IMPROVED COMPARATIVE MODELLING USING PROTEIN STRUCTURE PREDICTION

VICTORIA ANN MCKENNA
Molecules To Cells
2007

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Life Sciences

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

3D	THREE Dimensional
BLAST	Basic Local Alignment of Sequences Tool
BLOSUM	BLOcks Substitution Matrix
CASP	Critical Assessment of protein Structure Prediction
CDM	Consensus Data Mining
CE	Combinatorial Extension
CHARMM	Chemistry at HARvard Molecular Mechanics
COACH	Comparison Of Alignments by Constructing HMMs
CpFE	Cumulative Pseudo-Free Energy
DNA	DeoxyriboNucleic Acid
DSSP	Dictionary of Secondary Structures for Proteins
EC	Enzyme Commission
ELEPHANT	EmpiricaL Enhancement of Predicted HelicAl N-Termini
E-VALUE	Expectation VALUE
FRAGFOLD	FRAGment FOLDing
GOR	Garnier Osguthorpe and Robson
HMM	Hidden Markov Models
HMMSTR	Hidden Markov Models for connecting library of STRucture fragments
INDELS	Insertions and Deletions
JPRED	Jury PREDiction
MAMMOTH	MAtching Molecular Models Obtained from THeory
MSA	Multiple Sequence Alignment
Modelable	The amount of the alignment which MODELLER will build
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NC-IUBMB	NomenClature of the International Union of Biochemistry and Moleculai Biology
NiRMSD	Normalised Interaction RMSD
NMR	Nuclear Magnetic Resonance
NN	Neural Network
NNSSP	Nearest Neighbour Secondary Structure Prediction
PAM	Point Accepted Mutation
PDB	Protein Data Bank
PHD	Profile network HeiDlberg
PROCHECK	PROgrams to CHECK the stereochemical quality of protein structures
PSI	Protein Structure Initiative
PSI-BLAST	Position Specific Iterative BLAST
PSSM	Position Specific Scoring Matrix
Reference	The original target and template sequences before submitting to alignment programs
Retained	The amount of alignment each alignment method keeps
RMSD	Root Mean Square Deviation
SCOP	Structural Classification of Proteins
SOV	Segment OVerlap
SPACI	Summary PDB ASTRAL Check Index
SPTREMBL	Swiss-Prot TREMBL

SST	**S**econdary **St**ructure
SST1	The DSSP assignment of the target **S**econdary **St**ructure
SST2	No **S**econdary **ST**ructure restraints used, the default in MODELLER
SST3	The template containing gaps inserted into it
SSTRUC	**S**econdary **STRUC**ture
STALIN	protein **ST**ructural **ALI**g**N**ment
SVM	**S**upport **V**ector **M**achine
TREMBL	**Tr**anslation of **E**uropean **M**olecular **B**iology **L**aboratory nuclec sequence database

ABSTRACT

To understand the physiological role of proteins, a three-dimensional structure is a major asset. Given the expanding pool of sequences derived from genome sequencing projects, there is an increased demand to produce more accurate and effective comparative (homology) models of proteins. Comparative modelling exploits the concept that proteins with high sequence similarity adopt similar three-dimensional structures. The known protein structure can be used as a guide (or template) to predict the three-dimensional structure of the query protein (or target). The general accuracy of the model usually depends on the degree of similarity between the target-template sequences. The major limitation is that below approximately 30% sequence identity modelling becomes very difficult and errors in the sequence alignment become fatal. This would not prove to be a problem if most of the potential models existed at the higher percentage identity regions, but this is not the case. Secondary structure prediction can provide restraints to guide the model building process when an appropriate template cannot be found. The resulting models built with restraints were found to have lower RMSDs than those built without restraints, with more improvements seen in the region below 30% sequence identity. The accuracy of the starts of the helices was improved due to the ELEPHANT prediction algorithm, thus this increased the accuracy of some of the loops modelled. Various alignment protocols, including sequence and profile based, were assessed against structure based alignments. This revealed that profile and HMM methods outperformed sequence based methods, with the interface regions being more accurately aligned than the rest of the alignment, even with low sequence identity pairs. These alignments were then used in the comparative modelling protocol. The results suggested that the profile and HMM based methods could model the recognition region contacts more accurately than the sequence based methods, and the contacts could be modelled surprisingly accurately, even when the target-template pairs share modest similarity. This offered the potential for models to be built, using pairs stretching into the twilight zone, with modestly accurate interface areas even when the rest of the model may be deemed useless; providing predicted structures to extract functional information from.

DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

ACKNOWLEDGMENTS

Many people have helped me during the time I have spent completing this PhD and I am sure to miss someone out, it is not intentional and if it is you, I apologise now. I have really appreciated the support offered by many of my friends, family and co-workers. I would like to thank the following people:

- My PhD supervisor Dr. Simon Hubbard, whose expert advice and knowledge, and at difficult times, understanding, has made the production of this thesis possible.
- The BBSRC for funding this PhD.
- My Mum, for all her wisdom, humor, support and conversations to keep things in perspective when the going got tough. Mark, my brother Paul, and the rest of my family for making me smile when I needed it.
- My fellow research colleagues: Christian Cole, Ajanthah Sangaralingam, Paul Dobson, Julian Selley, Jennifer Siepen, Jennifer Lynch, Pedro Chan, Craig Lawless, Simon Williams, James Kitchen, Lauren Faulkner, Claire Wilson, Stephen Littler, and all of the Bioinformatics corridor, for providing computational support, knowledge, support in general, food and amongst other things, highly entertaining conversations. I will miss you all.
- All of my friends outside the University, but a few I would like to thank in particular for listening to me, helping me and generally being great friends. They include: Vanessa, Zoe, Thomas, Jamie, Gary, Phil, Chris, Sian, James A, Tim and Rachael.
- I would also like to thank Gail and Chris, Tim, Abi and Sue for all the understanding, interesting conversations and holidays they have given me.
- And finally James, whom without, I would quite literally have been unable to finish this thesis. He has provided me with so much support, love, understanding, comedy, patience and guidance; I can not thank him enough.

1. INTRODUCTION

This project is focused on improving and assessing the prediction of protein tertiary structure from its amino acid sequence, primarily targeted at comparative modelling of amino acid sequences using template structures.

This introduction will cover the basics of protein structure and the main concepts of protein structure prediction will be introduced and discussed. This will include a brief overview of the main levels of protein structure, an insight into the importance of structure prediction and its context in the scientific community and a description of three-dimensional structure prediction methods focusing on comparative modelling. The results chapters will contain more information relevant to each project and the methods used, so alignment techniques, secondary structure prediction and MODELLER will not be discussed here.

1.1 THE LEVELS OF PROTEIN STRUCTURE

Proteins are the most versatile materials used by the cell to support biological activity. It is no surprise that so much effort is devoted not only to decoding genomes in terms of the identification and annotation of the encoded proteins, but also more importantly to understanding the functions of the proteins, their regulation and their interactions in the cell (Bolognesi & Smith, 2006).

Proteins are polymers containing a main-chain of repeating peptides with a side-chain attached to each main-chain carbon. Natural proteins contain a basic repertoire of twenty amino acids. All of the twenty amino acids have in common a central, asymmetric ca rbon atom (Cα) to which are attached a hydrogen atom, an amino group (NH_2), and a carboxyl group (COOH), the only exception being proline which does not have an amino group. Side-chains are attached to the asymmetric Cα atom in all amino acids except glycine, which has no side-chain group. The unique sequence of the side-chains on the peptide units gives each protein its individual characteristics. Natural proteins contain only L-isomers and the D-isomeric form is not seen. Proline is special because its side-chain is linked to the backbone by link closure. Amino acids are joined end to end during protein synthesis by the formation of peptide bonds when the carboxyl group of one amino acid condenses with the amino group of the next to eliminate water.

Protein structure exists in a hierarchical nature (Figure 1.1a and 1.1b). A set of primary chemical bonds joins the constituent amino acids together into what is termed the primary sequence or structure. The secondary structure is formed from regular twists and turns in the backbone into structures known as helices and strands. Hydrogen-bonding of the main-chain amides and carbonyl groups is responsible for these two major structural types. The next level is the tertiary structure of a single polypeptide protein chain which is the assembly and interactions of the helices and sheets into what is often termed a protein "fold". For proteins composed of more than one subunit (more than one polypeptide chain) the overall assembly of the individual polypeptide monomers is termed the quaternary structure.

The hydrophobicity of an amino acid is a measure of the thermodynamic interaction between the side-chain and water. Hydrocarbon side-chains are electrically neutral and they interact unfavourably with water. This hydrophobic effect provides an important component of the driving force for protein folding, there is a tendency for the hydrophobic side-chains to sequester themselves in the interior of a protein away from contact with water, leaving polar residues on the surface of the protein. The accessible surface area of a protein is the area of molecular surface accessible to a water molecule (modelled as a sphere 1.4Å in radius) and helps rationalise the hydrophobic contribution to the thermodynamics of protein folding and interactions.

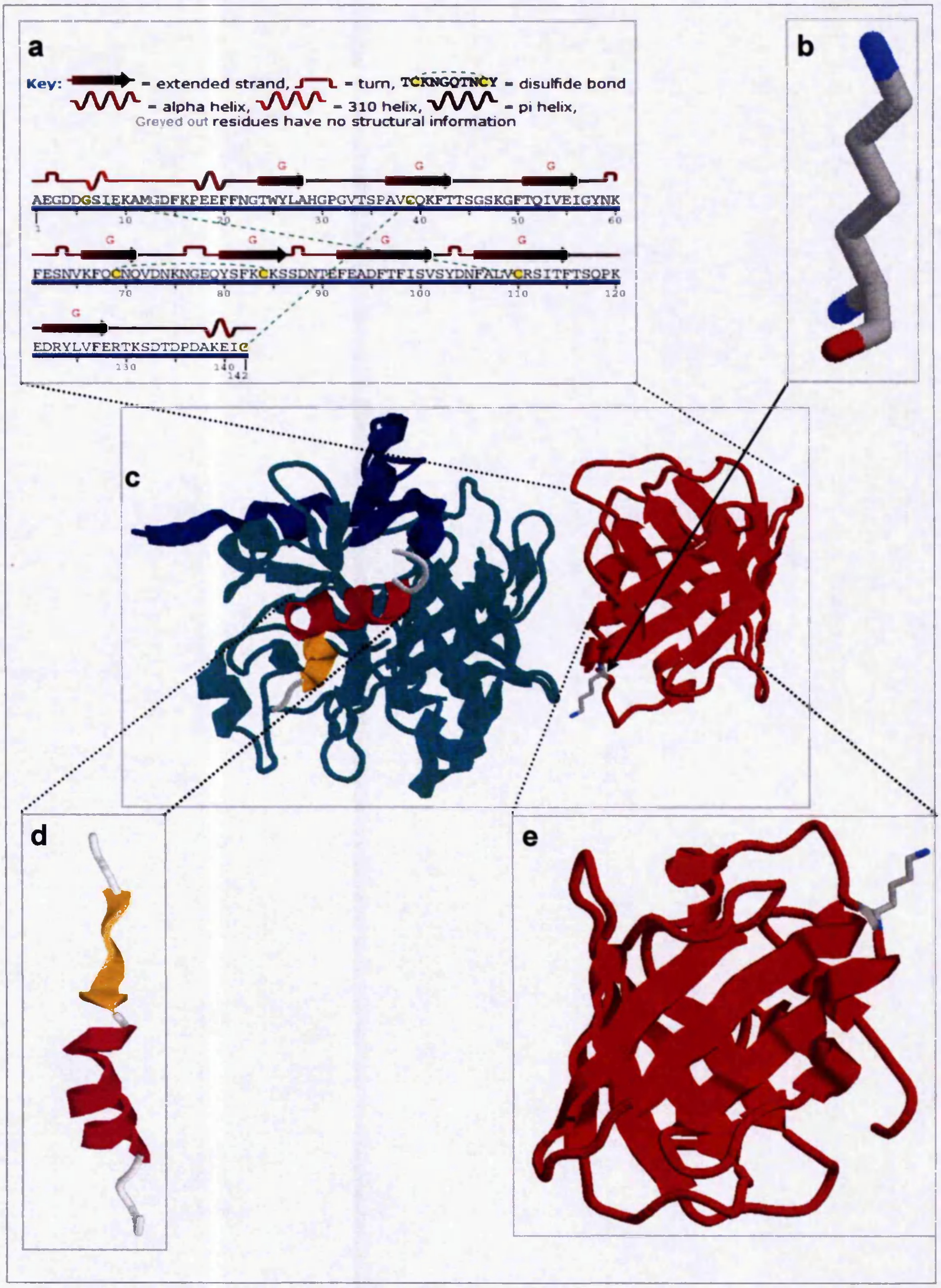

a
Key: = extended strand, = turn, TCKNGQTNCY = disulfide bond
= alpha helix, = 310 helix, = pi helix,
Greyed out residues have no structural information
G
G
G
AEGDDGSIEKAMGDFKPEEFFNGTWYLAHGPGVTSPAVCQKFTTSGSKGFTQIVEIGYNK
1
10
20
30
40
50
60
G
G
G
G
FESNVKFQCNQVDNKNGEQYSFKCKSSDNTEFEADFTFISVSYDNFALVCRSITFTSQPK
70
80
90
100
110
120
G
EDRYLVFERTKSDTDPDAKEIC
130
140
142
b
c
d
e

Figure 1.1a. The Different Levels of Protein Structure - A representation of the 1avg (Fuentes-Prior, et al., 1997) PDB structure. Shown at the centre (c) is the quaternary structure of all 4 domains of 1avg. The primary, linear amino acid sequence of Chain I (red) is (a) and the residue Lysine 10 of Chain I is shown (b). The secondary structure (d) of a helix (in pink) and sheet (in yellow) are displayed from chain H (green) and the tertiary structure of the domain from chain I is shown (e).

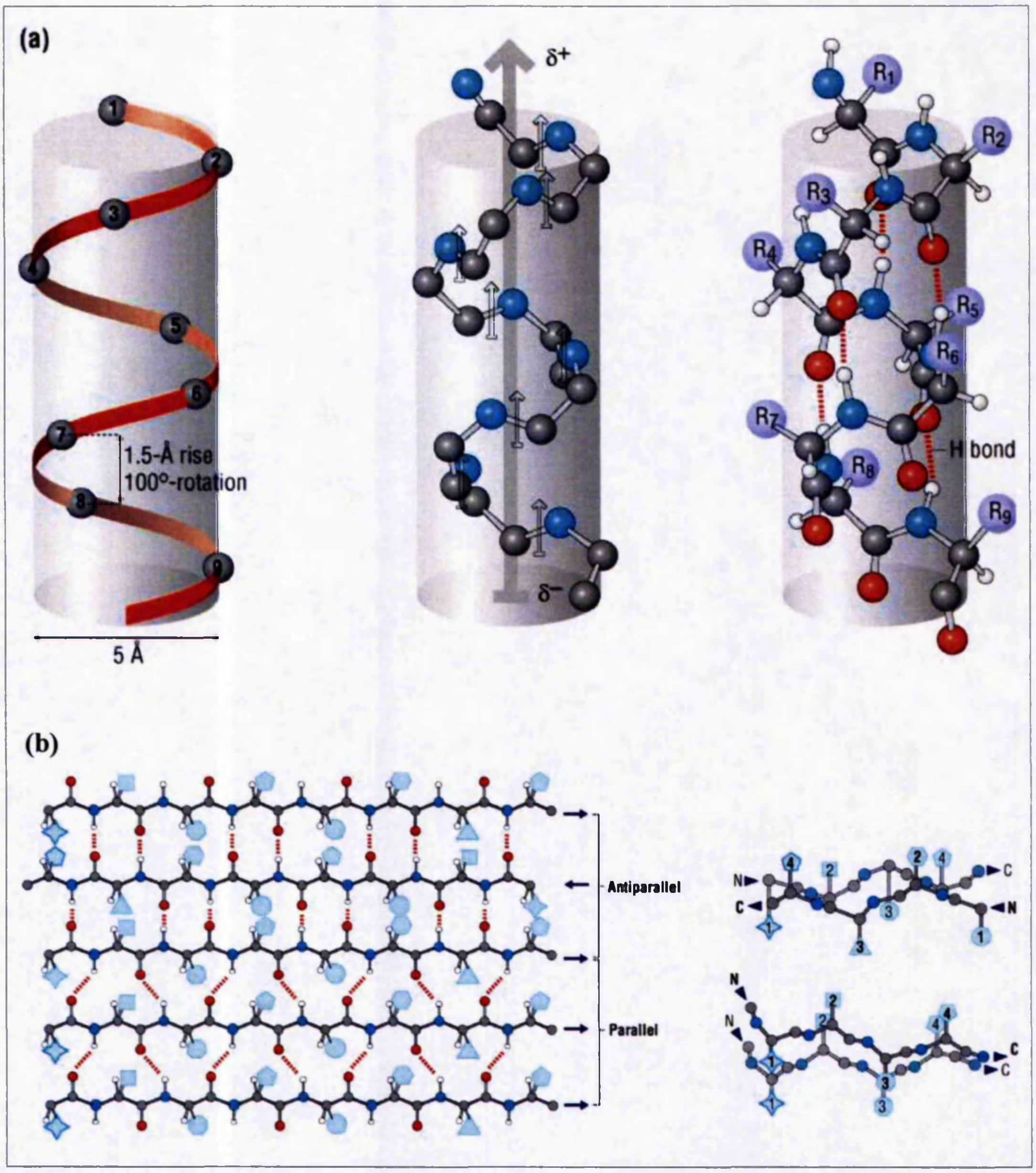


Figure 1.1b. Helix and Sheet Structure – The structure of the alpha-helix (a) and the beta-sheet (b) can be seen. The alpha-helix has 3.6 amino acids per turn. All main chain amino and carboxyl groups are hydrogen bonded, and the R groups stick out from the structure in a spiral arrangement. If the amino termini are on the same end of each chain, the beta-sheet is termed parallel, and if the chains run in the opposite direction (amino termini on opposite ends), the sheet is termed antiparallel. Image accessed from <http://wiz2.pharm.wayne.edu/biochem/prot.html>.

Any possible conformation of the polypeptide chain of a protein places different sets of residues in proximity. The interactions of the amino acid side-chains with the main-chain backbone and with other solvents and ligands, determines the energy of the conformation. Proteins have evolved so that one folding pattern of the chain produces a set of interactions that is significantly more energetically favourable than the others. This corresponds to the native state (Gto, 1976). It is this native structure that allows the protein to carry out its biochemical function (Floudas *et al.*, 2006). According to Anfinsen (1973) proteins are not assembled into their native structures by a biological process, but folding is a purely physical process that depends only on the specific amino acid sequence of the protein and the surrounding solvent. Anfinsen's hypothesis implies that in principle protein structure can be predicted if a model of the free energy is available, and if the global minimum of this function can be identified. This idea defines the protein structure prediction problem well. Protein structure prediction remains utterly complex, since even short amino acid sequences can form an abundant number of geometric structures among which the free energy minimum has to be defined (Floudas *et al.*, 2006).

1.2 THE IMPORTANCE OF STUCTURE PREDICTION

The enormous increase in the availability of data brought about by large scale genomic projects is paralleled by an equally unprecedented increase in the expectations for new medical, pharmacological, environmental and biotechnological discoveries (Tramontano, 2003). Although the draft sequence of the human genome has been published, the role of the gene products it encodes is far from being fully understood. Being able to read the linear sequence is not directly related to understanding its meaning. Therefore, the attention of many biologists is now focusing on the functional analysis of genomes (Peitsch, 2003).

1.2.1 Sequence, Structure and Function

Acquiring the function of a protein is important since the function and the physiological role of the protein can provide the basis for the discovery of novel medicines and protein-based products with medical, industrial or commercial applications (Peitsch, 2003). Biological function can be defined at several different levels, but in order to interfere with the function for therapeutic or investigative purposes, it needs to be characterised at the molecular level to identify the precise role of the specific amino acids and chemical groups. The problem is further complicated by the fact that function rather than being an attribute of a single protein, is determined by the plethora of interactions that it establishes with other proteins and with its surrounding environment (Tramontano, 2003). Thus, a protein's function is tightly linked to its three-dimensional structure, since this determines the spatial disposition of both the key chemical groups and the regions of the protein which can interact with partners. As residues located far apart in the primary sequence can be very close in space, and only a few residues are generally responsible for a protein's function, insights into the three-dimensional structure of a protein can represent a key component of the functional analysis process. Consequently, an atomic level three-dimensional representation to assign roles to specific residues is a major asset, both for planning experiments and explaining observations (Peitsch, 2003).

It is widely assumed that a structural resemblance between proteins implies a functional similarity. It is also widely assumed that structural features are closely related to sequence composition. Although a protein with a given sequence may potentially exist in different conformations, the chances that two close sequences will fold into distinctly different structures are so small that they are often neglected in research practice (Krissinel, 2007). Structure based transfer of functional information is preferred over sequence based as the similarity in the structure is generally more conserved than the similarity in the amino acid sequence and the protein structure allows a more informative transfer of functional description than the sequence alone (Sanchez, *et al.*, 2000).

Despite significant improvements in structure resolving methods, the gap between the number of known protein sequences and their resolved structures is rapidly increasing (Kazemian *et al,* 2007). This is mainly due to the increase in genome sequencing projects currently being undertaken, producing large numbers of sequences. However, the experimental process of structure determination is still relatively slow, despite recent significant advances in the field (Westhead & Thornton, 1998). This means there is an evident demand for more structures to be solved, which is increasing with each new genome project.

1.2.2 Structural Genomics

Structural genomics is a term that refers to high-throughput three-dimensional structure determination and analysis of biological macromolecules, which at this stage is primarily concerned with individual protein domains (Goldsmith-Fischman & Honig, 2003).

The ultimate aim of structural genomics is not to obtain the structures of all proteins, but to contribute to biology and medicine through functional annotation. Structural genomics focuses on delivering the st ructures of the complete protein repertoire of folds via X-ray crystallography and NMR experiments, so that all proteins are within homology-modelling range of one or more known experimental structures. Protein structure determination is an area of biology where both experimental and theoretical approaches complement each other (Sanchez, *et al.*, 2000). The major aims of structural genomics include (Watson *et al.*, 2007):

- High-throughput automation of protein production, structure determination and analysis;
- Increased coverage of protein fold space and hence the number of protein sequences amenable to homology modelling methods;
- Investigation of protein structure to elucidate function in health and disease;
- Reduction of the cost of structure determination.

The re are several current structural genomics efforts, and this includes the PSI (the Protein Structure Initiative). The PSI started seven years ago, with the long-term goal of making three-dimensional structures easily available after DNA sequence determination. The pilot phase aimed to streamline structure determination methods. Now, in its second year of phase 2, the emphasis has shifted to high-throughput structure determination, with a strong focus on improved bioinformatics-guided target selection. These efforts are already projected to outstrip the number of unique structures determined in the pilot phase (PSI consortium, http://www.nigms.nih.gov/Initiatives/PSI).

Even though the efforts of the structural genomic projects are great, they will not be sufficient to determine the structures of all the proteins of interest. This is when protein structure prediction can be used. Structural genomics aims to solve the key structures representing all folds, thus structure prediction methods such as comparative modelling will be essential to determine the structures for the remaining homologous proteins.

1.2.3 Structure Prediction and Bioinformatics

Over time, evolution has produced families of proteins whose members share the same three-dimensional architecture and frequently have detectably similar amino acid sequences. This conservation allows a structural description of all proteins in a family to be made even when only the structure of a single member is known. It is possible after the structure is acquired to then infer function from homology. This means that if a protein of unknown structure is related to a protein of known structure, and the relationship is detectable and more than nominal (above 10% sequence identity), the known structure can be used as a guide for predicting the unknown structure. For this to be a very successful method an accurate prediction technique and an accurate experimentally determined model that represents a protein in each of the families is required. To obtain these representative structures the scientific community is determining a vast amount of structures; however the community is dependent on experimental protein structure elucidation. The usual approaches, both x-ray diffraction and NMR, are hampered by technical hurdles and limitations, and so prediction of structures is required (Peitsch, 2003).

In the life sciences, the sheer volume of raw data that is being generated from the genome sequencing projects in need of annotation is unprecedented. Computational biology is being called upon, now more than ever, to process these data and provide us with biochemical, physiological, and evolutionary context. Even though experimental high-throughput functional annotations techniques have advanced, the time and cost of determining the function of every single gene and gene product are prohibitive. Therefore, most of the functional annotation will be done with computational tools (Friedberg *et al.*, 2006). Bioinformaticians must work in close relation with experimental scientists that determine protein structures in the lab to improve the accurate prediction of protein structures using known structures as an ultimate guide.

1.3 STRUCTURE PREDICTION METHODS

The prediction of the three-dimensional structure of a protein when only the amino acid sequence is known has been a problem of major interest for many years. Approaches have ranged from *ab initio* methods that use physical and chemical principles to model a protein from its raw amino acid sequence, to homology methods that are dependant primarily on the information available in sequence and structural databases. Threading methods and comparative modelling methods lie between these two extremes and involve the identification of a structural template that most closely resembles the structure of the query protein (Al-Lazikani *et al.*, 2001). *Ab initio* methods and threading will briefly be described before comparative modelling is introduced in greater detail later on.

1.3.1 *Ab Initio* Structure Prediction

The 'Holy Grail' of the protein modelling field has always been the construction of models of protein structures without the aid of a direct relationship to any experimentally known ones. The work of *ab initio* efforts is worthwhile for at least two reasons. Firstly, 'pure' structure prediction, even if only partially successful, provides a stringent test of our understanding of the principles of protein structure and energetics, and the role of folding pathways in attaining the functional conformation. Secondly, even when there is a database of structures representing all protein sequence families, *ab initio* techniques will still be required for modelling the differences between structures (Moult, 1999).

The native state of a protein represents the global free energy minimum that can be kinetically reached by the protein and, with rare exceptions, for example in chaperone-assisted folding (Feldman & Frydman (2000)), is solely determined by its amino acid sequence. The energetic terms that govern protein folding is not sufficiently understood to allow calculations of the minimum free energy structure for a given amino acid sequence. *Ab initio* methods carry out large-scale searches of conformational space for protein tertiary structures that are particularly low in free energy for the given amino acid sequence. The two key components of such methods are the procedure for efficiently carrying out the conformational search and the free energy function used for evaluating possible conformations. To allow rapid and efficient searching of conformational space, often only a subset of the atoms in the protein chain is represented explicitly; the potential functions must include terms that reflect the average effects of the omitted atoms and solvent molecules (Baker & Sali, 2001).

A particularly successful *ab initio* method is called Rosetta (Baker & Sali, 2001). Rosetta is based on the assumption that the distribution of conformations sampled for a given nine residue segment of the chain is reasonably well approximated by the sequence (and closely related sequences) in known protein structures. Fragment libraries for each three and nine residue segments of the chain are extracted from the protein structure database using a sequence profile-profile comparison method. The conformational space defined by these fragments is then searched using Monte Carlo procedure with an energy function that favours compact structures (Simons et al., 2001). This strategy resolves some of the problems with both the conformational search and the free energy function: the search is accelerated because the switching between different possible local structures can occur in a single step, and fewer demands are placed on the free energy function because the use of fragments of known structures ensures that the local interactions are close to optimal (Baker & Sali, 2001).

The precision needed in these calculations is not sufficient enough to discriminate between the energy of the native state and that of any other conformational states that the protein could assume. Leaving behind the idea of solving the problem on the basis of first principles, computational biology found other approaches based on the analysis of known protein structures (Peitsch, 2002).

1.3.2 Structure Prediction with Threading

While similar sequence implies similar structure, the converse is not necessarily true. In contrast, similar structures are often found for proteins for which no sequence similarity to any currently known protein structure can be detected. Fold recognition (or threading) methods are one class of structural modelling techniques that aim at predicting the three-dimensional folded structure for amino acid sequences for proteins which comparative modelling methods provide no reliable prediction (Floudas *et al.*, 2006).

Threading is the prediction that two proteins with no significant pair wise sequence identity will have similar folds. That is, given a library of known structures, determining which of them shares a folding pattern with a query protein of known sequence but unknown structure. The optimal alignment of the sequence onto a structure is found and then the likelihood that the unknown sequence adopts each fold is assessed and scored, with the fold having the highest score being inferred as the structure that is the most similar to the native fold of the query protein. The results are a nomination of a known structure that has the same fold as the query protein, or a statement that no protein in the library has the same fold as the unknown query protein (Lesk, 2002).

3D-PSSM (Bates *et al.*, 2001) is a threading method designed to take a protein sequence and attempt to predict its three-dimensional structure and its probable function. The sequence of unknown structure is "threaded" onto each structure in a library of known protein structures and a score for compatibility is calculated in each case.

Skolnick and co-workers (Skolnick *et al.*, 2004) developed an iterative approach that first aligns the target query protein and known structures in a database ignoring pair-wise residue interactions. In subsequent alignments, information from previous alignments is then used to evaluate pair-wise interaction energies. By identifying structurally similar regions in multiple template alignments, accurate regions of structure prediction can be distinguished from less accurate ones. They found that accurate fragments can be identified even if no template is convincing as a whole (Floudas *et al.*, 2006). This introduces the concept of fragment assembly.

Fragment assembly methods do not compare a target sequence to a known protein structure, they compare fragments of a target to fragments of known template structures obtained from the PDB (Protein Data Bank). Once appropriate fragments have been identified, they are assembled into a complete structure, often with the aid of scoring functions and optimisation algorithms (Floudas *et al.*, 2006). FRAGFOLD (Jones & McGuffin, 2003) is based on the assembly of super-secondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. In their new version of FRAGFOLD they attempt to greatly narrow the search of conformational space by pre-selecting super-secondary structural fragments from a library of highly resolved protein structures. FRAGFOLD selects favourable supersecondary structures of different lengths whereas ROSETTA () uses residues of nine amino acids in length. The generation of structures from these fragments is slightly different too.

1.3.3 Structure Prediction with Comparative Modelling

Comparative modelling is the method of choice for protein structure prediction, when the required known structural data is available, not only because of its higher accuracy compared to alternative methods, but also because it is possible to estimate *a priori* for the quality of the models that are produced, thereby allowing the usefulness of a model in a given functional role to be assessed beforehand (Cozzetto & Tramontano, 2005). Comparative modelling has already become one of the most effective computational approaches in facilitating structural/functional characterisation of many protein-coding sequences across genomes (Venclovas & Margelevicius, 2003). Approximately one half of all known sequences have at least one domain that is detectably related to at least one protein of known structure, and comparative modelling has the possibility to determine more structures than have been experimentally determined (Fiser & Sali, 2001). Comparative modelling aids functional classification of proteins without the use of experimentally determined structures and can reliably predict the three-dimensional structure of a protein with accuracy comparable to a low resolution experimentally determined structure. Errors are not a problem as such, since some aspects of function can be predicted from only coarse structural features of a model (Marti-Renom *et al.*, 2001).

Comparative modelling is also known as homology modelling and is based on the idea that if proteins display a high degree of similarity between their amino acid sequences they tend to adopt similar three-dimensional folds. Comparative modelling is the prediction of the three-dimensional structure of a protein from the known structure of one or more related proteins where the known (template) and unknown (target) proteins have significant and detectable pair wise sequence identity, usually above 25%. The assumption is that the unknown protein and the known protein(s) have nearly identical backbone structures in the aligned regions. Thus, the task is to place the side chains of the target correctly into the backbone of the template(s) (Rost & O'Donoghue, 1997). The results are a complete coordinate set for main chain and side chains intended to be a high quality model of the structure. The resulting models built using proteins with between 70% and 90% sequence identity are modestly accurate (comparable to medium resolution NMR structures and low resolution crystal structures).Th e accuracy of the model depends on the degree of similarity between the target and template sequence(s). The major limitation is that below 40% to 50% sequence identity modelling becomes very difficult and errors in the sequence alignment become fatal. Although comparative modelling is far from yielding perfect structures it is still seen as the most reliable method at present for three-dimensional prediction of proteins using homology.

The resulting models can be classified according to their correctness and accuracy, which in turn will impact their applicability and usefulness (Lesk, 2002). Figure 1.2 shows the applications of the models produced at different levels of accuracy. Within the low level accuracy range, models which are built using sequences sharing less than 30% sequence identity sometimes have less than 50% of their Cα atoms within 3.5Å of their true positions. In the mid accuracy range models have approximately 30-50% sequence identity corresponding to 85% of the Cα atoms being modelled within 3.5Å of their correct positions. The high accuracy range of models are based on target-template pairs with greater than 50% sequence identity, the average accuracy of these models will approach low resolution X-ray structures (3 Å resolution) or medium resolution NMR structures.

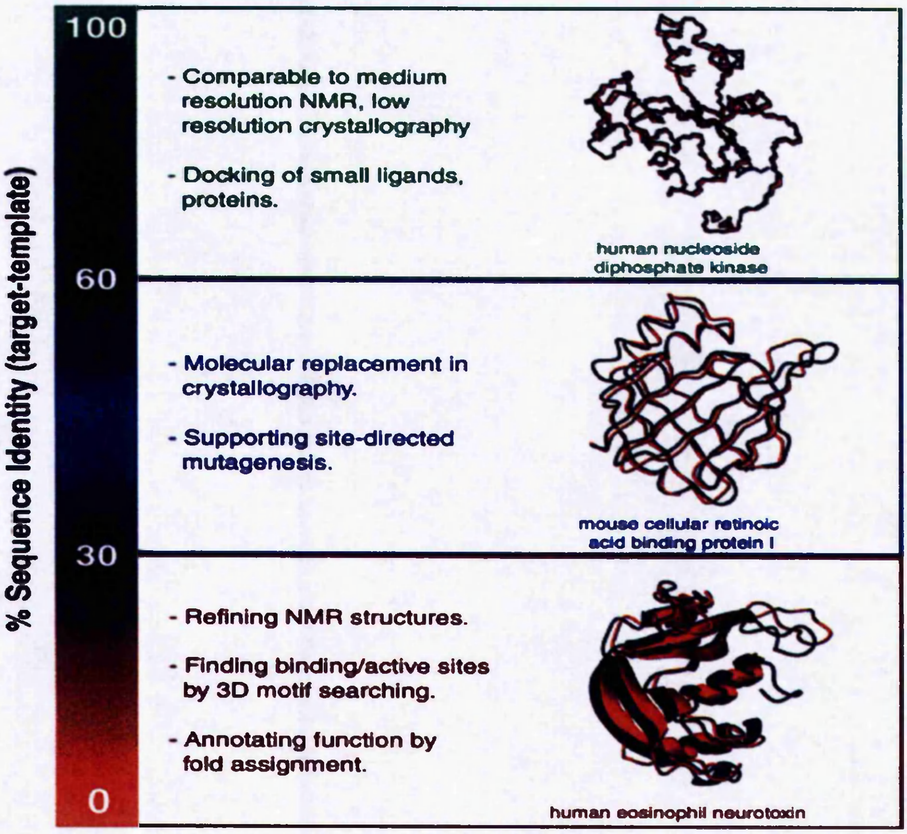

% Sequence Identity (target-template)
100
60
30
0
- Comparable to medium resolution NMR, low resolution crystallography
- Docking of small ligands, proteins.
human nucleoside diphosphate kinase
- Molecular replacement in crystallography.
- Supporting site-directed mutagenesis.
mouse cellular retinoic acid binding protein I
- Refining NMR structures.
- Finding binding/active sites by 3D motif searching.
- Annotating function by fold assignment.
human eosinophil neurotoxin

Figure 1.2. Applicability and Accuracy of Comparative Models (taken from Sanchez *et al.*, 2000). The accuracies, and thus, applications for the comparative models produced are usually dependent on the sequence percentage identities between targets and templates.

There are five main steps involved in comparative modelling (Figure 1.3).

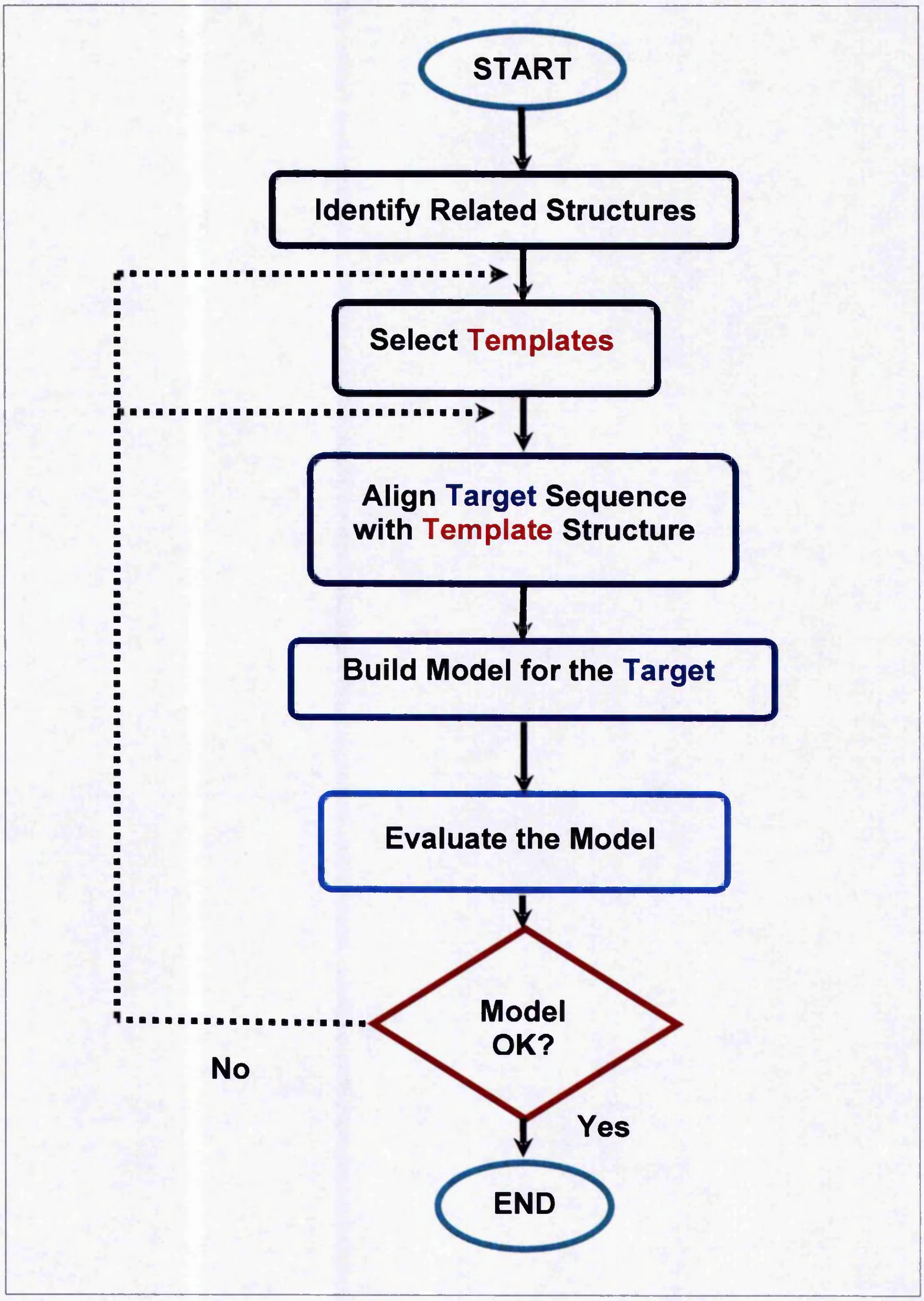

START
Identify Related Structures
Select Templates
Align Target Sequence with Template Structure
Build Model for the Target
Evaluate the Model
Model OK?
No
Yes
END

Figure 1.3. The Five Main Steps in Comparative Modelling. Above five main steps in comparative modelling are listed these steps will be described in detail within this Chapter. There is the option to iterate over the model building process until a satisfactory model has been built.

1.3.3.1 Fold Assignment

The first step in model building is finding related known protein structures for as many domains in the modelled query sequence of the unknown structure as possible. A search can be made of a structure database with the target sequence as the query. Alternatively, a search can be made of a sequence database. A useful and popular method that uses the target sequence to independently search the database using a pair wise sequence-sequence comparison is BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1990). PSI-BLAST (Position Specific Iterative – BLAST) relies on multiple sequence comparisons to improve the sensitivity of the search, usually finding more distant homologs than BLAST. Threading methods can be used when there are no sequences clearly related to the modelling target. These methods rely on pair-wise comparison of a protein sequence and a protein of known structure. The target sequence is threaded through a library of 3D folds (Marti-Renom *et al.*, 2000).

1.3.3.2 Template Selection

The selection of one or more templates is based on several factors. One factor is the environment of the target and the template, for instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, it is likely that the model will be incorrect irrespective of the target-template similarity or the accuracy of the template structure and the quality of the experimentally determined structure. The term environment is used in a broad sense and includes all factors that determine protein structure, except its sequence, for example, solvent, ligands and *p*H. The quality of the experimental template structure is another important factor. The resolution and the R-factor of a crystallographic structure are indicative of its accuracy. The priorities of the criteria for template selection depend upon the purpose of the model. For instance, if a protein-ligand model is to be constructed, the template should contain a similar ligand, this is more important than the resolution. However, if it is for the analysis of the geometry of the active site of an enzyme the high resolution template should be used first (Fiser *et al.*, 2001). The quality of the final model increases with the overall sequence similarity between the template and th e target, and decreases with the number and length of gaps in the alignment. A single target can be built based on the structure of several templates.The program MODELLER allows this in two ways, multiple template structures may be aligned with different domains of the target with little overlap, and the template structures may be aligned with the same part of the target. It is possible to generate and evaluate all the possible models and score them. The construction of a multiple sequence alignment and a phylogenetic tree can help in selecting a template (Sanchez & Sali, 1997).

Currently, the most widely used template selection methods involve the representation of templates as profiles. Profiles are a more accurate representation of the variability that can occur at individual positions of a protein sequence. This results in more sensitive detection of remote homologs (Petrey & Honig, 2005).

In theory, due to the greater coverage of conformational space, using more than one template should generate a model that is more accurate than any of the individual templates. However, CASP4 showed that only very occasionally were multi-template models more accurate than single-template models. The reasons for this are the choice of templates and sequence alignment errors (Tramontano *et al.*, 2001). Contreras-Moreira, Paul W. Fitzjohn and Paul A. Bates (2003) produced a study on template selection and whether multiple (up to five templates) or single (based on the highest percentage identity to the target) should be used. They applied techniques of a genetic algorithm, with crossover and mutation to select the different parts of the templates. It was concluded that current methodology is not taking full advantage of the possibility of using several templates to build comparative models, however in general, multiple-template models are no better than their corresponding ideal single-template models and can be considerably worse (Figure 1.4). Only in a marginal proportion of cases were multiple-template models found to improve over the ideal single-template model showing no preference for any region in the sequence identity range.

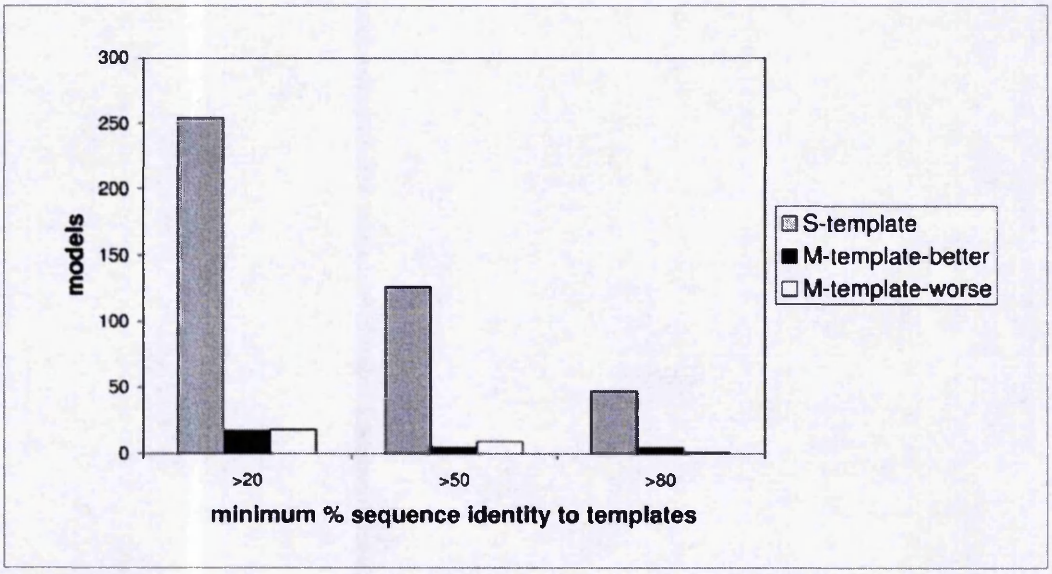


Figure 1.4. Single *Versus* Multiple Template Performance for Comparative Modelling (Contreras-Moreira *et al.*, 2003). Models were built using between one and five templates from the same SCOP family, with sequence identities ranging from 80–100%, 50–100% and 20–100% (*X*-axis). The *Y*-axis corresponds to the total number of models in each bin. Multiple-template models are compared to the best single-template model.

1.3.3.3 Target-Template Alignment

The results from the database search contain high and low regions of sequence similarity between the query and the database hits. The resulting alignment from the fold assignment method is usually not the optimal alignment for comparative modelling; searching methods are usually tuned for detecting remote relationships, not for optimal alignments. The correct alignment is the one in which the structurally equivalent positions are correctly aligned. This means that once the templates have been selected, the target sequence and template structure will have to be realigned using specialised methods such as MUSCLE (Edgar, 2004), to obtain a structurally relevant alignment (Marti-Renom *et al.*, 2002). More information on this step in the comparative modelling process can be found in the introduction section of Chapter 4.

1.3.3.4 Model Building and Side Chain Modelling

After the target-template alignment has been completed, the three-dimensional model can be built. A variety of methods can be used to construct a model for the target protein. The first and most widely used method is still rigid body assembly (Blundell *et al.*, 1987; Greer, 1990). This method assembles a model from a smaller number of rigid bodies obtained from aligned protein structures. The second technique is modelling by segment matching (Jones & Thirup, 1986; Levitt, 1992). This technique relies on the approximate positions of conserved atoms in the templates. The basis for this is models can be constructed using a subset of atomic positions from template structures as "guiding" positions. The third method is the satisfaction of spatial restraints of the target protein and local molecular geometry. This technique is used by MODELLER. Restraints are obtained by assuming that the corresponding distances and angles between aligned residues in the template and the target are similar. The model is then derived by minimising the violations of all the restraints (Marti-Renom *et al.*, 2003). More information on MODELLER can be found in chapter 2, section 2.17.

Other model building programs include SwissModel (Peitsch & Jongeneel, 1993), COMPOSER (Sutcliffe *et al.*, 1987), 3D-JIGSAW (Bates *et al.*, 2001), SegMod (Levitt, 1992), nest (Petrey *et al.*, 2003), Builder (Koehl & Delarue, 1994), and SCWRL (Bower et al., 1997).

The accuracy of the various model building techniques are relatively similar when used optimally (Marti-Renom *et al.*, 2002). It is important that the model building method allows a degree of flexibility and automation to enable easy recalculation of a model when a change is made in the alignment; it should be easy to calculate models based on several templates; and the method should provide tools to incorporate prior knowledge about the target (Marti-Renom *et al.*, 2000).

In a given fold family, structural variability is a result of substitutions, insertions, and deletions of residues during the evolution of members of the family. Such changes frequently correspond to exposed loop regions that connect elements of secondary structure in the protein fold. Thus, loops often determine the functional specificity of a given protein framework, and can contribute to active and binding sites. Consequently the modelling of loops is a major factor in determining the usefulness of comparative models in studying interactions between the protein and its ligands (Marti-Renom *et al.*, 2002). See Chapter 5, section 5.2 for more information on loop modelling.

Side-chain conformations can be predicted either from similar structures or from steric or energetic considerations. Two effects on side-chain conformation need to be considered. The first is the coupling between the main-chain and side chains, and the second is the trends in the distributions of the side-chain dihedral angles. Correlations between side-chain dihedral angle probabilities and backbone values are not dependent upon the secondary structure; rotamers (Summers & Karplus, 1989; Dunbrack & Karplus, 1993) can vary within the same secondary structure. As the sequence identity falls below 30% there are more variable conformations of side-chain packing, even though the fold is still the same, hence a backbone with less than 30% sequence identity to the sequence being modelled is not sufficient to produce the correct packing of buried side-chains (Marti-Renom *et al.*, 2000).

