Hobby into a Job: How Duplicated Genes Find New Functions. *Nature Reviews Genetics* **9**, 938-950 (2008).

Dayhoff MO. Survey of new data and computer methods of analysis. *In Atlas of protein sequence and structure.* vol. 5, suppl. 3. National Biomedical Research Foundation Georgetown University, Washington, D.C. (1978)

Del sol Mesa AD, Pazos F, and Valencia A. Automatic methods for predicting functionally important residues. *Journal of Molecular Biology* **326**, 1289-1302 (2003).

DeLano WL. The PyMol Molecular Graphics System. *DeLano Scientific LLC, Palo Alto, CA, USA.* http://www.pymol.org.

Devos D and Valencia A. Practical limits of function prediction. *Proteins-Structure Function and Genetics* **41**, 98-107 (2000).

Devos D and Valencia A. Intrinsic errors in genome annotation. *Trends in Genetics* **17**, 429-431 (2001).

Dorit RL and Ayala FJ. Adh Evolution and the Phylogenetic Footprint. *Journal of Molecular Evolution* **40**, 658-662 (1995).

Durbin R, Eddy S, Krogh A, and Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, *Cambridge University Press*, Cambridge, UK (1998)

Eddy SR. Hidden Markov models. *Current Opinion in Structural Biology* **6**, 361-365 (1996).

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32 (5)**, 1792-1797 (2004)

Efron B and Gong G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician* **37 (1)**, 36-48 (1983)

Eisen JA. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* **8**, 163-167 (1998).

Eisen JA and Wu M. Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theoretical Population Biology* **61**, 481-487 (2002).

Espadaler J, *et al.* Detecting remotely related proteins by their interactions and sequence similarity. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7151-7156 (2005).

Felsenstein J. Confidence-Limits on Phylogenies - An Approach Using the Bootstrap. *Evolution* **39**, 783-791 (1985).

Feng DF and Doolittle RF. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Computer Methods for Macromolecular Sequence Analysis* **266**, 368-382 (1996).

Fersht AR. Structure and Mechanism in Protein Science: a guide to enzyme catalysis and protein folding. *3rd Edition, W.H. Freeman & Co Ltd* (1999).

Filizola M and Weinstein H. The study of G-protein coupled receptor oligomerization with computational modelling and bioinformatics. *Febs Journal* **272**, 2926-2938 (2005).

Fitch WM. Distinguishing Homologous from Analogous Proteins. *Systematic Zoology* **19**, 99-& (1970).

Fleischmann W, Moller S, Gateau A and Apweiler R. A novel method for automatic functional annotation of proteins. *Bioinformatics* **15**, 228-233 (1999).

Friedberg I. Automated protein function prediction: the genomic challenge. *Briefings in Bioinformatics* **7**, 3, 225-242 (2006).

Frommlet F, Futschik A, and Bogdan M. On the significance of sequence alignments when using multiple scoring matrices. *Bioinformatics* **20**, 6, 881-887 (2004).

Gelbart WM, *et al.* Flybase - the Drosophila Database. *Nucleic Acids Research* **22**, 3456-3458 (1994).

Gilks WR, Audit B, De Angelis D, Tsoka S, and Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **18**, 1641-1649 (2002).

Glaser F, *et al.* ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **19**, 163-164 (2003).

Good P. Resampling Methods: A Practical Guide to Data Analysis. *Birkhauser Verlag AG* (1999)

Gonnet GH, Hallett MT, Korostensky C, and Bernardin L. Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* **16**, 2, 101-103 (2000).

Gotoh O. Multiple sequence alignment: Algorithms and applications. *Advances in Biophysics, Vol 36, 1999* **36**, 159-206 (1999).

Goward CR and Nicholls DJ. Malate dehydrogenase: A model for structure, evolution, and catalysis. *Protein Science* **3,** 1883-1888 (1994)

Gribskov M, Mclachlan AD, and Eisenberg D. Profile Analysis - Detection of Distantly Related Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 4355-4358 (1987).

Gutteridge A, Bartlett G, and Thornton J. Using a Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes. *Journal of Molecular Biology* **330**, 719-734 (2003).

Hannenhalli SS and Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *Journal of Molecular Biology* **303**, 61-76 (2000).

Hegyi H and Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology* **288**, 147-164 (1999).

Hegyi H and Gerstein M. Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. *Genome Research* **11**, 1632-1640 (2001).

Henikoff S and Henikoff JG. Amino acid substitution matrices from protein blocks. *PNAS* **89**, 10915-10919 (1992)

Hsu C, Chang C, and Lin C. A Practical Guide to Support Vector Classification. *http://www.csie.ntu.edu.tw/~cjlin/libsvm/*, (2008).