Programs that complete modelling by rigid body assembly are 3D-JIGSAW (Bates et al., 2001) and SWISS-MODEL (Schwede et al., 2004). SWISS-MODEL is a fully automated web server (http://swissmodel.expasy.org//SWISS-MODEL.html). The rigid body assembly approach is where a model is assembled from a small number of rigid bodies (large protein structure segments) obtained from the core of the aligned regions. The assembly involves fitting the rigid bodies onto the framework and rebuilding the non-conserved parts, i.e., loops and side chains. The main difference between the rigid body assembly programs lies in how side chains and loops are built.

1.3.3.5 Evaluating Models

As the similarity between the target and template sequences decreases, the errors that accumulate in the final model increase. Errors in comparative models can be divided into five main categories (Marti-Renom *et al.*, 2000):

- **Errors in side-chain packing.** As the sequences diverge, the packing of the side chains in the protein core changes. Side-chain errors are critical if they occur in the key regions of a protein.

- **Distortions and shifts in correctly aligned regions.** Even if the fold stays the same between target and template, the main-chain conformation can be different. Therefore, it is possible that in some correctly aligned regions of the model, the template is locally different from the target, resulting in errors in that region.

- **Errors in regions without a template.** Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are the most difficult regions to model.

- **Errors due to misalignments.** The largest source of errors in modelling is misalignments, especially when the target-template sequence identity falls below 30%.

- **Incorrect template.** When distantly related templates are used selecting the correct template can prove a problem. Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult.

One of the criticisms that comparative modelling receives is that the final model is usually closer to the template used than the target-experimental structure.

The model can be evaluated as a whole or as individual regions. The first step in model evaluation is to see if the model has the correct fold. If the correct template has been chosen and the template has been aligned accurately with the target sequence then the model will usually have the right fold. To assess whether the fold prediction is likely to be right the target will usually share high sequence similarity with the closest template (Sanchez & Sali, 1998).

After the fold has been proven correct, a more detailed evaluation of the overall model accuracy can be obtained based on the similarity between the target and template sequences. Above 30% sequence identity the similarity between the target and the template, due to the relationship between structural and sequence similarity of two proteins, is a relatively good predictor of the expected accuracy. In addition to target-template sequence identity, the environment and the target-template alignment can strongly influence the accuracy of the model (Sanchez & Sali, 1998).

Two types of evaluation can be carried out, internal and external. Internal evaluation consists of self-consistency checks to see whether a model satisfies the restraints used to calculate it. It tests whether the models have good stereochemistry; useful programs that perform this type of evaluation are PROCHECK (Laskowski *et al.*, 1998) and WHATCHECK (*Hooft et al.*, 1996). The features checked by these programs include bond lengths, bond angles, peptide bond and side-chain ring planarities, chirality, main-chain and side-chain torsion angles, and clashes between non-bonded pairs of atoms. External evaluation relies on information that was not used in the calculation of the model. When using a model of less than 30% sequence identity to the template the first purpose of the external evaluation is to see if the correct template was used. A way to predict if the template chosen is the correct one is to compare the Z-score of the model (a measure of the compatibility between the sequence and its structure which indicates how far and in what direction, that item deviates from its distribution's mean). Another external evaluation method is the prediction of unreliable regions in the model – the “pseudo energy” profile of a model, such as that produced by PROSALL (Sippl, 1993). These spatial features have been derived from high resolution protein structures and large deviations from these are usually interpreted as being a good indicator for errors in the model. The features include packing, formation of a hydrophobic core, residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom-atom distances, atomic volumes, and main-chain hydrogen bonding (Sanchez & Sali, 2000).

One of the main checks of model quality is to calculate the RMSD of the model. The RMSD is calculated between the alpha-carbons of the two protein structures unless otherwise stated. The RMSD is the root mean square deviation of the superposed atoms. It estimates the mean square distance between the equivalent alpha-carbons of the two superposed structures. The RMSD is a well established quantity to determine whether the model will be accurate enough to be used in applications such as protein-protein docking (Prasad *et al.*, 2003).

The aim of these evaluation techniques is to determine whether or not the model is acceptable. If it is not acceptable, that is if the current model violates some restraints, fails the profile tests, or simply does not appear satisfactory, these evaluations should help to re-align the target sequence and the templates for the next cycle of modelling.

1.3.3.6 Areas for Improvement

It has been noted that the areas where most of the errors in comparative modelling accumulate are when selecting the appropriate templates and obtaining the structurally relevant alignment of the template to the target (**Figure 2.4**). The twilight zone, and below, contain most of the errors in comparative modelling. This sequence similarity range will be targeted in this project in the hope to identify methods that may improve modelling in this difficult area.

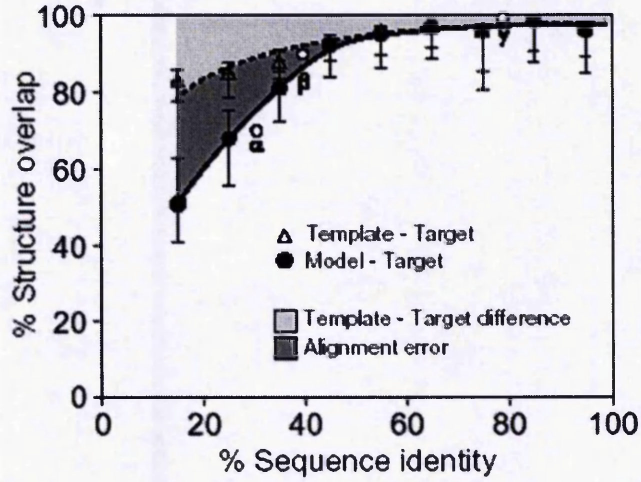


Figure 1.5. Errors in Comparative Modelling (Sanchez *et al.*, 2000a). The dotted line shows the actual experimentally determined structures and the solid line shows the modelled structures. The dark grey areas show the alignment errors and the light grey areas show errors due to template selection.

1.4 OVERVIEW

The work presented in this Thesis can be divided into three main areas all concerned with improving the process of comparative modelling in the twilight zone. Numerous programs and databases were used throughout this project and details can be found in Chapter 2. The first investigation (Chapter 3) includes insights into the improvement of the modelling of alpha-helices, in particular the N-termini of alpha-helices. This is aimed at improving the model building step of comparative modelling. Results suggest using the predicted secondary structure of the target improved the modelling of proteins within this non-trivial modelling area. Further experiments were completed for the target-template alignment step using peptidases as a test case. Chapter 4 evaluates different alignment protocols with respect to the overall alignment as well as evaluating residues at the interface between a protease and its inhibitor, concentrating on the results contained within the twilight zone. It was concluded that there are advantages of using some methods over others and that the majority of the alignment techniques could align the interface residues more accurately than the rest of the protein. These findings were then implemented in Chapter 5 in the model building step of comparative modelling. The conclusions drawn resulted in confirmation of certain alignment methods having increased alignment accuracy over other methods. The interface was more accurately modelled in the lower percentage identity areas when the contacts at these interfaces were tested. An example of a peptidase alignment pair and model is also presented in Chapter 5 to enhance understating of the results found in Chapters 4 and 5.

2. GENERAL RESOURCES AND DATABASES

A plethora of databases can be found that contain a wealth of biological sequence and structural information. The main databases and programs used within this thesis are described below. For a more detailed description about the implementation of the databases and programs please refer to the methods and materials section of each results chapter.

2.1 SEQUENCE DATABASE - *MEROPS*

MEROPS provided the sequence database of peptidases (28,445 peptidases) for Chapter 4. *MEROPS* has been in existence since 1996 and can be found at http://merops.sanger.ac.uk/. The *MEROPS* database provides a system of classification for protein functional groups which can be developed and used as an organisational framework around which to assemble a variety of related information. Most enzymes are named and classified on the basis of the enzyme reaction they catalyse but this has not proved possible for peptidases, because the specificities of enzymes hydrolysing proteins are almost impossible to determine rigorously or describe in a simple name. Trivial names have been used for most proteolytic enzymes but these can lead to confusion (Rawlings & Barrett, 1999). Thus the *MEROPS* database was developed. *MEROPS* (Rawlings *et al*, 2006) is a protein resource for information on peptidases (also termed proteases, proteinases and proteolytic enzymes) and their inhibitors. Around 3,000 individual peptidases and inhibitors are included in the database (Table 2.1). Information is assigned to each protein ranging from the peptidases' classification to literature references.

Catalytic Type	Sequences	Identifiers	Identifiers with EC Numbers	Identifiers with PDB entries
Aspartic	2915	191	31	36
Cysteine	8316	530	53	84
Glutamic	12	5	2	1
Metallo	15611	633	127	88
Serine	18812	880	112	152
Threonine	1744	64	21	22
Unknown	1560	22		
Grand Total	48970	2325	346	383
Inhibitors	4745	559		100

Table 2.1. Total Numbers for Catalytic Types. Numbers for the different types of peptidases and peptidase inhibitors in the MEROPS database, release 7.60. There are a total of 184 peptidase families and 50 peptidase clans and 52 inhibitor families and 33 inhibitor clans.

The three useful methods of grouping peptidases currently implemented in the *MEROPS* database are:

- by the chemical mechanism of catalysis
- by the details of the reaction catalysed
- by molecular structure and homology

Peptidases can be grouped by the chemical mechanism of catalysis and can be described as of serine, cysteine, threonine, aspartic, glutamic, or metallo catalytic type. The names of clans and families in *MEROPS* are built on the letters S, C, T, A, G, M and U (unknown) that stand for the catalytic type. One advantage of this classification system is that every serine peptidase contains a serine residue that acts as the nucleophile at the heart of the catalytic site, and as a result many are affected by generic inhibitors of serine peptidases.

Peptidases can also be grouped by the reaction type they catalyse. Different peptidases catalyse the hydrolysis of different peptide bonds, showing selectivity for the bonds they will hydrolyse. One form of selectivity is for a bond at a particular position in the polypeptide chain of the substrate molecule, on this basis they can be classified into groups such as endopeptidases, omegapeptidases, exopeptidases, aminopeptidases, carboxypeptidases, dipeptidyl-peptidases, tripeptidyl-peptidases, peptidyl-dipeptidases and dipeptidases. They provide an essential part of the description of the activity of any peptidase since the classification of enzymes in the EC list, the Nomenclature Committee of IUBMB, is classified by the reaction catalysed.

Peptidases can be grouped by molecular structure and homology. This is the newest method for catalysing peptidases since it depends on the availability of data for amino acid sequences and three-dimensional structures in larger quantities than previously held. Rawlings and Barrett (1993) described a system in which individual peptidases were assigned to families and the families were grouped into clans. This scheme was developed to provide the structure of the *MEROPS* database and has been extended to include the proteins that inhibit peptidases (Rawlings *et al*, 2004).

The *MEROPS* database classifies its contents using hierarchical, structure-based schemes. Each peptidase is assigned to a Family (each family is built around a 'type example') on the basis of statistically significant similarities in amino acid sequence to at least one other member of the family, and the relationship exists in the peptidase unit that is most directly responsible for catalytic activity. This is necessary because some peptidases are chimeric proteins and thus could potentially contain a catalytic domain related to another in a different family. The peptidase unit is generally a contiguous sequence of about 200 amino acids. Some families are divided into subfamilies because there is evidence of a very ancient divergence within the family. Families are grouped together into a clan if they are thought to have evolutionary diverged from a single peptidase origin, having similar tertiary folds, but diverged so far that their relationship can no longer be proved through the comparison of their primary structures. The peptidases in a clan have similar protein folds and it contains the whole of an evolutionary tree: the peptidases in a clan seem to be unrelated to those in any other clan. A non-redundant library of protein sequences of the peptidase units and inhibitor units of all the peptidases and peptidase inhibitors that are included in *MEROPS* was available for download. This "pepunit.lib" file formed the main peptidase dataset for the alignment chapter (Chapter 4). *MEROPS* version 7.30 released on the 22/12/2005 was the current version at the time of use.

2.2 STRUCTURAL DATABASES

The structural databases represented "gold standards" for this thesis and provided high resolution structures to be used in the modelling protocol.

2.2.1 PISCES

PISCES (Wang & Dunbrack, 2003) is a protein culling server and can be found at http://dunbrack.fccc.edu/PISCES.php. Some important features of this service are as follows:

- Sequence identities for PDB sequences are determined using PSI-BLAST against a non-redundant database (3 rounds, -h value of 0.0001) to create a position-specific substitution matrix which is then used to search the PDB to obtain alignments of all PDB sequences against the rest of the PDB.
- PISCES' alignments are therefore local, so that two proteins that share a common domain with sequence identity above a threshold will not be included in the output lists.
- PISCES can also provide meaningful results at low sequence identity (15-30%).
- PDB sequences, experiment type (X-ray, NMR and so on), resolutions and R-factors are obtained from the PDB's Data Uniformity Site. These fields have been curated by the RCSB to establish uniform representation of all structure data from the 1000s of legacy files from the Brookhaven PDB.
- Non-PDB sequences are culled with sequence identities from PSI-BLAST.

CullPDB and PDBaanr were obtained from PISCES.

2.2.1.1 PDBaanr

Only non-redundant sequences across all PDB files have unique entries in PDBaanr (Wang & Dunbrack, 2003), and the redundant chain identities from all other PDB files are added at the end of the title of the representative chain entries. Representative chains are selected based on the highest resolution structure available and then the best R-values. Non-X-ray structures are considered after X-ray structures. This set represents the whole of the PDB with known structures (possible templates). I n this project the query sequences (targets) would be used to search against this PDBaanr dataset. The PDBaanr was used instead of the entire PDB database to reduce redundancy and provide good quality structures.

2.2.1.2 cullPDB

CullPDB (Wang & Dunbrack, 2003) is a subset of the PDBaanr dataset and contains sequences sharing less than 20% sequence identity to another sequence within the subset, sequences that have below 2.0Å resolution (if it has been determined by X-Ray Crystallography), and sequences that have a R-factor below 0.25. This set represents the possible targets. In a real world situation they would be of unknown structure, but when training they are required to have known structures. CullPDB uses BLAST to produce alignments so the sequence identities are based on local alignments. Culled lists that pass the sequence percentage identity cut-off are sorted according to resolution from best to worst. Structures with the same resolution are then sorted according to R-factor.

2.2.2 ASTRAL Compendium

ASTRAL (Chandonia *et al*, 2002) is partially derived from, and augments the SCOP database of protein domains. ASTRAL contains structures derived using X-ray crystallography, excluding peptides. It provides subsets of selected representative domains created using different thresholds and measures of similarity. To choose the highest quality representatives for these subsets, Summary PDB ASTRAL Check Index (SPACI) scores are used to provide a first order guide to resolution, R-factor and stereochemical accuracy of each determined structure.

Version 1.67 of the ASTRAL SCOP genetic domain sequences was used for the Chapter 4 (the Alignment Chapter) and provided a representation of all PDB structures.

2.3 SEQUENCE DATABASE SEARCHING

To acquire target-template pairs to use in chapter 3 and chapter 4/5, BLAST and PSI-BLAST were implemented. These sequence-based searching methods were also used to provide sequence alignments for Chapter 4 and Chapter 5.

2.3.1 BLAST

Due to the vast amount of information contained within the different nucleic acid and protein databases a fast, qualitative and sensitive method to search and locate homologous sequences is needed. BLAST (Altschul *et al* 1990, 1997) is one of the most common and widely used algorithms to complete this search. BLAST is a heuristic (uses shortcuts to perform the search faster and the solution is not always the optimal answer) that attempts to optimise a specific similarity measure and performs local alignments, providing a service which searches a database for sequences similar to a submitted query sequence. It checks each entry in the database independently against a query sequence (pair-wise sequence searching based on the Smith-Waterman local alignment algorithm (Smith & Waterman, 1981)). It looks for well-matching local regions, then using a substitution matrix and allowing no gaps (initially no gaps were introduced), it identifies short matching contiguous regions between the database and the query sequence, which are extended as much as possible. BLAST uses statistical theory to produce a bit score and expectation value (E-value) for each alignment pair. The bit score gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced. A key element in this calculation is the 'substitution matrix', which assigns a score for aligning any possible pair of residues. The BLOSUM62 matrix is the default for most BLAST programs. Bit scores are normalised. The E-value (equation 2.1) gives an indication of the statistical significance of a given pair-wise alignment and reflects the size of the database and the scoring system used. The lower the E-value, the more significant the hit is.

$$\boxed{E = Kmn\, e^{-\lambda S}}$$

Equation 2.1. E-value. This would be the E-value for the score S of the local alignment in BLAST with sequence lengths m and n. The statistics of high scoring pairs are characterised by two parameters, K and lambda, which can be thought of as natural scales for the search space size and the scoring system respectively (Altschul *et al.,* 1990).

2.3.2 PSI-BLAST

In many cases PSI-BLAST (Altschul *et al*, 1997) is much more sensitive to weak but biologically relevant sequence similarities than the standard BLAST algorithm. Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (and more if required) BLAST search and the results of each 'iteration' used to refine the profile. This iterative searching strategy results in increased sensitivity.

2.4 SEQUENCE CLUSTERING WITH BLASTCLUST

In order to remove some of the redundancy within the selected databases used, the BLASTCLUST (Altschul *et al*, 1997) program was obtained from the main BLAST suite of programs and implemented locally.

BLASTCLUST automatically and systematically clusters protein or DNA sequences based on pair-wise matches found using the BLAST algorithm. BLASTCLUST finds pairs of sequences that have statistically significant matches and clusters them using single-linkage clustering (which puts a sequence into a cluster if the sequence is a neighbour to at least one sequence in the cluster). BLASTCLUST uses the default values for BLAST: matrix BLOSUM62; gap opening penalty 11; gap extension penalty 1; no low-complexity filtering. Taking in a database formatted in FASTA style, it outputs a cluster of sequence identities, sorted from the largest to the smallest cluster, and are sorted from the longest to the shortest sequence within each cluster.

2.5 ASSIGNING SECONDARY STRUCTURE USING DSSP

SSTRUC (Smith & Thornton, 1989) is an implementation of the DSSP (Kabsch & Sander, 1983**b**) method designed to replace DSSP and is used for determination of amino acid secondary structure. DSSP assigns the secondary structure to a protein sequence (by looking at the hydrogen bonding patterns indicative to secondary structure states), geometrical features and solvent exposure are given in the atomic coordinates in PDB format, in 8 states:

H	-	Alpha Helix
B	-	Isolated Beta Bridge
E	-	Extended Strand
G	-	3_{10} helix
I	-	Π helix
T	-	Hydrogen Bonded Turn
S	-	Bend
C	-	Coil

This can be converted into 3 states for ease of comparison. The CASP states are: H (H G I), E (E B) and C (C T S).

SSTRUC gives the secondary structure assignment, which is seen as the correct assignment since it is taken from the PDB files of the experimentally determined structures. The resulting output consists of: main-chain hydrogen bonding; phi, psi and chi angles; disulphide bonds and C_α-C_α distances.

2.6 PREDICTING SECONDARY STRUCTURE

The protein secondary structure prediction programs used in Chapter 3 are JPRED (Cuff *et al.*, 1998) and ELEPHANT (Wilson *et al*, 2004) .

2.6.1 JPRED

JPRED (http://www.compbio.dundee.ac.uk/~www-jpred/) is a web server that takes as input a protein sequence or a multiple alignment of protein sequences and from these predicts the secondary structure using a neural network called JNET. The application of multiple sequence alignment profiles is used to improve protein secondary structure prediction. The prediction is the definition of each residue into either: alpha-helix, beta-sheet, or random coil secondary structures. The server is the result of a large scale comparative analysis of secondary structure prediction algorithms (Cuff *et al*, 1998). More information on JPRED can be found in Chapter 3.

2.6.2 ELEPHANT

ELEPHANT (Wilson *et al*, 2004) has been designed to work primarily on secondary structures predicted by JPRED or PHD. It uses empirical and database derived data to improve the prediction of the beginning of alpha-helical regions. The start positions of all predicted helices are analysed by considering alternative start positions up to four positions either side of the original predicted start position. Each potential start position is represented as the sum of the pseudo-free energies (known as the cumulative pseudo-free energy, CpFE) for residues that correspond to the N-cap, N1, N2 and N3 residues (where N1 is the first residue in a helical conformation) as well as positional weighting term. Pseudo-free energies are derived from global propensities by transforming them using the Boltzmann equation. Again, more information can be found in Chapter 3.

2.7 ASSIGNING FOLDS USING SCOP

The SCOP (Lo Conte *et al*, 2002) database provides a manually curated set of domains from all PDB entries, classified in a hierarchy indicating different levels of structural and evolutionary relationships between the domains. A domain is an evolutionary unit, in the sense that it is either observed in isolation in nature, or in more than one context in multi-domain proteins. A SCOP domain may include fragments from different PDB chains. In most cases this appears to be the product of a single gene. Protein domains in SCOP (Table 2.2) are grouped into species and hierarchically classified into families, superfamilies, folds, and classes. Families have clear evolutionary relationships, superfamilies have common evolutionary origin and folds share majority structural similarity (Suhrer *et al*, 2007).

Class	Number of Folds	Number of Superfamilies	Number of Families
All alpha proteins	179	299	480
All beta proteins	126	248	462
Alpha and beta proteins (a/b)	121	199	542
Alpha and beta proteins (a+b)	234	349	567
Multi-domain proteins	38	38	53
Membrane and cell surface proteins	36	66	73
Small proteins	66	95	150
Total	800	1294	2327

Table 2.2. Total Numbers for the SCOP Database. Numbers for the classes in the SCOP database, release 1.65, August 2003. There are a total of 20,619 PDB entries and 54,745 domains.

Another major structural database is that of CATH (Orengo *et al.*, 1997). One of the main differences between CATH and SCOP is in the way they define the domain units. In CATH all of the classification is done on individual domains, whereas in SCOP sometimes it is organised by individual domains and other times by whole multi-domain proteins.

2.8 ACCESSIBILITY CALCULATIONS WITH NACCESS

NACCESS (S.J. Hubbard, personal communication, University of Manchester, Bioinformatics group) calculates the residue accessible surface defined by rolling a probe of given size (usually the size of a water molcule) around a van der Waals surface. The residue accessibility is simply the summed atomic accessible surface areas over each protein residue. A PDB structure file is submitted to NACCESS. The average over a seven residue sliding window was calculated and assigned to the fourth position of that window. The first and last 3 residues in the structure will not have residue accessibility values. A sliding window of seven residues was used to allow the comparison of the accessibility of the residues to the RMSD values calculated in the project which also used a seven residue window. Using a sliding window also enabled edge structures to be distinguished from buried structures.

2.9 ASSIGNING INTERFACE RESIDUES WITH DACCESS

DACCESS (S.J. Hubbard, personal communication, University of Manchester, Bioinformatics group) compares the PDB file without the inhibitor bound to the PDB file with the inhibitor bound and any changes noted must be due to the binding of the inhibitor. DACCESS first calculates the residue accessible surface area, as in NACCESS, and then calculates the differential residue accessible surface area between multiple chains in a PDB protein structure file (in this case the peptidase chain and the inhibitor chain).

2.10 SHANNON'S ENTROPY

The concept of Shannon's Entropy (Shannon, 1948) comes from the field of Information theory. It measures the degree of uncertainty that exists in a system. In the case of multiple alignments, the Shannon entropy of each protein site can be computed according to equation 2.1. If a column is completely conserved then the Shannon entropy is 0.

$$\boxed{H_n(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i.}$$

Equation 2.2. Shannon's Entropy. Where P*i* is the frequency of the amino acid *i* in that site. *n* is 20 (the total number of possible amino acids).

2.11 CORRELATION COEFFICIENT

The correlation coefficient is a measure of the correlation of two sets of variables *X* and *Y* (X and Y can be arrays of values), measured on the same object, that is, a measure of the tendency of the variables to increase or decrease together. Pearson's correlation coefficient (Pearson, 1896) is usually signified by *r* (rho), and can take on the values from -1.0 to 1.0. Where -1.0 is a perfect negative (inverse) correlation, 0.0 is no correlation, and 1.0 is a perfect positive correlation.

$$r=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{N})(\sum Y^2-\frac{(\sum Y)^2}{N})}}$$

Equation 2.3. Pearson's Correlation Coefficient. N is the number of variables.

2.12 BUILDING PROFILES USING THE HMMER SUITE

Profile hidden Markov models (Krough *et al*, 1994) are statistical models of the primary structure consensus of a sequence family. They use position-specific scores for amino acids (or nucleotides) and position-specific scores for opening and extending an insertion or deletion. Traditional pair-wise alignment methods (for example, BLAST (Altschul *et al*, 1990)) use position-independent scoring parameters. This property of profiles captures important information about the degree of conservation at various positions in the multiple sequence alignment, and the varying degree to which gaps and insertions are permitted.

HMMER (Eddy, 2001) is a freely distributable implementation of profile hidden Markov models (HMM) software for protein sequence analysis. One common use of HMMER is to search a sequence database for homologues of a protein family of interest which requires a multiple sequence alignment.

To construct a HMM using the HMMER package a profile needs to be built from the multiple sequence alignment using the hmmbuild program. The HMM is then calibrated using the program hmmcalibrate. This increases the sensitivity of the database search and returns an E-value. It is possible to search the sequence database for new homologues with hmmsearch giving a ranked list of the best scoring sequences, a list of the best scoring domains and alignments for these domains. Hmmalign enables one or multiple sequences to be aligned to a HMM. HMMER, version 2.3, was used in Chapter 4.

HMMER does not do local (Smith/Waterman) and global (Needleman/Wunsch) style alignments in the same way that most computational biology analysis programs do it. To HMMER, whether local or global alignments are allowed is part of the model, rather than being accomplished by running a different algorithm. By default, hmmbuild builds models which allow alignments that are global with respect to the HMM, local with respect to the sequence, and allows multiple domains to hit per sequence. Such models will only find complete domains.

Another use of profile HMMs is to create multiple sequence alignments of large numbers of sequences. A profile HMM can be built from a "seed" alignment of a small number of representative sequences, and this profile HMM can be used to efficiently align any number of additional sequences. This is in fact how the PFAM (Sonnhammer *et al*, 1997) database is updated as the main SPTREMBL database increases in size. The PFAM seed alignments are (relatively) stable from release to release; PFAM full alignments are created automatically by searching SPTREMBL with the seed model and aligning all the significant hits into a multiple alignment using hmmalign.

2.13 ALIGNING HMMs AND PROFILES

As part of the alignment investigation in Chapter 4 a profile-profile, sequence-profile and multiple sequence-profile method was used.

2.13.1 The Profile-Profile Method

This program produces a substitution matrix between two protein HMMs to be used in a pair-wise alignment. The matrix consists of Pearson's correlation of all against all residue positions. The result was a profile-based sequence alignment between the target-template pair.

2.13.2 HMMER

The hmmalign algorithm described above in section 2.12 was to align a sequence (the target) to the profile of the template previously built using the HMMER package.

2.13.3 COACH

COACH (Comparison Of Alignments by Constructing HMMs; Edgar & Sjolander, 2004) is one algorithm that the Lobster (http://www.drive5.com/lobster/; Edgar, 2004) software implements for analysing protein sequences. It is used to align two multiple sequence alignments to each other. This alignment produces a score that can be used as a relatedness measure. The basic ideal behind the method is to construct a profile HMM from one alignment and align the other multiple sequence alignment to that HMM.

2.14 SEQUENCE BASED ALIGNMENT METHODS

The sequence-based alignment methods were used as input for the investigation into alignment quality. They provided a target-template sequence alignment to be compared against the profile-based alignment methods. CLUSTALW (Thompson *et al*, 1994did not provide alignments for Chapter 4 as the other methods do but was only used for alignments in Chapter 3 only.

2.14.1 MUSCLE

MUSCLE (MUltiple Sequence Comparison by Log-Expectation; Edgar, 2004) creates multiple alignments of protein sequences. Following guide tree construction, the fundamental step is pair-wise profile alignment, which is used first for progressive alignment then refinement. MUSCLE was found to be more accurate than CLUSTALW (Thompson *et al*, 1994) whilst being the fastest of the tested methods, and achieves accuracy statistically indistinguishable from T-COFFEE (Edgar, 2004). MUSCLE uses a *k*mer distance (a contiguous subsequence of length *k*), where related sequences tend to have more *k*mers in common than expected by chance. The MUSCLE software, source code and test data are freely available at: http://www.drive5.com/muscle.

2.14.2 CLUSTALW

CLUSTALW (Thompson *et al*, 1994) is a general purpose multiple alignment program for DNA or proteins. All sequences are compared to each other, then a dendrogram is constructed describing the approximate groupings of the sequences by similarity and finally the multiple alignment is carried out, using the dendrogram as a guide. CLUSTALW was used in Chapter 3.

2.14.3 BLAST

BLAST (Altschul *et al* 1990, 1997), as discussed above in section 2.3.1, finds local sequence similarities, which might lead to evolutionary clues about the structure and/or function of the query sequence. It produces a local sequence-based alignment, using the target as the query sequence, from which the target-template pair can be extracted. BLAST permits a trade-off between speed and sensitivity and in many cases is not as sensitive to weak but biologically relevant sequence similarities (Altschul *et al*, 1997).

2.14.4 PSI-BLAST

PSI-BLAST (Altschul *et al*, 1997) is a sequence-profile method that aligns a query sequence to a profile, generating an alignment to be used in Chapter 4. PSI-BLAST has been acknowledged as one of the most powerful tools for detecting remote evolutionary relationships by sequence considerations alone. Several iterations were used based on the fact that several iterations obtain better alignments, with higher sensitivity and no significant negative effect on the specificity (Friedberg *et al*, 2000). PSI-BLAST is described in more detail in section 2.3.2.

2.15 STRUCTURE BASED ALIGNMENT METHODS

The structural alignment based methods provided gold standard alignments (alignments considered correct since they were based on the experimentally determined structures) to test the other sequence/profile based methods against in Chapter 4.

2.15.1 MAMMOTH

MAMMOTH (MAtching Molecular Models Obtained from THeory; Ortiz *et al*, 2002) is a method for sequence-independent structural alignment that allows comparison of modelled or experimental structures. MAMMOTH is sequence-independent, focusing on model Cα coordinates and avoiding references to sequence or contact maps. MAMMOTH is a heuristic method that computes the optimal similarity of the local backbone chain to establish residue correspondences between residues in both structures, it then computes the largest subset of residues found within a given distance threshold in Cartesian space. When the objective is the comparison of modelled structures with their experimental counterparts MAMMOTH can be an important tool due to its speed, insensitivity to differences in length, and rigorous evaluation score, particularly in those cases where partial or low–resolution models are of interest (Ortiz *et al*, 2002).

2.15.2 TM-align

TM-align (Zhang & Skolnick, 2005) is an algorithm developed to identify the best structural alignment between protein pairs that combines the TM-score rotation matrix (a rotation matrix which contains the best superposition of two structures as calculated using the TM-score) and dynamic programming. The TM-score weights the residue pairs at smaller distances relatively stronger than those at larger distances. Therefore, the TM-score is more sensitive to the global topology than the local structural variations and the value of the TM-score is normalised in a way that the score magnitude relative to random structures is not dependent on the protein's size. The algorithm's initial alignment is obtained by aligning the secondary structures of two proteins using dynamic programming, then, another type of initial alignment is generated based on the gapless matching of two structures and finally an initial alignment is obtained by dynamic programming using a gap-opening penalty where the score matrix is a half/half combination of the secondary structure score and the distance score selected in the gapless matching initial alignment.

2.15.3 CE

CE (Combinatorial Extension; Shindyalov, & Bourne, 1998) builds an alignment between two protein structures, obtaining an accurate three-dimensional structure alignment, including cases with low structure homology. The algorithm involves a combinatorial extension of an alignment path defined by alignment fragment pairs. Alignment fragment pairs are pairs of fragments, one from each protein, which confer structural similarity and are based on local geometry. Combinations of alignment fragment pairs that represent possible continuous alignment paths are selectively extended or discarded thereby leading to a single optimal alignment (Shindyalov & Bourne, 1998). In CE, the score is measured by the intra-structural distance of eight-residue fragments, and the alignment is built by gradually adding new eight-residue fragments to the existing alignment path (Zhang & Skolnick, 2005). Shorter alignments can be obtained with lower RMSD, or longer alignments with higher RMSD. It is also possible to take into account sequence information during the dynamic programming step (Shindyalov & Bourne, 2001).

2.16 ASSESSING THE SEQUENCE ALIGNMENTS USING NiRMSD

Armougom (2006) propose a new type of RMSD (part of the T-Coffee package (Notredame *et al*, 2000)), independent from any structure superposition and suitable for evaluating sequence alignments of proteins with known structures. They suggest that replacing the reference alignments (the gold standards) with an RMSD measure would be a more objective way to evaluate the sequence alignments of proteins, rather than setting one specific alignment as a reference. They imply that this method has two advantages of over standard methods: no dependence on a reference alignment and the possibility to quantify the structural correctness of any protein sequence alignments (provided the protein structures are known). The drawback of using the RMSD is the reliance on a structure superposition strategy: it offers many alternative solutions whose relative merits are difficult to estimate. Armougom *et al* redesigned the RMSD to make it independent from any structure superposition procedure, the iRMSD measure, which is based on intra-molecular distance comparisons. The iRMSD evaluates alignments for their compatibility with the structural superposition they imply. The iRMSD is not suitable for comparing alternative alignments, as it tends to give a better score to alignments with long gaps and few aligned residues. To take into account the superposition accuracy and the extent of the alignment the iRMSD is normalised. The NiRMSD measures the difference in distances between every aligned pair of residues against every other aligned pair of residues and measures the average RMSD of these differences of distances, thus the lower the NiRMSD the better the alignment.

2.17 COMPARATIVE MODELLING WITH MODELLER

MODELLER (Sanchez & Sali, 1997) is a comparative modelling method designed to find the most probable structure for a sequence given its alignment with related structures.

MODELLER implements an automated approach to comparative protein structure modelling by satisfaction of spatial restraints (figure 2.1). The core modelling procedure begins with an alignment of the sequence to be modelled (the target) with related known three-dimensional structures (the templates). This alignment is usually the input to the program. The output is a three-dimensional model for the target sequence containing all main-chain and side-chain non-hydrogen atoms. Given an alignment the model is obtained without any user intervention. First, many distance and dihedral angle restraints on the target sequence are calculated from its alignment with template three-dimensional structures. The form of these restraints was obtained from a statistical analysis of the relationships between many pairs of homologous structures. These relationships were expressed as conditional probability density functions (pdf's) and can be used directly as spatial restraints. The pdfs restrain C^{α}- C^{α} distances, main-chain N-O distances, main-chain and side-chain dihedral angles. The three-dimensional model of a protein is obtained by optimisation of the molecular pdf such that the model violates the input restraints as little as possible. The molecular pdf is derived as a combination of pdfs restraining individual spatial features of the whole molecule. An important feature of the method is that the spatial restraints are obtained empirically, from a database of protein structure alignments (Sali & Blundell, 1993). The spatial restraints include (i) homology-derived restraints on the distances and dihedral angles in the target sequence, extracted from its alignment with the template structures, (ii) stereochemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics force-field, (iii) statistical preferences for dihedral angles and non-bonded inter-atomic distances, obtained from a representative set of known structures, and (iv) optional manually curated restraints, such as those from NMR spectroscopy, rules of secondary structure packing (Jacobson & Sali, 2004). Next, the spatial restraints and CHARMM energy terms enforcing proper stereochemistry are combined into an objective function. Finally, the model is obtained by optimising the objective function in Cartesian space. The optimisation is carried out by the use of the variable target function method employing methods of conjugate gradients and molecular dynamics with simulated annealing. In this approach, the optimisation starts from a random initial conformation and initially uses only the sequentially local restraints. It then proceeds in a number of steps to increase the number of restraints, until, finally, all the restraints are included and their violations minimised. Several slightly different models can be calculated by varying the initial structure (Sali & Blundell, 1993).

More information on comparative modelling and MODELLER can be found in the Introduction chapter.

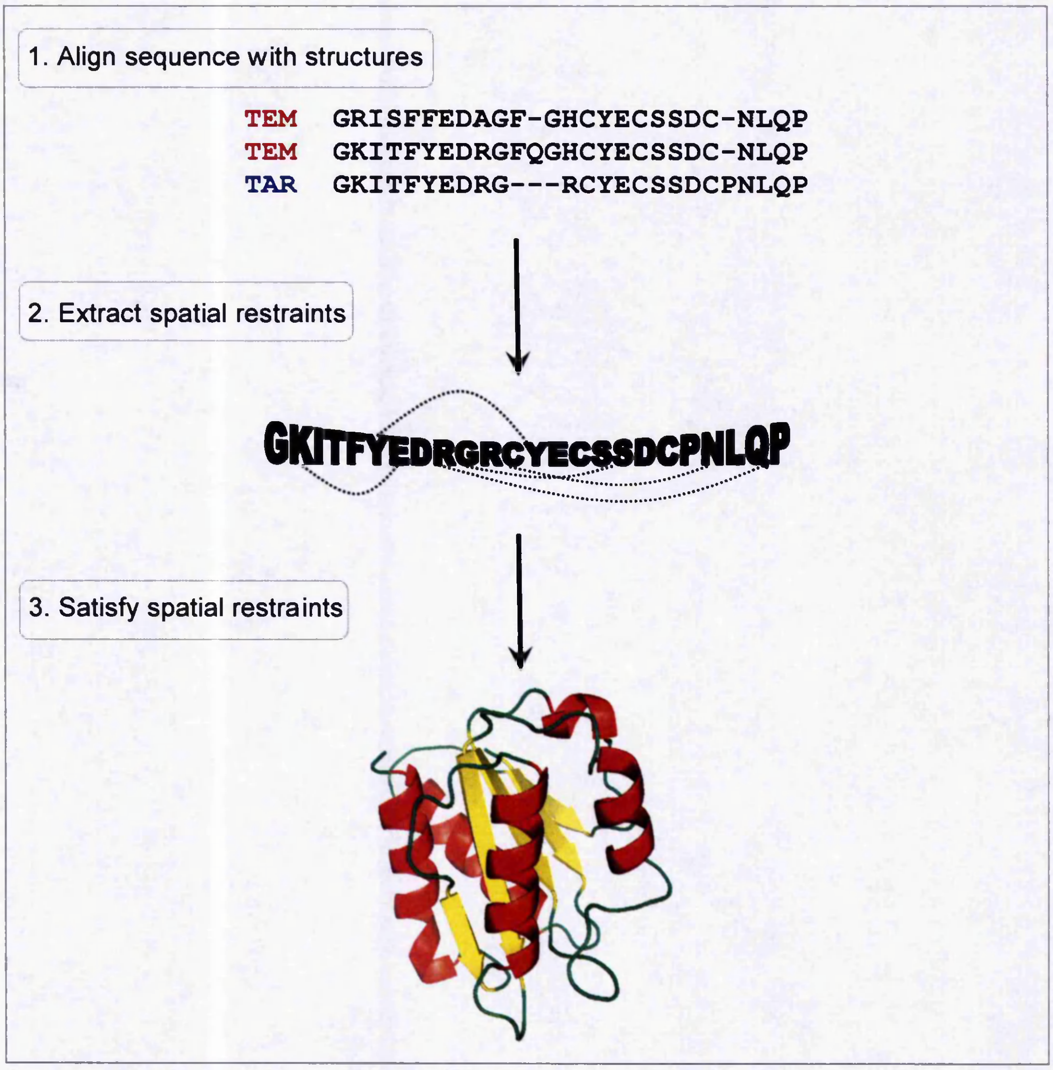

1. Align sequence with structures
TEM GRISFFEDAGF-GHCYECSSDC-NLQP
TEM GKITFYEDRGFQGHCYECSSDC-NLQP
TAR GKITFYEDRG---RCYECSSDCPNLQP
2. Extract spatial restraints
GKITFYEDRGRCYECSSDCPNLQP
3. Satisfy spatial restraints

Figure 2.1. Comparative Protein Modelling by Satisfaction of Spatial Restraints. First, the known template (TEM) structures are aligned with the target (TAR) sequence to be modelled. Second, spatial features, such as Cα-Cα distances, hydrogen bonds, and main-chain and side-chain dihedral angles, are transferred to the target from the template, thus a number of spatial restraints on its structure are obtained. Third, the three-dimensional model is obtained by satisfying all the restraints as well as possible (image reproduced from Sanchez & Sali, 2000).

2.18 STRUCTURAL SUPERIMPOSITION WITH STALIN

STALIN (S.J. Hubbard, personal communication, University of Manchester, Bioinformatics group) is a structural alignment program that aligns and superimposes two similar protein structures. The two structures are taken as input in the PDB format, and one PDB file is the resulting output. STALIN calculates an optimal alignment with the two proteins provided, and so it is considered as a global alignment method and therefore has no need for an alignment as an input. The alignment is done by least squares fitting of short segments of the backbone of each protein against all possible similarly sized segments of the other protein to obtain a two-dimensional matrix of root mean square deviations. The program then finds the maximum pathway through the matrix (dynamic programming) and then superimposes one protein on the other based on this match. STALIN was not used as a structural gold standard since it was only used to quickly align two similar proteins and would struggle when aligning proteins in and around the twilight zone. STALIN works by simply fitting short windows of backbone structure in one structure to all other windows of the same size in the other. This generates a matrix of RMSDs and STALIN then applies standard Needleman and Wunsch dynamic programming to find the optimal path (and hence alignment) through the matrix. Gap open and extension penalties can be applied in the usual way, although in this case a penalty of 0 (for both GapOpen and GapExtend) was used to allow large insertions. STALIN was applied only for superposition of models of the same sequences onto the original structure. As this is a relatively naive structural alignment program, it was deemed adequate for this task and gave good superpositions when viewed by the eye. Since this tool was already available in the research lab, and gave good results, we did not seek an alternative. Likewise, as it was unpublished and arguably inferior to tools such as SSAP and VAST as a structural alignment tool, its performance was not evaluated in the alignment chapter. We could have used PROFIT which also seems to work in a similar way, although unlike STALIN, it does not appear to offer the ability to produce simple structural alignments. STALIN is quite robust to small deletions and insertions, whereas PROFIT needs to be told which residues to fit together - STALIN can work this out independently.

3. SECONDARY STRUCTURE PREDICTION AND COMPARATIVE MODELLING

3.1 AIM

This study focuses on improving an aspect of protein structure prediction, namely comparative modelling of protein structure from its amino acid sequence. In particular, the focus is on model building when the target-template sequence identity falls close to, or in to, the twilight zone. The use of secondary structure prediction was evaluated to try and improve the models, especially the N-termini of alpha helices. Models were built with and without secondary structure restraints to see how much, if any, improvement in the accuracy of the final models was seen. The introduction of this chapter includes background information on secondary structure prediction and how it may be potentially useful to improve the comparative modelling process.

3.2 INTRODUCTION

As discussed in Chapter 1 (the Introduction), comparative modelling is seen as arguably the most reliable structure prediction technique when the sequence of a known structure is closely related to the target sequence of unknown structure. It is also accepted that as the percentage sequence identity between this target-template pair decreases, the ability of the comparative modelling protocol to produce accurate models also decreases (Sali *et al.*, 2001). One of the steps in comparative modelling which has the potential to be improved is the model building step. This is the step where the known structure provides a template to guide the prediction of the target structure, and the more structural information available for the target, the better the prediction. This is where secondary structure prediction is potentially useful.

3.2.1 Secondary Structure

Secondary structure prediction is essentially a one-dimensional prediction of the conformational state of an amino acid in a protein sequence, classifying each residue into one of three states, helix (H), extended (E) or coil

(C). After the primary amino acid sequence of a protein the first level of protein architecture is the secondary structure. The formation of regular secondary structure elements within the protein molecule solves the problem of exposing main-chain polar groups into a hydrophobic environment by the formation of hydrogen bonds. The secondary structure elements, formed this way and held together by the hydrophobic core, provide a rigid and stable framework, exhibiting relatively little flexibility with respect to each other. Such secondary structure is usually of two types: alpha he lices or beta sheets (Brandon & Tooze, 1999).

3.2.1.1 The Alpha Helix

The alpha helix (see figure 1.1b in the Introduction) is the classic element of protein structure and was first described by Linus Pauling in 1951. Alpha helices have 3.6 residues per turn with hydrogen bonds between the C'=O of residue *n* and NH of residue *n*+4, joining all NH and C'O groups with hydrogen bonds, except the first NH and the last C'O groups at the ends of the helix. This results in the ends of helices usually being polar and almost always on the surface of protein molecules (Brandon & Tooze, 1999). Many alpha helices present a hydrophilic face to the external aqueous solvent, and, on the opposite side, a hydrophobic face to the interior (Lesk, 2001).

3.2.1.2 The Beta Sheet

The second major structural element found in globular proteins is the beta sheet (see figure 1.1b in the Introduction). The alpha helix is built up from one continuous region of the polypeptide chain whereas the beta sheet is built up from several discontinuous, independent regions of the polypeptide chain, known as beta strands (Brandon & Tooze, 1999). Hydrogen bonds form between the carbonyl C'=O groups of one strand and amide NH groups on an adjacent strand. Parallel or anti-parallel beta sheets can be formed, forming a pleated sheet. Parallel beta sheets are formed by strands which run in the same direction and anti-parallel from successive strands which have alternating directions. Beta strands occupy a broadly linear three dimensional path, so that the main chain backbone is broadly in an extended conformation; hence, the secondary structure state for both parallel and anti-parallel beta strands is usually classed as “extended” (E).

3.3 SECONDARY STRUCTURE PREDICTION

The prediction of the three-dimensional structure of a protein from its amino acid sequence is one of the oldest, yet still one of the most important, problems in structural bioinformatics. As a first step to solving this problem, many algorithms for predicting the local secondary structure, instead of the full global tertiary structure, have been developed (Lee, 2006).

3.3.1 Its Importance in Protein Structure Prediction

The prediction of the secondary structure of a protein from its amino acid sequence remains a key element in many different approaches to tackle the protein folding problem and to bridge the sequence-structure gap (Schulz, 1978). It is often used to provide constraints for comparative modelling or can be used as a starting point for fold recognition (Rost, 1997). Accurate secondary structure information provides a useful baseline for fold recognition (McGuffin, 2001). Indeed, in CASP6 predicted secondary structure information was an integral part of the best performing schemes for the comparative modelling, fold recognition and new fold predictions tasks (Karypis, 2006). Over the years after its introduction in the late 1960s (Guzzo, 1965), secondary structure prediction has gone through generations of improvements in accuracy and algorithms, combining some of the state of the art techniques available at the time.

3.3.2 The Early Methods

Attempts were first made to predict secondary structure more than four decades ago. These early methods were based on either simple stereochemical principles (Lim, 1974) or statistics (Chou & Fasman, 1974; Garnier *et al*, 1978). Secondary structure prediction methods can be classified into one of the following categories accordingly (Ouali & King, 2000):

- Simple, linear statistics based either on residue or physicochemical properties or even both;
- Nearest neighbour approaches;

- Machine learning and methods employing complex, non-linear statistics, including the application of neural networks and hidden Markov models.

3.3.3 *Ab Initio* or Linear Statistic Methods

This type of method predicts the secondary structure based on a single query sequence. It measures the relative propensity of each amino acid belonging to a certain secondary structure element (Xiong, 2006).

One of the most widely recognised early simple linear statistic methods is that of Chou and Fasman (Chou & Fasman, 1978). The predictions are based on differences in residue composition for three states of secondary structure: alpha helix, beta strand and turn (Higgins & Taylor, 2001). It determines the propensity or intrinsic tendency of each residue to be in the helix, strand and beta turn conformation based on the observations from the crystal structures. The propensity scores are derived from known crystal structures and a score of less than one indicates that the residue has less chance of being found (for example) in helices. The disadvantages to this method includes the statistics being used are naïve and the prediction rules being somewhat arbitrary (Sternberg, 1996).

Another early method was that of Lim. Lim's (Lim, 1974) method was stereochemically orientated, relying on conserved hydrophobic patterns and thus a set of stereochemical prediction rules being developed for alpha helices and beta sheets based on their packaging as observed in globular proteins. The actual prediction rules developed are quite complicated but computer implementations now also exist (Sternberg, 1996).

The GOR (Garnier-Osguthorpe-Robson; Garnier *et al*, 1978) method has been particular popular due to the simplicity of implementing the method in software (Jones, 1999). GOR considers the statistics of flanking residues and the conformational state of a selected amino acid to be predicted. The GOR method has been shown to be more accurate than Chou-Fasman because it takes the neighbouring effect of residues into consideration (Xiong, 2006). A more recent GOR method is GOR V (Sen et al., 2006) which uses the evolutionary information provided by multiple sequence alignments.

These early *ab initio* methods were developed in the 1970s when protein structural data was very limited. The statistics derived from the limited data sets can therefore be rather inaccurate and the predictions solely rely on local sequence information and fail to take into account long range interactions; this limits the prediction accuracy to around 50%, with random predictions being around 40%. However, the methods are simple enough that they are often used to illustrate the basics of secondary structure prediction (Xiong, 2006). (had to remove this sentence as I couldn't find the original citation)

3.3.4 Nearest Neighbour Methods

In nearest neighbour methods the secondary structure of a new primary sequence is classified to be the same as that of the closest primary sequence to it of known secondary structure, based on the idea that similar primary sequences will adopt similar secondary structures (Sternberg, 1996).

Yi and Lander (Yi & Lander, 1993) use substitution matrices and neural network methods to incorporate environmental factors and produce a probability distribution over the three secondary structure states. They exploit the underlying structural similarity between segments of different proteins to aid in the prediction of secondary structure resulting in a peak prediction accuracy of 68%. The NNSSP (Salamov & Solovyev, 1995) method adopts the nearest neighbour approach of Yi and Lander for single sequences using multiple alignments (Higgins & Taylor, 2001).

The PREDATOR algorithm (Frishman & Argos, 1996) incorporates long-range interactions for beta strand prediction and employs propensities for all types of helices, improving the nearest neighbour method.

Methods of predicting protein secondary structure have improved substantially in the 1990s through the use of machine learning methods (Pollastri *et al*, 2002).

3.3.5 Hidden Markov Models and Machine Learning Methods

Within the machine learning category are neural networks. These are used to generalise rules that relate protein primary structure to its secondary structure and then apply these rules to predict secondary structure. Neural networks are considered a popular method for predicting secondary structure; the network is trained on known examples and predicts the secondary structure for a central residue in a window of specific size. They have been applied to a variety of pattern recognition, classification and decision problems. One of the first simple neural network based method was NNPred (Kneller *et al.*, 1990). A summary of a few key neural network and hidden Markov models prediction methods are described.This is by no means an exhaustive list of techniques but seeks to include some of prediction protocols that are more commonly used or were influential in the development of secondary structure prediction.

PSIPRED (Jones, 1999) is a neural network method that can be used to predict protein secondary structure based on the position specific scoring matrices generated by Psi-Blast. The prediction method is split into three stages: generation of a sequence profile, prediction of initial secondary structure, and finally the filtering of the predicted structure. According to the results of CASP3 the PSIPRED method is deemed to be superior to other methods, including PHD. PSIPRED has an upper accuracy limit of 78%.

SSpro (Pollastri *et al*, 2002) uses an ensemble of neural network architectures, Psi-Blast derived profiles, and a large non redundant training set to derive SSpro, a program for secondary structure classification into three categories with an accuracy reaching about 78%.

PHD (Rost, 1996) includes a machine learning technique to compensate for the well known composition biases of large low-similarity databases (Baldi *et al*, 1999). It is a web-based program that combines neural network with multiple sequence alignment. It takes into account flanking residues and makes final filtering by deleting extremely short helices and converting them into coils (Xiong, 1999).

Hidden Markov models use probabilistic models based on the probability of the preceding residue being in a particular type of secondary structure (Krogh *et al*, 1994). Hidden Markov models have been incorporated into secondary structure prediction due to their strong statistical background (Martin *et al.*, 2006).

HMMSTR (Bystroff, *et al*, 2000) builds a library of local stretches of residues with basic structure motifs and then assembles these local motifs (common to all protein families) through hidden Markov models, introducing structural context on the level of super-secondary structure (Rost, 2001).

YASSPP (Karypis, 2006) is a more recent method which incorporates SVMs (support vector machines) and claims to be more accurate than other widely used schemes such as PSIPRED and SSpro, achieving up to 78% accuracy. YASSPP uses an input that combines both position-specific and non-position-specific information and captures the sequence conservation signals around the local window of each residue.

It is thought that combining predictors usually improves prediction accuracy (Pollastri *et al*, 2002). JPred (Cuff *et al.*, 1998) is able to combine its predictions.

3.3.6 Consensus and Hybrid Approaches

Combinations of independent prediction methods seem to yield levels of accuracy higher than that of the single best method. However, for every protein, one method tends to be clearly superior to the combined prediction. It seems that choosing the combination of methods is not trivial; indeed, using inferior methods decreases the accuracy over the best methods, and when to include a method and when not to seems unclear (Rost, 2001).

The JPred server accepts a multiple sequence alignment and predicts the secondary structure of the sequence on top of the alignment. JPred takes, as input, a multiple sequence alignment, a hidden Markov model, PSI-BLAST profiles and position-specific scoring matrices. It runs prediction programs such as: PHD; PREDATOR; Jnet and NNSSP. The query sequence is used to search the databases with PSI-BLAST for three iterations and redundant sequence hits are removed. The resulting sequence homologues are used to build a multiple sequence alignment from which a profile is extracted. The profile information is submitted to the six prediction programs. If sufficient methods predict an identical secondary structure for a given alignment position then that is the structure taken. If there is no majority in agreement in the prediction outputs, the PHD prediction is taken (Xiong, 2006). It produces predictions that are over 76% accurate.

Sen and colleagues (Sen *et al*, 2006) proposed a novel hybrid algorithm called Consensus Data Mining (CDM) that combines their two previous methods. These were Fragment Database Mining, which exploited PDB structures, and GOR V, which is based on information theory, Bayesian statistics and multiple sequence alignments. In CDM, the target sequence is dissected into smaller fragments that are compared with fragments obtained from related sequences in the PDB. For fragments with a sequence identity above a certain threshold the Fragment Database Mining method is applied for the prediction. The remainder of the fragments are predicted by GOR V. The accuracy of CDM measured by Q_3 (the equation for Q_3 can be found in equation 3.1) which ranges from 67.5% to 93.2% and depends on the availability of known structural fragments with sufficiently high sequence identity.

More recently, Pollastri and colleagues (Pollastri *et al*, 2007) proposed a high-throughput machine learning system for the prediction of protein secondary structure that exploits homology to proteins of known structure in the form of structural frequency profiles extracted from sets of PDB templates. They showed that structural information from templates greatly improved secondary structure and found that for sequence similarity exceeding 30%, secondary structure prediction quality was approximately 90%.

3.3.7 Limitations in Secondary Structure Prediction

The effectiveness of local secondary structure prediction depends upon the extent to which a protein's structure, particularly the secondary structure, is determined by local, short-ranged interactions between residues closely spaced along the backbone, as opposed to non-local or long-ranged tertiary interactions (Crooks & Brenner, 2004).

Kihara (2005) discusses how earlier works mostly relied on the propensities of amino acids for three states of secondary structure and that one of the main reasons for the limitations in prediction accuracy comes from long-range interactions. These distant interactions may override the local sequence propensity of secondary structures, since most of the current methods assign a secondary structure to a window of a local segment and thus usually do not explicitly consider long-range interactions of amino acids. Using sliding windows in the prediction of secondary structure is thought to incorporate medium-range interactions to some limited extent.

The difficulty in considering long-range interactions in the prediction of secondary structure from the primary sequence explains the increased prediction accuracy of the alpha helix compared to the beta sheet, which contains more long-ranged interactions; regardless of different approaches used in prediction engines, beta strands have been predicted with less accuracy (Kazemian *et al*, 2007).

3.3.8 Advances in Secondary Structure Prediction

Although there are many different secondary structure prediction methods available in the literature, their cross-validated prediction accuracy is generally below 80% (Sen *et al*, 2006*)*, yet the improvement of prediction accuracy of the protein secondary structure is deemed essential for further developments of the whole field of protein research (Kazemian et al, 2007).

Major improvements in secondary structure prediction came about when the methods began to include multiple sequence alignments; the predictions improved by 9% compared to single sequence predictions. Multiple sequence alignments include evolutionary information through patterns of sequence variability and locations of insertions and deletions (Jones, 1999). Since major changes occur at the boundaries of secondary structure, this is where most errors are accumulated in contrast to the cores of the structures which can be predicted to quite a high accuracy. Most state-of-the-art methods include multiple sequence alignments to obtain higher accuracies. The main source of information in this approach to secondary structure prediction is obtained by observing that the most conserved regions of a protein sequence are those regions which are either functionally important, and/or buried in the protein core. By clustering the sequences in an aligned family, and assessing the degree of sequence variability observed between very similar pairs, the degree of solvent accessibility of an amino acid can be predicted with reasonable accuracy. Secondary structure can then be predicted by comparing the accessibility patterns generally associated with specific secondary structures when packed against a hydrophobic protein core. The prediction accuracy of methods based on multiple sequence alignments has been found to correlate with the degree of divergence present in the aligned set of sequences. Alignments which incorporate sequences with significantly low sequence similarity to a target protein produce more accurate predictions than those which incorporate sequences which are very closely related to the target (Jones, 1999).

Montgomerie's lab (Montgomerie et al., 2006) state that the accuracy of protein secondary structure prediction could be further improved by including structure (as opposed to sequence) database comparisons as part of the prediction process. They developed a method (PROTEUS) that performs structure-based sequence alignments as part of the prediction process. By mapping the structure of a known homologue onto the query protein's sequence, it is possible to predict at least a portion of that query protein's secondary structure. They find that by using both sequence and structure databases and by exploiting the latest techniques in machine learning it is possible to routinely predict protein secondary structure with an accuracy well above 80%.

The ability to produce profiles that include increasingly remote homologues using Psi-Blast has also contributed to performance improvement (Pollastri *et al*, 2002; Jones, 1999). Pollastri and colleagues (Pollastri *et al*, 2007) later suggest that the use of larger training sets, the use of multiple predictors trained independently, and a more sophisticated machine learning techniques have all added to the slow, but steady improvement of secondary structure prediction methods.

3.3.9 Assessing the Accuracy of Secondary Structure Prediction

Various evaluation measures have been used to assess accuracy of secondary structure prediction, such as the Q_3 score and segment overlap (SOV score).

The Q_3 score (Equation 3.1) is a popular way to measure the accuracy of a secondary structure prediction and is conventionally used to score secondary prediction accuracy (Orengo *et al*, 1999). It is the percentage of helix, strand and coil correctly predicted compared to actual number of helix, strand and coil assigned.

$$\boxed{\mathbf{Q_3} = [(\mathrm{PH} + \mathrm{PE} + \mathrm{PC}) / \mathrm{N}] \times 100\%}$$

Equation 3.1. The Q_3 Score. Where N is the total number of residues predicted and P is the number of correctly predicted residues in state H, E or C. The Q_3 score deals with the prediction accuracy of the whole secondary structure content of proteins regardless of the prediction accuracy of each secondary structure class. This is a global measurement (Kazemain *et al*, 2007).

Segment overlap (SOV, equation 3.2) values attempt to capture segment prediction quality rather than just individual residue-level prediction, and vary from an "ignorance" level of 37% (random protein pairs) to an average 90% level for homologous protein pairs.

$$\boxed{\mathrm{SOV} = \mathrm{SOV}_3 = \frac{1}{\mathrm{N}} \cdot \sum_{\mathrm{S(i)}} \frac{\mathrm{MINOV(S1;S2)} + \mathrm{DELTA(S1;S2)}}{\mathrm{MAXOV(S1;S2)}} \cdot \mathrm{LEN(S1)}}$$

Equation 3.2. The SOV Score. Where N is the total number of residues, the SUM is taken over S all the pairs of segments (S1;S2), S(i) is the number of all the pairs of segments (S1;S2), where S1 and S2 have at least one residue in state i in common, MINOV is the actual overlap, with MAXOV is the extent of the segment. LEN is the length of segment 1. Delta is the accepted variation which assures a ratio of 1.0 where there are only minor deviations at the ends of segments.

The usual hindrance in comparing the accuracy of secondary structure prediction methods is that each group uses different datasets used in the training and testing stages, affecting the accuracy. Indeed, early methods generally tested the prediction accuracy on the sequences the rules were derived from. Fortunately, this was changed when Kabsch and Sander (Kabsch & Sander, 1983) tested all methods on different datasets, resulting in a decreased accuracy of all methods.

CASP (Critical Assessment of Protein Structure) meetings are held every two years. Structure predictions are submitted using proteins of unknown structure and released after being evaluated. JPred is one of the programs that is continually assessed and the methods are tested on new structures in the PDB.

Kazemian's lab (Kazemian *et al*, 2007) introduced an index to evaluate the prediction accuracy of each secondary structure class based on an amino acid index, hoping to lead the groups to enhance the methods more objectively and expose more facts of prediction methods. An amino acid index is a set of numerical values representing any of the different physicochemical and biological properties of amino acids. Kazemian *et al* state that the "expertness" of the secondary structure prediction engines have been studied in three levels: the Q_3 score, a global measurement of the whole secondary structure content of proteins regardless of prediction accuracy of each secondary structure class, the Q_H, Q_E and Q_C criteria which evaluates the prediction accuracy of each secondary structure class (helix, strand and coil) separately, and the third level, which they introduce, evaluates the accuracy based on the amino acid index. They introduce this evaluation method to assess the overall strength or weakness of prediction regarding the type of amino acids, reiterating that beta strands are predicted with less accuracy, finding that despite the different prediction accuracy of different amino acids achieved by a certain engine, the orders of prediction accuracy of amino acids are almost the same in all prediction engines.

3.3.10 Improving Secondary Structure Prediction

Secondary structure prediction methods are superior when long-range effects are minimal, hence are better at predicting helices. Beta strands involve more long-range interactions which are difficult to model. Including multiple sequence alignments mean that areas of poor sequence identity can be identified, which tend to represent loop regions connecting secondary structure elements; conserved positions are more likely to occur in regions of secondary structure with increased variation seen in loops and gaps in coil conformation. This means methods tend to fail when predicting the N-termini of alpha helices (Wilson *et al*, 2002). The study of Wilson and colleagues led to the development of an improved N-termini alpha helix secondary structure prediction method called ELEPHANT (Wilson *et al*, 2004).

It is clear to see why correct identification of true N-termini of helices is important, since this will potentially improve predictions of loops that border the secondary structure. These are a key component for successful fold recognition and modelling, and it is anticipated that improved prediction of the N-termini of the helix would lead to better comparative models. An attempt using the empirical information regarding specific residue preferences at the N-termini of alpha helices has been employed to improve secondary structure prediction *via* the ELEPHANT algorithm (Wilson *et al.,* 2004). The termini of alpha helices show unique structural and energetic properties with distinct preferences for the fringes of the helices allowing this study to improve the prediction of these fringes. Indeed, the accuracy for predicting the N-termini of alpha helices rose from 30% to 36% whilst the overall prediction accuracy (Q_3) remained the same. ELEPHANT calculates the energies over a sliding window of four residues. It finds the most energetically favourable one to be in helical conformation (using the results of the previous study which found certain residues to have a preference for the N-cap of an alpha-helix) and assigns it to be the residue immediately preceding the first helical residue, the N-cap (Wilson *et al*, 2004). It should be noted that prediction accuracies at the fringes of secondary structure remain modest, and well below those obtained for proteins globally.

3.3.11 Protein Secondary Structure Prediction and Comparative Modelling

Secondary structure prediction can be used to provide constraints for comparative modelling, aiding usually in the refinement of the model or to help search for distantly related proteins.

One previous study looked at using secondary structure prediction in comparative modelling with respect to the refinement procedure (Aloy *et al*, 2000). Knowledge-based energy profiles combined with secondary structure prediction were applied to molecular modelling refinement. Aloy *et al* state that the methodology can be used to distinguish regions where comparative modelling may fail or to choose the best conformation when more than one model is considered. Regions miss-modelled in the experiment were detected and modified afterwards according to the secondary structure prediction. They describe an example of a case where comparative modelling by homology fails to predict the secondary structure of a connecting segment region, whilst the predictive methods of secondary structure are more accurate.

Contreras-Moreira and colleagues (Contreras-Moreira *et al*, 2003) include predicted secondary structure assignments for distantly related proteins of unknown structure. This is used to improve the alignments of the proteins. They offer an effective way to exploit the variability of templates and sequence alignments to produce populations of optimised models by artificial selection. Their method simulates artificial genetic selection on a population of single-template models created from different templates and different sequence alignments per template. Fitness for each member of the population is defined as a simple function of solvent accessibility and residue-residue pair potentials on a simplified side-chain representation. This new methods permits the identification of more favourable alignments and tertiary structure conformations.

Using the predicted secondary structure in the modelling process has generally been confined to aiding the search for distantly related proteins, refining the model once built and helping the target-template alignment to be improved. Therefore, using the predicted secondary structure of the target to guide the model building process is an under-exploited technique. It has always been thought previously that taking the actual structure of the template was more likely to provide a better starting point for comparative modelling.

Secondary structure prediction is not normally used with comparative modelling, however, with the program called MODELLER, it can provide restraints when an appropriate template cannot be found. Similarly, MODELLER supports the use of secondary structure constraints in the general modelling process. However, the areas where predicted secondary structure may well be of assistance is in the so-called "twilight zone", where the target and template sequence are relatively poorly conserved. If the predicted secondary structure of the target is to prove useful it will need to be closer in structure to the gold standard (i.e. the *actual* secondary structure of the target) than the assigned secondary structure of the template is to the gold standard (Figure 3.1).

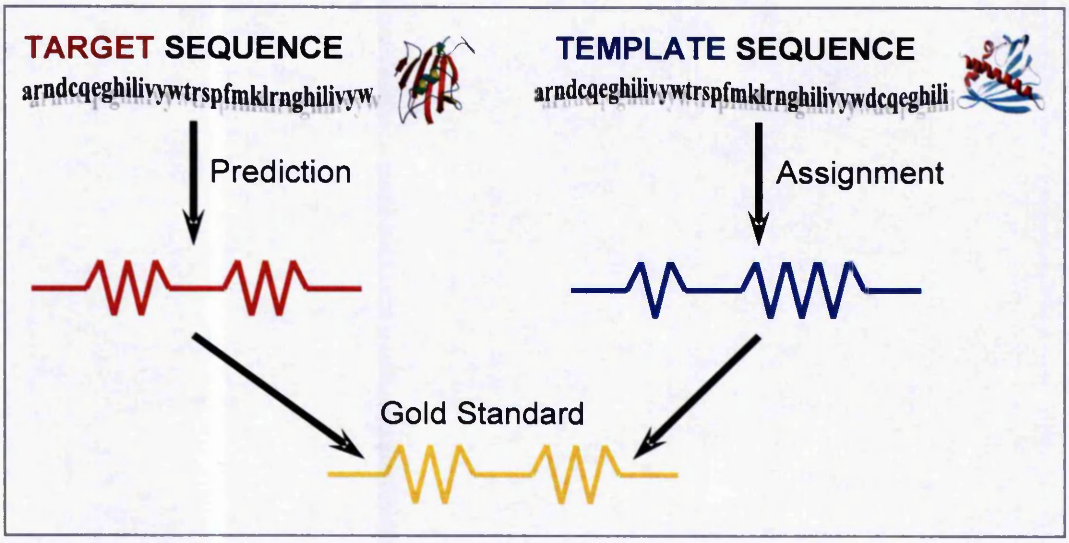

TARGET SEQUENCE
arndcqeghilivyywtrspfmklrnghilivyw
TEMPLATE SEQUENCE
arndcqeghilivyywtrspfmklrnghilivyywdcqeghili
Prediction
Assignment
Gold Standard

Figure 3.1. Prediction and Assignment of Secondary Structure. The predicted secondary structure (shown in red) of the target will need to be closer in structure to the gold standard (displayed in gold) of the target than the secondary structure assignment of the template (the blue structure) for it to be useful in modelling.

3.4 METHODS AND MATERIALS

The first stage in this project was to obtain a dataset of target-template pairs that would provide input for comparative modelling, focusing on pairs with sequence identity stretching into the twilight zone.

3.4.1 Fold Assignment

An initial dataset of template sequences, *pdbaanr,* contained 14,677 protein sequences, representing a non-redundant set of all the PDB sequences available (obtained early in 2005). A subset of the *pdbaanr* dataset, *cullpdb,* was obtained representing potential targets. It is a precompiled dataset from PISCES (please refer to section 2.2.1 of Chapter 2 for an explanation of PISCES) containing sequences sharing less than 20% sequence identity, with structures better than 2.0Å resolution and with R-factors below 0.25.

After removing the structures containing only Cα traces, proteins less than 50 residues in length and proteins less than 3.0Å in resolution (from the *pdbaanr* dataset), the sets were reduced to the following numbers: 12,817 sequences in the *pdbaanr* dataset and *1,448* in the *cullpdb* subset.

To reduce the sets further (making viable sets in which to complete modelling experiments) the *pdbaanr* set was submitted to BLASTCLUST. After applying BLASTCLUST to the *cullpdb* set, it was found that it could not be further reduced, even with the most stringent conditions set (pairs sharing more than 80% sequence identity). The *pdbaanr* dataset was reduced again to a set of 6,202 sequences, clustering similar sequences with 95% or greater pairwise identity to reduce redundancy.

Gen erally, the first step in comparative modelling is fold assignment, which requires finding related structures (templates) for as many domains in the modelled query sequence of unknown structure (target) as possible. A common way of finding homologous protein sequences is to use the BLAST tool (Altschul *et al*, 1997). The *cullpdb* sequences (targets) were searched with BLAST against the *pdbaanr* database (the templates) using an E-value of 0.001, filtering out low complexity regions and resulting in over 2,000,000 hits (total hits between all targets and all templates). Templates containing ligands were removed.

3.4.2 Selecting Templates

To select the most appropriate template(s) for each target, the BLAST hits were clustered into different percentage identity bins, over the 0-100% range in 10% intervals, allowing the effect of percentage sequence identity on improving comparative modelling using secondary structure prediction to be examined. From each bin, the six highest resolution structures were taken, giving potentially sixty templates f or each target (the six highest resolution structures in each of the ten 10% identity bins). Removing redundancy resulted in 1,756 pairs of templates and targets, for a total of 567 targets (not all targets had a template partner, or had one in each 10% bin).

3.4.3 Target-Template Alignment

The optimal alignment between target and template is the one in which the structurally equivalent positions are correctly aligned (most search methods are usually tuned for detecting remote relationships and not for optimal alignment). This means that once the templates have been selected, the target sequence and template structure will have to be realigned using specialised methods such as CLUSTALW, a dynamic programming algorithm, to obtain an alignment. Alignment pairs which did not have above 50% of the target sequence retained were removed (some alignment methods chop off parts of the target sequence thus if more than half of the target was lost in this process the alignment was discarded). In total, 293 protein sequence alignment pairs were left with high quality targets and varying percentage identities between the target and template pairs.

3.4.4 Model Building and Secondary Structure Restraints

Two methods were used to apply the secondary structure restraints of the alpha helices and the beta sheets to the alignments of the 293 pairs. Both techniques were applied to the four different alignments (SST1, SST2, JPred and ELEPHANT) for each target-template alignment pair (see figure 3.2 and figure 3.3, or the abbreviations section for explanations of SST1, SST2 and SST3). These four alignments contained different secondary structure restraints on the target. MODELLER can use restraints calculated from the target's predicted or assigned secondary structure in the hope to build a more accurate model.

```
SEQ:        vywtrspfmklrnghilivywmklrng   TARGET
SST1:       --HHHHH---EEE-E-HHH---HH---   TARGET
JPRED:      ---HHHH----EEEE-HH----HH---   TARGET
ELEPHANT:   --HHHHH----EEEE-HHH---HH---   TARGET
SEQ:        vwrfmklghiiywlnglgilhywwrft   TEMPLATE
SST2:       --HH---EEE-E-HHH-HHH--HH---   TEMPLATE
```

- SST1: the DSSP assignment of the target secondary structure;
- JPred: the prediction of the target secondary structure;
- ELEPHANT: the improved N-termini prediction of the target;
- SST2: no secondary structure restraints used, essentially the secondary structure of the template.

Figure 3.2. Variations of Restraints: JPred, SST1, ELEPHANT and SST2 . The different secondary structure restraints of the target (in green) are shown. The actual secondary structure of the target assigned by DSSP is shown in gold (SST1) and the secondary structure predictions by JPred and ELEPHANT shown in red. The SST2 (also in gold) is the actual secondary structure of the template (blue) which is equivalent to using no secondary structure restraints on the target.

3.4.4.1 Description of SST1

The first method (SST1) contained restraints from the target secondary structure supplied by DSSP (Kabsch & Sander, 1983) run on the target structure (i.e. the actual secondary structure restraints of the target). Normally, this is not known and hence this was used as the control. The third alignment method (JPred) used target restraints obtained from the JPred prediction program (Cuff *et al*, 1993) and the fourth (ELEPHANT) from the ELEPHANT prediction program. To impose the restraints, the starts and stops of every secondary structure element in the target was obtained and included in the MODELLER program.

3.4.4.2 Description of SST2

The second of the four different alignments used no specific secondary structure restraints (SST2). Instead, the actual secondary structure of the template was implicitly used – this is the default behaviour in MODELLER.

3.4.4.3 Description of SST3

Since the inclusion of gaps in the target-template sequence alignment usually results in models with higher RMSDs, a fifth target-template alignment provided a potential control for inserting these gaps into the alignment: SST3 (see figure 3.3 for a an example of SST3). SST2 was just the alignment of the target to the template with no secondary structure restraints applied (the default in MODELLER which will apply these restraints). SST3 contained the explicit secondary structure restraints of the template in the same way as the explicit restraints of the target; this was achieved by 'masking' out the template secondary structures and then reapplying the restraints in MODELLER. This was completed in the same way as the restraints of the target were applied for JPred, SST1 and ELEPHANT (again see figure 3.2 and figure 3.3 for examples and explanations). A summary of SST1-SST3 can be found in the abbreviations.

(a)

```
SEQ:  vywtrspfmklrnghilivy  TARGET
SEQ:  vwrfmklghiiywlnglgil  TEMPLATE
SST2: --HH---EEEEE--HHHHH-  TEMPLATE
```

(b)

```
SEQ:   vy--wtrsp-----fmklrng-----hilivy  TARGET
SEQ:   vwrf--mklghiiy-----wlnglgi-----l  TEMPLATE
SST3:  --HH-----EEEEE-------HHHHH------  TEMPLATE
```

- **SST2:** no secondary structure restraints used, essentially the secondary structure of the template.

- **SST3**: the template containing gaps inserted into it to remove the secondary structure restraints of the template

Figure 3.3. Insertion of Gaps in the Target and Definition of SST2 and SST3. (a) shows the original target-template alignment and the SST2 restraints, (b) shows the resulting alignment when gaps are inserted into the target (and the template to keep the alignment of the target-template residues the same) to 'mask' out the secondary structure restraints of the template, SST3.

3.4.4.4 The Explicit Method

The first technique used to apply secondary structure restraints of the four different target-template alignments (Figure 3.2) for each protein pair included 'masking' out the secondary structure restraints of the template. In order for MODELLER (Sali, 1993) to explicitly use the secondary structure restraints from the target, and not a combination of the target and the template restraints, the residues in the template that contained secondary structure information had to be aligned to gaps in the target (for an example see figure 3.3) and then the secondary structure restraints of the target (either predicted or known) could be applied. Aligning the target residues to gaps in the template removed the secondary structure constraints from the template and allowed the secondary structure restraints of the target to be used in the model building stage. Removing the secondary structure restraints from the template meant that the all of the constraints from the template would be lost, which would probably result in models with lower accuracy.

3.4.4.5 The Combined Method

The second technique used to apply secondary structure restraints to the four different target-template alignments included a combination of the secondary structure restraints from the template and from the target. This was achieved by not explicitly stating the target restraints (that is, no gaps were inserted into the target when the corresponding template residues contained secondary structure information) but by just including the secondary structure information of the target; MODELLER would apply the restraints of both the template and the target. This meant for each target-template protein sequence alignment a total of nine models would be built: five for the explicitly defined target secondary structure restraints set (SST1, JPred, ELEPHANT, SST2 and SST3) and four for the combined set (SST1, JPred, ELEPHANT and SST2). This meant there was a potential for 2637 (9x293) models to be built.

3.4.4.6 Secondary Structure Assignment

The most commonly used secondary structure assignment method, DSSP (Kabsch and Sander, 1983), was used to assign the secondary structure to the targets and the templates. SSTRUC is an implementation of the DSSP method used for the determination of amino acid secondary structure. The eight states from the DSSP assignments were converted to the usual three states; alpha helices (H), beta sheets (E), and coil (C). These are considered the correct assignments for the secondary structure since they were taken from the PDB files of the experimentally determined structures.

3.4.4.7 Secondary Structure Prediction

The secondary structure predictions for the targets were carried out using the JPred (Cuff *et al.*, 1998) server and the improved prediction method ELEPHANT (Wilson *et al*, 2004) . For more information on the prediction protocols please refer to chapter, section 2.6. The JPred predictions of the targets were predicted using the consensus method. Sequences were submitted in batches to the JPred server, which can be found at http://www.compbio.dundee.ac.uk/~www-JPred/. The ELEPHANT target secondary structure predictions were completed using a local version.

3.4.4.8 Understanding how Fold Class Affects the Results

An investigation into the improvement of using the predicted secondary structure of the target rather than using the actual secondary structure of the template in terms of the fold class was completed. The protein targets were split into the secondary structure SCOP classes (all alpha helix, all beta sheet, alpha helix / beta sheet, alpha helix + beta sheet and 'others') and percentage identity (between the target-template pair) bins to reveal how the different fold levels affect the improvement (if at all) in the Q_3 score, at different percentage identities.

3.4.4.9 Assessing the Secondary Structure Predictions

The qualities of the secondary structure predictions were assessed using the Q_3 and Q_N (the N-termini residues predicted correctly) scores to calculate the overall accuracy and the N-terminal accuracy of the secondary structures. Obtaining these scores would reveal how much, if any, the ELEPHANT algorithm improves the quality of the predictions, and whether using the predicted secondary structure of the target might prove useful, specifically at the N-termini of the secondary structures (alpha helices in this case).

3.4.5 Model Evaluation

The RMSDs of the alpha-carbons and main-chains were used as a guide to the model accuracy. As stated before, a highly successful model is usually considered to be one having an overall RMSD value of less than, or equal to, 2Å over the whole protein fold; although, a model with lower accuracy (higher RMSD) can still prove useful. The models built using the predicted secondary structure restraints were compared to the models built using the secondary structure restraints for the template provided by DSSP. This enabled a comparison to be made to investigate whether using the predicted secondary structure of the target in the model building process resulted in more accurate models than using o nly the actual secondary structure of the template (no secondary structure restraints). These results were presented over various percentage sequence identities to find out where, if anywhere, the secondary structure predictions were most useful.

To evaluate the accuracy of the explicit method (the models built with the secondary structure restraints from the template 'masked' out, and the target secondary structure restraints applied) the models built with the secondary structure restraints from the target were compared to the models built with the secondary structure restraints from the template (the explicit way: SST3, see figure 3.3 for an explanation of SST3). They were also compared to the models built with the secondary structure restraints from the template (no restraints 'masked' out or applied). The combined method (the restraints from the target combined with the restraints from the template) was evaluated in the same way, but with only the models with the target restraints being compared to the models with the template restraints (no restraints were applied; essentially just the template). To assess the ability of ELEPHANT to improve the helix N-termini, the residues in a helix conformation (or strand conformation for comparison) were assigned to be in a helical region if more than four helical residues existed in a seven residue window. The average RMSD from these residues was obtained, and a final average 'helical' RMSD was obtained for all of the helical regions in that sequence.

3.5 RESULTS AND DISCUSSION

To assess how much improvement was made when using the predicted secondary structure of the target instead of the actual structure of the template during the model building stage in comparative modelling, the amount of improvement made by the predicted secondary structure over the actual secondary structure was obtained in by observing the differences in the Q scores. The results from the secondary structure prediction method JPred and ELEPHANT were compared to one another and also to the secondary structure assignment results of DSSP to determine how much, if any, the ELEPHANT prediction method had improved the models. The different models were assessed using the RMSD values to determine whether using the secondary structure prediction actually improved the quality of the models.

3.5.1 Assessing the Use of the JPred Algorithm

By calculating the percentage of correctly predicted secondary structure (obtained by JPred) states of the target compared to the correctly assigned states (provided by DSSP) of the template, with respect to the target-template alignment, the results would reveal whether using the predicted secondary structure of the target was beneficial (Figure 3.4). It was noted that the improvement in accuracy increases as the percentage identity of the pair decreases. It is possible to see from graph (a) in figure 3.4 that, on average, using the predicted secondary structure of the target (when the percentage identity is above 50%) results in a lower Q_3 score, and therefore actually makes the "prediction" worse in most cases, than when using the actual secondary structure of the template. This mean that using the assigned secondary structure restraints through alignment is more accurate than using the predicted secondary structure restraints when the alignment is easy (the sequences share high sequence similarity). T he situation is quite different for target-template pairs closer to and beyond the twilight zone in graphs b)-d). Here, there is an increasing proportion of target-template pairs where the predicted target secondary structure is closer to the true, target DSSP assignments than the template. Indeed, graph (d) shows the most improvement when using the predicted secondary structure, where this is upheld for the majority of the pairs.

This is encouraging as a platform to improve the quality of models built at this similarity (below ~40% identity), where using the predicted secondary structure appears to outperform using the template's implicit secondary structure as a basis for modelling. The improvement in the quality of the models could be due to real differences between the secondary structure of the target and the secondary structure of the template in pairs sharing low percentage sequence identity or it may be because the transfer of secondary structure information depends on the quality of the alignment and errors in alignments are more frequent and significant in these pairs. It is difficult to distinguish between these reasons for improvement but it would be interesting to optimise the alignments more before applying the secondary structure restraints to understand the reason for this improvement.

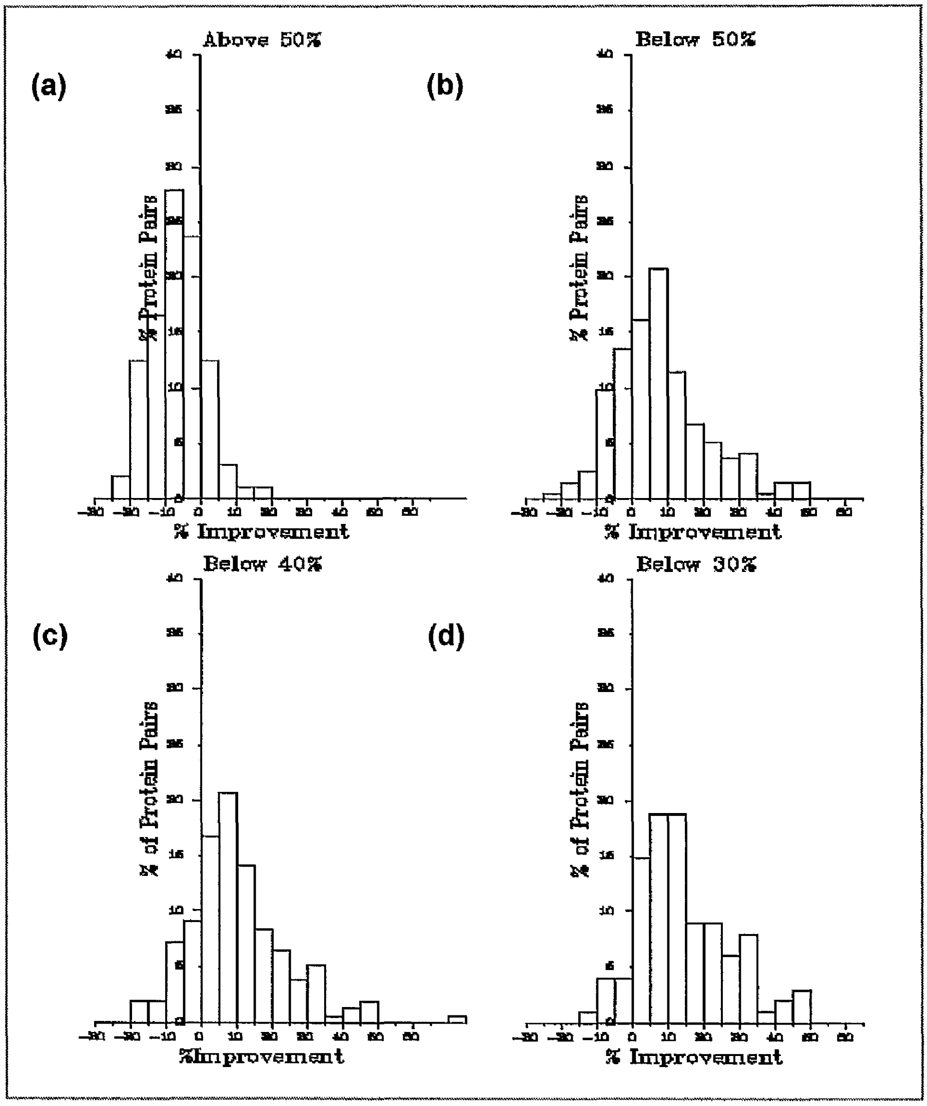


Figure 3.4. How well does Secondary Structure Prediction do? The X-axis shows the percentage improvement in Q_3 made when using the predicted states of the target from JPred compared to using the assigned states of the template (made by DSSP). The predicted JPred secondary structure of the target and the actual secondary structure of the template were compared to the actual (DSSP assigned) secondary structure of the template. The Y-axis is the percentage of protein pairs that this holds true for. These were split up into bins according to the percentage sequence identity between the target and template. The percentage identity bins are inclusive except for graph (a), all target-template pairs above 50% sequence identity. The others correspond to all pairs below 50% in (b), below 40% in (c) and below 30% in (d). Negative numbers indicate pairs where the predicted secondary structure of the target was less accurate than using the actual structure of the template – the norm for pairs above 50%.

It is worth noting that the majority of typical sequence similarities for homologous protein pairings from a simple BLAST search against PDB are in and around the twilight zone (Figure 3.5).

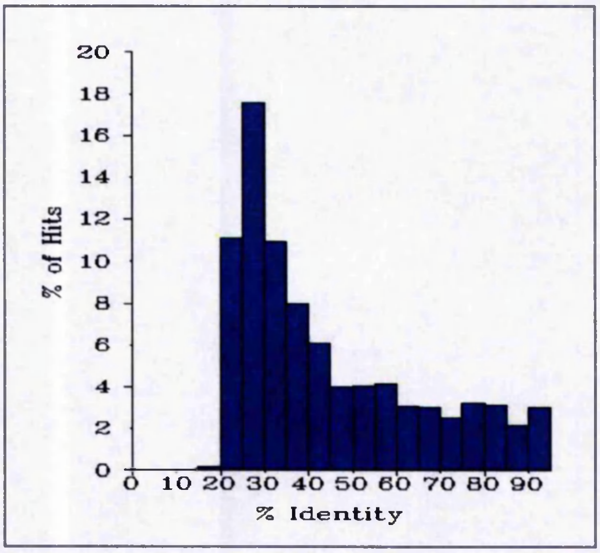


Figure 3.5. Potential Models and Percentage Sequence Identity. The plot shows that most of the models that can be built are in the lower region of percentage sequence identities (between 20-40%). A search of the PDB against itself, removing self-self hits, revealed potential homologues.

3.5.2 Assessing the Use of the Improved Algorithm, ELEPHANT

Once it was established that using the predicted secondary structure of the target was more accurate than using the actual secondary structure of the template, at least for target-template pairs below 40% identity, the improved N-termini prediction program ELEPHANT was evaluated. It was hoped that this could increase the accuracy of the secondary structure prediction of the target further. The hypothesis was that ELEPHANT should improve the Q_N score (the N-termini residues predicted correctly) and the scores of the secondary structure whilst not affecting the Q_3 score (the overall accuracy of the three predicted states - Helix, Strand, Coil). The results can be seen in Table 3.1.

Percentage Identity	Number of Pairs	SST2		JPred		ELEPHANT	
		Q_3	Q_N	Q_3	Q_N	Q_3	Q_N
<=100	293	84.59	58.19	76.71	32.40	76.66	37.28
<=50	196	80.45	49.78	76.18	32.68	76.13	37.68
<=40	162	78.85	46.59	75.72	34.09	75.62	38.76
<=30	114	74.91	41.54	74.96	33.11	74.87	37.98
<=20	56	69.45	31.29	74.68	34.26	74.64	38.77

Table 3.1. Q_3 and Q_N Scores. The resulting average Q_3 and Q_N scores for the DSSP assignments of the secondary structure for the template and for the secondary structure predictions of the target from JPred and ELEPHANT are shown. Comparisons were made to the DSSP assignments of the target structure as the standard of truth. They are grouped into different percentage identity bins which are inclusive.

Table 3.1 displays the Q_3 and the Q_N scores of the assigned, actual secondary structures of the template (SST2) and the predicted secondary structures of the target (JPred and ELEPHANT) when being compared to the actual assigned secondary structures of the target. Overall, the accuracy of all of the methods (more evident with SST2) decreases as the percentage identity between the target and the template decreases; this holds true for the N-cap accuracy (Q_N) as well. The predictions from JPred and ELEPHANT (regarding the N-cap and the overall accuracy) fail to improve upon the actual assignments from SST2 within all of the percentage identity bins (inclusive bins were used due to the relatively low numbers of proteins in each individual bin if the bin was split. Using exclusive bins would have meant fewer protein pairs in each bin making it harder to observe trends), except for below 20% identity. At this low sequence similarity level the predictions are more accurate by around 5%, on average, for the Q_3 scores and around 3%-7% for the Q_N scores. Indeed, the Q_N scores improve by over 7% when the ELEPHANT algorithm was used as the prediction method. It is indeed true that the N-caps of the ELEPHANT predictions were improved on from the JPred predictions, whilst keeping the overall accuracy of the predictions the same. Improving the accuracy of secondary structure prediction at the N-termini of helices using the ELEPHANT algorithm meant that regions at the start of loops should be better modelled, improving the trajectory of the end of the loop, leading to more accurate models being built.

3.5.3 Fold Class and the Predictions

The ELEPHANT predictions showed the most improvement over the DSSP assignments in alpha helices, with below 50% sequence identity to the template (Figure 3.6), and the least improvement was seen in the beta sheet classes. This is partly because beta sheet classes are where most prediction algorithms fail to predict with high accuracy. More obviously, the ELEPHANT algorithm only attempts to improve predictions at the fringes of alpha helices.

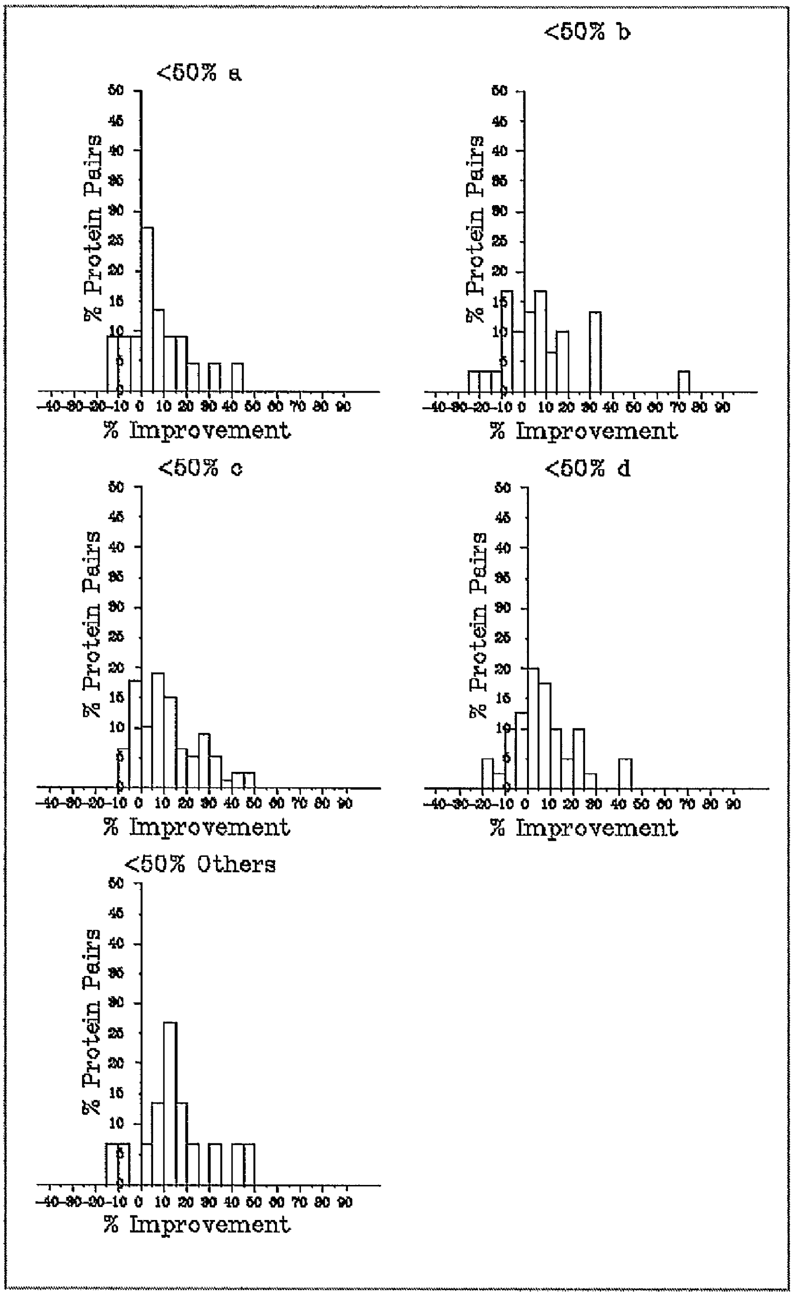


Figure 3.6. The Improvement in Different SCOP Classes. The X-axis is the percentage of improvement made when using the predicted states of the target by ELEPHANT compared to using the assigned states of the template at below 50% sequence identity in the Q_3 score. The Y-axis is the percentage of protein pairs (a total of 293 pairs were used) that this holds true for. 'a' is the all alpha helix class, 'b' is all beta, 'c' is the alpha/beta class, 'd' is alpha+beta and the 'Others' class includes all other folds. Total number of pairs was 193 as the cut-off is below 50%. Inclusive bins were used to smooth the graph results.

3.5.4 Model Evaluation

The explicit method was evaluated by comparing the models built with the explicit secondary structure restraints of the target (SST1, JPred and ELEPHANT) to the models built using the explicit secondary structure restraints of the template (SST3) and the model built using just template sequence (no restraints explicitly assigned). Here, explicit means no other secondary structure restraints were used, that is, those restraints from the template were removed, and then those of the target were applied.

3.5.4.1 The Explicit Method

For the explicit method, five models per target-template pair were built: SST1, JPred, ELEPHANT, SST2 and SST3. The number of models actually built in each of these cases can be seen in table 3.2. No method built the full 293 alignment pairs, however SST2 built the most: 290 models, and SST1 the least: 287 models. For some of the more difficult models (target-template pairs sharing below 20% sequence identity) MODELLER exceeded the maximum number of errors allowed during model building, thus the models did not build.

Percentage Identity	SST1	JPred	ELEPHANT	SST2	SST3
<=100	287	288	288	290	289
<=80	249	250	250	252	251
<=60	214	215	215	217	216
<=40	157	158	158	160	159
<=30	110	111	111	112	112
<=20	53	54	54	54	54

Table 3.2. Number of Pairs Built for the Explicit Method. The total number of successful models built in the percentage identity bins (which are inclusive) using the different secondary structure restraints can be seen.

The average alpha-carbon RMSD over the different percentage identity bins for the different secondary structure restraints is shown in table 3.3.

Percentage Identity	SST1 (Å)	JPred (Å)	ELEPHANT (Å)	SST2 (Å)	SST3 (Å)
<=100	11.17	11.40	11.35	5.71	12.01
<=80	11.60	11.85	11.78	6.15	12.36
<=60	12.17	12.41	12.33	6.82	12.74
<=40	13.32	13.38	13.27	8.08	13.72
<=30	14.36	14.40	14.34	10.05	14.66
<=20	16.07	15.80	15.86	12.68	16.31

Table 3.3. Average RMSDs for the Explicit Method. The average alpha-carbon RMSD over all of the models built using the different secondary structure restraints is displayed in inclusive bins.