Hulo N, *et al.* Recent improvements to the PROSITE database. *Nucleic Acids Research* **32**, D134-D137 (2004).

Hunter S, *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Research* **37 (suppl 1)**, D211-D215 (2009).

Iliopoulos I, *et al.* Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* **19**, 717-726 (2003).

Jensen LJ, Gupta R, Staerfeldt HH, and Brunak S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* **19**, 635-642 (2003).

Joachims T. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press*, (1999).

Joachims T. Training Linear SVMs in Linear Time, *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, (2006).

Johnson JM and Church GM. Predicting ligand-binding function in families of bacterial receptors. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 3965-3970 (2000).

Jones DT, Taylor WR, and Thornton JM. The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275-282 (1992).

Jones DT. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of Molecular Biology* **292**, 195-202 (1999).

Jones DT and Swindells MB. Getting the most from PSI-BLAST. *Trends in Biochemical Sciences* **27**, 161-164 (2002).

Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

Katoh K, Misawa K, Kuma K, and Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002).

Kretschmann E, Fleischmann W, and Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* **17**, 920-926 (2001).

Lander ES, *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

Landgraf R, Xenarios I, and Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of Molecular Biology* **307**, 1487-1502 (2001).

Larkin MA, *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21), 2947-2948 (2007).

Li W and Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 13, 1658-1659 (2006).

Lichtarge O, Bourne HR, and Cohen FE. Evolutionarily conserved G(alpha beta gamma) binding surfaces support a model of the G protein-receptor complex. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 7507-7511 (1996).

Lichtarge O and Sowa ME. Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology* **12**, 21-27 (2002).

Livingstone CD and Barton GJ. Protein-Sequence Alignments - A Strategy for the Hierarchical Analysis of Residue Conservation. *Computer Applications in the Biosciences* **9**, 745-756 (1993).

Madabushi S, *et al.* Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology* **316**, 139-154 (2002).

Marcotte EM. Computational genetics: finding protein function by nonhomology methods. *Current Opinion in Structural Biology* **10**, 359-365 (2000).

Martin ACR, *et al*. Protein Folds and Functions. *Structure* **6**, 875-884 (1998).

Martin DMA, Berriman M, and Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* **5**, art-178 (2004).

Moller S, Leser U, Fleischmann W, and Apweiler R. EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics* **15**, 219-227 (1999).

Mount DW. Bioinformatics: sequence and genome analysis. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York, 2nd Edition (2004)

Murzin AG, Brenner SE, Hubbard T, and Chothia C. Scop - A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* **247**, 536-540 (1995).

Needleman SB and Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* **48,** 443-453 (1970)

Notredame C, Higgins DG, and Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217 (2000).

Ohno S. Evolution by Gene Duplication. Springer, New York (1970).

Orengo CA, *et al.* CATH - a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108 (1997).

Orengo CA and Thornton JM. Protein families and their evolution - A structural perspective. *Annual Review of Biochemistry* **74**, 867-900 (2005).

Ouzounis C and Karp P. The past, present and future of genome-wide re-annotation. *Genome Biology* **3**, comment2001 (2002).

Ouzounis CA, Coulson RMR, Enright AJ, Kunin V, and Pereira-Leal JB. Classification schemes for protein structure and function. *Nature Reviews Genetics* **4**, 508-519 (2003).

Patthy L. Protein Evolution. Blackwell Science, Oxford (1999).

Pawlowski K, Jaroszewski L, Rychlewski L, and Godzik A. Sensitive Sequence Comparison As Protein Function Predictor. *Pacific Symposium of Biocomputing* (2000).

Pazos F, *et al.* Comparative Analysis of Different Methods for the Detection of Specificity Regions in Protein Families. 132-145. 1997. Singapore, World Scientific. Biocomputing and Emergent Computation. Lundh D, Olsson B and Narayan A. (1997).

Pazos F and Sternberg MJE. Automated prediction of protein function and detection of functional sites from structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14754-14759 (2004).

Pazos F, Rausell A, and Valencia A. Phylogeny-independent Detection of Functional Residues. *Bioinformatics* **15**, 22(12), 1440-1448 (2006).

Pearson WR and Lipman DJ. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 2444-2448 (1988).

Pei J and Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 8, 700-712 (2001)

Petrova NV and Wu CH. Prediction of Catalytic Residues using Support Vector Machine with Selected Protein Sequence and Structural Properties. *BMC Bioinformatics* **7**:312, (2006).

Platt J. Probabilities for Support Vector Machines in Advances in Large Margin Classifiers. Smola A, Bartlett P, Scholkopf B, Schuurmans D, eds. MIT Press, pp. 61-74 (1999).