The RMSD of the models increases as the percentage identity decreases (Table 3.3), most dramatically in the twilight zone where model building becomes less than trivial. Indeed, it increases from an average of 12.01Å (for the alpha-carbon explicit secondary structure of the template: SST3) in the below 100% sequence identity bin (inclusive bins) to an average of 16.31Å, when the sequence identity between the target and the template drops below 20%. Improvement in the average RMSD is seen in all percentage identity bins when comparing the models built using predicted secondary structure of the target made by ELEPHANT to the models built using the actual secondary structure of the template (SST3). However, when the average RMSD of the models for the explicit method using the template restraints was compared to the average RMSD of the models built using only the template (no restraints applied: SST2), SST2 is more accurate. SST2 achieves an RMSD of 5.71Å for the 100% percentage identity inclusive bin, whereas SST3 achieves an RMSD of 12.01Å in the same percentage identity inclusive bin. The results for the average alpha-carbon helical and sheet RMSD over all of the models built using the different secondary structure restraints can be seen in Appendix 1. Appendix 1 also contains the figures showing the number of pairs which have lower RMSDs than SST3, and a figure showing improvement in RMSD when comparing the resulting models built using the different secondary structure restraints to those models built using the restraints explicitly from the template.

Evidently, the explicit method holds no improvement over just using the template structure because of the numerous gaps introduced when trying to explicitly assign the secondary structure of the target. This leads to the evaluation of the combined method.

3.5.4.2 The Combined Method

In the combined method, four models per target-template pair were built: SST1, JPred, ELEPHANT and SST2. The combined method refers to an approach where the template restraints were combined with either the predicted or actual restraints of the target (see section 3.4.4 for a more detailed explanation of the combined method). The number of models built in each of these cases can be seen in Table 3.4. JPred built the full 293 alignment pairs, with the remaining methods building 292 alignment pairs; an improvement over the number of successful models built with the explicit method.

Percentage Identity	SST1	JPred	ELEPHANT	SST2
<=100	292	293	292	292
<=80	254	255	254	254
<=60	219	220	219	219
<=40	161	162	161	161
<=30	113	114	113	113
<=20	55	56	55	55

Table 3.4. Number of Pairs Built for the Combined Method. The total number of successful models built in the percentage identity inclusive bins using the different secondary structure restraints can be seen.

Decreasing percentage identities of the target-template pairs was coupled with increasing RMSDs (Table 3.5). The average RMSD values for the pairs in the 100% bin increased from 5.69Å to 12.19Å for the SST2 models. SST1 and JPred consistently have lower average RMSDs than SST2. ELEPHANT improves over SST2 when the percentage sequence identity of the pairs drops below 60%. The SST1 results display the potential for improvement by using the actual secondary structure of the target combined with the actual secondary structure from the template, compared to only using the actual restraints from the template (SST2). The limiting factor was due to the introduction of errors (hence higher RMSDs) by the secondary structure prediction methods JPred and ELEPHANT, thus JPred and ELEPHANT do not obtain the same, lower RMSDs that SST1 achieves. However, one must remember that SST1 represents a theoretical "optimal" performance, given that these are the known secondary structure restraints of the true structure.

Percentage Identity	SST1(Å)	JPred(Å)	ELEPHANT(Å)	SST2(Å)
<=100	5.41	5.67	5.75	5.69
<=80	5.85	6.11	6.21	6.17
<=60	6.52	6.77	6.77	6.86
<=40	7.73	7.93	7.93	8.13
<=30	9.24	9.46	9.47	9.78
<=20	11.65	11.72	11.79	12.19

Table 3.5. Average RMSDs (in Ångstroms) for the Combined Method. The average alpha-carbon RMSD over all of the models built using the different secondary structure restraints is displayed in inclusive bins.

To examine the ability of ELEPHANT to improve the modeling of the N-termini of alpha helices, the average RMSD of all of the helical regions in the target structure was calculated and compared to the average RMSD of the helical regions of JPred and SST2 (Table 3.6). ELEPHANT obtains lower average RMSDs than JPred for the helical regions when the percentage sequence identity falls below 30%; as an average it is marginal (0.02Å and 0.03Å) but the improvement of ELEPHANT over SST2 is found within all of the percentage identity inclusive bins. Below 20% sequence identity ELEPHANT obtains a 0.39Å improvement over SST2.

Percentage Identity	SST1(Å)	JPred(Å)	ELEPHANT(Å)	SST2(Å)
<=100	0.48	0.62	0.62	0.71
<=80	0.51	0.66	0.66	0.77
<=60	0.55	0.71	0.71	0.85
<=40	0.64	0.81	0.81	0.99
<=30	0.72	0.90	0.88	1.13
<=20	0.84	1.07	1.04	1.43

Table 3.6. Average Helical RMSDs (in Ångstroms) for the Combined Method. The average alpha-carbon helical RMSD over all of the models built using the different secondary structure restraints is displayed in inclusive bins.

The strand regions are modeled to similar accuracies for JPred and ELEPHANT (Table 3.7) as ELEPHANT does not alter the residues in beta strand conformation. Improvements over using the template secondary structure (SST2) can be seen only when incorporating the actual secondary structure restraints from the target (SST1). This emphasises the limitations of secondary structure prediction programs when predicting beta strand conformations (Levin & Garnier, 1988).

Percentage Identity	SST1	JPred	ELEPHANT	SST2
<=100	0.72	0.80	0.80	0.73
<=80	0.79	0.86	0.86	0.79
<=60	0.82	0.93	0.93	0.88
<=40	0.93	1.07	1.06	1.01
<=30	1.18	1.26	1.25	1.19
<=20	1.51	1.57	1.56	1.52

Table 3.7. Average Strand RMSDs (in ngstroms) for the Combined Method. The average alpha-carbon strand RMSD over all of the models built using the different secondary structure restraints is displayed in inclusive bins.

The number of pairs improved when using the restraints from JPred, ELEPHANT or SST1 (combined with using the restraints from the template) compared to using the restraints from the template alone (SST2) over the different percentage identity inclusive bins can be seen in figure 3.7. The overall number of pairs does vary, however this is only by one or two pairs (figure 3.4). For the global, helical and strand RMSD averages, the number of pairs where an improvement is seen decreases as the percentage identity of the pair decreases, this holds true for all of the predicted secondary structure programs. As the percentage identity decreases, the number of pairs with an improvement in the RMSD when using the JPred or ELEPHANT secondary structure prediction programs, becomes almost equal to the number of improved pairs when using the actual secondary structure of the target (SST1). The number of pairs with lower global RMSDs built using the ELEPHANT secondary structure restraints (combined with the template secondary structure restraints) rather than using the template restraints alone, ranges from around 140 pairs in the 100% identity bin to around 50 pairs in the 20% identity bin. To examine the extent of the improvement, the average improvement in the RMSD of the models was calculated (Figure 3.8).

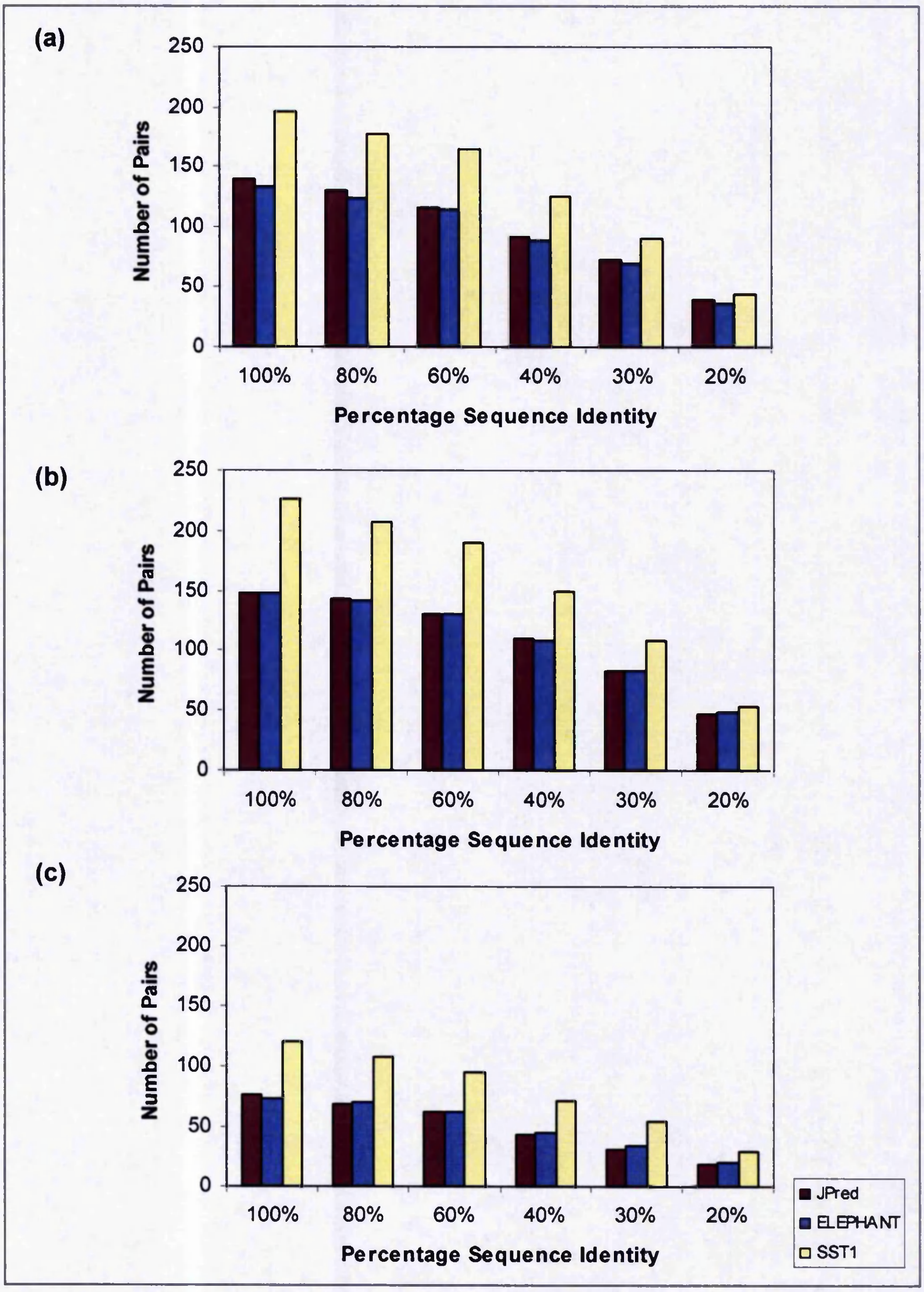


Figure 3.7. The Number of Pairs with Lower RMSDs than SST2. (a) shows the number of pairs which have lower RMSDs (an improvement) than SST2 (the template with the secondary structure restraints obtained from DSSP), (b) displays the results for the regions which are in helical conformation and the graph in (c) shows the results for the regions in beta strand conformation. All RMSDs are between the alpha-carbons of the target and template and reflect the results when using the combined target and template restraints method.

The average improvement in global RMSD (Figure 3.8a) increases as the percentage sequence identity decreases. For ELEPHANT, the RMSD ranges from -0.1Å (a minus indicates an increase in RMSD) to 0.6Å, for JPred it ranges from 0Å to 0.4Å and for SST1, from 0.3Å to 0.7Å. The largest improvement is found in the lowest percentage identity bin; 20%, this holds true for all of the models that have been built from a combination of target secondary structure restraints and template secondary structure restraints. For the helical regions (Figure 3.8b) a similar trend is observed; as the percentage sequence identity decreases, larger improvements are seen. Larger improvements in the helical RMSDs exist in the models built using ELEPHANT restraints, rather than those built using JPred restraints sharing below 40% sequence identity (Figure 3.8b). Although more pairs are improved in the higher percentage identity bins (Figure 3.7), greater improvements are seen in the lower percentage identity bins (Figure 3.8). Only slight improvements are seen for the strand regions when the models have applied the SST1 secondary restraints combined with the template restraints over using the template restraints alone (Figure 3.8c); ELEPHANT and JPred reduce the accuracy of the modelled beta strands in all instances. This may be because ELEPHANT only works on helices so the stands will not be improved by this method, it may also be due to beta strands being more difficult to predict (Ganier & Levin, 1988) and so the prediction methods apply restraints which are not as accurate as using the template actual restraints, even when the target-template sequence identities fall into the twilight zone. An example input for MODELLER can be seen in Appendix 1, Figure A1.1.

Overall, these results d emonstrate that comparative models can be improved by careful and judicial use of secondary structure restraints derived from predicted secondary structure. This can improve the overall quality of models, as well as local regions around the secondary structures themselves. In this case, particularly helical segments were improved. Although this effect is general for all evolutionary distances of the target-template pairs, it is clear to see that those closer to the twilight zone stand to benefit from the most improvement.

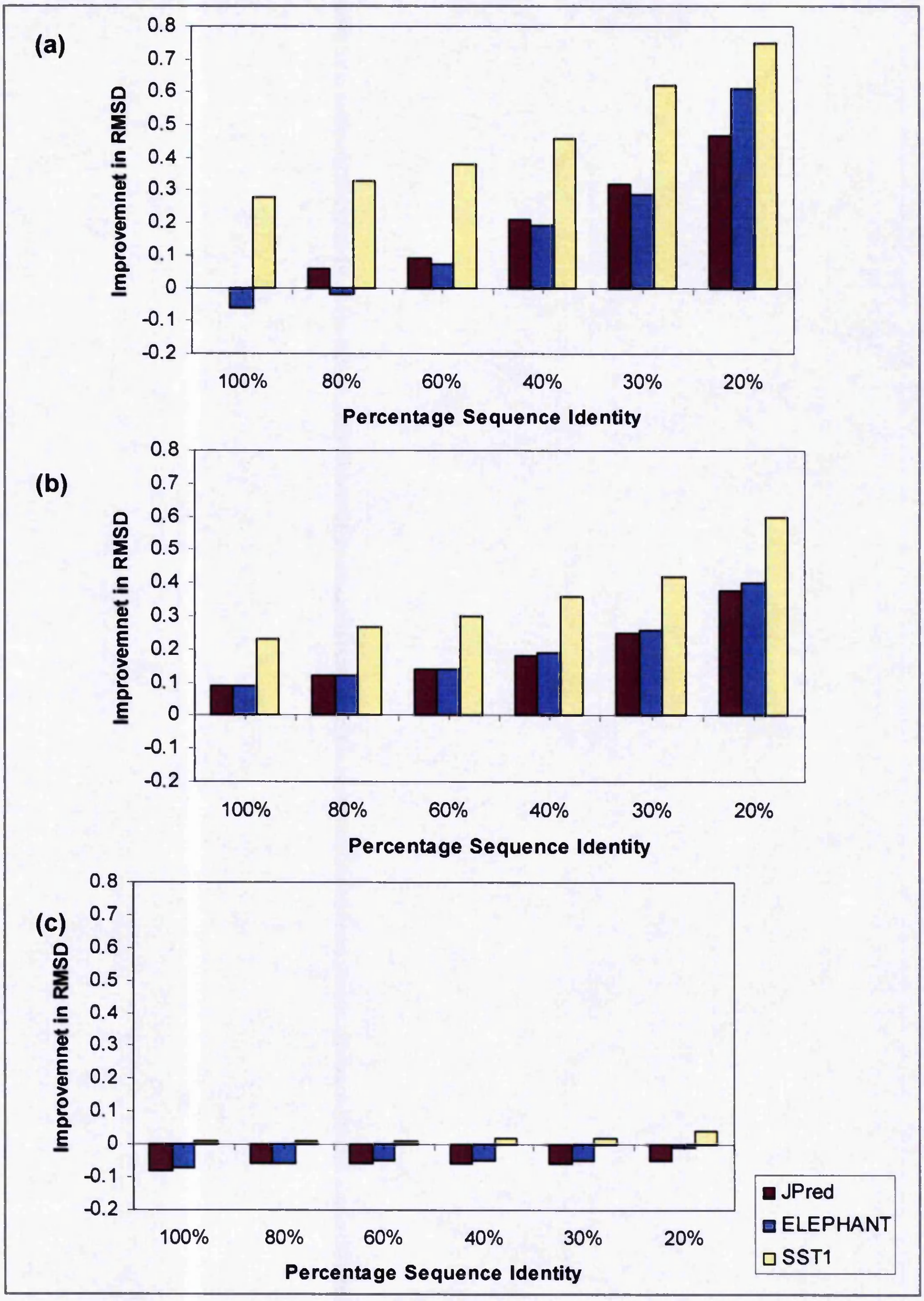


Figure 3.8. The Improvement in RMSD. The graphs show the average improvement in RMSD when comparing the resulting models built using the different secondary structure restraints (combined with the template restraints) to those models built using just the restraints obtained from the template (SST2). (a) shows the average improvement in RMSD between the alpha-carbons of the target and template, (b) displays the RMSD results for the helical regions and (c) shows the average RMSD results for the beta strand regions.

As noted, the use of predicted secondary structure (over using the template structure alone) does have some benefits for structure prediction and some specific examples are presented here. An example of a protein target-template pair (1m2x_A+1dxk_A) is shown in figure 3.9.

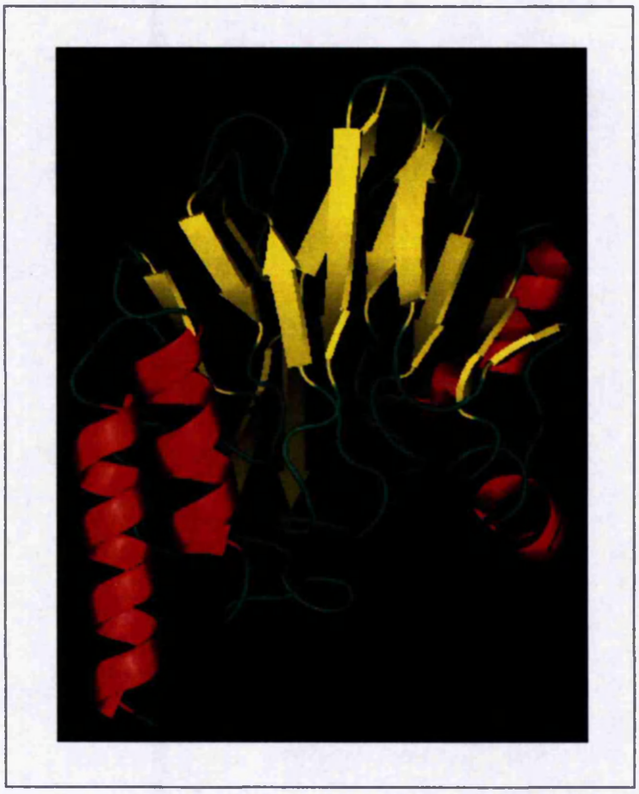


Figure 3.9. 1m2x_A PDB. The structure of the target of 1m2x chain A (taken from the PDB) is shown. Alpha helices are shown in red and beta sheets in yellow. The figure was produced using PyMOL (DeLano 2004).

The protein pair 1m2xA+1dxkA share 35% sequence identity. The resulting model built by ELEPHANT has an alpha-carbon RMSD of 1.62Å. The protein shown in figure 3.9 (1m2xA) is a hydrolase containing a four layer sandwich of 219 residues. The model built using the predicted target secondary structure restraints made by ELEPHANT (using the combined method) was more accurate than the model built using only the secondary restraints from the template. The alpha-carbon RMSD of the model built using the template restraints was 2.30Å compared to 1.62Å when using the target restraints as well.

Including secondary structure prediction in the model building process has managed to improve some of the alpha helical structures (figure 3.10). The model built using the target secondary structure restraints made by ELEPHANT (right hand image in pink) is closer to the actual structure of the target (shown in blue in both images) than the model built just using the template secondary structure restraints (the left hand image, shown in pink).

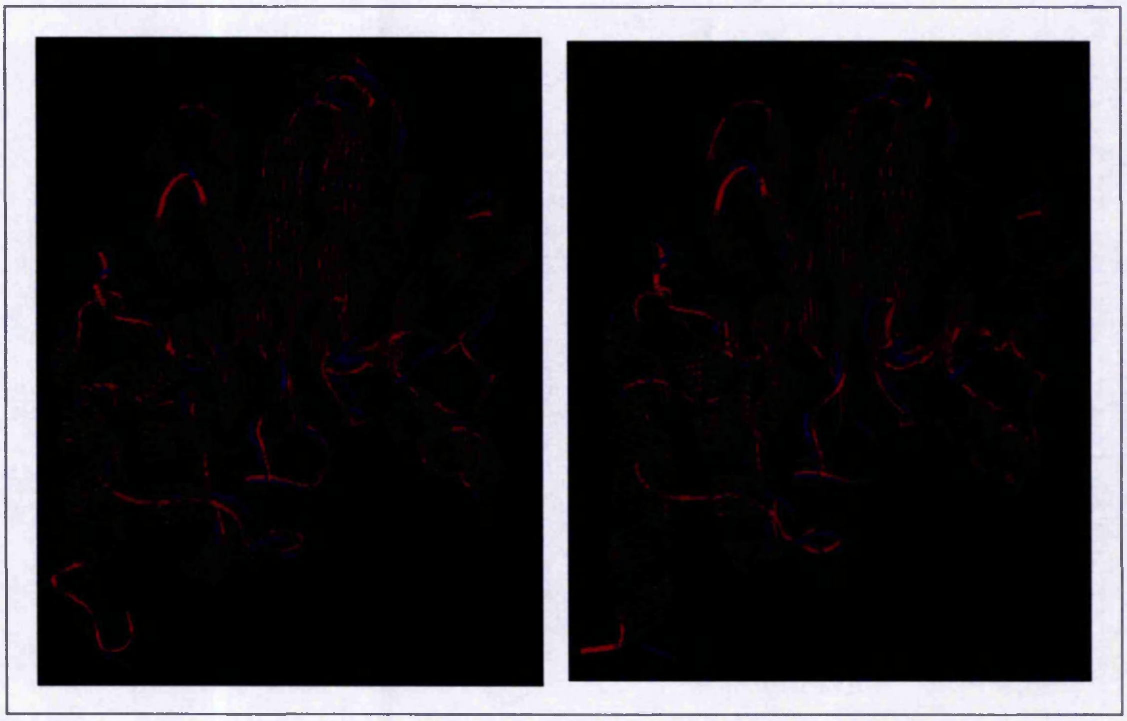


Figure 3.10. 1m2xA Models. The PDB structure of the target of 1m2x, chain A, is shown in blue in both images. The left hand image contains the model built with the template structure (without any secondary structure restraints and shown in pink) superimposed on the actual structure of the target (in blue). The right hand image contains the model built with the secondary structure restraints of ELEPHANT in pink, superimposed onto the PDB structure of the target, again in blue. Notice the helix in the lower left-hand corner of the image is closer in structure to the actual structure in the ELEPHANT image. The N-termini of the alpha-helix which has been improved is the bottom of the helix in the left hand bottom corner of both figures. The image was created using Rasmol (Sayle, 1993).

The helix of the model (target 1m2xA) has been isolated to display the difference in accuracy of the methods more clearly (Figure 3.11). The helix built using the ELEPHANT restraints of the target and the restraints of the template, is closer in structure to the PDB of the target, than the model built using only the restraints of the template.

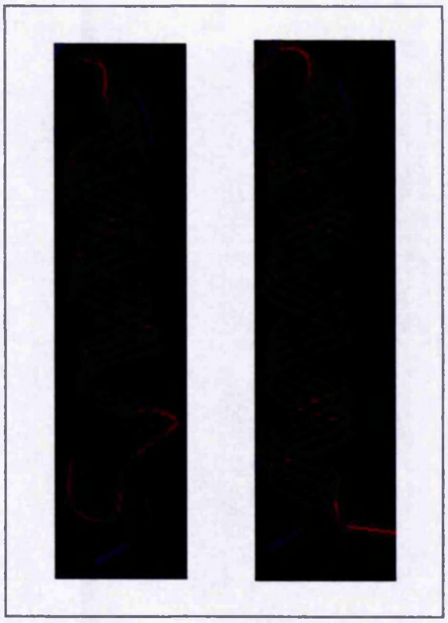


Figure 3.11. 1m2xA Alpha helices. The PDB structure of the target of 1m2x chain A is shown in blue. The left hand image contains the model built with the template structure, in red, (without any secondary structure restraints) superimposed on the actual structure of the target (in blue). The right hand image contains the model built with the secondary structure restraints of ELEPHANT in red, superimposed onto the PDB structure of the target, again in blue. The N-termini of the alpha-helix is at the bottom of this figure.

In some cases, there is an improvement using the predicted secondary structure restraints from ELEPHANT over using the predicted secondary structure restraints from JPred. Figure 3.12 shows one of these examples. 1gy7A shares 39% sequence similarity with the template 1qmaA, and the model built using restraints from JPred achieves an RMSD of 1.61Å, the ELEPHANT model obtains a lower RMSD of 1.41Å.

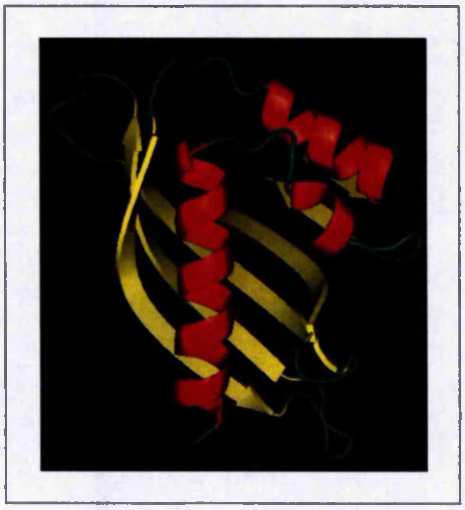


Figure 3.12. 1gy7_A PDB. The structure of the target of 1gy7 chain A (taken from the PDB) is shown. Alpha helices are shown in red and beta sheets in yellow. The figure was produced using PyMOL (DeLano 2004).

Figure 3.13 shows the improvement made when using ELEPHANT over JPred in the model building process, particularly the start of the helix. For visual ease, this helix is isolated in figure 3.14.

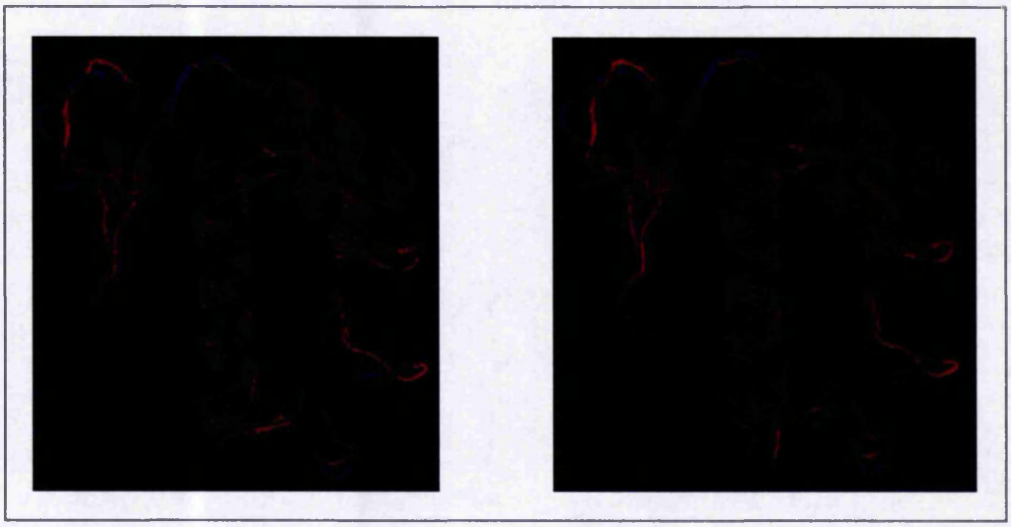


Figure 3.13. 1gy7A Models. The PDB structure of the target of 1gy7 chain A is shown in blue in both images. The left hand image contains the model built with JPred secondary structure restraints shown in pink, which has been superimposed on the actual structure of the target (in blue). The right hand image contains the model built with the secondary structure restraints of ELEPHANT in pink, superimposed onto the PDB structure of the target, again in blue. Notice the N-termini of the centre helix The image was created using Rasmol (Sayle, 1993).

The N-terminus of this particular helix was closer in structure to the target when the secondary structure restraints were obtained from ELEPHANT rather than JPred. This also improved the overall accuracy of the helix. The average RMSD value of the helices in the 1gy7A model when using the JPred restraints was 0.3 Å and for ELEPHANT it was 0.24Å.

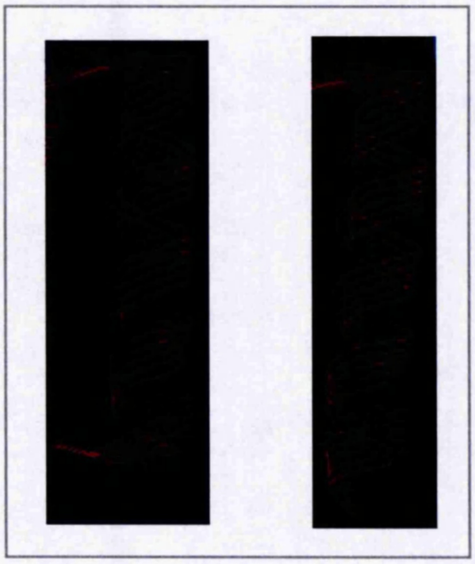


Figure 3.14. 1gy7A Alpha helices. The PDB structure of the target of 1gy7 chain A is shown in blue. The left hand image contains the model built using the restraints of JPred, shown in red, superimposed on the actual structure of the helix from the target PDB (in blue). The right hand image contains the model built with the secondary structure restraints of ELEPHANT in red, superimposed onto the PDB structure of the helix, again in blue. Notice the improvement of the N-termini of the helix in the ELEPHANT model. The N-termini of the alpha-helix is at the bottom of this figure.

Another example where ELEPHANT was found to be an improvement over JPred was 1ik2A (Figure 3.15). 1ik2A and 1a0gA shared a modest 20% sequence identity. The average RMSD for the helix residues in the model produced when using the restraints from JPred was 0.83Å, which was lowered to 0.76Å when using the restraints from ELEPAHNT to build the model.

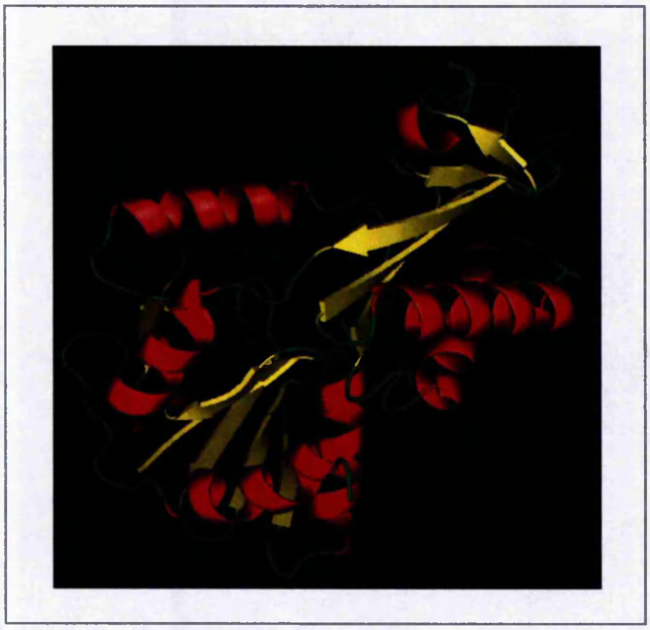


Figure 3.15. 1i2k_A PDB. The structure of the target of 1i2k chain A (taken from the PDB) is shown. Alpha helices are shown in red and beta sheets in yellow. The figure was produced using PyMOL (DeLano 2004).

The models produced by JPred and ELEPHANT superimposed onto the actual structure of the target can be seen in figure 3.16. The improvement in the N-termini of the helix in the top left hand corner can be seen (the model built using the ELEPHANT restraints is closer to the actual structure than the model built using JPred's restraints).

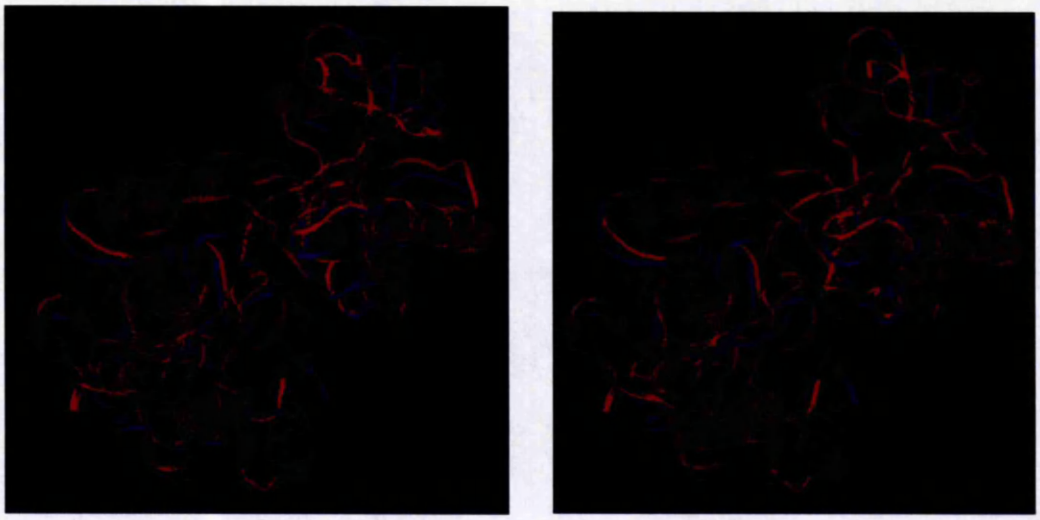


Figure 3.16. 1i2kA Models. The PDB structure of the target of 1ik2 chain A is shown in blue in both images. The left hand image contains the model built with JPred secondary structure restraints shown in pink, which has been superimposed on the actual structure of the target (in blue). The right hand image contains the model built with the secondary structure restraints of ELEPHANT in pink, superimposed onto the PDB structure of the target, again in blue. Notice the N-terminus of the top left helix. The image was created using Rasmol (Sayle, 1993).

The top left hand helix of the 1i2kA structure has been improved when using the restraints from ELEPHANT rather than the restraints from JPred (Figure 3.17).

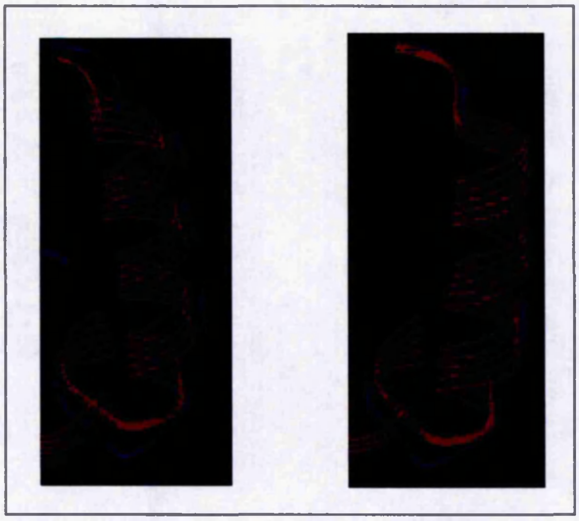


Figure 3.17. 1i2kA Alpha helices. The PDB structure of the target of 1i2k chain A is shown in blue. The left hand image contains the model built using the restraints of JPred, shown in red, superimposed on the actual structure of the helix from the target PDB (in blue). The right hand image contains the model built with the secondary structure restraints of ELEPHANT in red, superimposed onto the PDB structure of the helix, again in blue. Notice the improvement of the N-termini of the helix in the ELEPHANT model. It appears here that the structures have the same N-termini here however ELEPHANT has produced a model where the structure is closer to that of the actual structure and thus is still an improvement.

From the examples it has be shown that using the secondary structure restraints of the target combined with the secondary structure restraints of the template, more accurate models can be built than when using the restraints of the template alone. This has been possible even when the sequence identity of the pairs reaches into the twilight zone. The modelled N-termini of alpha helices was improved when the secondary structure restraints were obtained from the ELEPHANT algorithm rather than from the JPred algorithm.

3.6 CONCLUSIONS AND FUTURE WORK

The aim of this project was to improve structure prediction, more precisely comparative modelling. It is a non-trivial task to produce accurate comparative models when the target-template sequence identity falls below 30%. It was proposed that by using the secondary structure of the target to guide the model building process it could be possible to improve the overall model accuracy. JPred provided secondary structure predictions of the target and ELEPHANT provided secondary structure predictions with improved N-caps. These predictions were assessed using the Q_3 a nd Q_N sco res against the assigned secondary structure of the template provided by DSSP to discern whether using the predicted secondary structure of the target offered more accurate results than using the actual secondary structure of the template. Models were built with the secondary structure restraints obtained from the JPred, ELEPHANT and DSSP algorithms across a range of percentage sequence identities between the target and template.

The results of this project revealed that models built using the predicted secondary structure of the target and the combined secondary structure of the template, were closer to the actual structure of the target than those models built when using the actual secondary structure of the template alone. This held true for a large number of candidate modelling pairs. This worked using the combined method rather than the explicit method. For an explanation of the explicit and combined methods please refer to section 3.4.4.4 and 3.4.4.5. Additionally, ELEPHANT improved some predictions of the N-caps of alpha helices over JPred. Furthermore, the actual niche for this improvement was established; sec ondary structure prediction improved the r egions where the target-template sequence identity was below 30%, this is indeed where comparative modelling tends to struggle at producing accurate models. Using the predicted secondary structure restraints in comparative modelling aimed to improve only the model building step of this process, it may be worth optimising the alignment between the target and template before applying these restraints.

In summary, this study has presented a method to increase the accuracy of comparative protein modelling when the target-template identity falls below 30-40%. It not only decreases the RMSD of the model but also increases the accuracy of the starts of the helices, thus potentially increasing the accuracy of the attendant loops modelled (this has not been shown directly in this project, but would be worth investigating in the future). This improvement is considered to be significant when trying to improve the comparative modelling of any protein which shares below 30% sequence similarity to its template. It is also plausible that this method could optimize the modelling of key loop regions where a misplacement of one residue can have significant detrimental effects on the region being modelled.

The next stage of this project would include learning how and when to ignore the actual secondary structure restraints of the template and just use the predicted restraints of the target. At present, the combined method uses the actual restraints of the template and then uses the predicted restraints additionally. A rather simplistic approach has been applied, but a more intelligent one would seek to find patterns or rules to decide when to use template structural restraints and when to use predicted properties from the target sequence itself.

This general approach, of using target-predicted features to guide comparative modelling, is not widely used in protein structure prediction, and this work suggests novel applications in comparative modelling protocols. Previously it has always been thought better to use the actual structure of the template rather than the predicted secondary structure of the target, and it is hoped that these proof-of-principle results can be fully exploited to improve the modelling process.

4. ALIGNMENT PROTOCOLS

4.1 AIM

The aim of this study was to investigate various alignment techniques which included sequence-sequence based methods, sequence-profile based methods and profile based methods using peptidase sequences and structures. Alignments were assessed as a precursor to comparative model building, hence accuracy was expected to influence the overall quality of the models built. The accuracy of the overall alignment was assessed, as well as the accuracy of the interface regions, in order to extend the investigation into whether these methods can align the interface residues more accurately than they can align the rest of the sequence, and if so, which method can do this best. This chapter covers the introduction of the alignment methods used and a brief introduction on the importance of using peptidases as a test case. It also puts the use of alignments in the comparative modelling process in perspective. A more detailed explanation of comparative modelling can be found in the main introduction section (Chapter 1).

4.2 INTRODUCTION

The growth of experimentally determined protein structures is heavily outpaced by the growth of available protein sequences. One important role in reducing disparity between the volume of sequence and structural information belongs to computational methods. Of these, comparative modelling is usually the method of choice when it comes to structure prediction from protein sequences that are related to known structures (Venclovas & Margelevicius, 2005).

4.2.1 The Importance of a High Quality Model

Acquiring the three-dimensional structure of a protein is an important asset to understanding its biological function and in turn aids drug discovery. Even when comparative modelling produces protein models of only modest accuracy, significant information about the ligand it binds can be extracted, as long as the interface region has been modelled to an acceptable level. This is because the quality of the alignment between target and template sequences is the single most important factor in determining the accuracy of the 3D model (Fiser *et al.*, 2001), it is of substantial interest to develop methods that can both provide highly accurate sequence alignments. Accuracy in the alignment of protein sequences is key to a number of biological problems, including those of gene annotation, phylogeny determination, protein structure modelling and protein function annotation (Mahusudhan *et al.*, 2006).

4.2.2 Sequence Identity and Alignment Errors

Modelling a target-template pair sharing high sequence similarity (usually considered above 40% pair-wise identity) is relatively trivial; it is the low percentage sequence identity, below 20%, (the twilight zone) pairs where major difficulties are encountered. In this range of sequence similarity, the largest errors in comparative modelling due to misalignments begin to appear (Sanchez & Sali, 1997). When proteins share 30-50% identity, significant shifts between different alignments emerge, mostly in loop regions (Jaroszewski *et al.*, 2002). When sequence identity is below 30%, sequence alignments become very unstable, changing dramatically with scoring matrices and gap penalties (Vogt *et al.*, 1995); they essentially become random for structurally similar proteins with undetectable sequence similarity (Holm *et al.*, 1992). At very low sequence identity, structures diverge significantly enough so that some parts of the sequence al ignment lose meaning, thus there is a limit to how accurate a sequence alignment can be (Grishin, 2001). Studies have shown that protein sequence alignment methods often fail to align sequences accurately (Saqi *et al.*, 1998), and that none of these techniques produce consistently good solutions for all cases (Rai & Fiser, 2006). CASP5 re-iterated the significance and difficulty of alignment errors: a number of structurally conserved regions in submitted models of remote evolutionary distance from the template were misaligned. Analysis of these errors indicates that the absolute majority of them occurred in regions deemed unreliable in the course of model building (Venclovas, 2003). An average misalignment of only one residue position could result in an error of approximately 4Å in the model (Fiser & Sali, 2003) and any change in distance as large as this will result in a loss of function. One or two misplaced residues can substantially reduce a protein's predicted affinity for its substrate (Kimura *et al.,* 2001). This proves that structural errors seriously limit the value of the models in biological applications (Prasad *et al.,* 2003).

4.2.3 Alignments in Comparative Modelling

Searching methods used during the target-template alignment stage of comparative modelling (step one in the modelling protocol) are usually tuned for detection of remote relationships, not for optimal alignments (Marti-Renom *et al.,* 2003), and often include only regions of high similarity between the query sequence and the database hits. A simple and classic example is the BLAST algorithm, which finds homologues based on local alignment properties only. This means that it is usually necessary to realign the selected template to the target sequence (Tramontano, 1998), step two in the comparative modelling protocol. The "correct" sequence alignment is the alignment in which structurally equivalent positions are correctly aligned. Of course, in reality, the information about the structure of one of the proteins is not available and the alignment must be inferred from sequence alone (Tramontano, 1998).

4.2.4 Different Alignment Techniques

Alignments can be generated by a pair-wise alignment (aligning two sequences) or a multiple sequence alignment (aligning more than two sequences which share homology). Alignments may be generated by increasingly complex methods, which are usually coupled with increasing accuracy, in the form of a sequence-sequence alignment, a sequence-profile alignment, or a profile-profile alignment. The standard profile or position-specific sequence matrix (PSSM) can be replaced with a hidden Markov model (HMM) to produce a sequence-HMM alignment, or even a HMM-HMM alignment. Alignments can be global or local and can incorporate the use of dynamic programming. These protocols are discussed in the following sections.

4.2.4.1 Dynamic programming

Dynamic programming (Needleman & Wunsch, 1970; Smith & Waterman, 1981) calculates the score of the optimal alignment between two protein sequences and provides a single alignment with this score (Jaroszewski *et al.,* 2002). Dynamic programming optimises a scoring function that depends on residue-residue substitution scores and penalties for the creation and extension of gaps (Madhusudhan *et al.,* 2006). It does not provide information about how many different alignments have scores close to the optimal one, and how different these alignments are. In principle this information is easily available in alignment algorithms based on high-scoring segment pairs, such as those used in BLAST (Jaroszewski *et al.,* 2002). There is a great variety of protein sequence alignment methods, many of which are based on dynamic programming techniques (Barton & Sternberg, 1987; Taylor *et al.,* 1994). Dynamic programming algorithms that use standard substitution matrices, such as PAM (Dayhoff & Eck, 1968) or BLOSUM (Henikoff & Henikoff, 1992) were the initial methods of choice. Although dynamic programming guarantees the optimal solution, the insensitivity and generality of the substitution matrices limited the usefulness of such methods to cases of high sequence identity (Marti-Renom *et al.,* 2003).

4.2.4.2 Global Alignments, Needleman and Wunsch

The original and arguably still most popular method for sequence alignment is based on the dynamic programming algorithm of Needleman and Wunsch (Needleman & Wunsch, 1970). The Needleman and Wunsch global alignment algorithm (figure 4.1) determines the alignment and the alignment score of a pair of sequences by finding the highest score and the maximum match pathway that leads to the accumulation of the highest score in the two-dimensional array (Yang, 2002). It has since been built upon to improve its accuracy and speed (Thompson et al., 1994; Myers & Miller, 1988).

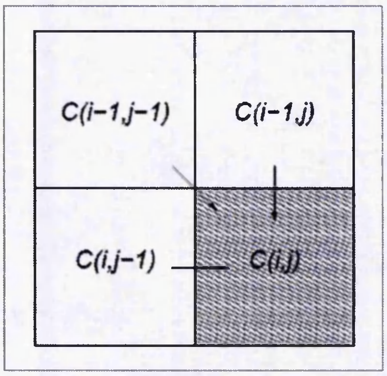


Figure 4.1. The Needleman-Wunsch Algorithm. The cells of the score matrix are labelled C(i; j) where i = 1, 2,...N and j = 1, 2,...M. The value of the cell C(i; j) depends only on the values of the immediately adjacent northwest diagonal, up, and left cells.

4.2.4.3 Local Alignments, the Smith-Waterman Algorithm

The Smith-Waterman algorithm (1981) is based on the Needleman and Wunsch global alignment algorithm. Instead of looking at each sequence in its entirety, this compares segments of all possible lengths and reports sub-sequences which optimise the similarity measure hence, not all of the sequence is always retained.

4.2.4.4 Pair-wise and Sequence-Sequence Alignments

Pair-wise sequence alignment programs consider all possible alignments and gap positions and create the alignment with the highest score and the fewest gaps. In general, they use the alignment methods of Needleman and Wunsch (Needleman & Wunsch, 1970) or some modification of it (Tramontano, 1998). At 40% sequence identity, alignments by pair-wise methods are only 80% correct on average and this number drops sharply at lower similarity ranges; they especially become random for structurally similar proteins with "undetectable" sequence similarity (Holm *et al.,* 1992; Orengo *et al.,* 1997).

4.2.4.5 The use of Multiple Sequence Alignments

The use of multiple sequence alignments (Gribskov *et al.,* 1987; Gribskov *et al.,* 1990; Gribskov, 1994) has improved alignment accuracy considerably. This is because a multiple sequence alignment of homologous sequences contains more information about the sequence family than a single sequence (Soding, 2005). It hopefully aligns residues in a given column that share homology and may share a common functional role.

4.2.4.6 Progressive Alignments and Iterative Optimisation

ClustalW and T-Coffee remain popular choices for multiple sequence alignment. These programs employ progressive alignment (Dunbrack, 2006). The progressive algorithm (Hogeweg & Hesper, 1984), used for example in ClustalW (a global alignment program) (Thompson *et al.,* 1994), T-Coffee (Notredame *et al.,* 1998) and MUSCLE (Edgar, 1994) starts with the alignment of two sequences and then, adds other sequences one by one according to a predetermined order (Zhou & Zhou, 2005). This order is based on a guide tree of the sequences to be aligned. They employ a global alignment algorithm to construct an alignment over the entire length of the sequences and differ mainly in the procedure employed to determine the order of the alignment of the sequences (Chkrabarti *et al.,* 2004). Often amino acids are misaligned by such methods because of small misalignments early in the process (Dunbrack, 2006). More recent studies focused on iterative optimisation, for example, MUSCLE (Zhou & Zhou, 2005). Dunbrack found the MUSCLE program to have improved performance over other methods and that MUSCLE is more accurate than those when subjected to benchmarks and generally faster (Dunbrack, 2006). Nuin (2006) also found MUSCLE to be superior to other alignment methods. They compared nine of the most often used protein alignment programs and found that sequence length did not affect alignment accuracy. However, they found that ClustalW had the steepest decline in accuracy when the sequence length was increased, especially when INDELs were present and had the worst accuracy. MUSCLE was found to be in the intermediary group, along with T-COFFEE. They determine that T-COFFEE generates good alignments but the processing time is the worst for every sequence size, whereas MUSCLE produced good quality alignments and was very fast. The MUSCLE authors also compared multiple sequence alignment methods (Edgar & Batzoglou, 2006). They considered the best current programs that are directly comparable to CLUSTALW in the global alignment tools to include MUSCLE and T-COFFEE. They add that MUSCLE was found to offer significant improvements in scalability with comparable accuracy and T-COFFEE had limiting factors of computation time and memory usage for larger alignment problems. This explains the choice of using MUSCLE in this project over the other T-COFFEE.

4.2.4.7 Sequence-Profile Alignments

The development of sequence-profile comparison methods such as PSI-BLAST has led to a great improvement in sensitivity over sequence-sequence comparison methods such as BLAST (Soding, 2005).

A profile is a representation of a group of related protein sequences, usually based on a multiple alignment of those sequences. Once the multiple alignment has been defined, the profile is constructed by counting the numbers of each amino acid at each position along the multiple alignment. These counts are transformed into probabilities by normalising the counts by the total number of amino acids and gaps observed at that position. These empirical probabilities reflect the likelihood of observing any amino acid *k* at position *i*. Since the counts are based on a finite set of sequences it can happen that not all 20 amino acids are observed at each position. Therefore, pseudo counts are introduced so that no amino acid has a zero probability to occur at position *i* (Yona & Levitt, 2002) as a zero probability would result in an error in the calculation when trying to divide zero by a whole number. Profiles can be variations of the 20 x *L* matrix used in PSI-BLAST, where *L* is the length of the generating sequence. These variations might include a gap character, for example (Dunbrack, 2006). Wang and Dunbrack (2004) found that removing positions in the profile with gaps in the query sequence results in better alignments. Tan (2006) found that a better amino acid similarity matrix can improve a profile itself.

Profile alignment methods allow efficient recognition of remotely related sequences. These methods "outperform" the ability of comparative modelling in a sense that they are able to locate remotely related template-target sequence pairs, that are sometimes identified only by a few short conserved segments, and for which no reliable comparative model can be built (Rai & Fiser, 2006).

The sequence alignment between the target and template is usually derived from a multiple sequence alignment using as many proteins of the family as possible. Its accuracy depends on the number and similarity distribution of the sequences of the protein family. Homology is transitive; therefore if two proteins are evolutionary related to a third protein, they are also evolutionary related to each other. This can be used to detect more distant evolutionary relationships in database searching strategies, by 'hopping' in sequence space from one homologous protein to the next and thus increasing the number of proteins that can be included in the family, a concept applied by the PSI-BLAST algorithm (Altschul *et al.*, 1997; Cozzetto & Tramontano, 2005).

PSI-BLAST aligns the target sequence to a sequence profile constructed from a multiple sequence alignment of members of a protein family. PSI-BLAST uses the core BLAST algorithm to collect related sequences to the query sequence and iteratively scan a sequence database for more homologues to then construct its profile (Altschul *et al.,* 1997). Sequence alignment profiles have been shown to be very powerful in creating accurate sequence alignments. More accurate and longer alignments have been obtained with profile-profile comparison (Wang & Dunbrack, 2004).

4.2.4.8 Profile-Profile Alignments

Ohlson and co-workers showed that profile-profile based methods perform at least 30% better than standard sequence-profile methods in the quality of the obtained alignments. This is probably because profile-profile scoring methods are better at distinguishing evolutionary related positions from non-related positions. For each alignment, a model of the query protein was created and compared with the structure of this model with the correct structure. The quality of the alignments was measured by MaxSub (Siew *et al.*, 2000), a measure that should be one for a perfect model and zero for a completely wrong model. MaxSub finds the largest subset of atoms of a model that superimposes well over the experimental model (Ohlson *et* al., 2004).

Marti-Renom (2004) tested thirteen different protocols for creating and comparing profiles, including the correlation coefficient and the dot product, and eight different protocols for aligning sequences and/or profiles, including LOBSTER (Edgar, 2004), BLAST and PSI-BLAST. They found that in general the smaller the fraction of target modelled, the more accurate the model. Additionally, algorithms that are local (BLAST and PSI-BLAST) generally do not align whole sequences, but only regions that are quite similar to each other. Global protocols ensure an optimal alignment that is forced to cover whole sequences. Their results showed PSI-BLAST to out-perform BLAST and LOBSTER to out-perform PSI-BLAST (compared to their structural gold standard, generated using the CE algorithm), emphasising the improvement in alignment accuracy due to the inclusion of multiple sequence alignments and profiles. The correlation coefficient comparison scheme (Marti-Renom *et al.*, (2004) was found to be the best.

4.2.4.9 The use of Hidden Markov Models

Hidden Markov models (HMM), a class of probabilistic models, can also be used to represent a multiple sequence alignment. Profile HMMs are similar to sequence profiles, but in addition to amino acid frequencies in the columns of a multiple sequence alignment they contain the position-specific probabilities for insertions and deletions along the alignment. Profile HMMs perform better than sequence profiles in the quality of alignments (Krough *et al.*, 1994; Eddy, 1998). The higher sensitivity is due to the fact that position-specific gap penalties penalise the chance hits more than true positives which tend to have insertions or deletions at the same positions as the sequences from which the HMM was built (Soding, 2005). Soding showed that by aligning profile HMMs instead of simple sequence profiles the sensitivity of the alignment could be improved. By comparing BLAST and PSI-BLAST as popular representatives of sequence-sequence and sequence-profile methods, the sequence-HMM comparison package HMMer (Eddy, 2001) and the profile-profile methods COMPASS (Sadreyev & Grishin, 2003) and PROF_SIM (Yona & Levitt, 2002), Soding notes that PSI-BLAST produces much better alignments than BLAST, the profile-profile methods perform better than PSI-BLAST and aligning profile HMMs instead of simple profiles improves the alignment quality significantly. COACH is a hybrid method that compares a multiple sequence alignment with an HMM and is not strictly a 'true' HMM-HMM alignment method (Dunbrack, 2006). Wistrand and Sonnhammer (Wistrand & Sonnhammer, 2005) compared the performance of HMMer and SAM (Hughey & Krogh, 1996) and found that although SAM models were better, HMMer model scoring was better. It is thought that including as many sequences of remote homology as possible in the multiple sequence alignment that generates the profile or HMM would be preferable, however, retaining sequences that are more divergent from the target than the chosen template might decrease the alignment accuracy. Johnston and Shields (Johnston & Shields, 2005) found that combining HMMs from multiple sequence subsets of a larger set of sequences performed better than using the single HMM built from an alignment of all of the sequences.

4.2.5 Including Structural information

In more difficult cases, it is frequently beneficial to rely on multiple structures and sequence information (Barton & Sternberg, 1987; Taylor *et al.*, 1994). As with multiple sequence alignments, better profiles and HMMs can be built using structural alignments of remote homologues and by adding sequences of unknown structure that can be easily aligned with each structure (Dunbrack, 2006). The use of structural information for one of the sequences in a pair-wise alignment improves the accuracy of the alignment in the low sequence similarity range. Methods that employ this approach include threading and 3D template matching (Bowie *et al.*, 1991; Godzik & Skolnick, 1992; Jones *et al.*, 1992; Kelley *et al.*, 2000; Shi *et al.*, 2001 Fischer, 2003). Using structure alignment in combination with sequence alignment methods is more powerful, hence T-COFFEE being improved by the use of pair-wise structural alignments in 3D-COFFEE (O'Sullivan *et al.*, 2004). Dunbrack states that it is more important to combine sequence and structure information, rather than using structural information alone, due to the inherent ambiguities of deriving a sequence alignment from a structural superposition (Dunbrack, 2006). In contrast to a pair-wise sequence alignment that simply compares two strings of characters, a pair-wise structure alignment is a more difficult problem that optimally superimposes two sets of coordinates and finds the regions of closest overlap in the three-dimensional space (Chen *et al.,* 2005).

There are also practical problems to consider when selecting algorithms for aligning sequences for comparative modelling. Some sequence-structure based methods can't be used straightforwardly because they are implemented as web servers or are not generally available from the authors (Marti-Renom *et al.,* 2004). Similar practical considerations in implementing them might also persuade users to select alternatives.

4.2.6 Improving Alignments

Jaroszewski and co-workers set up an experiment that sampled a huge conformational space of alternative alignments by combining an approach of varying parameters (for example, gap penalties) with an iterative approach that penalises regions of the sample space that have already been visited. The study states that there are alignments that exist that are of better quality than the original alignments for about 50% of the protein pairs with moderate-to-low sequence similarity, less than 45% identical (Jaroszewski *et al.,* 2002).

There are ways to improve the alignments of sequences for comparative modelling, one such example is from the Sali lab (John & Sali, 2003). They refine an initial target-template alignment using a genetic algorithm protocol that starts with the initial alignments and then iterates through the re-alignment, model building and model assessment to optimise a model assessment score. They found the average CE overlaps of their genetic algorithm protocol higher than those produced by PSI-BLAST. Another, more recent example, is that of Rai and Fiser (Rai & Fiser, 2006). These authors developed an optimal combination of alignments produced by alternative methods which are superior in certain segments but inferior in others when compared to each other.

This project does not consider the ways in which to improve the alignments or selection of the alignments, but just evaluates various protocols for aligning sequences in terms of comparative modelling.

4.2.7 Selecting the Best Alignments

Other groups use the multiple model approach to test 16 pairs of distantly related proteins to focus on the question of whether a good (an alignment which contains useable information in it) alignment exists in a set of alternative alignments using a given method (Jaroszewski *et al.,* 2002). They state that there is strong evidence that recognising the correct alignment is possible by building a protein model and evaluating it, however they take their research in this article no further and stop at obtaining a relatively small set of alignments that contains at least one significantly better alignment and do not address the issue of how to select this alignment.

Saqi and colleagues reveal that short stretches of high local identity may not always be reflected in the structure based alignment (Saqi *et al.,* 1998). Previously it was assumed that any misaligned regions, compared to the structure based alignment, occur in regions where the local sequence identity is lower than the global, usually in sequences where the global sequence identity is less than 40%. This means that sequence similarity may not always give clear indication of the resulting comparative model and that high local sequence identity can result in lower quality regions of the model. Following on from this, Contreras-Moreira *et al* prove that the optimal sequence alignment is not always the best for modelling (Contreras-Moreira *et al.,* 2003). They analysed how often the optimal sequence alignment corresponded to the model with the lowest RMSD. They found that the alignment with highest sequence identity provided the lowest RMSD model in 42 cases (out of 58) but that the other 16 would have been modelled more accurately using a suboptimal alignment. This suggests other alternative alignments should be considered in model construction. Contreras-Moreira and colleagues also propose the use of genetic algorithms for constructing a large number of alternative alignments by recombining an initial set of alignments (John & Sali, 2003). A common problem of these approaches however, is the selection of the "best" alignment to construct the final model.

4.2.8 The Alignment Methods Used

The alignment methods included in this current investigation were:

- BLAST to represent a standard sequence-sequence based method;
- PSI-BLAST to provide a sequence-profile protocol;
- MUSCLE to represent an accurate multiple sequence alignment method;
- MAMMOTH, CE and TM-align to provide gold standard structural alignments;
- COACH, a hybrid method similar to a MSA-HMM technique;
- a locally implemented HMM-HMM method;
- HMMer, as a sequence-HMM method.

Information on each of these methods, including versions used, can be found in Chapter 2, Resources.

4.2.9 Assessing Alignment Accuracy

Sequence alignment accuracy can be measured in different ways. Generally, the most common way involves structure-based alignment of the target-template pair, deriving a sequence alignment from this structure alignment, and comparing the predicted sequence alignment with the structure based alignment (Dunbrack, 2006).

Despite some ambiguities in the definitions of structural alignments (Godzik, 1996), structural alignments are often treated as the "standards of truth" in evaluating sequence alignments because it is generally accepted that, with increasing evolutionary distance, structures change less than do sequences (Vogt *et al.*, 1995).

Protein structure comparisons are employed in almost all branches of contemporary structural biology, ranging from protein structure modelling to structure-based protein function annotation (Zhang & Skolnick, 2005). Many structure-based alignment methods have been developed (SSAP, Orengo et al., 1996; CE, Shindyalov & Bourne, 1998; Dali, Holm & Sander, 1993). Generally there are two types of approaches to the structural alignment: coordinate-based and environment-based. In the coordinate-based approach, an alignment is just like aligning two sets of points, and the similarity is evaluated based on how well the two sets can be superimposed in 3D space. In the environment-based approach, structure-derived descriptors (for example, solvent accessibility and hydrogen bond strengths) rather than explicit Cartesian coordinate-based distances are used to generate the structure-based alignment (Chen & Crippen, 2005).

A further example is MAMMOTH (Ortiz *et al.,* 2002). The authors conclude that MAMMOTH shows performance consistent with other structural alignment methods when comparing experimental protein structures - in this case, including, Dali (Holm & Sander, 1993). Another popular tool is the TM-align algorithm developed by Skolnick, which was concluded to be around four times faster than CE and 20 times faster than Dali and is considered to have resulting alignments with higher accuracy and coverage than those provided by these methods (Zhang & Skolnick, 2005).

TM-align and MAMMOTH were chosen to represent the gold standard structural alignment tools due to their higher accuracy and speed compared with other structural alignment methods available at the time. CE was also used, as it is a common local structural alignment method and different from the TM-align and MAMMOTH which produce structural alignments which are effectively global. TM-align was chosen as an 'overall' gold standard to compare alignments to as CE is a local alignment method and it was therefore difficult to compare the global alignment methods to CE and MAMMOTH chopped off end residues in the alignments.

In this project the sequence based alignments were tested by observing the similarity against the known structures (our standards of truth) and the use of iRMSD (see Chapter 2, the Resources chapter, for more information on the

NiRMSD algorithm).

The datasets were generated using sequences and structures of peptidases. The importance and significance of using peptidases in this chapter are discussed below.

4.2.10 Importance of Peptidases

A peptidase is an enzyme that hydrolyses peptide bonds in proteins and peptides. They are ubiquitous, constituting around 2% of the genome proteins in all kinds of organisms. It has been estimated that 14% of the five hundred human peptidases are under investigation as drug targets and there are over 550 active and putative peptidases in the human genome (Rawlings & Morton, 2006). Peptidases are perhaps the largest class of enzyme to be used as targets for structure-based drug design (Mittl & Grutter, 2006). A well known example is the HIV protease, an aspartic protease responsible for the processing of the HIV viral proteins and the basis for the design of a variety of lead compounds against the virus. Peptidases cause irreversible modification or destruction of their substances that may be of biological importance in many different ways. Peptidases have been of interest to mankind for hundreds of years because of the many ways in which they are involved in human physiology, pathology and technology (Barrett *et al.,* 2001). They are involved in blood clotting, apoptosis, pre/pro-hormone processing and digestion amongst many other biological functions.

Considering the functional relevance of peptidases it is not difficult to understand that a deficiency of these enzymes underlies several pathological conditions such as cancer, arthritis, neurodegenerative and cardiovascular disease. Moreover, many infectious micro organisms, viruses and parasites use peptidases as virulence factors. Accordingly, many peptidases or their substrates are an important focus of attention for the pharmaceutical industry as potential drug targets (Lopez-Otin & Overall, 2002).

Peptidases regulate the fate and activity of many proteins by controlling many appropriate intra- or extra-cellular localisation; shedding from cell surfaces; activation or inactivation of peptidases and other enzymes, cytokines, hormones or growth factors; conversion of receptor agonists to antagonists; and exposure of cryptic neoproteins (which is when the proteolytic cleavage products are functional proteins with roles that are distinct from the parent molecule). Hence peptidases initiate, modulate and terminate a wide range of important cellular functions by processing bioactive molecules and thereby directly controlling essential biological processes, such as DNA replication and cell proliferation (Lopez-Otin & Overall, 2002).

4.2.11 Peptidase inhibitors

Since the regulation of the activities of peptidases is crucial, the hundreds of proteins that inhibit them are equally relevant. The concept that peptidase inhibitors can make effective drugs has been validated most dramatically for retropePSIn, the processing endopeptidase of the human inmmunodeficiency virus, several inhibitors of which have been proved to be potent antiviral agents (Wlodawer & Vondrasek, 1998).

4.2.12 Why Named Peptidases

Peptidase is the most correct scientific term for the proteolytic enzymes that are colloquially called proteases or proteinases. Amongst the reasons for using the term peptidases is that this is the word recommended by the NC-IUBMB (Nomenclature of the International Union of Biochemistry and Molecular Biology, NC-IUBMB, 1992), as well as *MEROPS*, as well as the fact that it is the word that already forms the root of the names of the many different sub-types of peptidases: aminopeptidase, carboxypeptidase, and so on, and thus leads to a very rational and intuitive system of terminology.

4.2.13 Different Types of Peptidases

Peptidases are grouped by the chemical mechanism of catalysis. Peptidases can be described as of serine, cysteine, threonine, aspartic, glutamic, or metallo catalytic type. In peptidases of serine, threonine and cysteine type, the catalytic nucleophile is the reactive group of an amino acid side chain, either a hydroxyl group (serine and threonine peptidases) or sullfhydryl group (cysteine peptidases). As far as is known, the activity of all cysteine peptidases depends on a catalytic dyad of cysteine and histidine. The order of the cysteine and histidine residues in the linear sequence differs between families, and this is among the lines of evidence suggesting that cysteine peptidases have had many separate evolutionary origins (Rawlings & Barrett, 1994). In aspartic and metallo- peptidases, the nucleophile is commonly an activated water molecule. In aspartic peptidases, the water molecule is directly bound by the side chains of aspartic residues. In metallopeptidases, one or two metal ions hold the water molecule in place, and charged amino acid side chains are ligands for the metal ions. The metal may be zinc, cobalt or manganese, and a single metal ion is usually bound by three amino acid ligands. Metallopeptidases form the most diverse of the catalytic types of peptidases. About half of the families comprise enzymes containing the His-Glu-Xaa-Xaa-His (or HEXXH) motif that has been shown by X-ray crystallography to form part of the site for binding of the metal (normally zinc) atom in some families. Proline is never found in this region; all of the available tertiary structures for metallopeptidases containing HEXXH show the motif in a helix, which would be broken by proline (Rawlings & Barrett, 1995). The glutamic peptidases seem to employ a Glu/Gln catalytic dyad.

4.3 METHODS AND MATERIALS

To allow the assessment of different alignment protocols using peptidases as a test case, a set of target-template pairs was required. Each of these pairs needed corresponding structures, to enable assessment of the quality of any subsequent comparative modelling, and importantly for this chapter, to obtain 'gold standard' alignment results, for both the target and the template.

4.3.1 The Dataset

The primary source of sequence information was the MEROPS database (Rawlings *et al.*, 2006). The MEROPS "pepunit.lib" file formed the main peptidase dataset containing 28,445 peptidase sequences. All characters that would prove a problem being recognised by any of the alignment methods were removed; these included all Bs, Zs, J and Os, which were replaced with Xs. MEROPS sequences were then cross-referenced with structural information. In order to find structures for all of the sequences in the MEROPS database the astral dataset (Version 1.67, current release at the time, http://dunbrack.fccc.edu/PISCES.php) was used which contained all genetic domain sequences based on PDB ATOM records (versions and more information on the databases stated here can be found in the Resources chapter).

4.3.1.1 Selecting the Targets

Each sequence in the astral database (50,495 structures) PDB provided a query sequence to be used in standard BLAST search (e-value cutoff 0.001) against the whole of the pepunit.lib (MEROPS) database (28,445 sequences), to match PDB structures to a MEROPS sequence. This resulted in 5,337 top hits of which 4,314 were enzymes with no inhibitors, 1,023 were enzymes with bound inhibitors, corresponding to 5,337 PDB entries with at least one MEROPS sequence.

Potential matches between the MEROPS sequences and PDB domains were further filtered to ensure only true peptidase homologues were identified in the PDB. Different levels of coverage between the MEROPS domain sequence and the PDB sequence was tested. A trade-off between the coverage of the MEROPS domain sequence and the PDB sequence existed, for example, 75% coverage resulted in a total of 2,582 hits, whereas 90% coverage resulted in 4,109 hits. To maximise the dataset, 90% coverage of the peptidase domain sequence by the PDB sequence was required in order for it to be considered a match and be chosen as a possible target or template candidate. After the PDB ATOM sequences (the hits) were found to share at least 96% sequence identity and 90% coverage to the MEROPS sequence, the astral PDB hit with the best e-value score was taken to represent the peptidase domain sequence. This was to ensure all the sequences had a corresponding structure.

The database was split into two groups, those peptidase domains that had an inhibitor bound and those that did not. The “enzyme only” set (peptidase domain without an inhibitor bound) contained 3,495 matches to PDB, and the "enzyme + inhibitor" set just 614 hits. Metal ions were retained in the metallopeptidases if they were in the active site, or were part of the inhibitor. Occasionally a MEROPS sequence was assigned with multiple PDB files. The one with the best e-value score from the original BLAST hits, the longest sequence, and the best coverage was chosen. Finally, NMR structures were removed along with PDB files with only carbon alpha traces or with missing electron density (chain breaks – missing residues).

4.3.1.2 Generating the Target-Template Pairs

Two datasets of target-template pairs were generated (Figure 4.2), one with peptidases with an inhibitor bound as the target and with peptidases with inhibitors bound as the template (the I-vs-I set), and another one with peptidases with an inhibitor bound as the target and with peptidases with no inhibitor bound as the template (the I-vs-S set). The two datasets were used to see how much having an inhibitor bound to the template affected the modelling results. A set with no inhibitor bound to the target was not used since the aim of this investigation was to find out how well methods could predict the structure of the interface region between the peptidase and the bound inhibitor. These were also generated using BLAST using astral-MEROPS cross-referenced data set. For candidate target-template pairs duplicates were removed. This left candidate 5,979 I-vs-I pairs and 62,501 I-vs-S pairs.

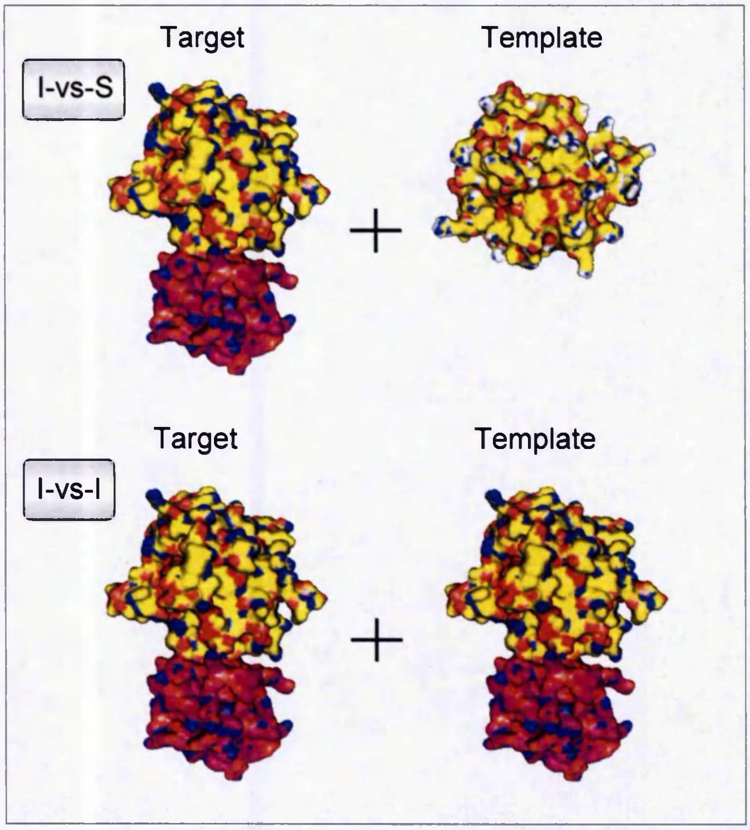


Figure 4.2. The Datasets. The I-vs-S dataset contains a target which consists of a peptidase chain (yellow) and an inhibitor chain (pink) and a template which only has a peptidase chain. In the I-vs-I set both the target and the template contain a peptidase (yellow) chain and inhibitor chain (pink). Produced using Pymol (DeLano 2004).

To remove trivial modelling pairs, those above 80% sequence similarity between the target and template were removed. This left 3,083 pairs (191 unique PDB entries) in the I-vs-I set and 38,629 pairs (2,002 unique PDB entries) in the I-vs-S set.

For modelling purposes (Chapter 5), only PDB structures with 2.5Å resolution or better were used. The number of pairs at different resolutions was calculated to enable a substantial dataset to be created. The majority of pairs existed at equal to or better than 2.5 Å (2,440 in the I-vs-I set and 9,652 in the I-vs-S set) resolution. Too many pairs were lost when the resolution was better than 2.0A (1,280 in the I-vs-I set and 4,705 in the I-vs-S set).

The redundancy at the 95% level between all the candidate templates for each target was removed and the PDB template with the best resolution was selected. The first hit (i.e. the longest sequence and the highest sequence identity) was not chosen since in this case for modelling purposes the template with the best resolution was more important than the template with the highest sequence identity. This left 3,789 pairs or 33 targets in the I-vs-S set and 609 pairs or 29 targets in the I-vs-I set.

The targets were also clustered at the 95% sequence identity to remove similar pairs. The top hit in each cluster was chosen to leave 26 target-template pairs in the I-vs-I set and 144 pairs in the I-vs-S set.

4.3.2 Constructing the Target-Template Alignments

Alignments were retained as candidate target-template pairs if the sequence coverage between the target and template was >50%. Pair -wise alignments from multiple sequence-based protocols were extracted and cleaned up by removing "double gaps" (indicating insertion/deletion relative to other sequences in the multiple sequence alignment and not the target or template).

4.3.2.1 Calculating Percentage Sequence Identity

The percentage sequence identities of the alignments were calculated in the following ways:

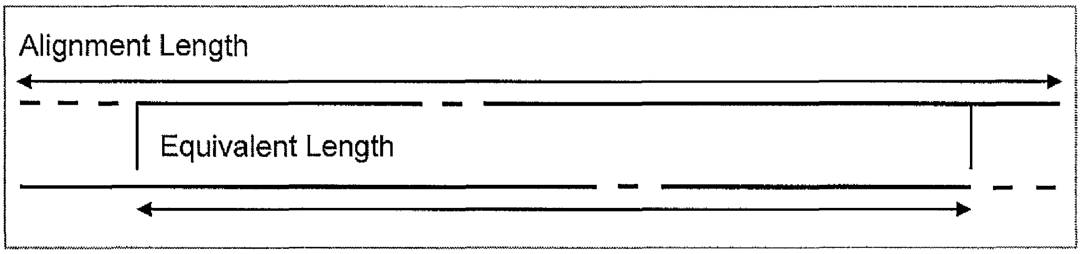


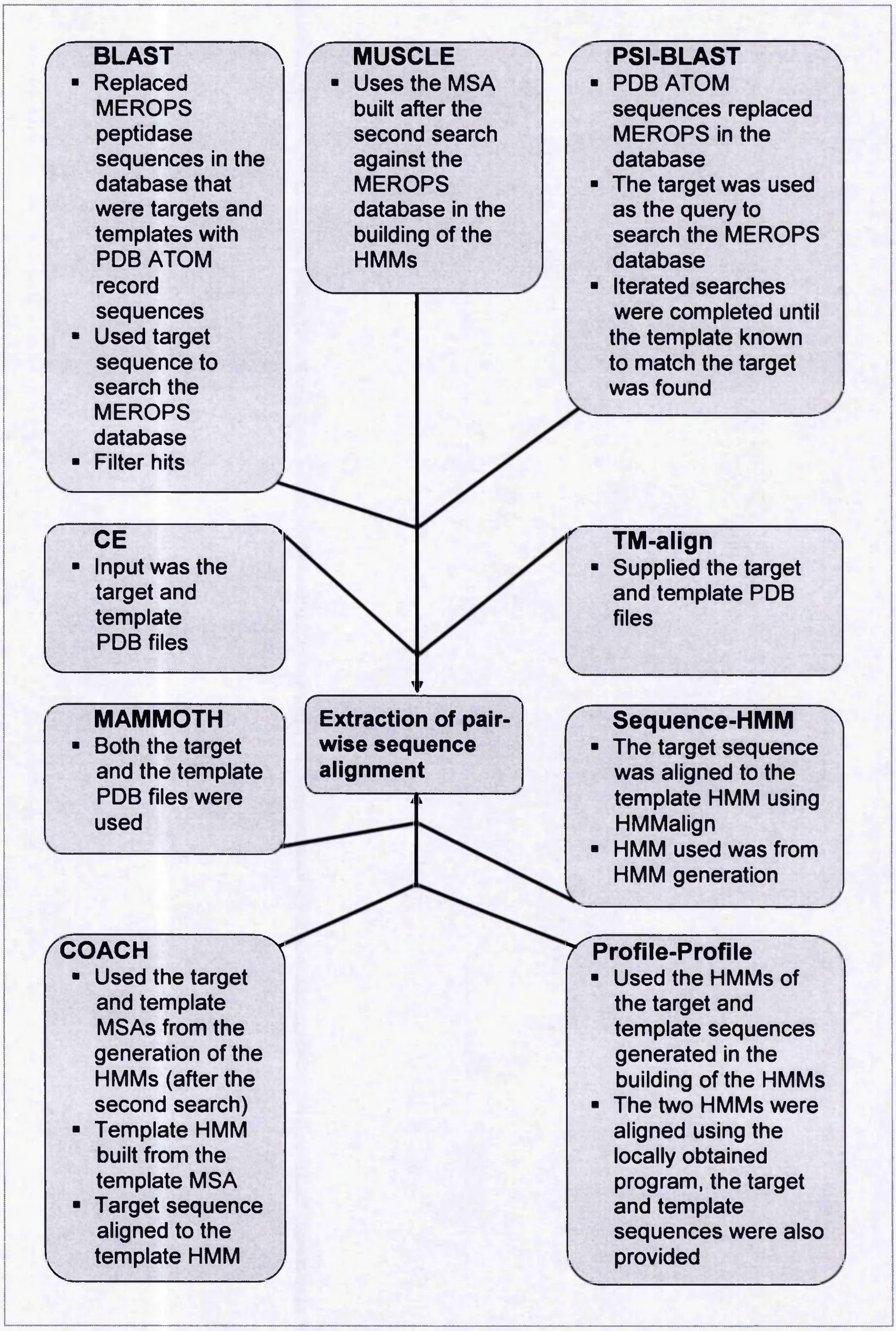
Figure 4.3. The Different Lengths of the Sequence Used. An example of a target-template pair is shown with gaps displayed as dashes and the definition of alignment and equivalent lengths depicted. The method used to calculate the percentage identity was the "Length of the alignment method".

- *Shortest Sequence:* This is the number of identical residues in the target and the template divided by the length of the shortest sequence.

- *Length of alignment:* T he number of identical residues divided by the alignment length (Figure 4.3).
- *Mean length of the two sequences:* This is calculated by dividing the number of identical residues by the mean length of the target and template sequences.
- *Equivalent positions:* The identical residues divided by the equivalent length (Figure 4.3).

The method used to calculate the percentage identity was the "Length of the alignment method". The percentage identity of TM-align alignment (the chosen gold standard method) was used.

A summary of the protocols described in this section which were used to generate the target-template alignment pairs can be seen in figure 4.4.



BLAST

- Replaced MEROPS peptidase sequences in the database that were targets and templates with PDB ATOM record sequences
- Used target sequence to search the MEROPS database
- Filter hits

MUSCLE

- Uses the MSA built after the second search against the MEROPS database in the building of the HMMs

PSI-BLAST

- PDB ATOM sequences replaced MEROPS in the database
- The target was used as the query to search the MEROPS database
- Iterated searches were completed until the template known to match the target was found

CE

- Input was the target and template PDB files

TM-align

- Supplied the target and template PDB files

MAMMOTH

- Both the target and the template PDB files were used

Extraction of pair-wise sequence alignment

Sequence-HMM

- The target sequence was aligned to the template HMM using HMMalign
- HMM used was from HMM generation

COACH

- Used the target and template MSAs from the generation of the HMMs (after the second search)
- Template HMM built from the template MSA
- Target sequence aligned to the template HMM

Profile-Profile

- Used the HMMs of the target and template sequences generated in the building of the HMMs
- The two HMMs were aligned using the locally obtained program, the target and template sequences were also provided

Figure 4.4. The Construction of the Alignments. The generation of the alignments using the different protocols described.

4.3.2.2 MAMMOTH

On occasion MAMMOTH missed residues off from the end of the alignment, in these cases the alignment was examined manually and the residue was added back on to the end of the alignment. For a few PDB files, e.g. 1ton_.pdb, 1h8dH, 1ueaA, 1h7lP, 4htcH and 4cpa_ one or two residues had side-chain coordinates missing. This caused MAMMOTH to miss these residues in the final alignment. MAMMOTH took the PDB files of the target and template as input. The different side-chain conformations displayed in the PDB file of the target and template were removed since MAMMOTH would leave out these residues in the final alignment. The default parameters were used.

4.3.2.3 TM-align

TM-align was run using the renumbered target and template PDB files (accepting these as input) which resulted in a structure-based sequence alignment between the target and template. The default parameters were used.

4.3.2.4 CE

The PDB files of the target and template provided the input for CE. Chain breaks were removed and the PDB files contained only the relevant peptidase chains were submitted to the CE algorithm. A structure-based sequence alignment of the target-template was obtained.

4.3.2.5 BLAST

The target sequences provided the query to be used to search the "pepunit.lib" database. Redundancy was removed at the 95% sequence similarity level. The MEROPS sequences in the "pepunit.lib" database were replaced with the PDB ATOM sequences of the targets and templates used in the study, as slight discrepancies exist between the MEROPS sequence database and the sequence of the PDB ATOM sequences. This meant that when extracting the alignment pairs, the sequences would be the PDB ATOM record structural ones rather than the MEROPS peptidase ones. Standard BLAST filtering was used, with the standard e-value cut-off of ten. T arget-template pairs were extracted from the output.

4.3.2.6 PSI-BLAST

Each target sequence was searched against the MEROPS "pepunit.lib" database which included the template sequence for that given target-template pair. PSI-BLAST was run with an e-value cut-off of 1E-10 for up to 4 rounds. If the search did not generate a match to the chosen template after 4 rounds, the round number was lowered by one and the results examined. This was repeated until the template was found. If it was not, the e-value was increased to include more hits. The majority of searches produced hits at round 4, with a 1E-8 cut-off. Once the template had been found, the iterations with lower e-values ceased and the target and template sequences were extracted as an alignment. The sequences of the target and template of the MEROPS database were replaced with the sequences derived from the ATOM records of the PDB file.

PSI-BLAST alignment accuracy was evaluated in comparison to structural alignments by Friedberg *et al* (Friedberg *et al.,* 2000). They used 123 pairs of proteins that were structurally similar but sequentially dissimilar, and evaluated them by determining the percentage of residues correctly aligned in the sequence alignment with respect to the structural alignment. They found it worthwhile to continue for several iterations to obtain better alignments, with higher sensitivity and no significant effect on the specificity.

4.3.2.7 MUSCLE

Multiple sequence alignments for target and template sequences were generated using MUSCLE after the second search against the MEROPS database, as described above in the building of the HMM. The target and template sequences were included in the multiple sequence alignment, and the target-template pair was then extracted from the MSA as a final step prior to assessment as a pair-wise alignment.

4.3.2.8 Building the Hidden Markov Models

HMMs were created for both target and template sequences. An initial set of homologues was generated using BLAST by searching the *MEROPS* database with an E-value cut-off of 1E-60. If fewer than 10 hits were obtained, the search was repeated with an E-value cut-off of E-55, or subsequently increased values until 10 or more hits were found. The target/template MEROPS sequences were replaced with the PDB ATOM sequence for consistency with structural mapping. A seed multiple sequence alignment was then generated using MUSCLE, prior to removal of redundancy (hits above 95% sequence identity). The HMMer suite (Eddy, 1998) was then used to build and calibrate the HMMs, prior to a second search of the MEROPS sequences using hmmsearch with an e-value of 0.0001. New sequence hits were then aligned to the HMM using hmmalign and the new HMM was recalibrated. This method generated HMMs or multiple sequence alignments that were used in the Profile-Profile method, the Sequence-Profile method, the MUSCLE algorithm and the COACH method. This protocol for building the HMMs can be seen in figure 4.5.

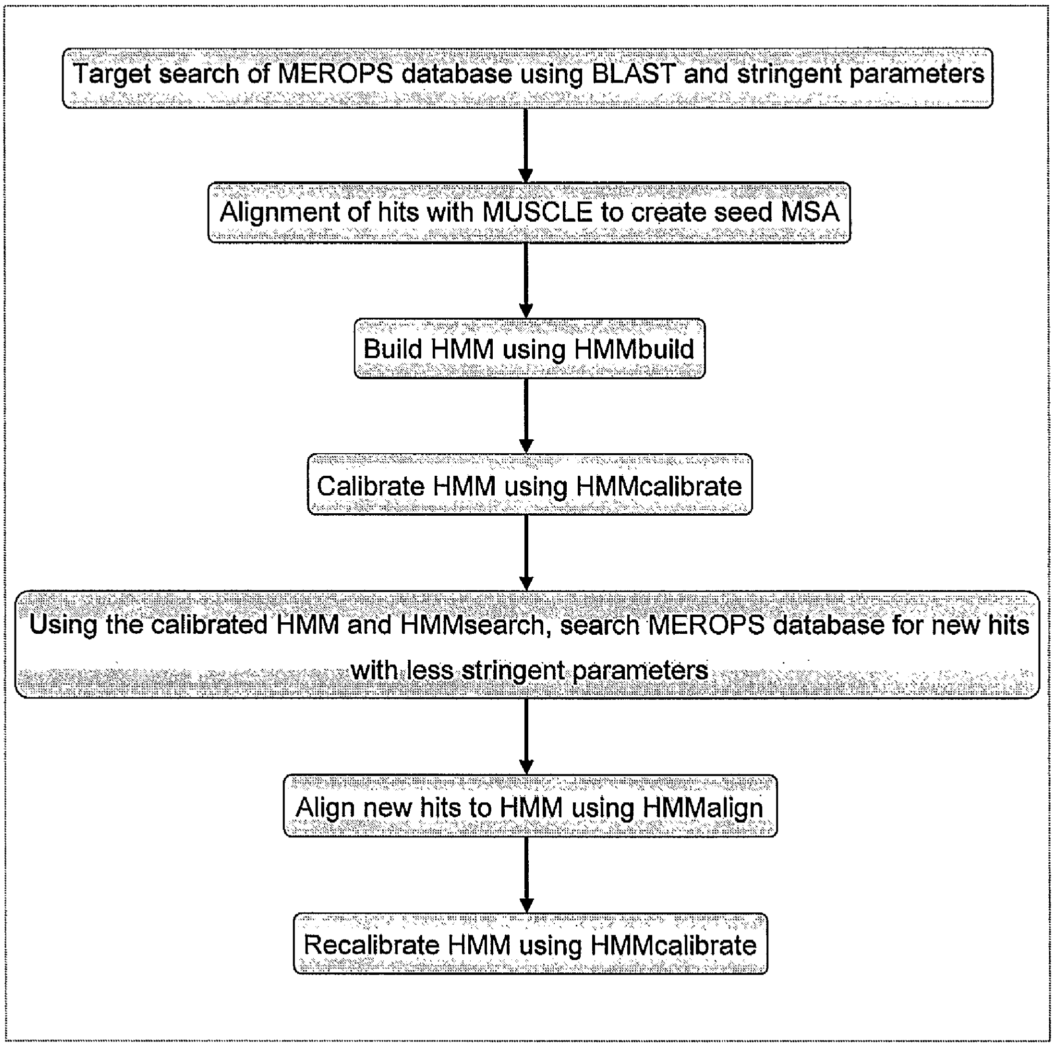

Target search of MEROPS database using BLAST and stringent parameters
Alignment of hits with MUSCLE to create seed MSA
Build HMM using HMMbuild
Calibrate HMM using HMMcalibrate
Using the calibrated HMM and HMMsearch, search MEROPS database for new hits with less stringent parameters
Align new hits to HMM using HMMalign
Recalibrate HMM using HMMcalibrate

Figure 4.5. The Construction of the HMMs. The building of the HMM is shown, the different steps and programs in the HMMer used are also displayed.

4.3.2.9 Sequence-Profile

The target sequence was aligned to the template HMM, constructed as above, using hmmalign from the HMMer suite. The target and template were then extracted as a pair-wise alignment.

4.3.2.10 Profile-Profile

A program was provided by Craig Lawless (Personal communication, University of Manchester, Bioinformatics group) that aligned two profiles using the method described by Sali and colleagues as optimal (Marti-Renom *et al.,* 2004). This uses the correlation coefficient calculated between the columns of the HMMs to populate a distance matrix prior to dynamic programming to find the best match pathway. The target-template pair-wise alignment is then extracted from the trace back path in the dynamic programming algorithm. The HMMs of the target were generated in the same way as the template HMM, created as above in figure 4.4, and were submitted to this program as well the sequences of the target and template.

4.3.2.11 COACH

Again, the MSAs generated for the HMMer suite were used to construct HMMs for template sequences using the COACH algorithm (Edgar & Sjolander, 2004). The target multiple sequence alignments were then aligned to this HMM using COACH, extracting the pair-wise alignment between the given target and template for consideration. COACH was used to also represent profile-profile method as well as the above method obtained from personal communication.

4.3.3 Assessing the Alignments

The different pair-wise sequence and profile-based alignment methods for each set (set I-vs-S and set I-vs-I) were assessed against each of the gold standards (TM-align, CE and MAMMOTH). These gold standard structural alignments were also assessed against each other to measure the level of agreement (and hence estimate a structure-based error) between them. The alignments were also assessed using the NiRMSD measure (Armougom *et al.*, 2006) using each target-template sequence pair, together with the target and template corresponding PDB protein structures. This is a normalised measure of alignment accuracy which also accounts for length and gaps, and provides a better metric for inter-alignment comparisons.

4.3.3.1 Obtaining Equivalent Residues

When each alignment was assessed against a gold standard it was important to make sure that equivalent amino acids were compared, for example that Serine 123 of the alignment method was compared to Serine 123 (or it's equivalent residue, the number may differ if some of the alignment has been lost) of the gold standard.

TM-align Alignment

```
TARGET       A1 C2 D3 G4 A5 L6 L7 M8 - - -
TEMPLATE     -  -  A1 G2 Y3 L4 M5 L6 - - -
```

BLAST Alignment

```
TARGET       A1 C2 D3 G4 A1 L2 L3 M4 - - -
TEMPLATE     -  -  A1 G2 Y1 L2 M3 L4 - - -
```

Figure 4.6. Calculating Equivalent Residues: an Example. The sequence residues of the target and template are represented as a single letter and are numbered to enable equivalent residues of the BLAST alignment compared to the TM-align alignment to be found. The green residues show the residues which BLAST has chopped off with respect to the gold standard TM-align.

As shown in figure 4.6 above, A1 (shown in blue) of the target in the BLAST alignment is equivalent to A5 (shown in blue) in the target in the TM-align alignment (both are followed by LLM...) but the sequence numbering does not correspond (the BLAST position is 1, because the beginning of the alignment has been chopped off by BLAST: the green residues, but the TM-align position is 5). Therefore, to find equivalent residues, MUSCLE was used to align the target sequence of the alignment protocol against the target sequence of the gold standard protocol, the same method was applied to the template.

4.3.3.2 Treatment of Gaps

Gaps in the target or template sequence in the alignments were considered correct if there was also a gap in the same position in the gold standard structural alignment.

4.3.3.3 Percentage Identity Calculations

For comparative purposes, target-template pairs were separated into different percentage identity bins. The percentage identity of the gold standard TM-align pair was used for this, since this more accurately reflects the similarity between the two proteins, binning all pairs less than a given cut-off (<50%,

<40%, <30% etc).

4.3.3.4 Amount of the Alignment Retained by the Different Methods

Some of the alignment protocols produced local alignments, and hence some sequence was "lost". Therefore, the target and template sequences retained (See figure 4.7) by the various alignments were calculated by comparison to the reference target/template sequences (the actual sequence of the target and template before being submitted to an alignment method). This is an important step since it is undesirable to obtain 100% accuracy in an alignment if only 20% of the original sequence is retained. More importantly for this study, the alignment section that has been lost should not contain a portion of the recognition (interface) region. The assessment of the alignment methods thus included the number of correctly predicted residues in terms of how much of the original alignment sequences were retained.

4.3.3.5 Calculating the Sensitivity and Specificity

The number of correct pairs divided by the number of pairs in the predicted sequence alignment is effectively a measure of the specificity of the alignment, and the number of correct pairs divided by the number of pairs in the structure-based alignment, is a measure of the sensitivity of the predicted alignment.

4.3.3.6 Amount of the Structural Overlap

The structural overlap was obtained by comparing a given alignment with the structural alignment obtained from the Combinatorial Extension (CE) algorithm, the MAMMOTH algorithm and the TM-align algorithm. The percentage overlap between the two alignments was calculated as the percent of residue pairs aligned the same way in both alignments. To allow the different algorithms to be compared and general trends to be observed, the average values over the entire method was calculated.

4.3.3.7 Amount of the Alignment that is "Model-able"

To allow the alignments to be assessed in terms of how well they fare with respect to comparative modelling, the percentage of the sequence that was considered 'model-able' (See figure 4.7) was calculated. When MODELLER produces models based on the template structure, if the target has sequence present at the ends of the alignment but the template does not then MODELLER will just build loops with basic angles since there is no template structure to use. If only the template has sequence at the ends of the alignment MODELLER will not build anything, the model will start from where the target sequence starts. Figure 4.7 illustrates this in more detail.

Gold Standard Structural Alignment.

```
>Target
IVEGQDAEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTVDDLLVR
>Template
LIDGKMTRRGDSPWQVVLL-DSKKKLACGAVLIHPSWVLTAAHCM-----DESIRK------
```

Alignment produced by a given method (The "Retained" Alignment).

```
>Target
------AEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTVDDLLVR
>Template
LIDGKMTRRGDSPWQVVLL-DSKKKLACGAVLIHPSWVLTAAHCM-----DES---------
```

"Model-able" Alignment.

```
>Target
------AEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTVDDLLVR
>Template
LIDGKMTRRGDSPWQVVLL-DSKKKLACGAVLIHPSWVLTAAHCM-----DES---------
```

Figure 4.7. The Different Ways to Consider the Alignments. The gold standard alignment contains the full, original sequences of the target and template. The alignment method is the resulting alignment of, for example, one of the local alignment methods that has lost some of its sequence from the target and template. The model-able alignment is the amount of the alignment that is model-able resulting from the alignment method. The red residues in the alignment method sequence correspond to the residues that the alignment method has retained compared to the gold standard. The blue residues in the model-able sequence show the residues which are model-able (no overhanging gaps) in the alignment. For a definition of the "retained", "model-able" and "reference" alignments please see the abbreviations list.

4.3.3.8 Amount of the Alignment that is Unaligned Gaps

Gaps in alignments are usually portrayed as a disadvantage; affecting the alignment step and the modelling step negatively. To enable a greater understanding of the models produced, and any reasons for poorer quality models, the percentage of gaps and number of gapped instances in the target and the template sequence was calculated. The number of gapped regions was assessed as shown in figure 4.8.

```
>Target
------AEVG---FRKS—ELASLISD-RWVLTAAHCL--YPPVDDLLVR---
```

Figure 4.8. Counting Gap Instances
The above target example would be considered to have a total of 16 gaps and 6 gapped instances.

4.3.4 Obtaining and Defining the Interface Regions

For each dataset the interface regions on the enzyme were defined using the DACCESS program. DACCESS calculates the differential residue accessible surface area between multiple chains in a PDB protein structure file (in this case the peptidase chain and the inhibitor chain). Differential residue accessible surface areas greater than 5Å^2 were considered to be interacting. The residues in these protein surface areas were later assigned to the actual interface category, represented by "2" (See Figure 4.9 and Figure 4.10). In the I-vs-I set, it is possible for both the target and the template to have residues in the actual interface category "2", since the target peptidase PDB file contains an inhibitor chain, as does the template PDB file. However, for the I-vs-S set it is only possible for residues in the target to be assigned as an actual interface (category "2"). The template can not have any residues in this category as its PDB file only contains a single peptidase chain. Residues either side of an actual interface (category "2") were considered to be part of a wider interface region (category "1"). All other residues were defined as non-interface residues (category "0"). Variations of how to combine the interface regions from each aligned pairing of the target and the template were tested. It was decided that an interface between residues in the target and template would be assigned if the target residue position was an actual interface, category "2", regardless of what category the template residue position was. An example can be seen in figure 4.9, where the "Y"s indicate that a residue position in the target and the corresponding residue in the template would be considered as an interface between the target and template.