Porter CT, Bartlett GJ, and Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research* **32**, D129-D133 (2004).

Press WH, Teukolsky SA, Vetterling WT, and Flannery BP. Numerical Recipes in C: The art of scientific computing. *Cambridge University Press, Cambridge* (1992)

Reese JT and Pearson WR. Empirical Determination of Effective Gap Penalties for Sequence Comparison. *Bioinformatics* **18**, 11, 1500-1507 (2002).

Rice P, Longden I, and Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16,** 6, 276-277 (2000)

Riley M. Systems for categorizing functions of gene products. *Current Opinion in Structural Biology* **8**, 388-392 (1998).

Riley ML, Schmidt T, Wagner C, Mewes HW, and Frishman D. The PEDANT genome database in 2005. *Nucleic Acids Research* **33**, D308-D310 (2005).

Rost B. Enzyme function less conserved than anticipated. *Journal of Molecular Biology* **318**, 595-608 (2002).

Rost B, Liu J, Nair R, Wrzeszczynski KO, and Ofran Y. Automatic prediction of protein function. *Cellular and Molecular Life Sciences* **60**, 2637-2650 (2003).

Saghatelian A and Cravatt BF. Assignment of protein function in the postgenomic era. *Nature Chemical Biology* **1,** 3, 130-142 (2005)

Schaffer AA, *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**, 2994-3005 (2001).

Sethi A, O'Donoghue P, and Luthey-Schulten Z. Evolutionary profiles from the QR factorization of multiple sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4045-4050 (2005).

Sjolander K. Phylogenetic inference in protein superfamilies: analysis of SH2 domains. Proc Int Conf Intell Syst Mol Biol.1998 6, 165-174 (1998).

Sjolander K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* **20**, 170-179 (2004).

Storm CEV and Sonnhammer ELL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18**, 92-99 (2002).

Tang Y, Sheng Z, Chen Y, and Zhang Z. An Improved Prediction of Catalytic Residues in Enzyme Structures. *Protein Engineering Design & Selection* **21**, 5, 295-302 (2008).

Tatusov RL, Koonin EV, and Lipman DJ. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

Tatusov RL, *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, art-41 (2003).

Taylor JS and Raes J. Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics* **38**, 615-643 (2004).

Taylor WR. The Classification of Amino-Acid Conservation. *Journal of Theoretical Biology* **119**, 205-& (1986).

Tetko IV, *et al.* MIPS bacterial genomes functional annotation benchmark dataset. *Bioinformatics* **21**, 2520-2521 (2005).

Thompson JD, Higgins DG, and Gibson TJ. Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* **22**, 4673-4680 (1994).

Tian WD and Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology* **333**, 863-882 (2003).

Tian WD, Arakaki AK, and Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Research* **32**, 6226-6239 (2004).

Todd AE, Orengo CA, and Thornton JM. Evolution of protein function, from a structural perspective. *Current Opinion in Chemical Biology* **3**, 548-556 (1999).

Todd AE, Orengo CA, and Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology* **307**, 1113-1143 (2001).

Tress ML, Jones D, and Valencia A. Predicting reliable regions in protein alignments from sequence profiles. *Journal of Molecular Biology* **330**, 705-718 (2003).

Tucker CL, *et al.*Two amino acid substitutions convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 11, 5993-5997 (1998).

Valdar WSJ. Scoring residue conservation. *Proteins-Structure Function and Genetics* **48**, 227-241 (2002).

Valencia A. Automatic annotation of protein function. *Current Opinion in Structural Biology* **15**, 267-274 (2005).

Vapnik V, The Nature of Statistical Learning Theory. *Springer* (1995).

Venter JC. The sequence of the human genome (vol 292, pg 1304, 2001). *Science* **292**, 1838 (2001).

Waterhouse AM, *et al.* Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25** (9), 1189-1191 (2009).

Watson JD, Laskowski RA, and Thornton JM. Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology* **15**, 275-284 (2005).

Whisstock JC and Lesk AM. Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics* **36**, 307-340 (2003).

Wilks HM, *et al*. A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science* **242,** 4885, 1541-1544 (1988)

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. *Journal of Molecular Biology* **297**, 233-249 (2000).

Wootton JC and Federhen S. Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.* **266,** 554-571 (1996)

Xie HQ, *et al.* Large-scale protein annotation through gene ontology. *Genome Research* **12**, 785-794 (2002).

Youn E, *et al.* Evaluation of Features for Catalytic Residue Prediction in Novel Folds. *Protein Science* **16**, 216-226 (2007).