The I-vs-I set.

	N	N	N	Y	Y	Y	Y	N	N	N	N
Target	0	0	1	2	2	2	2	1	0	0	0
Template	0	0	0	0	1	2	2	1	0	0	0

The I-vs-S set.

	N	Y	Y	Y	Y	N	N	N	N	N	N
Target	1	2	2	2	2	1	0	0	0	0	0
Template	0	0	0	0	0	0	0	0	0	0	0

Figure 4.9. Defining the Interface. A single categrory was assigned to each residue in the target and template. Category "2" was assigned if there was an interaction between enzyme and inhibitor at a distance less than 5Å, and category "1" w as assigned to adjacent positions. Category "0" indicates a non-interface residue. If more than three "2"s were found in the seven residue window about any given position, then that position was assigned as an interface (a "Y").

A seven residue sliding window was used to determine the extended interface regions using the target sequence (the "Y"s). If there were more than three "2"s in the seven residue window, then the fourth residue in that window was assigned as an interface residue (a "Y"). The rest of the residues are considered as non-interface residues ("N"s). An example using the first five sliding windows is shown in figure 4.10.

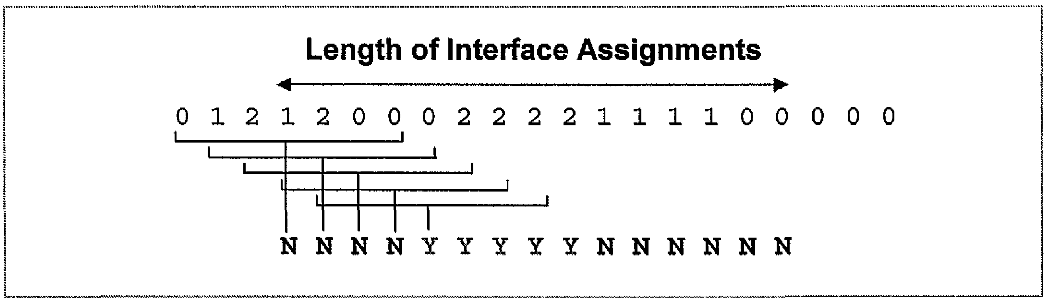

Length of Interface Assignments
0 1 2 1 2 0 0 0 2 2 2 2 1 1 1 1 0 0 0 0 0
N N N N Y Y Y Y Y N N N N N N

Figure 4 .10. Example of Assigning the Interface. The interface/non-interface is assigned at position four in the centre of the seven residue sliding window.

A seven residue window was chosen because six is similar to the size of protease inihbitor recognition loop and the number of residues that is structurally conserved in such a motif in serine protease (Hubbard *et al.*, 1991). The sliding window provided consistency as the RMSDs used the same sliding window. This was extended by one residue to seven since this is an odd number and symmetrical about the window centre. The '1's were used in the initial assessment of the interface but were later discarded as they did not provide any extra information about the interface. We used a window to define the interface to test whether the interface is aligned and modelled better than the non-interface. However, if we just took the structurally defined regions of the interface then we would have an imblance, both in terms of numbers of residues, and the discontinuity (in sequence terms) of the regions. So to normalise we use fixed segments of sequence (7-mers). Structures were used to define the interfaces and this was then mapped back to the structures.

4.3.4.1 Assessing the Interface Regions

As some alignment methods lost part of the target or template sequence, it was important to check how much of the interface-assigned residues, if any, had been lost. Hence, the percentage of the extended interface regions (categories "1"s and "2"s) for the target and/or template was calculated for each method, for the full alignment and also for the "model-able" regions.

All alignment methods were compared to the three (structural) gold standards. The following percentages were calculated: the percentage correct (with respect to the gold standards) of the actual interface residues ("2"s) assigned to the target and template for both the I-vs-I and the I-vs-S set, and the percentage correct of the extended regions (using the seven residue window) for both sets. The percentage correct in terms of the amount of sequence retained, the original sequences and the model-able alignment was calculated as well.

4.3.4.2 Gaps in the Interface

The number of gaps and gapped instances that were introduced into the target interface regions was calculated as in figure 4.11. Gaps were considered to interrupt the interface region if they came between any "1"s or "2"s.

```
Target Sequence   AEVG---FRSKS-ELLCGAD-RVLHCL--LYPPWDKNFTV
Interface Code    0012---22100-0000122-122221--21000000000
```

Figure 4.11. Counting Gap Instances. The above target example would be considered to have a total of 7 gaps and 4 gapped instances, with 3 gapped instances in the interface region.

4.4 RESULTS AND DISCUSSION

When assessing the different alignment protocols the amount of sequence retained (since some local methods frequently align only sub-sequences, please see the abbreviations list or Figure 4.6 for an explanation of the retained and modelable alignments) was determined. In addition, the amount of "model-able" sequence and the number of gaps introduced into the target or template was also calculated. The resulting alignments from the different methods were assessed in comparison with the different gold standards to determine how much of the alignment, and/or interface region, was predicted correctly, again, in terms of the amount of sequence retained and the amount of sequence model-able.

4.4.1 The Amount of Sequence Retained

Initially it was established whether each alignment protocol, when compared to all of the gold standard structural alignment protocols, retained 100% of the submitted sequence. The gold standards were also compared to each other (for retention). The results can be seen in Table 4.1 where each method is split into two boxes, one for the target and one for the template.

I-vs-S Set. Gold Standards.

	CE		TM-align		MAMMOTH		Reference	
CE			N	N	N	N	N	N
TM-align	N	N			Y	Y	Y	Y
MAMMOTH	N	N	Y	Y			Y	Y

I-vs-S Set. Other Methods.

	BLAST		COACH		MUSCLE		Profile-Profile		PSI-BLAST		Sequence-Profile	
Reference	N	N	Y	Y	Y	Y	N	N	Y	N	Y	Y

I-vs-I Set. Gold Standards.

	CE		TM-align		MAMMOTH		Reference	
CE			N	N	N	N	N	N
TM-align	N	N			Y	Y	Y	Y
MAMMOTH	N	N	Y	Y			Y	Y

I-vs-I Set. Other Methods.

	BLAST		COACH		MUSCLE		Profile-Profile		PSI-BLAST		Sequence-Profile	
Reference	N	N	Y	Y	Y	Y	N	N	Y	N	Y	Y

Table 4.1. Methods that Retain 100% of their Sequences. For both of the datasets, the amount of sequence retained after submitting the target-template sequence pairs to the methods was calculated against the reference sequence (the original sequence before submission to the method). The gold standards were also assessed against each other. Each method result is split into two boxes, the first box is the result for the target and the second box is the result for the template. A "Y" represents that that particular method did retain 100% of its sequence, an "N" indicates that it did not.

Referring to table 4.1, the gold standards retained 100% of both the target and the template sequence compared to the reference sequence and to one another, except CE. This is understandable since it is a local alignment method and MAMMOTH and TM-align are not. The gold standards TM-align and MAMMOTH retained 100% of the submitted reference sequences and so it is expected that they retained 100% with respect to each other. COACH, MUSCLE and the Sequence-Profile method all retained 100% of both the target and the template sequence. BLAST, which effectively produces a local alignment, and the Profile-Profile method retained less than 100% for both the target and the template sequences. PSI-BLAST retained 100% of all of its target sequences but not its template sequences. These observations hold true for both the I-vs-I set and the I-vs-S set. These results highlight the assessment issues required when comparing modelling protocols, since some will truncate the target and/or template sequences and not all the target sequence can therefore be modelled.

4.4.2 The Retained Alignment and Percentage Identity

For a greater understanding of the amount of sequence each alignment protocol retained compared to the original, full sequence, the percentage retained of the sequence was split into ten percent bins and the number of sequences in each bin was determined, figure 4.12.

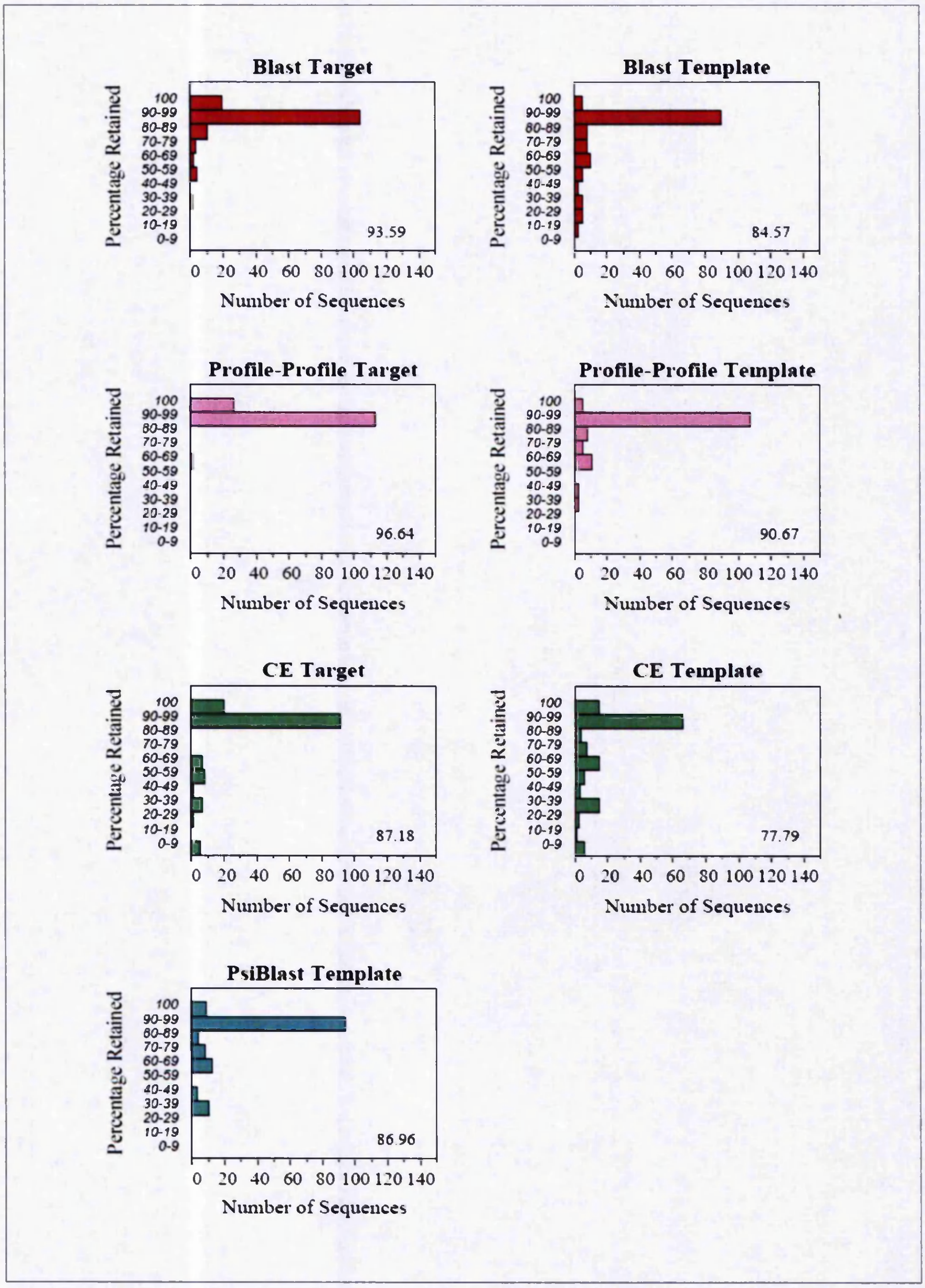


Figure 4.12. The Percentage of Sequence Retained Between the Target and the Template by the Different Protocols. The percentage of sequence retained for the targets and templates of the different alignment protocols for the I-vs-S set are shown, with the average percentage retained over all the sequences displayed in the lower right corner of each graph. Only those methods with average values below 100% sequence retained are shown.

Figure 4.12 shows the percentage retained of the target or template sequence below 100%, and demonstrates the local aspect of some alignment techniques; some occasionally retain little of the original sequence submitted to the alignment protocol. This would be no cause for concern if the goal was only to build a model based on a local alignment. However, this is not the case and indeed, on average, less template sequence is retained than target with these methods. Despite this, in the majority of cases, above 90% of the sequence is retained. The most sequence "lost" is from CE where around 8 sequences (out of 144 sequences) have between 0 and 9 percentage sequence retained for both the target and the template. CE uses aligned fragment pairs and thus tends to opt for shorter alignments with lower RMSDs. The target-template pair which loses 80% or more of its target and template sequence in the BLAST method includes the target 2kaiA and different templates. 2kaiA is a relatively small sequence (80 residues in length) and the templates are larger (around 200 residues). BLAST sometimes chops residues from the template as it is a local alignment search method. This may be the reason for the loss of sequence in the templates for the other pairs with different targets and templates. For the I-vs-I set, all of the different alignment methods retained over 60% of their target and template sequences, with averages over all the alignments above 90%.

It was expected that the local alignment methods retained less of the sequence as the percentage identity of the target-template pair decreased, that is, as the alignment becomes increasingly difficult, figure 4.13. The number of sequences that are in each percentage identity bin for the I-vs-S set used are displayed in table 4.2.

	Number of Sequence Pairs	
Percentage Identity Bin	**I-vs-S**	**I-vs-I**
<20	24	0
<30	51	1
<40	121	20
<50	135	21
<60	138	21
<70	142	23
<80	144	26
<90	144	26
<=100	144	26

Table 4.2. Number of Sequences in Each Percentage Identity Bin. For the I-vs-S and I-vs-I set, the number of protein sequence pairs in each of the percentage identity bins are shown, the bins are inclusive. The pairs with percentage identities above 80%, thus seen as trivial pairs to align, were removed originally.

Since the I-vs-I set contains less sequences than the I-vs-S set (Table 4.2) and no pairs with percentage sequence identities below 20%, the majority of the alignment assessment takes place on the I-vs-S set; the aim of this study is to try and assess how well alignment methods compare in the twilight zone.

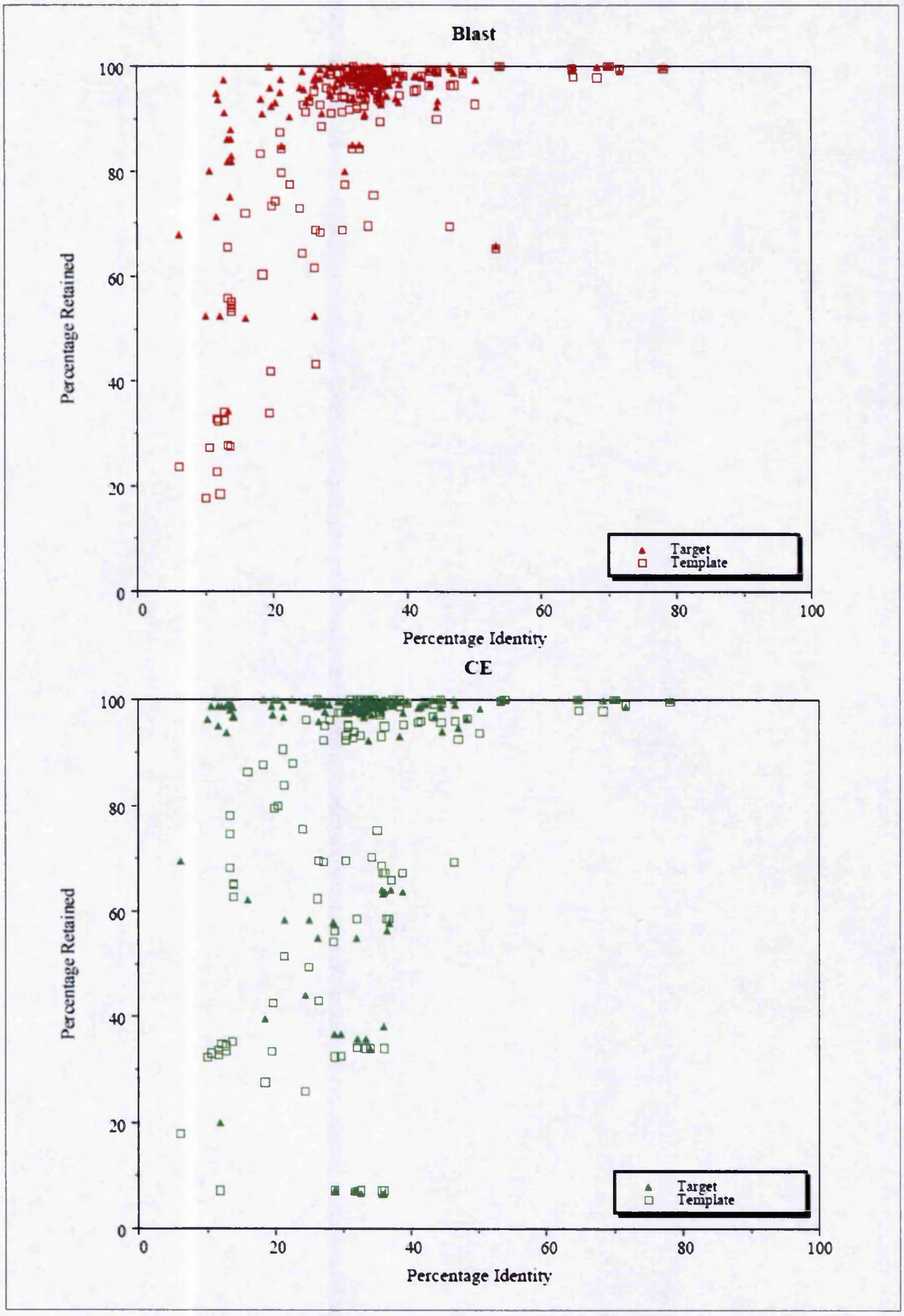


Figure 4.13. The Percentage of Sequence Retained at Different Percentage Identities. The percentage retained of the target and template sequences *versus* the percentage identity of the aligned pair (from structural alignment). Data from the CE and BLAST alignment methods are displayed only from the I-vs-S set.

Referring to figure 4.13, BLAST, as expected, retains less of the sequence as the percentage sequence identity decreases. The more difficult the alignment, the more of the sequence is not aligned. This holds true for all of the local alignment methods (including Profile-Profile, CE, BLAST and PSI-BLAST) in both sets. Only BLAST and CE from the I-vs-S set are shown as they contain the most sequences with below 20% of the alignment retained (see appendix 2, figure A2.1, for the graphs for the Profile-Profile method and for the PSI-BLAST alignment method). A larger fraction of the target tends to be retained compared to the template and more of the target and template sequence is lost from the CE alignments, in comparison to the other methods, at higher sequence percentage identity.

4.4.3 The Amount of Interface Retained

Loss of target or template sequence may prove problematic when assessing alignment methods for comparative modelling if a part, or the entire interface region, has been discarded by local alignment methods.

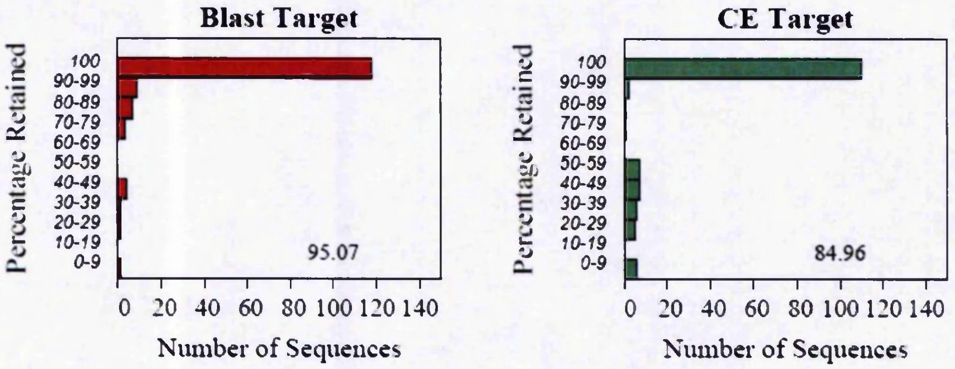


Figure 4.14. The Percentage of the Interface Retained. The percentage of the interface sequence retained for the targets and templates of the different alignment protocols for the I-vs-S set, with the average percentage retained over all the sequences are shown. Only the positions assigned category "2" were considered. Only those methods with below 100% interface retained are displayed. There are no interface residues for the templates in the I-vs-S set.

Retaining interface residues is important for this study, and figure 4.14

shows that in general the interface residues are retained, even in the cases where the whole sequence is not retained 100%. There are some exceptions for the BLAST and the CE target sequences, where worryingly in a few instances (more with the CE alignment method), less than 10% of the target interface residues remain. In the I-vs-I set all of the alignment methods retain over 70% (except CE, retaining above 40%) having an average interface retained of over 90%.

4.4.4 The Number of Gaps Introduced

Each alignment method will introduce INDELs (insertions and deletions) into the target-template sequence alignment, which were characterised for this project. Modelling a target-template pair becomes more of a challenge if there are a large number of gaps in the pair. Gaps in the template sequence would mean those parts of the target do not use the restraints provided by the template. For the target sequences in the I-vs-S set, MUSCLE, COACH, the Sequence-Profile method and the Profile-Profile method all introduce more than 20% (and less than 50%) gaps with respect to the target sequence (see appendix 2, figure A2.2, for the graphs). BLAST and PSI-BLAST alignments do not contain any target sequences having greater than 20% gaps, with an average for BLAST of 4.35% and PSI-BLAST of 4.23%. On average, Sequence-Profile introduced the most gaps with 18.19%, followed by COACH with 16.71%, TM-align with 15.70%, MUSCLE with 15.22% and Profile-Profile with the least at 11.02%.

In the I-vs-S set, the templates generally contain fewer gaps than the targets, with the single exception of PSI-BLAST, containing more than 20% gaps in some of its template sequences (see appendix 2, figure A2.3). The Sequence-Profile method has the largest number of gaps on average for the template sequence with 6.77%, followed by Profile-Profile with 5.31%, COACH 5.16%, TM-align 4.10%, MUSCLE 3.50% and finally with BLAST introducing 2.90% of gaps on average into the template sequence.

The gold standards CE and MAMMOTH alignments never contain more than 20% of gaps in their target or template sequences, with the majority containing less than 10%. On average CE contains 5.31% and 3.94% in their target and template sequences respectively. MAMMOTH contains 15.26% and 3.61% gaps in its target and template sequences respectively. TM-align, however, contains up to 50% gaps in some of its target and template sequences for the I-vs-S set.

The I-vs-I data set target-template alignments contain fewer gaps in general, with none of the methods (including the gold standard methods) introducing gaps constituting greater than 15% of the alignment into either the target or the template sequence. There are fewer sequence pairs in the I-vs-I set and on average they have higher percentage sequence identities, which explains this result. A target-template pair sharing remote homology will, in general, be more difficult and tend to accumulate more INDELs than a pair sharing modest sequence similarity.

4.4.5 Gaps and Percentage Identity

As expected, there is a negative correlation between the percentage sequence identity of the target-template pair and the percentage of gaps in the target or template. As the sequence identity decreases, the number of gaps increases. This trend can be seen for TM-align in the I-vs-S set in figure 4.15. The correlation is much stronger for the templates for every method, which have a higher percentage of gaps as described above. The I-vs-I set was not represented here as there are too few pairs having lower percentage sequence identities and this is the area of interest.

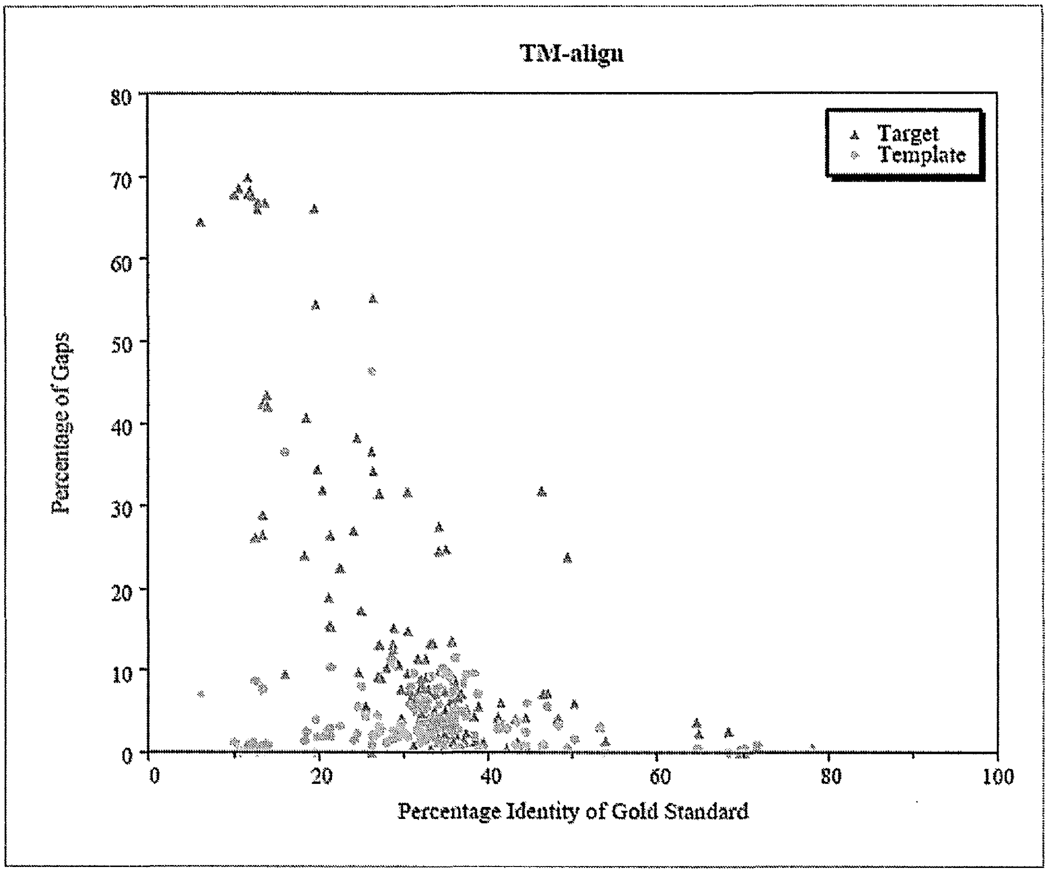


Figure 4.15. The Percentage of Gaps Introduced at Different Percentage Identities. The percentage of gaps introduced into the target and the template sequences (calculated as a fraction of the sequence length) plotted against the percent identity of the target-template pair..

4.4.6 Gapped Instances

The number of gapped instances (number of INDELs) was obtained for the target of each method, shown in figure 4.16. COACH, MUSCLE, Profile-Profile, Sequence-Profile, CE and TM-align all contain more than eleven gapped instances (a gapped instance is not how many INDELs there are individually, but how many INDEL regions exist in total) in their target sequences in the I-vs-S set. The Profile-Profile method introduces up to 19 gapped instances, sharing the highest number on average, of seven gapped instances with the Sequence-Profile method. There are less gapped instances introduced into the templates than the target sequences. COACH, Profile-Profile, Sequence-Profile and TM-align contain more than eleven INDELs in their templates in the I-vs-S set (Figure 4.17). In the I-vs-I s et, for all the methods and both the target and template sequences, n o more than nine gapped instances were ever introduced (except for the TM-align target sequences which contained up to thirteen gapped instances). The alignments in this set all share a similar average of five or six instances.

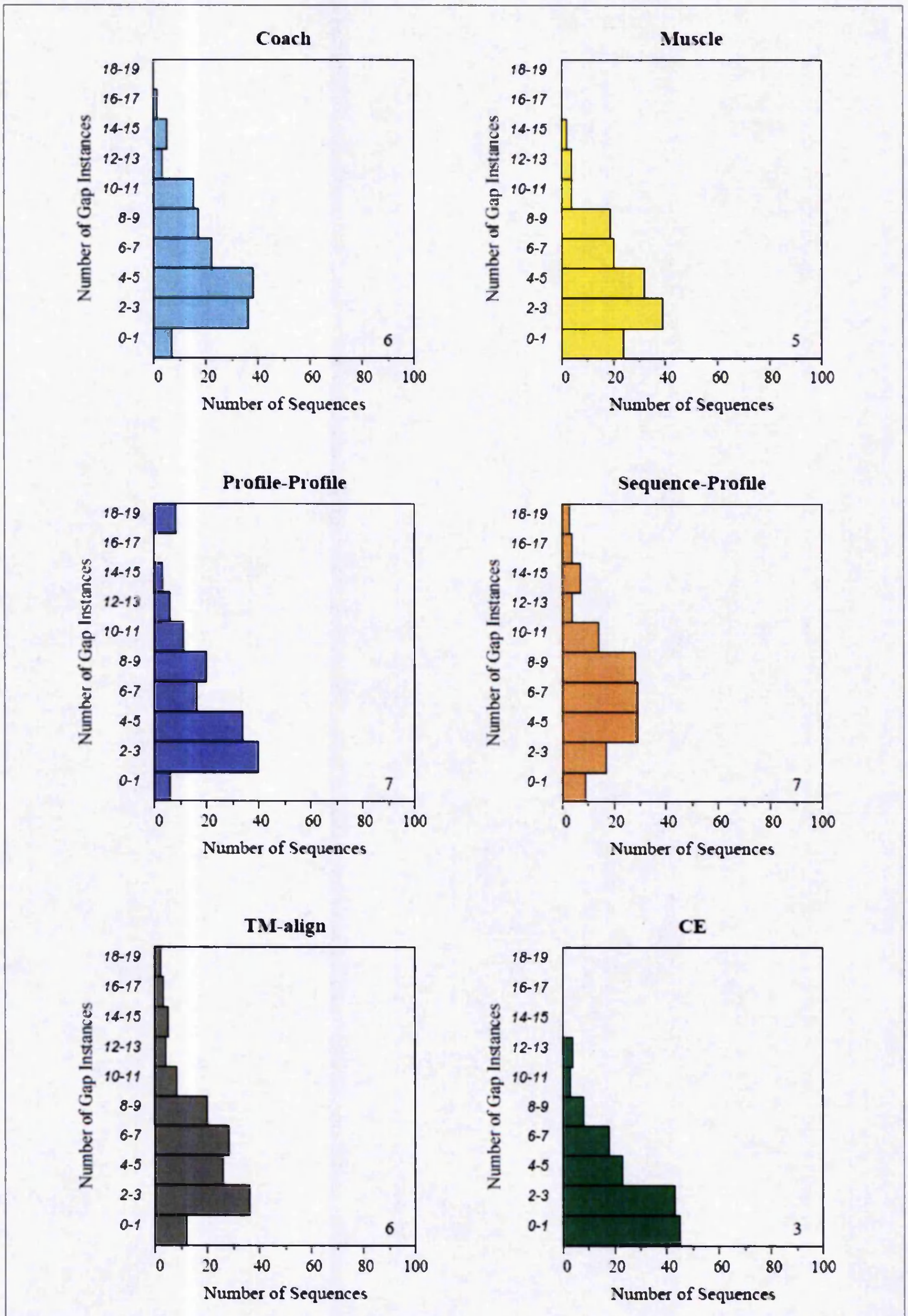


Figure 4.17. The Number of Gap Instances Introduced into the Target. The number of gapped instances introduced by the alignment methods, for the target of the I-vs-S set for the methods, containing over ten gap instances, with the average number of gapped instances (in the lower right hand corner of each graph) are shown.

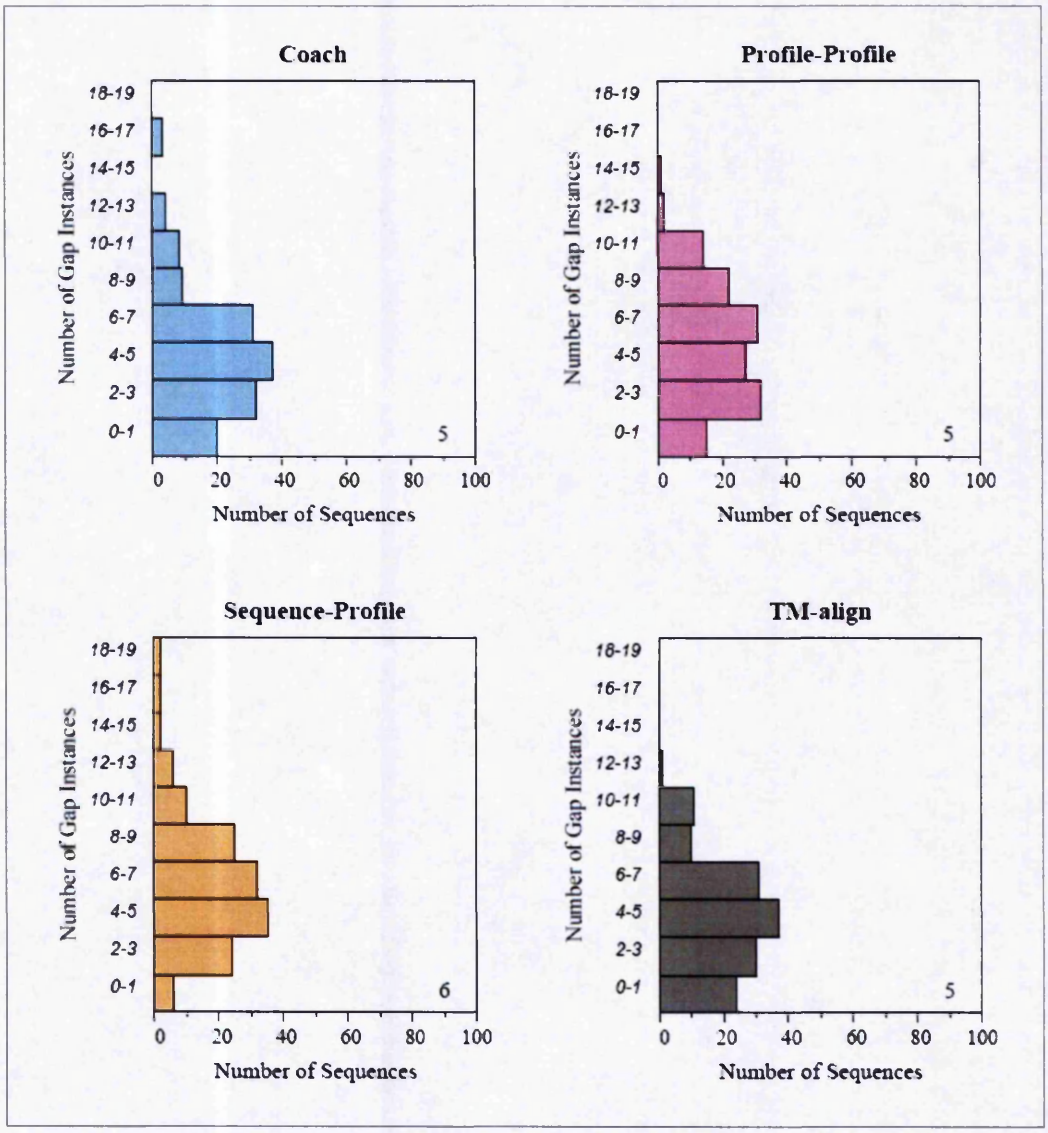


Figure 4.18. The Number of Gap Instances Introduced into the Template. The number of gap instances introduced by the alignment methods for the template of the I-vs-S set for the methods containing over ten gap instances, with the average number of gap instances shown.

A key concern was that particular alignment methods introduced large numbers of INDELs into the interface regions. This could prove disadvantageous when modelling the target, if unnecessary gapped blocks were introduced, forcing extra loop modelling to be carried out. Therefore, to assess this, the percentage of gaps in the interface regions alone was obtained.

4.4.7 The Amount of Interface Regions Containing Gapped Instances

The percentage of interface regions that contain a gapped instance is shown in table 4.3. The least insertions, as an average over all the pairs, is made by PSI-BLAST at 9.49% and the most made by Sequence-Profile at 21.79%. It is not surprising that even though they are high quality alignment techniques, COACH and Profile-Profile seem to insert gapped instances into the interface more regularly than the supposedly inferior methods of BLAST and PSI-BLAST. This may be due to the fact these are local alignment algorithms and because local alignment methods tend to chop parts of sequences off it is more likely that they will include more gaps. Of course this may be due to the use of different gap penalties and scoring schemes.

Method	Gaps (%)
Sequence-Profile	21.79
Profile-Profile	20.76
COACH	17.24
TM-align	16.09
CE	14.27
MUSCLE	12.19
BLAST	11.71
MAMMOTH	11.00
PSI-BLAST	9.49

Table 4.3. Gaps and Interfaces. The results of the I-vs-S set for the target only are displayed. Here, the percentage of interfaces that contain gapped instances are shown as an average percentage for all the target-template pairs of that method.

4.4.8 The Model-able Part of the Alignment

An alignment that results in less than 100% of the target being aligned to the template, and therefore not capable of being modelled in its entirety, should be regarded as inferior to an alignment that would have its entire target length modelled. This holds if the purpose is to model a full alignment (rather than concentrating on the specificity of the interface region). Sometimes, even though a method may retain 100% of its sequence it is possible for it to produce an alignment where all the sequence may not be modelled. This usually happens when the target or template contains gaps at the start or end of the alignment. In the I-vs-S set COACH, MUSCLE, PSI-BLAST, Profile-Profile, Sequence-Profile, MAMMOTH and TM-align all contain alignments with less than 90% of the target being model-able (Table 4.4). I n contrast, CE and BLAST methods produce alignments with 100% of the aligned target being model-able. However, BLAST and CE are likely to have alignments that are 100% model-able due to the reduced amount of the alignment that is retained. TM-align and MAMMOTH seem to have a more varied amount that is model-able. The I-vs-I set methods all produce alignments with more than 90% of it being model-able.

Method	**Percentage Model-able**							
	100	**100-90**	**90-80**	**80-70**	**70-60**	**60-50**	**50-40**	**40-30**
BLAST	144	0	0	0	0	0	0	0
PSI-BLAST	37	105	0	0	1	1	0	0
MUSCLE	54	84	1	3	0	0	1	1
Sequence-Profile	32	104	0	5	0	2	1	0
Profile-Profile	51	87	5	0	0	1	0	0
COACH	16	116	10	1	0	0	1	0
TM-align	33	88	2	3	3	2	4	8
MAMMOTH	28	92	2	3	5	2	3	9
CE	144	0	0	0	0	0	0	0

Table 4.4. The Model-able Part of the Alignment. For the I-vs-S set, the number of sequences in that percentage model-able range. The one hundred percent column is the number of sequences obtaining exactly one hundred percent model-able, the other columns contain up to the highest value in that range but not including it.

To discern whether there was a trend between the amount of alignment that was model-able and the percentage sequence identity of the pair, figure 4.18 was produced. The Pearson correlation coefficient was the highest for the

TM-align method (0.56) compared to COACH (0.31), Profile-Profile (0.15), Sequence-Profile (0.33), PSI-BLAST (0.15), MUSCLE (0.22) and BLAST (no correlation since all sequences were 100% model-able). The amount of sequence that was model-able in the alignments produced by TM-align share a more varied distribution than the rest.

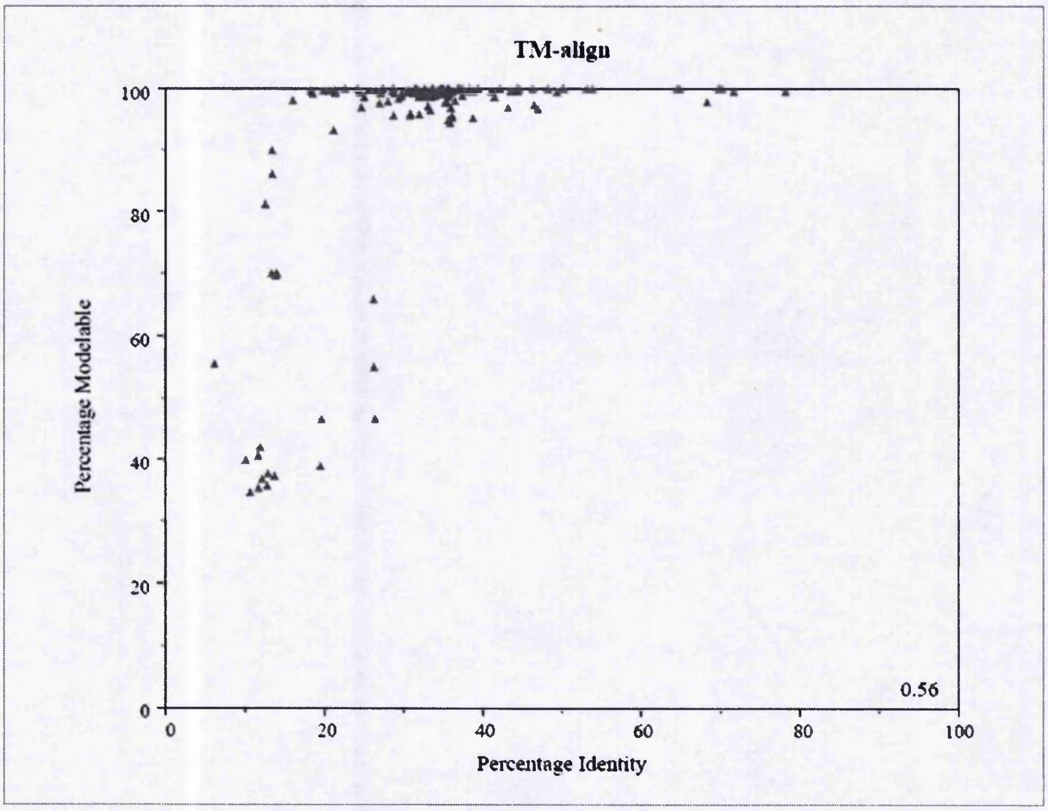


Figure 4.18. The Percentage Identity and Percentage Model-able. For the I-vs-S set and the method TM-align, the correlation (lower right hand corner of graph) between the percentage identity of the alignment pair and the percentage model-able is shown.

4.4.9 The NiRMSD of the Alignments

The normalised iRMSD was calculated for all of the alignments and averaged for each method. The results for both sets can be seen in table 4.5. In both cases TM-align has the lowest NiRMSD, which indicates the best score for the alignments. The NiRMSD figures are virtually impossible to distinguish with the exception of MUSCLE (in the I-vs-S set) and BLAST (in the I-vs-I set), being slightly worse than the other sequence-based alignments, although PSI-BLAST seems to have a small edge over the other methods.

	NiRMSD Å	
Alignment Method	**I-vs-S Set**	**I-vs-I Set**
TM-align	0.68	0.60
PSI-BLAST	0.76	0.66
BLAST	0.79	0.73
Sequence-Profile	0.80	0.67
COACH	0.81	0.69
Profile-Profile	0.82	0.67
MUSCLE	0.88	0.69

Table 4.5. The NiRMSD of the Alignments. For the I-vs-S and I-vs-I set the NiRMSD was calculated for the alignments.

4.4.10 NiRMSD and Percentage Identity

The correlation between the NiRMSD of the alignment and the percentage identity of the alignment is a negative one; as the percentage identity of the alignment pair increases, the accuracy (NiRMSD) decreases. The scores for each of the methods are as follows: BLAST -0.69, COACH -0.28, MUSCLE -0.67, Profile-Profile -0.19, PSI-BLAST -0.61, Sequence-Profile -0.61, TM-align -0.59. Since TM-align is the principal gold standard, the graph for the NiRMSD *versus* the percentage identity of the TM-align aligned pairs is shown in figure 4.19 and the graphs for the other methods can be seen in appendix 2, figures A2.4 – A2.9.

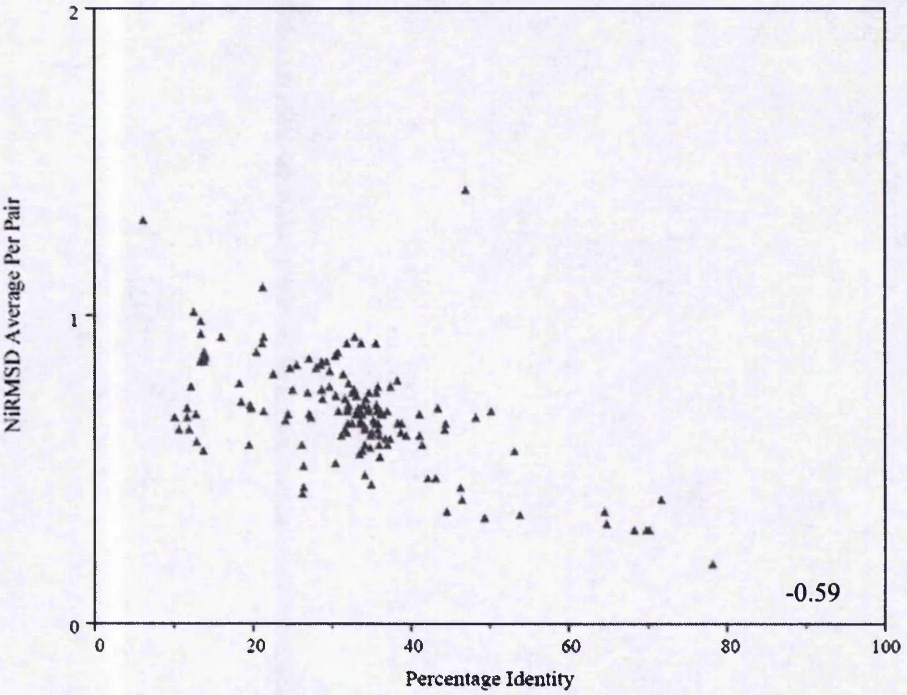


Figure 4.19. The Percentage Identity and the NiRMSD. For the I-vs-S s et, the average NiRMSD (the y-axis is measured in Ångstroms) per pair was plotted for the TM-align method against the percentage identity of that alignment pair.

4.4.11 The Accuracy of the Alignments

The percentage of correctly aligned residues compared to the gold standard TM-align alignments is shown in figure 4.20 for the different methods, over varying percentage identity bins. The results of the alignment methods compared to CE and MAMMOTH graphs are in appendix 2, figure A2.10 and A2.11. It is worth noting that since CE is a local alignment method this prevented straightforward comparisons with global alignment methods. The MAMMOTH results were very similar to the TM-align results and the latter were chosen, as the initial results were more comparable with previous investigations and MAMMOTH displayed some minor inconsistencies, deleting terminal residues from some alignments. T he percentage of correctly predicted alignment residues was calculated three different ways: the percentage of correctly predicted residues across the whole alignment, the percentage of correctly predicted residues from the retained sequence in the alignment, and the percentage of correctly predicted residues as a subset of the model-able sequence (for a more detailed explanation see t he Methods and Materials Chapter, section 4.3.3).

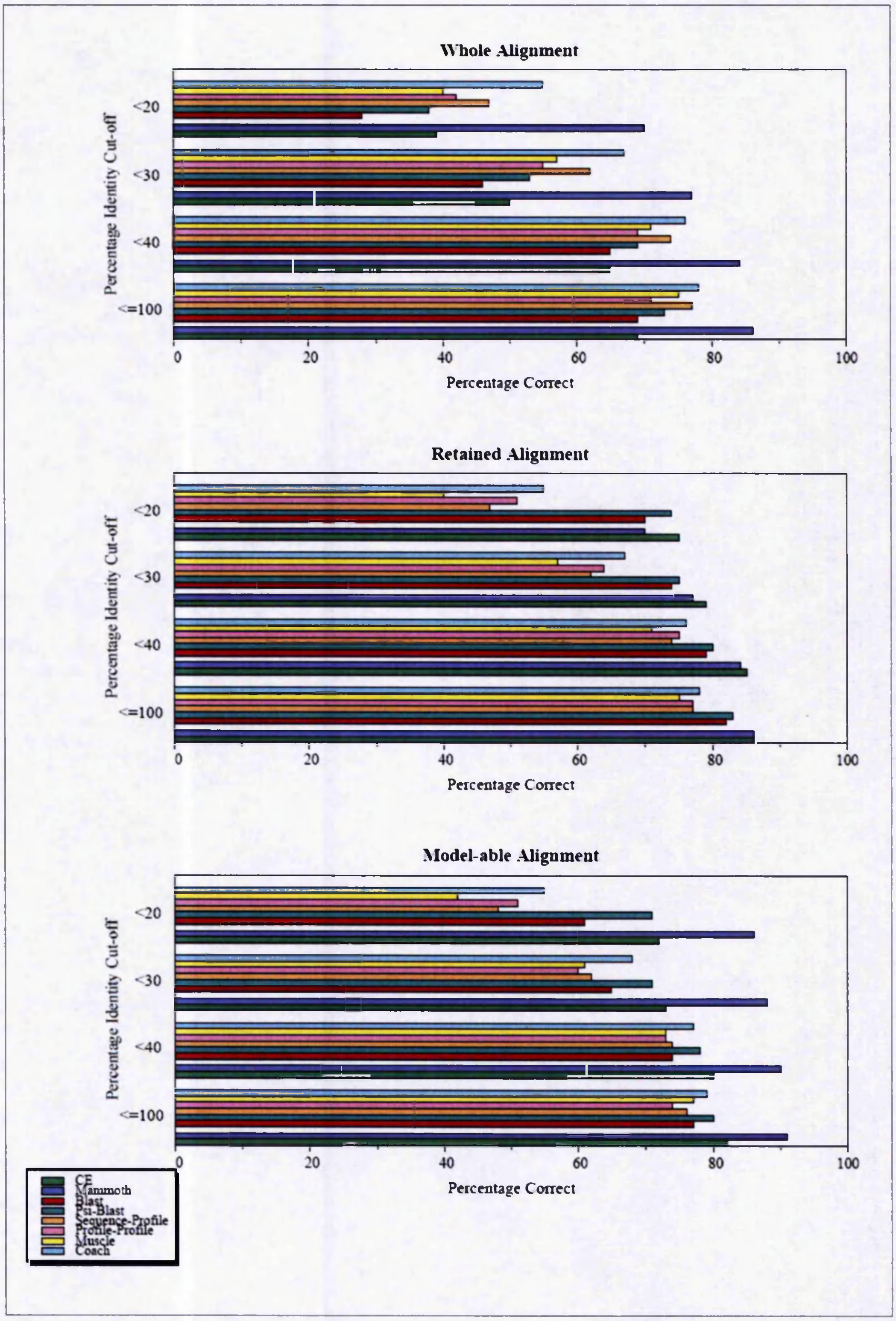


Figure 4.20. Percentage of Correctly Predicted Residues. Results for the percentage of correctly aligned residues by each alignment method, as an average across all pairs in the I-vs-S set, assessed against the gold standard TM-align are shown. The percentage identity bins are inclusive as indicated by the "less than" signs.

In figure 4 .20, for the I-vs-S set, the percentage correct for all the alignment methods (assessed against TM-align, refer to section 4.4.11 for an explanation of why TM-align was chosen) decreases as the percentage sequence identity between the target and the template decreases. The percentage correct is the amount of alignment positions in the alignment method which are correct compared to the gold standard TM-align. This also includes the gold standard structural methods. This applies to the three different techniques for calculating the percentage of correctly aligned residues (please refer to figure 3.6 for a description of the three techniques, "whole alignment/reference", "model-able" and "reference". The gold standards, MAMMOTH and CE, are placed slightly apart from the other methods to aid visualisation. It is interesting to note that the gold standards CE and MAMMOTH represent the best possible alignments that the sequence-based methods could arguably achieve since they are derived from known structures. Indeed, in the majority of cases for all percentage identity ranges (albeit, all are below 80% sequence identity) they still only assign a maximum of 80% of residues correct when being assessed against TM-align. CE fares less well since it is a local alignment method. However, MAMMOTH might be expected to achieve near parity with TM-align and it is clear here that the 2 methods typically agree on only 80% of the aligned positions. This represents a theoretical "maximum" that the sequence-based methods might achieve, as well as the theoretical error or uncertainty in aligning two structures.

Only the Profile-Profile, CE, PSI-BLAST and BLAST results are affected by considering percentage retained assessments. Although CE's, BLAST's, PSI-BLAST's and Profile-Profile's score increases, this potentially could be a problem in future modelling assessments if the segments removed were part of the recognition interface (which is only the case for targets of the BLAST and CE methods, see figure 4.14). In essence, this means that the local alignment protocols do better than expected, with this caveat in mind. In the model-able graph (Figure 4.20) the number of correctly aligned residues is represented with the amount of the sequence that is considered model-able to hopefully assess each method in terms of modelling the correctly aligned residue positions. The most noticeable difference in the third graph and the second graph is that PSI-BLAST and BLAST seem to have a marked improvement, since the amount of sequence that is model-able is similar to other methods regardless of how much of the sequences has been retained. Unsurprisingly, MAMMOTH assigns more correct residues than all other methods in the percentage model-able, meaning more of the alignment can be potentially modelled compared to the abilities of the other alignment methods. Above the 20% identity bin cut-off there is a less varied distribution of correctly predicted residues on average, with the method order (the method with the most correct residues being listed first) of correctly predicted residues staying generally the same for the whole alignment: COACH, Sequence-Profile, MUSCLE, Profile-Profile, PSI-BLAST and BLAST, and for the retained alignment: PSI-BLAST, BLAST, Profile-Profile, Sequence-Profile and MUSCLE, and for the model-able alignment: PSI-BLAST, COACH, BLAST, Sequence-Profile, MUSCLE and Profile-Profile. It is clear that MAMMOTH always out-performs the other methods (with CE usually following) as these are the other gold standards. . Between the sequence-based methods BLAST is usually the worst. Below 20% there is more variation between the methods and the percentage of correctly predicted residues drops at least 10%, but still they stay in the same order of which methods predict the most residues correctly in the alignment. The methods can predict around 70% correct residues (Table 4.6) in the retained and model-able assessment stage, which is surprisingly good.

It is generally assumed that profile or HMM-based alignment protocols will outperform the single sequence based methods. Indeed, this seems to be the case here when the alignment is assessed as a whole, but take into consideration the amount of alignment that is retained and is actually deemed model-able and the results seem to surprisingly suggest PSI-BLAST is a worthy candidate for predicting alignments, even into the twilight zone, of course the Profile or HMM methods not affected by the amount of alignment retained still remain the best alignments to use over the sequence based alignments that are also not affected by the amount of sequence retained.

	Whole %	Retained %	Model-able %
MAMMOTH	70	70	86
COACH	55	55	55
Sequence-Profile	47	44	48
Profile-Profile	42	51	51
MUSCLE	40	40	42
CE	39	75	72
PSI-BLAST	38	74	71
BLAST	28	70	61

Table 4.6. The Percentage Correct for the Different Alignment Types. For the I-vs-S below 20% sequence identity, the percentage of correctly predicted residues assessed against the gold standard TM-align is shown.

These results and general findings concur with the findings of Jaroszewski's lab (2002) who also found in their assessment of alignment methods that the distribution of alignment accuracy is very broad. Some of the alignment methods for this range of sequences were very accurate despite low sequence identity. However, none of the alignments were completely incorrect, even at low sequence identity and so all alignments were in agreement with the structural alignments to some extent. There is much less alignment accuracy variation with protein pairs with identities above the 30% threshold.

The same assessments were applied to the I-vs-I set. However, the I-vs-I set has only one sequence below 30-20% identity bin and none below 20% bin. Hence, it is difficult to observe any clear trends from such a small data set. The graphs for this set can be found in appendix 2, figure A2.12 - A2.14.

Once the accuracy of the alignments had been found, it was necessary to assess how well the interface residues were aligned compared to the rest of the alignment.

4.4.12 Alignment Accuracy of the Interface Regions

In all cases bar one (BLAST in the retained alignment), interface segments were found to be more accurately aligned than the rest of the alignment (Figure 4.21). As the percentage identity of the target-template pair decreases the difference in the amount of correctly aligned residues in the rest of the alignment compared to the interface decreases, suggesting that the interface is more conserved than the rest of the alignment, especially in the more challenging alignments (lower percentage identity). This effect is slightly reduced when the alignment being considered is the amount retained or model-able. In general, the method with the largest difference in accuracy of the alignment compared to the interface is the Profile-Profile method, with up to a 40% increase in accuracy going from the alignment to the interface. This is encouraging, since this protocol was observed to be the best for alignment in general for comparative modelling (Marti-Renom *et al.*, 2004). This current investigation, however, is biased in some respect since the amount of interface residues is far smaller than the amount of non-interface residues in the alignment and with the local alignments benefiting from an increase in accuracy over the whole alignment assessment method. However, this is lost when looking at the retained and model-able alignment.

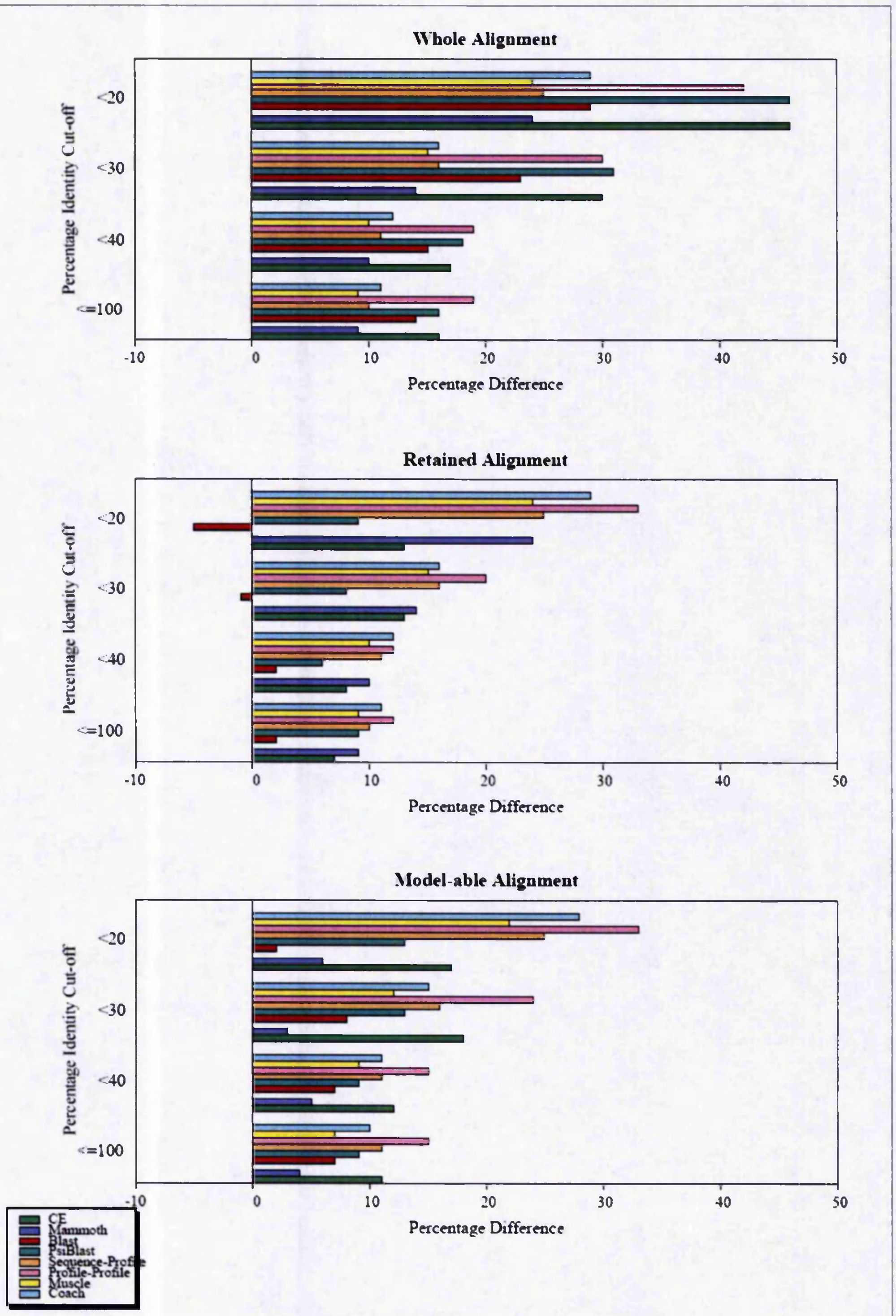


Figure 4.21. Difference in the Accuracy of the Interface Residues and Non-interface Residues. The different methods were displayed on the graph for the I-vs-S set with the average percentage identity of that bin plotted against the difference in the percentage of correctly predicted residues of the alignment (residues not included in the interface) compared to the interface region, as an average over all the pairs in that identity bin.

A clearer representation that the interface residues are consistently more accurately aligned than the non-interface region can be seen in figure 4.22. All of the methods more accurately predict the interface regions (except BLAST in the retained alignment, shown by the red coloured bar being added onto the empty, white bar). The larger the white bar is compared to the solid coloured bar, the more accurately aligned the interface is than the non-interface. It is easier to see that with more challenging alignments, the accuracy of the interface outweighs the accuracy of the non-interface residues, with the profile and HMM methods outperforming the sequence based methods. The graphs for the other gold standards CE and MAMMOTH, together with the I-vs-I set can be found in appendix 2, figure A2.15 - A2.19.

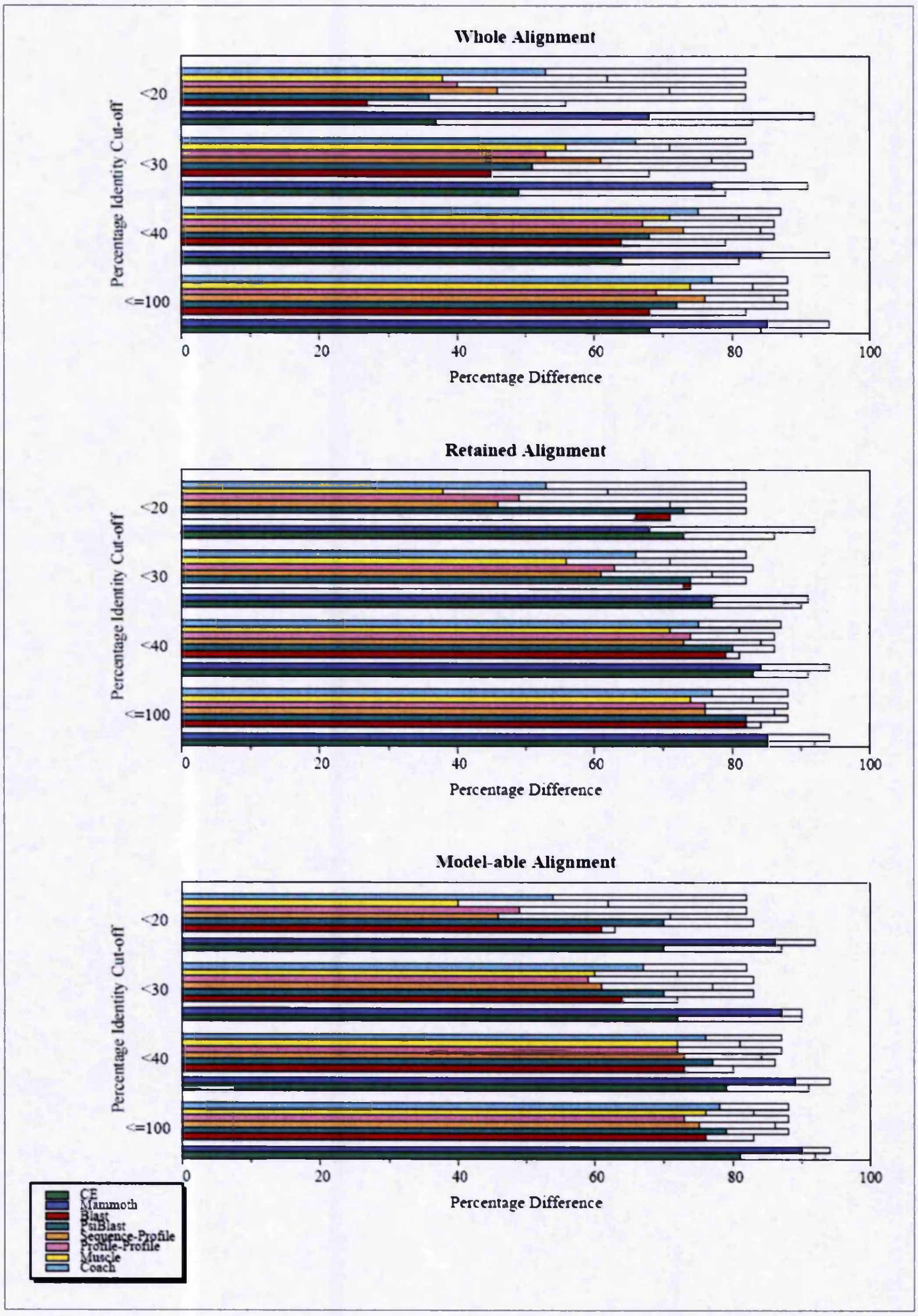


Figure 4.22. The Accuracy of the Non-interface Residues and Interface Residues. The different methods were displayed on the graph with the percentage identity bins for the I-vs-S set, plotted against the percentage of correctly predicted residues of the alignment (non-interface residues; the solid coloured bars) with the percentage of correctly predicted interface residues (the white, empty bars).

COACH aligned both the interface and the non-interface regions more accurately than the rest of the methods, not including the gold standards, and BLAST was the worst. The amount of correctly aligned residues in the interface and the non-interface decreases as the percentage identity of the alignment pair decreases, with the alignment residues being less accurately aligned than the interface residues (Figure 4.23). Althou gh the target sequence may not be aligned with a high accuracy, overall, the interface regions often are. Even at low percentage identities the interface can be aligned quite accurately. For example, using the COACH method and a target-template pair having below 20% sequence identity, the interface positions are aligned with up to 80% correct. The corresponding non-interface residues are aligned with around 50% accuracy.

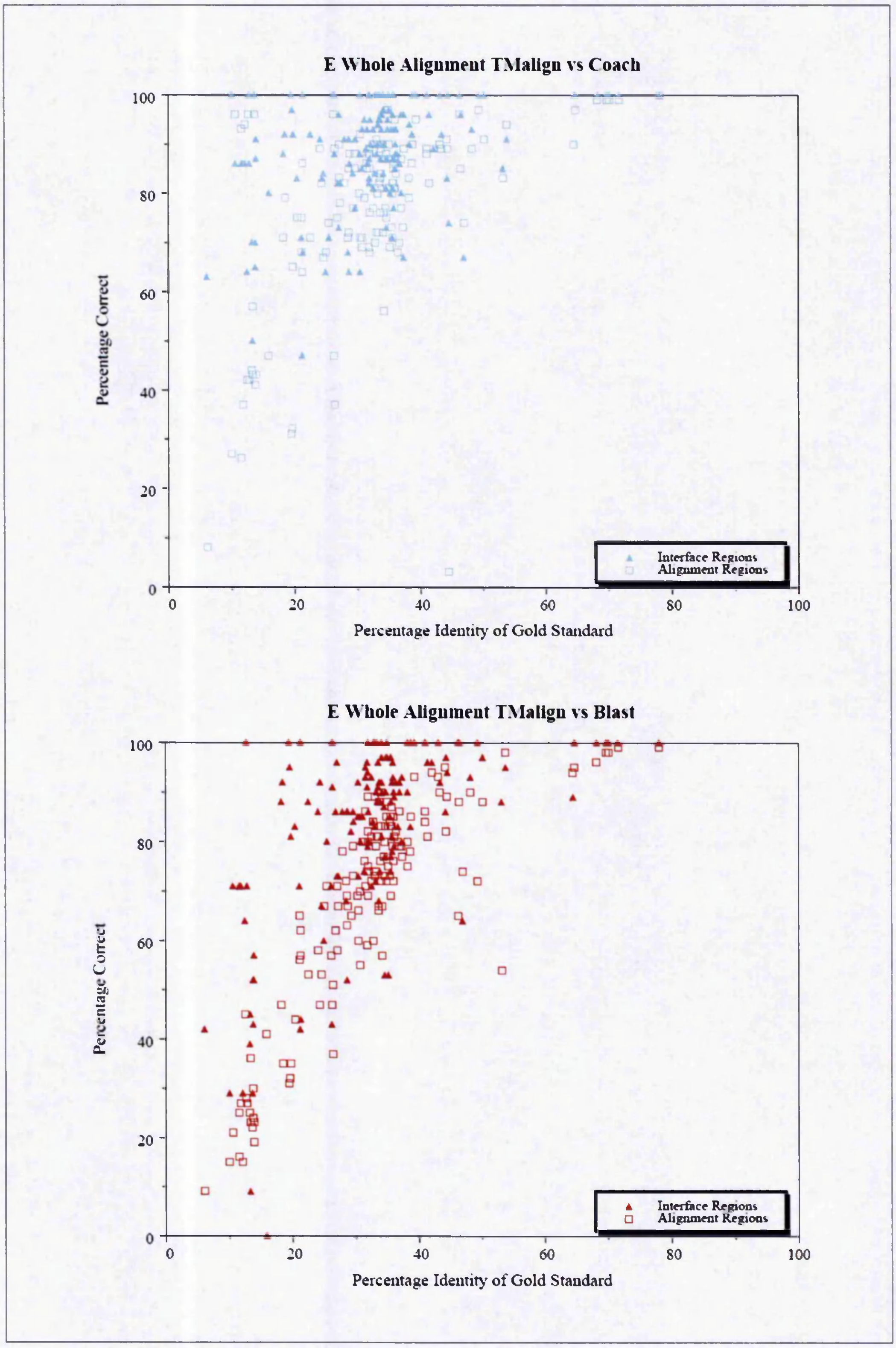


Figure 4.23. The Percentage Identity and the Percent Correct of BLAST and COACH. The individual pairs can be seen in the plots with the interface regions and the non-interface regions (alignment regions) for the I-vs-S set.

4.5 CONCLUSIONS AND FUTURE WORK

A study on the alignment accuracy of various methods, including sequence-based methods, profile-based methods and HMM-based methods, was performed on two sets of peptidase proteins. One set included pairs where the targets consisted of a protein chain bound to an inhibitor chain and the templates contained a peptidase chain only (I-vs-S). The other set included a target and a template, both with inhibitors bound (I-vs-I). Unfortunately, the I-vs-I set did not contain enough target-template pairs with percentage identities stretching into the twilight zone (below 20% sequence similarity, where this project was aimed at investigating), thus, most of the presented results reflect the I-vs-S set. The challenge was to determine whether the different alignment protocols could align the interface residues between the peptidase and the inhibitor more accurately than they could align the non-interface residues. If the interface could be aligned more accurately, this offered the potential to model the recognition regions of these proteins with higher accuracy than is normally expected in and around the twilight zone.

As the percentage identity decreased, the alignment accuracy decreased. The more distantly related the proteins are, the less conserved the structures are, making it harder to align their corresponding sequences through sequence similarity, usually incorporating more gaps into the alignment. The amount of gaps and number of gapped instances introduced into the alignment increased, as noted by other groups (Prasad *et al.,* 2003). Prasad's group also state that as the percentage identity decreases, gaps and localised regions of total dissimilarity increase in size and number, making the modelling protocol of highly divergent sequences more complex and less accurate. The global alignment methods tend to insert more gaps and retain less of the alignment. Understandably the global methods produce alignments with more gaps since they are accommodating for the difference in length between the target and the template whilst retaining the entire sequences. The local methods will discard parts of the sequence if the alignment is more challenging and the lengths of the two sequences differ greatly. CE, on average, chopped off the most residues in the alignments. Lo cal alignment methods improved in accuracy when being assessed in terms of the amount of sequence retained and the amount model-able, allowing for the loss of sequence. Problems could occur if the discarded sequence included interface residues; the important residues needed for specificity and required for modelling. Only in a few instances CE and BLAST did not retain 100% of the interface residues, with the Profile-Profile and the HMM methods containing the most gapped instances in the interface region, this could be due to the need for more gap optimisation.

Increasingly complex alignments produced more accuracy variation between the methods and a smaller model-able portion of the alignment. With the more trivial alignments, the accuracy of the different methods is, in general, only marginal. At lower percentage identities the more advanced methods are tuned for remote homology detection and aligning the sequences, and outperform the other methods, giving a broader range of alignment accuracies between the protocols.

Overall, and as expected (because of their ability to contain more evolutionary information), the profile and the HMM methods were more accurate at aligning sequences. None of the methods spanning all the sequence identities were completely different than the gold standards, and even though the structural alignment algorithms represented the gold standards (and the maximum possible accuracy achievable by the other methods) they were still not 100% correct. Interestingly, when most hope for accurate alignments diminishes, the alignments are still fairly accurate. For example, COACH predicted up to 50% of the alignment correctly, which is not ideal, but the accuracy of the interface residues was much better.

The interface was more accurately aligned than the non-interface residues in the alignment, even below 20% sequence identity. A greater difference in accuracy between the interface and non-interface at lower percentage identities existed. Even though the overall accuracy of the alignment may not be high, the accuracy of the interface region surpassed the rest of the alignment. For example, below 20% sequence identity the accuracy of the interface residues can be up to 80%, and the non-interface residues up to 50%, whereas when the alignment was assessed overall, the accuracy (made by COACH in this case) only reached a maximum of 60% accuracy. The profile and HMM based alignment protocols were better at predicting the interface than the sequence-based methods. This increase in accuracy is because the interface is more conserved than the rest of the accessible surface of the peptidase, due to the additional evolutionary pressures exerted on them, and the profile/HMM methods are better at distinguishing evolutionary related positions from non-related positions.

Alignment of a target-template pair where only the template has a structure is of great importance in comparative modelling. The alignment step is seen as one of the most important steps where errors cannot be rectified at a later date. However, if the aim of the comparative modelling of a particular target was not to achieve an overall accurate model, but to enable specificity predictions through highly accurate aligned and modelled portions of the model, it would be plausible to envisage a target-template pair with homology reaching into the twilight zone achieving this. Of course, the alignment methods should be able to produce a satisfactory alignment with accurate interface residues regardless of the accuracy of the rest of the alignment. From these results, this seems to be the case.

Despite the progress in this chapter a few issues exist; since the writing of this chapter other alignment protocols may have been developed which can replicate the structural-based alignment better than the alignment techniques used here (refer to the Introduction of this chapter), also the parameters of these alignment methods were all set to default, so no optimisation of the individual methods was obtained (this was to ensure the results reflected the novice and general outcome of the methods). The gap optimisation plays an important role in the production of profiles and more optimisation could be beneficial. The definition and assignment of the interface residues and regions was investigated, yet more studies could be done. It would also be advantageous to obtain more pairs for the I-vs-I set to understand how both of the target and template sequences having an inhibitor bound would affect the alignment results. It may have been wise to use CE as the gold standards for the local alignment methods only, rather than for both the global and local methods since CE is a local alignment method and so it may be fairer to assess local alignment methods against a local gold standard method.

The next chapter looks at the modelling process of the alignment and tries to determine if the interface portion of the alignment is modelled accurately, even at lower identities, and to what extent can this information be used.

5. Alignment Protocols and Comparative Modelling

5.1 AIM

The aim of this work was to assess the quality of alignment methods (the alignments built in chapter 4) with respect to comparative modelling, in particular evaluating the predictive qualities of the models for the purposes of protease molecular recognition. The previously built alignments provided target-template pairs as input for comparative modelling, and a variety of models were built with different refinement levels to investigate the accuracy of these levels. The accuracy of the models was also assessed by comparing the models to the known structures of the targets in the PDB, as well as considering the accuracy of the interface regions independently. Various properties were considered including the RMSD of the models, the sequence conservation of the alignments, the accessibility of the residues in the structures, and the differences in distances in the contacts made between the modelled structures and the actual structures. The results were obtained in order to assess the specificity of the alignment protocols in terms of modelling at different percentage sequence identities, and whether the alignment methods could model the interface more accurately than the rest of the alignment. This chapter covers the introduction of loop modelling using MODELLER, whilst building, refining and evaluating comparative models in MODELLER and structural superposition can be found in the main introduction: Chapter 1.

5.2 INTRODUCTION

The introduction to this chapter is by no means an exhaustive description of comparative modelling, but aims to introduce in greater detail concepts unique or more relevant to this particular study.

The accuracy of a protein model is directly linked to the usefulness of the model, so it is of great importance that the resulting predicted structure is of as high accuracy as possible. One of the major limitations, and a potentially highly error prone step, of comparative modelling is the loop modelling step.

5.2.1 Loop Modelling

For the basics of comparative modelling please refer to the Introduction Chapter. It should be noted that strictly no specialist loop modelling was done, but the building and refinement of the loops were completed in MODELLER.

Currently, around 60% of all protein sequences can have at least one domain modelled on a related, known protein structure (Fernandez-Fuentes *et al.*, 2006). At least two-thirds of the comparative modelling cases are based on below 40% sequence identity between the target and the templates, and thus generally require loop modelling (Fernandez-Fuentes *et al.*, 2006).

In comparative modelling, target sequences often have residues inserted relative to the template structures or have regions that are structurally different from the corresponding regions in the template. Thus, no structural information can be extracted from the template structures. These regions frequently correspond to surface loops, and show the greatest variation in the amino acid sequence. Loops often play an important role in defining the functional specificity of a given protein, forming the active and binding sites. The accuracy of loop modelling can be a major factor determining the usefulness of comparative models in applications such as ligand docking. Loop modelling can be seen as a mini protein folding problem because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold. Some additional restraints are provided by the core anchor regions that span the loop and by the structure of the rest of the protein that cradles the loop (Jacobson & Sali, 2004).

There are two main techniques of loop modelling procedures: the *ab initio* me thods, an d the database search protocols. There are also "hybrid" methods that combine these two approaches. These will be described briefly in the next sections.

5.2.1.1 *Ab Initio* Methods for Loop Modelling

The *ab initio* (CONGEN: Bruccoleri & Karplus 1987) loop prediction is based on a conformational search or enumeration of conformations in a given environment, guided by a scoring or energy function. There are many such methods: exploiting different protein representations, energy function terms, and optimisation algorithms. Loop prediction by optimisation is in principle applicable to simultaneous modelling of several loops and loops interacting with ligands, which is not straightforward for the database search approaches (Marti-Renom *et al.*, 2000).

ModLoop (Fiser & Sali, 2003b) is a web server for automated modelling of loops in protein structures. The server relies on the loop modelling routine in MODELLER that predicts the loop conformations by satisfaction of spatial restraints, without relying on a database of known protein structures. The method optimises the positions of all non-hydrogen atoms of a loop in a fixed environment.

5.2.1.2 Database Search Techniques for Loop Modelling

The database approach to loop prediction (COMPOSER: Sutcliffe *et al.*, 1987; SLoop: Burke *et al.*, 2000; CAMAL: Martin *et al.*, 1989) consists of finding a segment of main-chain that fits the two stem regions of a loop. The stems are defined as the main-chain atoms that precede and follow the loop but are not part of it; they span the loop and are part of the core of the fold. The search is performed through a database of many known protein structures, not only homologues of the modelled protein. Usually, many different alternative segments that fit the stem residues are obtained, and possibly sorted according to sequence similarity, for example. The selected segments are then superposed and annealed onto the stem regions and then refined (Marti-Renom *et al.*, 2000). The database approach is limited to the size of the loops; the more residues, the more the possible conformations, reducing the size to seven residues or less for their conformations to be present in the database.

5.2.1.3 Combined Approaches

Combined methods use both database search techniques and *ab initio* methods. The underlying idea is the use of database search methods to find candidate loops for a given target and subsequently evaluate and re-optimise it in the target protein (Fernandez-Fuentes *et al.*, 2006)

The ArchPRED (Fernandez-Fuentes *et al.*, 2006) server implements a fragment-search based method for predicting loop conformations. The inputs to the server are the atomic coordinates of the query protein and the position of the loop. The algorithm selects candidate loop fragments from a loop library by matching the length, the types of bracing secondary structures of the query and by satisfying the geometrical restraints imposed by the stem residues. Candidate loops are then inserted in the query protein framework where their side-chains are rebuilt and their fit is assessed by the RMSD of stem regions and by the number of rigid body clashes with the environment. The remaining candidate loops are ranked by a Z-score that combines information on sequence similarity and observed main-chain dihedral angle propensities. The final loop conformation is built in the protein structure and annealed in the environment. This method was benchmarked and it was found possible to predict loops of length 4, 8 and 12 with coverage of 98, 78 and 28% with at least 0.22, 1.38 and 2.47 of RMSD accuracy, respectively.

5.2.1.3 Loop Refining in MODELLER

The loop refining method first takes the generated model, and selects all standard residues around gaps in the alignment for additional loop modelling. An initial loop conformation is then generated by simply positioning the atoms of the loop with uniform spacing on the line that connects the main-chain carbonyl oxygen and amide nitrogen atoms of the N- and C-terminal anchor regions respectively (this model is written to a file with the extension .IL). Next, a number of loop models are generated, each taking the initial loop conformation and randomising it by +/-5Å in each of the Cartesian directions. The model is then optimised thoroughly twice, firstly considering only the loop atoms and secondly with these atoms ‘feeling’ the rest of the system. Non-homology derived restraints are used in this procedure. Each loop model is written out with the .BL extension (Fiser *et al.*, 2000).

As this project focused on the comparative modelling of the peptidases, and MODELLER was used to build the comparative models, for ease and time factors, MODELLER was chosen to complete refinement of the loops.

5.3 METHODS AND MATERIALS

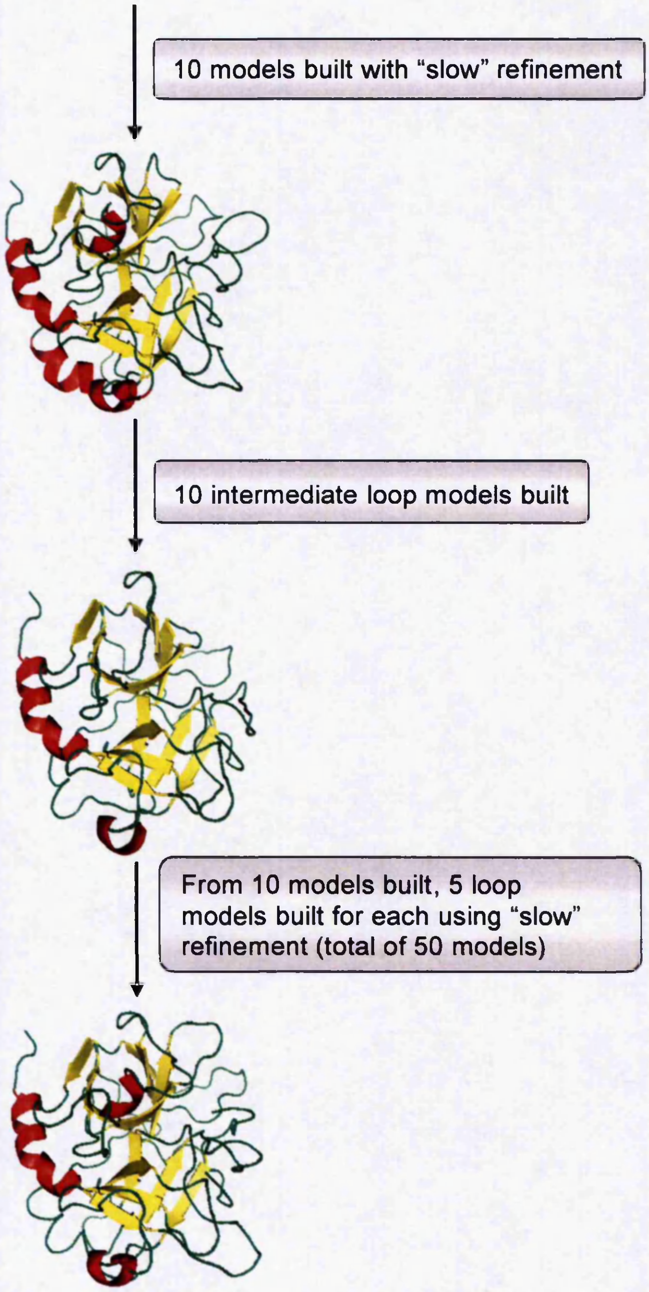
The target-template alignment pairs generated in the previous chapter provided input for the comparative modelling protocol: MODELLER. Models with as high accuracy as attainable needed to be built for each of these pairs. After the construction of the model, refinement of the model and the loops was required to enable the important interface regions to be modelled as close to the actual target structure as conceivable. Evaluation methods would have to distinguish how precise each alignment methods' model was, as well as the modelled specificity of the interface regions.

5.3.1 Developing the Model Building Protocol

When building the comparative models in MODELLER, it is possible to construct multiple models for each target-template alignment pair. These models are able to have different refinement levels applied during the model building stage. MODELLER also allows the refinement of the loops once these models have been produced, with the possibility of more than one model being built with various loop refinements. The refinement levels in MODELLER refer to the optimisation approach of molecular dynamics with simulated annealing. More refinement contains more cycles of molecular dynamics and slower schedule for simulated annealing. The optimisation of the models and the loop models can be done to different degrees, with levels ranging from no refinement to maximum refinement (termed none, very_fast, fast, slow and very_slow in MODELLER).

To determine the best refinement levels plausible within certain time restraints, a small subset of the alignment pairs from both of the datasets (I-vs-I and I-vs-S) with low percentage sequence identities was taken, and different levels of refinement applied whilst building various numbers of models. A trade-off between the accuracy of the resulting models and the amount of time it took to complete the refinement levels and build the multiple models existed. The chosen protocol used for each target-template pair consisted of building an initial ten models with "slow" refinement of the model, but no refinement of the loops, then building ten models which provided intermediate files for the loop refinement (file extension .IL), and then for each of the final ten models built, five loop models were built with "slow" loop refinement. Thus, for each alignment pair a total of seventy models were produced (figure 5.1).

```
>Target
IVEGQDAEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPWDKNFTVDDLLVR
>Template
LIDGKMTRRGDSPWQVVLL-DSKKKLACGAVLIHPSWVLTAAHCM-----DESIRK------
```



10 models built with "slow" refinement

10 intermediate loop models built

From 10 models built, 5 loop models built for each using "slow" refinement (total of 50 models)

Figure 5.1. The Modelling Protocol. From the target-template alignment of each pair an initial 10 models were built using the "slow" refinement protocol in MODELLER, then 10 intermediate loop models were built from each of the first 10 initial models. Finally the 10 intermediate loop models produced 5 loop models each, with "slow" refinement of the loops being applied.

5.3.2 Building the Comparative Models

For all of the alignment methods (MUSCLE, PSI-BLAST, BLAST, Profile-Profile, Sequence-Profile, COACH and TM-align) and both datasets (I-vs-S and I-vs-I) a total of seventy models were built for every protein target-template pair. Only the gold standard TM-align alignments were built by comparative modelling, since TM-align was the chosen gold standard against which the alignments were assessed in chapter 4. Building the models of the gold standard alignment method TM-align meant that the other sequence and profile-based methods could be compared to the TM-align model as well as the actual structure of the target. This enabled the resulting models from each of the alignment techniques to be assessed against the gold standard of "truth" as well as the actual structure of the target PDB (this is important since the TM-align modelled structure represents the maximum "quality" the other alignment models could achieve – since it is based on aligning the known target structure *a priori* to the template).

For the modelling process the template PDB structure files were the input with the target-template sequence alignment to MODELLER. For both datasets, the template structure file only contained information on the peptidase chain being modelled.

5.3.3 Assessing the Alignments and Comparative Models

To assess the final models, global and local RMSDs were calculated by comparing each of the models against the actual target structures. The local RMSDs were calculated over a seven residue sliding window, placing the average RMSD of the window at the fourth residue position of that window (see figure 5.2 for more detail on the seven residue sliding window and the reasons for using seven residues).. The local RMSDs of the structures were used as well as the global, since global RMSDs can be misleading; it is possible for the global RMSD to be quite poor and the local areas to have better RMSDs. Even if only a small local portion of the model has an error in it and the rest of the model has been modelled accurately it could result in an overall poor quality model. It is important in this project to be able to observe local RMSDs, to determine how well contiguous sections corresponding to interface residues have been modelled. The RMSD values between the backbone Cα-Cα, main chain – main chain and for all of the atoms were obtained. It is of course possible that the opposite can be true, and that several well modelled local regions might not share the correct relative disposition to each other as in the true structure – and hence the global RMSD should not be ignored as a quality assessment measure.

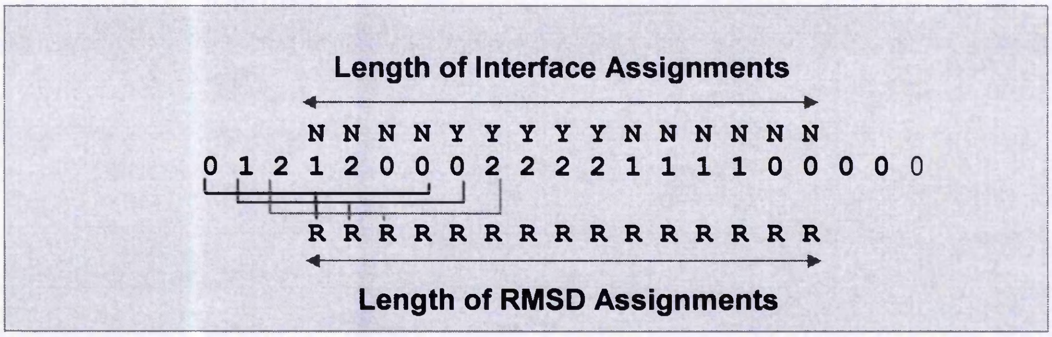

Length of Interface Assignments
N N N N Y Y Y Y Y N N N N N N
0 1 2 1 2 0 0 0 2 2 2 2 1 1 1 1 0 0 0 0 0
R R R R R R R R R R R R R R R
Length of RMSD Assignments

Figure 5.2. Example of Averaging RMSDs. The interface/non-interface status is assigned at position four of the seven residue sliding window, as is the average (local) RMSD value.

5.3.4 Assessing the Interface Regions and Comparative Models

To assess the modelling of the interface regions, the RMSDs of the residues considered to be part of the interface were found (please refer to chapter 4, section 4.3.4 for a more detailed description on how the interface regions were defined).

5.3.5 The Accessibility Calculations

The NACCESS program (described in chapter 2, section 2.9) provided the residue accessible surface for all the residues in each of the targets in the target-template alignment pair. This made it possible to determine if a correlation existed between the RMSDs of the residues in the model and the accessibility of each residue.

5.3.6 Calculating the Sequence Entropy

The multiple sequence alignments generated for the hidden Markov models built in chapter 4 were used in the calculation of the sequence entropies as a measure of conservation. The normalised Shannon entropy (Shannon, 1948) was calculated at each column in the alignment and averaged over a seven residue window (to provide consistency when comparing to the RMSD results which were also averaged over a seven residue window) , assigning the average value to the fourth position. The lower the value, the more conserved that column is relative to the submitted multiple sequence alignment. Calculating this would enable any correlation (if one existed) to be found between the resulting local residue RMSDs of the models and the conservation of that residue position. Details of the Shannon entropy can be found in the chapter 2, section 2.10.

5.3.7 The Correlation Coefficient Calculations

The Pearson correlation coefficient (Pearson, 1896) determined the correlation between various results, including the accessibility of the residues and the RMSD of the residues, as well as the correlation between the conservation entropy of the residues and the RMSD of the residues. Details of Pearson's correlation coefficient can be found in the chapter 2, section 2.11.

5.3.8 Structural Superposition

To enable contacts of the target PDB structure file (containing the contacts between the peptidase chain and the inhibitor chain) to be compared to the contacts between the modelled peptidase chain and inhibitor chain, a simple structural superposition "docking" procedure was used. STALIN (a program provided by Dr SJ Hubbard, University of Manchester, Bioinformatics group, see chapter 2, section 2.18) completed the structural superposition; the modelled peptidase chain was superposed onto the PDB peptidase chain, which had the inhibitor chain bound to it. The modelled peptidase chain and the original inhibitor chain were then copied from the STALIN output file to a separate file, to compare with the actual PDB peptidase and inhibitor chain structure. The structural alignments created during the superposition were also checked manually, to ensure STALIN had made a plausible superposition from which to base any comparisons.

5.3.9 Assessing the Model Specificity

Using MODCONTA (a program provided by Dr SJ Hubbard, University of Manchester, Bioinformatics group) the contacts between both the true and the modelled target peptidase and inhibitor chain were calculated. When five contacts are found within five Angstroms an interacting pair of residues is defined. The actual PDB file was submitted to MODCONTA first, with a deliberately large distance cut-off of 8Å (to allow all possible contacts to be found), reporting all contacts between the peptidase and inhibitor chain below this value. Secondly, the modelled file was submitted to the program with a threshold of 10Å (this was used to allow all possible contacts to be found that were obtained with the actual PDB file and allows an error of 2 Å above the 8 Å). Both the main chain-main chain and side chain–side chain contacts were calculated. All the contacts found using the PDB file below 5Å (this was seen as a large enough distance to allow for errors but small enough to be contacting residues) were considered "correct" contacts. These contacts were noted as entirely "missed" by the modelled structure if the same contact could not be found below 10Å distance within the models' output. This allowed the model to have an error of +/- 5Å. If the same contact from the PDB file (of below 5Å) was found in the models' contact output, then a correct match is assumed, and the difference in distance was computed. The number of correct matches was also calculated for three distance e rror ranges (difference in the distances); the number of contacts correctly found within +/- 1Å, +/- 2Å and +/- 3Å (to test the specificity of the alignment method) of the PDB structure contacts was summed.

The sensitivity of the predicted contacts for each alignment method was also found (Equation 5.1). In this case a true positive was assigned when a contact found below 5Å in the PDB file was below 6Å (for the +/- 1Å difference distance example) in the model file. A false negative was assigned for all of the contacts that the PDB file made below 5Å that the model file did not find for a given cut-off (i.e. below 6Å and above 4Å, for the distance range of +/- 1Å). ROC plots could not be calculated due to the difficulty of obtaining the number of true negatives which would be very large.

$$\textbf{Sensitivity} = \frac{TP}{TP+FN}$$

Equation 5.1. Calculating the Sensitivity. The equation used for calculating the sensitivity of the contacts. Where TP = true positives and FN = false negatives.

The positive predicted value (PPV, equation 5.2) contains false positives; if the model files contained contacts below 6Å (again, +/- 1Å, for example) but the same contact could not be found in the PDB file of contacts below 5Å.

$$\textbf{PPV} = \frac{TP}{TP+FP}$$

Equation 5.2. Calculating the Positive Predicted Value (PPV). The equation used for calculating the PPV of the contacts. Where TP is true positives and FP is false positives.

5.4 RESULTS AND DISCUSSION

As a first assessment of which alignment method produced the most accurate models, the global RMSDs between the modelled and actual structures were obtained. Further, to determine whether the protease interface could be modelled more accurately than the non-interface residues, the RMSDs for these residue subsets was also calculated. To aid understanding of the results, the sequence entropy of the residues and their accessibilities was also computed. The different model refinement methods were al so a ssessed by comparing RMSD results. Correlations between the RMSDs of the models and their target-template alignments, amongst others, were found. Since the I-vs-I dataset contained a limited number of alignment pairs (and a smaller number of pairs provides less reliable statistics), the results in this chapter reflect the I-vs-S set.

5.4.1 The Models Built

Out of the 144 pairs in the I-vs-S set, all but the following pairs produced successful (models actually built by MODELLER) models under the MODELLER protocol:

- 1jiwP+1af0A, BLAST
- 1nw9B+1bmqA, COACH, Sequence-Profile, Profile-Profile

For each alignment pair a total of seventy models could be built, ten in the initial stage, another ten in the loop intermediate stage and a final fifty in the last stage. The initial ten will be referred to as the "standard" model set and the fifty referred to as the "loop" set (since these models contain loop refinement), the intermediate loop models will not be discussed, but will be known as the "intermediate" group. The intermediate group consists of models built by MODELLER as a stepping stone to the refined models, hence they are expected to be of lower quality that the refined "loop" set. The percentage sequence identity of the alignment pair is included in the results to help understand the potential difficulty of the modelling process and in this case the bins are inclusive. The percentage identity of the alignment pair was calculated as the number of positions containing matching amino acids, in terms of the alignment length. Some of the MODELLER methods did not produce all of the models (some of the low quality alignments exceeded the maximum violation of restraints value used in MODELLER). The numbers of models and pairs for the I-vs-S set in the different percentage identity bins can be seen in table 5.2. Please refer to table 4.2 in chapter 4 for the number of pairs in each percentage identity bin. The wide range of conformations built allowed the different model building techniques to be assessed. Difficulty in choosing the best conformation ensued and this would depend on what the resulting model was to be used for. For example, if it were to be used to investigate ligand docking the model with the highest resolution would be chosen.

(a)

PID	TM-align	PSI-BLAST	BLAST	MUSCLE	Sequence-Profile	Profile-Profile	COACH
<20	239	240	240	240	232	235	235
<30	509	510	510	510	502	505	505
<40	1199	1200	1200	1200	1182	1195	1195
<60	1369	1369	1360	1370	1352	1363	1365
<100	1439	1439	1430	1440	1422	1433	1435

(b)

PID	TM-align	PSI-BLAST	BLAST	MUSCLE	Sequence-Profile	Profile-Profile	COACH
<20	1175	1200	1200	1200	1150	1155	1150
<30	2525	2550	2550	2550	2550	2505	2500
<40	5975	5975	6000	6000	6000	5955	5950
<60	6825	6825	6800	6850	6850	6755	6800
<100	7175	7175	7150	7200	7200	7105	7150

Table 5.1. Number of Pairs and Models. The PID is the percentage sequence identity of the alignment pair, the models are split into the PID bins which are inclusive (indicated by the below "<" symbol). Table (a) represents the "standard" set and table (b) the "loop" set.

5.4.2 The Accuracy of the Models

For all of the pairs in each method in the I-vs-S dataset, the average global Cα-Cα RMSD and the average global interface Cα-Cα RMSD of all the models were obtained (Table 5.4a). The main chain – main chain and side chain – side chain results are not shown. To see these results please refer to Appendix 3 (tables A3.1 and A3.2). The average local residue RMSDs were also computed (Table 5.4b).

(a)

Methods	Standard		Loop	
	All	I	All	I
TM-align	2.66	2.31	2.91	2.65
PSI-BLAST	3.94	3.31	4.05	3.38
BLAST	3.09	2.97	3.18	3.07
MUSCLE	4.50	3.71	4.59	3.80
Sequence-Profile	3.85	3.34	3.92	3.42
Profile-Profile	3.92	3.36	4.07	3.54
COACH	3.91	3.24	4.04	3.38

(b)

Methods	Standard		Loop	
	NI	I	NI	I
TM-align	0.67	0.89	0.72	0.97
PSI-BLAST	0.82	1.07	0.83	1.09
BLAST	0.81	1.00	0.82	1.02
MUSCLE	0.92	1.15	0.93	1.16
Sequence-Profile	0.85	1.09	0.85	1.09
Profile-Profile	0.82	0.96	0.86	0.99
COACH	0.83	1.08	0.86	1.10

Table 5.2. The Global and Local RMSD Results. For the I-vs-S set, the global (table a) and the local (table b) carbon alpha - carbon alpha RMSDs are shown. "Standard" and "loop" refer to the ten or fifty models, respectively, and the refinement level these results were averaged over. "All" indicates the global RMSD value of all of the residues and "I" indicates the interface global RMSD result. "NI" indicates the non-interface RMSD value of all of the residues that were not part of the interface and "I" indicates the interface local RMSD result. The green boxes highlight the best (lowest) RMSD result, the red, the worst – ignoring the benchmarking TM-align which is a structural alignment method. The differences in these results are occasionally very small and become insignificant unfortunately no error statistics were completed on this set.

The Needleman & Wunsch (1970) algorithm was not used in this project as it has since been built upon to improve its accuracy and speed (Thompson et al., 1994; Myers & Miller, 1988) and so MUSCLE is seen as an improvement over CLUSTALW which is an improvement over the Needleman & Wunsch algorithm (Thompson *et al.,* 1994).

The results shown in table 5.4(a) seem to suggest, on average and based on the global RMSD, BLAST produces the most accurate model and MUSCLE the worst performance. This is also reflected in the local RMSD results as well (table 5.4(b)). Table 5.4(a) reveals that the interface (I) is more accurately modelled than the whole structure (All) in terms of global results, again with BLAST producing the best models and MUSCLE the worst. BLAST seems to produce models with better RMSDs than PSI-BLAST even though PSI-BLAST produced more accurate alignments than BLAST, this may be due to BLAST chopping off the more difficult regions to model of the target sequence than PSI-BLAST does (figure 4.11). When the local RMSD results (the interface and non-interface residues are placed into separate categories to be evaluated) are observed the non-interface residues consistently result in lower RMSDs. Accessibility calculations and conservation calculations were completed in order to try and explain these results, the graphs and tables of which can be found later on in this section. The explanation for this is that interface regions would be expected to be more accessible, closer to the surface, and perhaps therefore modelled less accurately on average than non-interface which might contain more buried residues. Of course, none of the alignment methods produce a model with precision and accuracy as high as with the gold-standard alignment method, TM-align. This does confirm how important it is to have accurate alignments in the comparative model building process. The "standard" models seem to outperform the "loop" models consistently, this suggests it may not always be advantageous to refine the loops in the models. This may be because the loop refinement step was not optimised to a very high level to allow for time constraints. The limited "loop" modelling we implemented in Modeller was not able to improve the models at the modest levels of sequence identity we were investigating. There is not a significant improvement in RMSDs. However, no full loop modelling or any optimisation of sidechains was completed in the modelling. The preliminary conclusion is that such high level modelling and refinement appears to be of limited use when modelling such divergent protein pairs.

5.4.3 The RMSD and the Percentage Identity

As the sequence similarity between the target-template alignment pairs decreases, the RMSD of the comparative model increases; the model becomes less reliable (Table 5.3). The RMSDs of the models within the "<20%" sequence identity bin are almost double than those of the sequences in the "<100%" sequence identity bin. Again, on average BLAST seems to produce the most reliable models and MUSCLE the least. There seems to be a slight decrease in RMSD values for the "loop" set compared to the "standard" set, suggesting loop refinement does improve the accuracy. For both of the "standard" and "loop" sets the trend is very similar; as the percentage sequence identity decreases, the profile and HMM based methods become more accurate than the sequence based methods. For the higher percentage identity bins (<100 and <60%) the lowest average RMSD is found to be BLAST followed by the Sequence-Profile method, COACH, the Profile-Profile method, PSI-BLAST, and finally MUSCLE. However, for the 20% bin ("loop" set), the method order changes to: BLAST, COACH, PSI-BLAST, Profile-Profile, Sequence-Profile and MUSCLE. The best and the worst methods remained the same, but the profile and HMM methods become more accurate. As the percentage identity decreases, the more sophisticated methods, containing more evolutionary information, outperform the sequence based methods. BLAST and PSI-BLAST may appear more accurate because they do not retain 100% of their sequence alignment (BLAST on average retains ~93% of its target sequences and PSI-BLAST ~86%, figure 4.11, chapter 4), making the remaining sequence alignment simpler to model. This is because it might be expected to concentrate on "easier-to-align" sections of the model with strong local conservation, a consequence of using local alignment-based protocols. Results for the main chain - main chain results and for all atoms (Appendix 3, tables A3.1 and A3.2) display generally poorer RMSD values but with TM-align having the lowest RMSD values in all but one case. MUSCLE still has the highest RMSD values.

(a)

PID	TM-align	PSI-BLAST	BLAST	MUSCLE	Sequence-Profile	Profile-Profile	COACH
<20	3.12	5.79	5.07	10.26	6.72	6.36	6.22
<30	3.21	5.22	4.25	7.30	5.50	5.18	5.16
<40	2.85	4.31	3.39	5.01	4.17	3.58	4.19
<60	2.76	4.09	3.21	4.58	4.01	4.08	4.07
<100	2.66	3.94	3.09	4.50	3.85	3.92	3.91

(b)

PID	TM-align	PSI-BLAST	BLAST	MUSCLE	Sequence-Profile	Profile-Profile	COACH
<20	3.11	5.98	5.12	10.28	6.65	6.29	5.92
<30	3.37	5.36	4.31	7.34	5.49	5.30	5.13
<40	3.12	4.42	3.49	5.10	4.22	3.90	4.31
<60	3.02	4.20	3.30	4.78	4.08	4.24	4.20
<100	2.91	4.05	3.18	4.59	3.92	4.07	4.04

Table 5.3. RMSD and Percentage Sequence Identity. PID is the percentage sequence identity of the alignment pair. The models were split into inclusive PID bins. The average Cα-Cα RMSD for each method is displayed. Table (a) refers to the results for the "standard" set, and table (b) for the "loop" set. Again, the green boxes indicate the method with the lowest RMSD in that PID bin, the red indicates the highest RMSD.

Figure 5.3 demonstrates the correlation that exists between the RMSD of the models and the percentage sequence identity of the target-template alignment; as the percentage sequence identity decreases, the RMSD increases. Surprisingly, BLAST is able to produce models sharing below 20% sequence identity with the template with RMSDs below 3Å, although, it is expected these alignment pairs will not have retained 100% of their sequences, investigations into this are completed later in this chapter. Although the trends are quite clear visually, the actual correlations are modest (and negative, as expected), again reflecting the possibility to produce both high and low quality models are low sequence identities. This is clearly shown in Figure 5.3 below and s similar graph can be seen in Martin et al., 1997.

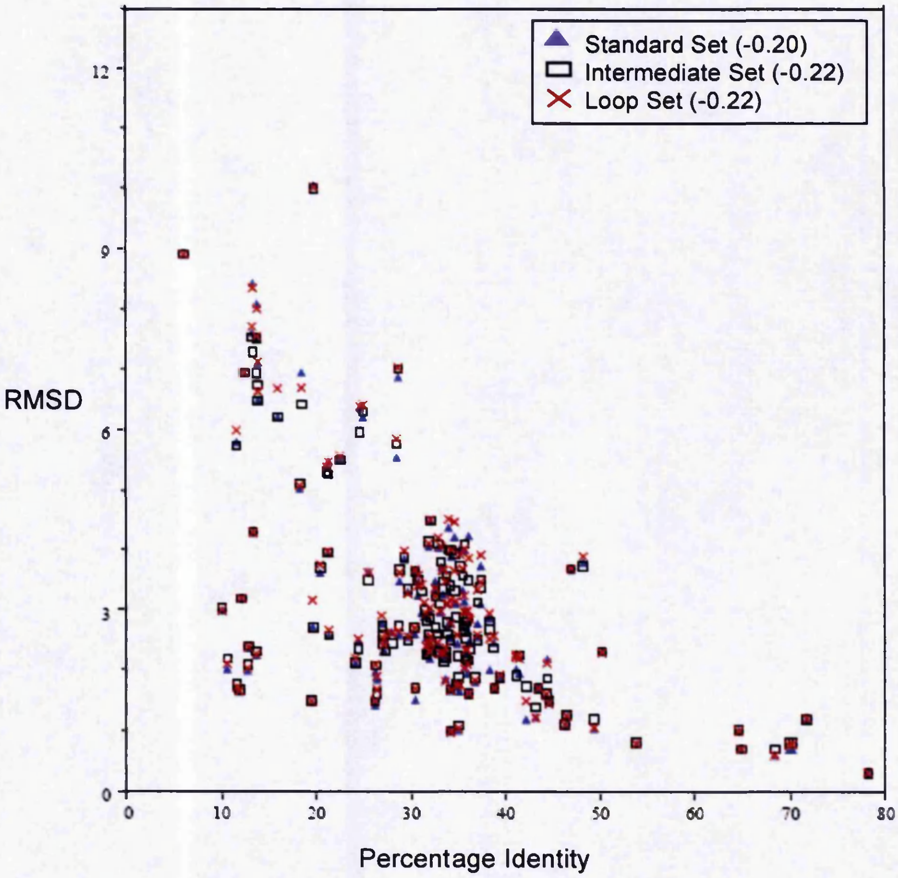


Figure 5.3. RMSD and Percentage Sequence Identity. The average RMSD and percentage identity results for all of the pairs, using the BLAST alignment method are shown for each of the refinement level sets ("standard, "intermediate" and "loop"). The Pearson's correlation coefficient has been placed in the key

5.4.4 The RMSD and the Percentage of Gaps

As expected, the more gaps introduced into the alignment, the higher the RMSD becomes. The correlation between the amount of gaps in the alignment and the RMSD of the model is higher for the target (Figure 5.4) than the template (on average), although this is probably due to the fact that the target tends to contain more gaps than the template (Section 4.4.4, chapter 4) and gaps in the target are more of a problem since MODELLER will produce a loop in the target model where there is no template structure to use the restraints from.

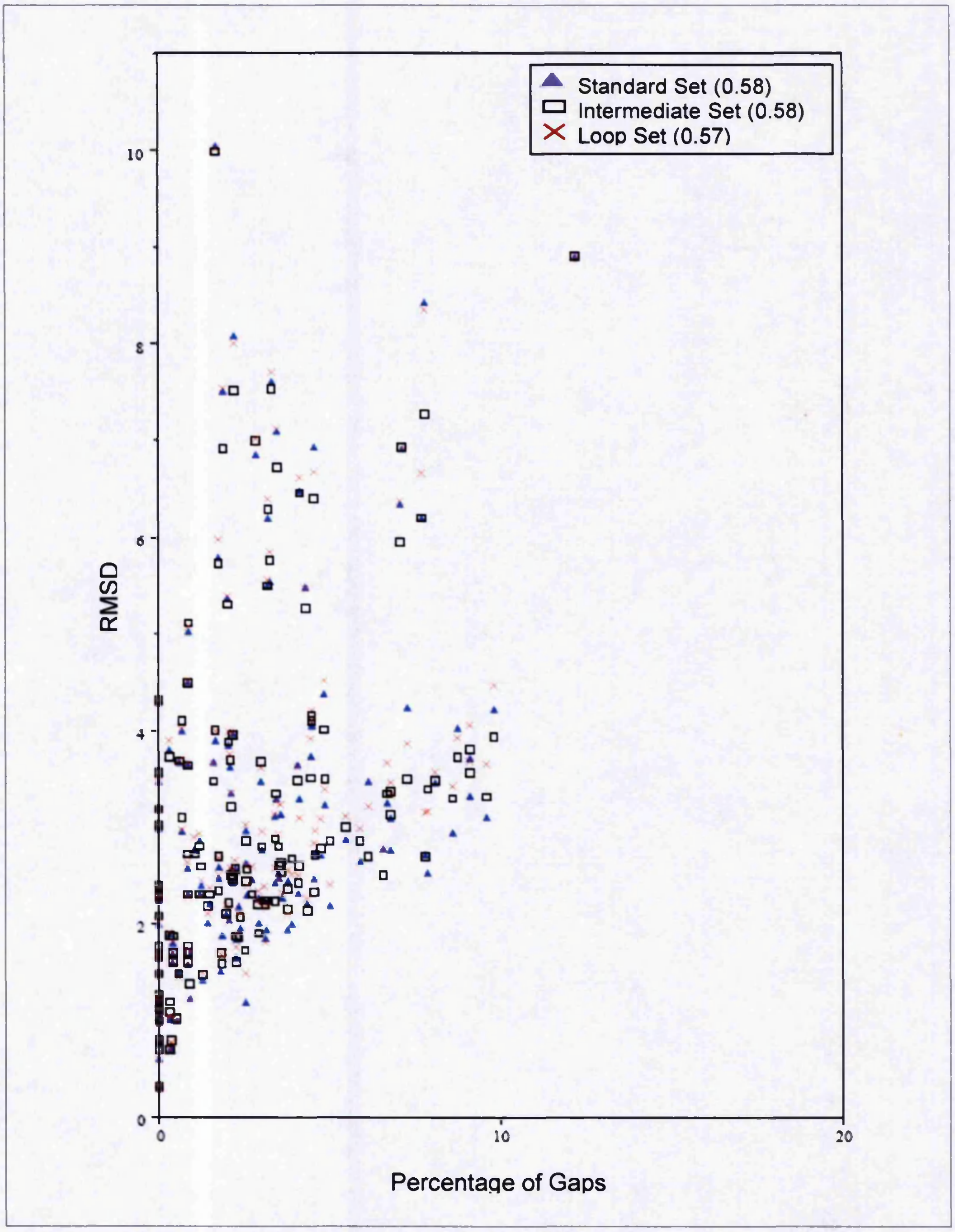


Figure 5.4. RMSD and Percentage of Gaps, Target. The average RMSD for each pair in the different refinement sets, “standard”, “intermediate” and “loop” plotted against the percentage of gaps contained within the alignment pair.

5.4.5 Six Example Pairs

For some of the results, an example subset of the I-vs-S set was chosen to enable more thorough observations to be made and to further understand the results. Six pairs were picked, essentially randomly, representing a range of percentage sequence identities. The sequence identities have been calculated as in figure 4.2 in chapter 4. A summary of the global RMSDs and the lengths of the targets and templates are also shown in table 5.5. Usually, if the target sequence length is very different to the template sequence length INDELs become a problem and the alignment and modelling steps suffer. The first three pairs sharing below 30% sequence identity have large differences in the lengths of the target compared to the template lengths, which hinder the modelling process. A single model (from the seventy generated for each alignment pair) with the lowest global RMSD was chosen to use in the calculation of the results.

I-vs-S Set Pairs	Percentage Sequence Identity Between the TM-align Alignment	Percentage Sequence Identity Over the Target	Sequence Length	
			Target	Template
2kaiA+1cvwH	10.6	33.8	80	254
2sicE+1r64A	13.8	24.4	275	481
1stfE+1cs8A	26.4	40.1	212	316
1ppfE+1pytD	30.5	35.8	218	251
1avgH+1autC	37.4	38.2	259	240
1f34A+1htrB	48.3	50.3	326	329

Table 5.4. The Six Example Pairs. The six pairs chosen over a range of percentage sequence identities from the I-vs-S set to be used in the results. The length in number of residues for the target and template is shown. The percentage sequence identity calculated over the target is shown as well.

The global RMSDs for the six pairs range from 1.51Å (1stfE+1cs8A, Profile-Profile) to 8.09Å (2sicE+1r64A, BLAST), as shown in Table 5.6. The RMSDs for the lower percentage sequence identity pairs, and the global alignment methods, are usually higher than those pairs which share more sequence similarity. Surprisingly, 2kaiA+1cvwH which shares a modest 10.63% sequence identity can achieve a global RMSD of 3.52Å with the Profile-Profile alignment method and 2.01Å with BLAST. Equally as surprising, is 1f34A+1htrB (48.28% sequence identity) which has an RMSD of 5.65Å when using TM-align. 2kaiA does however, have a relatively short length of 80 residues and 1f34A, 326 residues. The TM-align alignment between 1f34A and 1htrB has a limited amount of gaps and shares similar lengths to one another (the target and the template), hence it would be seen as a fairly easy model to build. The alignment between 2kaiA and 1cvwH generated by BLAST is a short alignment (most of the template sequence has been removed to minimise gaps) with only a few gaps in the alignment, whereas the Profile-Profile alignment has retained the whole template sequence and so has many more gaps inserted into the shorter target sequence, making it a much harder model to build.

Pairs	TM-align	BLAST	MUSCLE	PSI-BLAST	Sequence-Profile	Profile-Profile	COACH
2kaiA+1cvwH	1.61	2.01	7.88	1.78	3.89	3.52	3.87
2sicE+1r64A	1.97	8.09	7.31	3.93	4.81	5.80	6.93
1stfE+1cs8A	1.36	1.84	1.92	2.00	1.66	1.51	1.98
1ppfE+1pytD	2.43	3.66	3.59	2.60	2.54	2.78	2.66
1avgH+1autC	2.79	3.10	3.13	3.26	3.53	3.71	4.03
1f34A+1htrB	5.65	3.80	4.60	4.62	4.70	3.50	4.70

Table 5.5. The RMSD of the Six Example Pairs. Shown are the global RMSDs for the six pairs for the standard set.

The six target-template pairs, and for all the various methods, all achieved alignments with over 90% (90% or above of the target-template alignment) classified as being "model-able" (please refer to the abbreviations list or chapter4 section 4.3.3.7 for an explanation of 'model-able'), except in the following small number of cases:

- TM-align: 2kaiA+1cvwH (34.65%) and 2sicE+1r64A (69.69%)
- Sequence-Profile: 2sicE+1r64A (73.12%)
- COACH: 2kaiA+1cvwH (84.25%)

The TM-align pair 2kaiA+1cvwH only has 34.65% of its alignment that is model-able, this means that the starts or the ends of the target or template sequence are aligned to gaps in the other sequence. This is because all of the gaps in the target are placed at the end of the target sequence, and the difference between the target and template sequence length is great. This does mean that MODELLER will ignore the gaps placed at the end and so the target model will not suffer. However, the other methods, for the same alignment pair, place the gaps within the target sequence, resulting in an alignment that has a higher model-able percentage but more gaps within the target sequence, and hence lower quality models (higher RMSDs).

The percentage of gaps (the number of gaps placed within the target sequence compared to the length of the target sequence) in the aligned target sequence for the six pairs can be seen in Table 5.7.

Pairs	TM-align	BLAST	MUSCLE	PSI-BLAST	Sequence-Profile	Profile-Profile	COACH
2kaiA+1cvwH	68.50	8.57	68.63	5.88	68.63	66.67	68.50
2sicE+1r64A	43.30	10.70	43.65	14.33	50.72	25.37	45.97
1stfE+1cs8A	34.16	6.31	34.16	5.78	35.37	7.11	2.77
1ppfE+1pytD	14.84	8.79	13.83	9.17	15.15	10.29	15.83
1avgH+1autC	2.26	0.40	0.38	0.38	0.38	3.54	2.26
1f34A+1htrB	4.12	0.92	1.21	0.91	3.83	2.15	1.51

Table 5.6. The Percentage of Gaps in the Target for the Six Pairs. The percentage of gaps placed in the target sequence, for each of the chosen six pairs is displayed.

The global alignment methods: (which are unable to discard any of the target or template sequence) TM-align, MUSCLE, Sequence-Profile and COACH all result in a much higher percentage of gaps (gap penalties were not optimised though) being introduced into the target, increasing in number as the percentage sequence identity decreases. The local alignment methods, which are capable of removing some of the alignment, tend to contain fewer gaps than the global methods, especially in poor alignment quality cases. The same holds true for the template, but with far less gaps being inserted (below 15%, except for COACH: 2kaiA+1cvwH 34.77%).

The percentage of the target and template sequences retained compared to the sequences before submission to the alignment methods is shown in Table 5.8 for the six pairs. Only those methods with below 100% sequence retention were included. The lower percentage sequence identity pairs lose more sequence, with most lost in the template sequence for the pair

2kaiA+1cvwH. More importantly, above 95% of the target is retained for all of the methods and pairs except for BLAST: 2kaiA+1cvwH and 2sicE+1r64A, the pairs sharing below 20% sequence similarity. Losing sequence from the template might not be a problem if only non-interface regions are being removed.

Pairs	BLAST		PSI-BLAST		Profile-Profile	
	Target	Template	Target	Template	Target	Template
2kaiA+1cvwH	80.00	27.17	100	33.07	100	93.70
2sicE+1r64A	88.00	55.09	100	64.86	92.00	65.90
1stfE+1cs8A	98.11	68.67	100	68.99	98.58	68.67
1ppfE+1pytD	100	94.42	100	94.42	100	93.23
1avgH+1autC	96.53	97.50	100	94.49	94.59	98.75
1f34A+1htrB	99.08	98.78	100	99.70	97.55	97.26

Table 5.7. The Percentage Retained for the Six Pairs. The percentage of sequence retained for the chosen six pairs.

5.4.6 The Accuracy of the Different Refinement Levels

The refinement levels of the models (standard, intermediate and loop) seem to make only marginal differences to the quality of the models in our hands. Indeed, generally, they only slightly reduce the model quality when more refinement is used, particularly in the global quality of the models (Figure 5.5). BLAST and MUSCLE are chosen since they have the best and worst overall RMSD values, respectively (Table 5.4). For the pairs with low sequence identity, the "loop" set (having the most refinement on the loops) have lower RMSDs, with a broader range of RMSD values (indicated by longer bars in figure 5.5). The loop set tends to have a more varied range of RMSD values and the model with the lowest RMSD is more accurate than any of the models in the other sets. The spread of RMSDs is not particularly useful unless the best one can be selected correctly which is not a trivial task. Ho wever, the model with the highest RMSD is also generally higher than any of the models in the other sets.

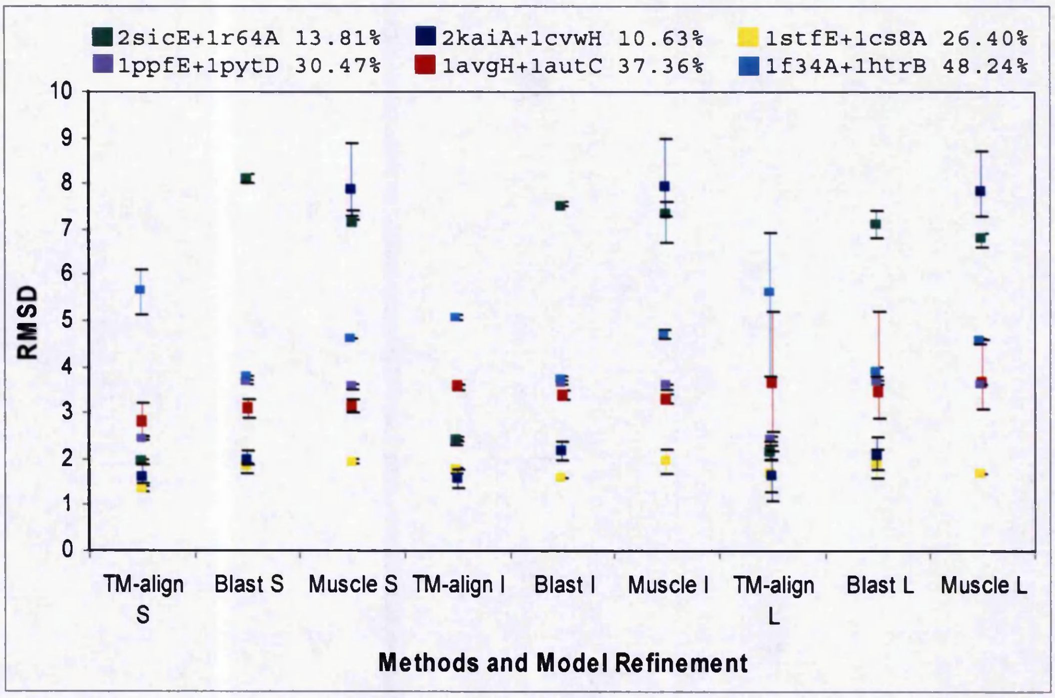


Figure 5.5. The RMSD and the Model Refinements. The minimum, maximum and average (shown as squares) values for the RMSD for all of the models in that set for the six pairs are shown as error bars. "S" indicates the standard set, "I" is the intermediate set and "L" is the loop set. The pairs and percentage identity can be found at the top of the graph.

5.4.7 The RMSD and the Conservation Entropy.

The more conserved a region of a sequence is in a multiple sequence alignment, the more accurate the model built from that sequence will be in general (Figure 5.6). For each target, the average sequence entropy for each position in the target sequence in the multiple sequence alignment (the multiple sequence alignment used in the generation of the HMMs used in Chapter 4) was calculated. If a target sequence was relatively more conserved in the multiple sequence alignment, compared to another target in a different multiple sequence alignment, it would be likely that the more conserved target would have a template closer in structure than the less conserved one, and so would produce a higher quality model. Valdar (2002) review and compare different conservation scoring methods which may suggest entropy is not the optimal method since it does not account for amino acid similarity and suggest the use of Scorecons which calculates the degree of amino acid variability in each column of the alignment.

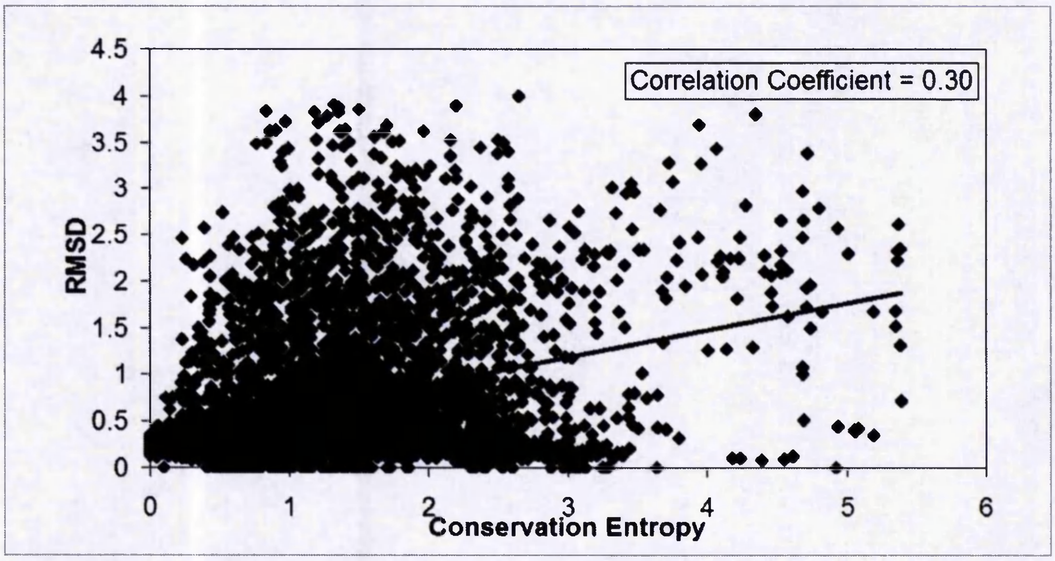


Figure 5.6. The RMSD and Sequence Entropy for all Models. The average (local) RMSD over a seven residue window for all of the models in the different refinement levels was calculated for the TM-align method and plotted against the average sequence entropy for the same seven residue window in the same target sequence.

The interface regions of the six pairs generally contained residue segments with both high RMSDs and low RMSDs, coupled with complementary low and high conservation entropies (Figure 5.7). This meant that whilst some regions of the interface are highly conserved and are well modelled (with low RMSDs), the other parts of the interface are more variable and contribute to higher local RMSDs. This may explain the results in Table 5.3; the local interface RMSDs are higher than the local RMSDs of the non-interface regions (the rest of the alignment which is not interface). This may be a result of the fact that some of the interface is necessarily highly variable in order to give different specificities to each peptidase. The correlations between RMSD and sequence entropy for the other four pairs (TM-align method) are:

▪ 1avgH+1autC	= 0.19
▪ 2kaiA+1cvwH	= 0.16
▪ 1f34A+1htrB	= 0.35
▪ 1ppfE+1pytD	= 0.42

All of these correlations are positive, and although they are modest it does suggest that regions of the proteins which are well conserved across the family are better modelled.

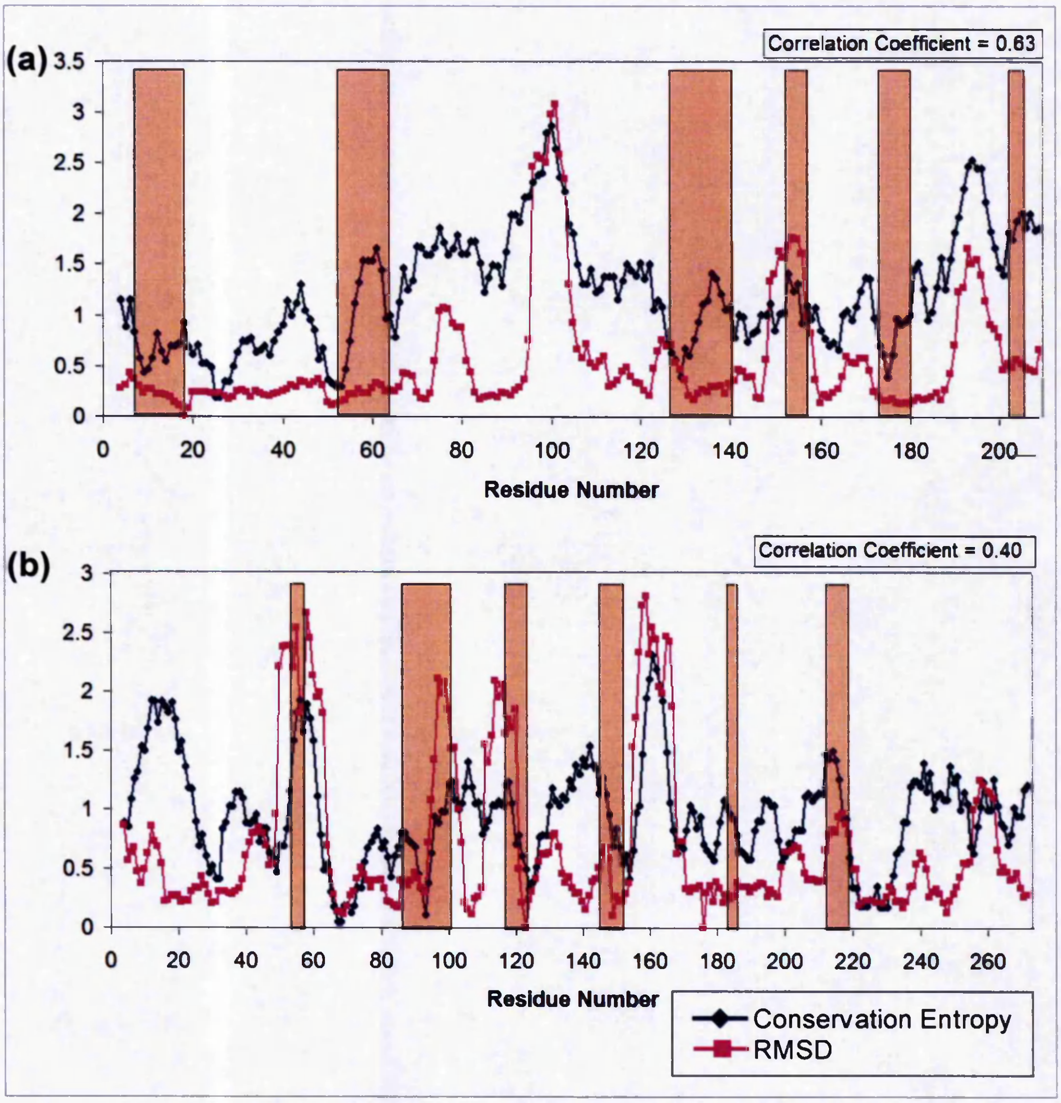


Figure 5.7. The RMSD and Conservation Entropy for the Interfaces. The average RMSD and conservation entropy for each residue is shown (both averaged over a seven residue window) for TM-align. The pink rectangles show the positions of the interface. Graph (a) refers to pair 1stfE+1cs8A sharing 26.40% sequence identity and graph (b) to pair 2sicE+1r64A sharing 13.81%.

5.4.8 The RMSD and the Accessibility

Figure 5.8 displays the accessibility of the residues (again averaged over a seven residue window) and the RMSDs of the residues (for the example pair 1stfE+1cs8A). The interface has relatively highly accessible residues and quite high RMSDs. A correlation between the RMSD and the accessibility of the residues was also found, shown graphically for one example (Figure 5.9) and for the pairs considered here (Table 5.9).

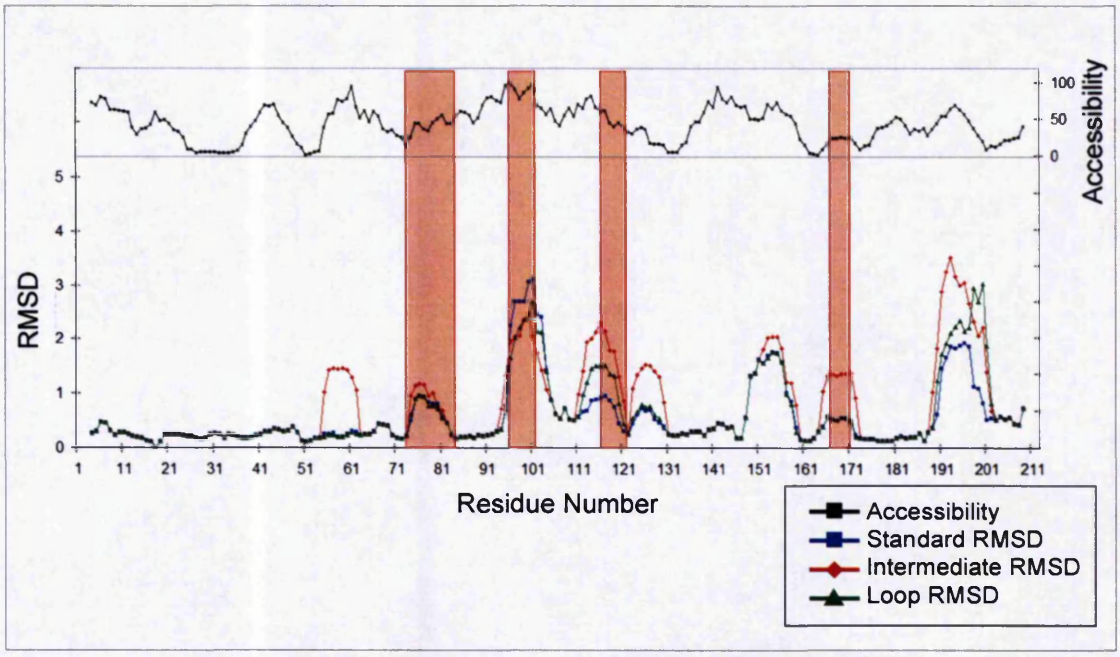


Figure 5.8. The RMSD and Accessibility for the Interfaces. For the method TM-align of the 1stfE+1cs8A pair and the different refinement levels, the accessibility and RMSD are shown. The pink rectangles show the positions of the interface.

The interface residues for the pair 1stfE+1cs8A (method TM-align) were found to have a lower RMSD, on average, than the non-interface residues (Figure 5.9). This would support the idea that the interface should be better modelled than the non-interface residues; however, out of the six pairs for the method TM-align, this is the only case where this applies.

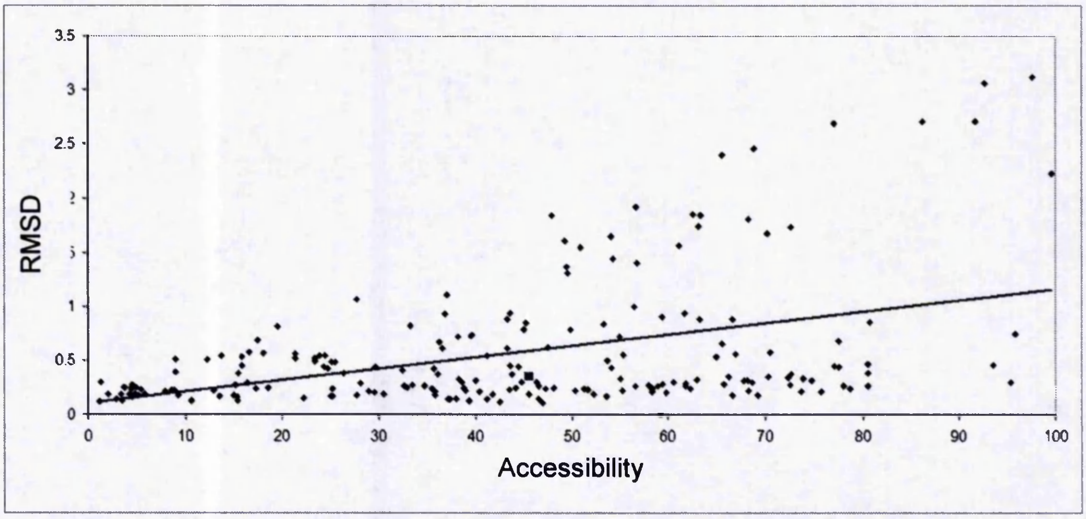


Figure 5.9. The RMSD and Accessibility. For the method TM-align of the 1stfE+1cs8A pair, the accessibility and RMSD are shown. The pink square shows the average RMSD and accessibility for the interface residues.

The residue accessible surface for each of the six pairs and the correlation between the accessible surface and the local RMSDs of the pair was calculated (Table 5.8). There is a correlation; the more accessible the surface, the higher the RMSD. This is more pronounced with the method TM-align, which would be expected to produce the best alignments between target and template. Hence, the trend is most evident when the alignment is closest to the one expected to be correct, and the trend is partly reduced by poorer alignments. The BLAST correlation is negative this may be because residues are counted as accessible in the partial model but should not be.

Pairs	TM-align	BLAST	MUSCLE
2kaiA+1cvwH	0.40	-0.27	0.05
2sicE+1r64A	0.41	0.35	0.43
1stfE+1cs8A	0.43	0.37	0.37
1ppfE+1pytD	0.40	0.39	0.35
1avgH+1autC	0.45	0.41	0.43
1f34A+1htrB	0.24	0.22	0.22

Table 5.8. The Correlation Between the RMSD and Accessibility. The Pearson's correlation coefficient between the RMSD and the accessibility of the six pairs for the standard set is shown.

5.4.9 The Specificity of the Methods

In order to assess how well the alignment methods could model the interface regions, the contacts between the peptidase chain and the inhibitor chain in the target model output file were assessed in comparison with those present in the actual PDB structure file of the target with an inhibitor bound. Comparisons were made to the contacts between the peptidase and inhibitor chain found in the actual PDB structure below 5Å. It is worth noting that no contacts were lost when the local alignment methods did not retain 100% of their sequences.

The percentage of correct contacts for the PDB obtained through the modelling process is displayed in Table 5.9. This is dependent on the alignment method used. In this table, the correct contacts are defined as contacts made in the target PDB structure file between the peptidase chain and the inhibitor chain below 5Å which are also found in the model (built using the alignment methods) below 10Å; this allows an error of +/- 5Å. A large tolerance on the contact distance was used in the first instance, as the modelling process is deliberately quite crude, with a simplistic “docking” step and no attempt has been made to refine the “docked” co-ordinates with any energy minimisation or similar. The results shown are for side chain – side chain contacts, because more specificity is defined by the side chain contacts rather than just main chain contacts. The main chain results can be found in Appendix 3, Table A3.3.

For the six pairs in Table 5.9, the total number of contacts found in the PDB structure file varies between 227 (1f34A+1htrB) and 36 (2kaiA+1cvwH). There is a question of the reliability of the results from only viewing six of the results this is taken into consideration and the results were chosen at random. The percentage of correct contacts does not fall below 63%, even in the low percentage sequence identity pairs (2sicE+1r64A and 2kaiA+1cvwH), although 2kaiA+1cvwH only has 36 contacts. Unsurprisingly, the sequence pairs with higher percentage identities achieve better results than the lower sequence identity pairs, except 2kaiA+1cvwH (probably due to the small number of total contacts) and 1stfE+1cs8A, where at least three of the methods achieve 100% of the contacts. The pairs above 20% sequence identity seem to have a varied range of which methods are the best at predicting the most correct contacts, with BLAST apparently being the best method for predicting the most correct contacts and COACH being the worst. In the two lower sequence identity pairs, BLAST is consistently the worst method and it could be deduced that PSI-BLAST is the best method. In all but one case (1f34A+1htrB) the gold standard method TM-align achieves a higher percentage of correct contacts than the other methods. This however, will have to be put into context by viewing the distances of these correct contacts, and which methods can predict them with more specificity. It is expected that if BLAST obtains more correct contacts then PSI-BLAST should as the PSI-BLAST algorithm is based on the BLAST algorithm.

Methods	1f34A +1htrB	1avgH +1autC	1ppfE +1pytD	1stfE +1cs8A	2sicE +1r64A	2kaiA +1cvwH
PDB	227	192	147	151	176	36
BLAST	85.46	89.58	73.47	100	65.34	63.89
COACH	85.02	76.56	74.83	99.34	76.70	100
MUSCLE	86.78	83.85	72.79	99.34	77.27	100
Profile-Profile	86.34	82.29	73.47	95.36	79.55	100
PSI-BLAST	88.55	83.85	72.79	100	89.20	100
Sequence-Profile	85.02	78.13	75.51	100	73.86	100
TM-align	85.30	95.31	76.87	100	90.34	100

Table 5.9. The Percentage of Correct Contacts. The percentage of correct contacts made by each method for the six pairs. The total number of contacts made in the PDB structure file below 5Å distance is shown in the top row. The highest (best) score for each pair (column) is shown in green, the worst in red. The results are for side chain – side chain contacts.

Once the total number of correct contacts had been established, the difference between the distances of these correct contacts in the models built by the various alignment protocols and the distances of the contacts in the PDB structure file was computed (Table 5.10).

Methods	1f34A +1htrB	1avgH +1autC	1ppfE +1pytD	1stfE +1cs8A	2sicE +1r64A	2kaiA +1cvwH
BLAST	1.89	1.48	1.22	0.53	1.26	2.45
COACH	1.95	1.58	1.30	1.05	1.01	0.76
MUSCLE	1.87	1.62	1.12	0.68	1.44	1.60
Profile-Profile	1.90	1.83	1.41	0.72	1.41	0.89
PSI-BLAST	1.80	1.48	0.74	0.70	1.53	1.08
Sequence-Profile	1.95	1.90	0.98	0.61	1.24	1.02
TM-align	1.76	1.86	0.86	0.63	1.04	0.57

Table 5.10. The Average Difference in Distances of Correct Contacts. The average difference in distance, in Angstroms, (between the method and the PDB structure file) for the contacts made below 5Å distance is shown. The lowest (best) score for each pair (column) is shown in green, the worst in red. The results are for side chain – side chain contacts.

Table 5.10 reveals the differences in distances between the contacts modelled by the methods and the actual contacts in the PDB structure file; the smaller the distances, the more accurate the modelling. The pairs sharing over 30% sequence similarity have marginal difference in distances between the methods, probably because there is less variation in the different alignments produced. There appears to be less importance in the choice of alignment method for these pairs (notably 1f34A+1htrB and 1avgH+1autC). PSI-BLAST however, has a small edge over the other methods. 1stfE+1cs8A (sharing ~26% sequence identity) is unusual; BLAST outperforms the other methods with COACH being the method worst by far. The pairs sharing below 20% sequence identity have a broader range of differences in the distances and have a clear favourite: COACH, differing by distances of 1.01Å and 0.76Å (pairs 2sicE+1r64A and 2kaiA+1cvwH, respectively). The worst methods are BLAST (2sicE+1r64A, 1.53Å and PSI-BLAST (2kaiA+1cvwH, 2.45Å).

The specificity of the methods was also tested, as the contacts of the models were assessed within ranges of +/- 1Å, +/- 2Å and +/- 3Å distances from the correct contacts in the structure PDB file. The results for the difference in distance in the contacts which fall into the range of +/- 1Å are listed in Table 5.11, which shows the percentage of correctly identified contacts within +/ 1Å.

Methods	1f34A +1htrB	1avgH +1autC	1ppfE +1pytD	1stfE +1cs8A	2sicE +1r64A	2kaiA +1cvwH
PDB	227	192	147	151	176	36
BLAST	37.00	42.19	41.50	87.42	33.52	13.89
COACH	32.60	37.50	51.02	68.21	58.82	75.00
MUSCLE	34.80	39.06	46.94	83.44	49.43	38.89
Profile-Profile	33.92	32.81	40.14	73.51	46.59	75.00
PSI-BLAST	36.12	46.35	54.42	86.09	47.16	52.78
Sequence-Profile	32.60	30.21	51.02	78.15	52.84	69.44
TM-align	36.56	33.33	54.42	79.47	58.52	88.89

Table 5.11. The Percentage of Correct Contacts Within +/- 1Å. The percentage of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 1Å from the correct contact in the PDB file. The total number of contacts made in the PDB structure file below 5Å distance is shown in the top row. The highest (best) score for each pair (column) is shown in green, the worst in red. The results are for side chain – side chain contacts.

Rather surprisingly, the percentage of "correct" contacts (made within +/- 1Å, Table 5.11) appears to decrease with increasing percentage identity. PSI-BLAST alignments produce models with highest number of correct contacts for two of the pairs (above 30% sequence identity) and Profile-Profile the least. COACH and the Sequence-Profile methods predict the least correct contacts for 1f34A+1htrB, managing to obtain only 32.60% of the correct contacts. For the two pairs under 20% sequence identity COACH, in both instances, predicts the most correct contacts and BLAST the least (in the 2kaiA+1cvwH instance, only predicting a small 13.89% correct contacts), with the Profile-Profile method also predicting 75% of the contacts correct in the pair 2kaiA+1cvwH.

It should be noted here that these distances refer to side chain-side chain contacts, and that little attempt to optimise side chain geometry has been made in the modelling protocol. In practice it would be advantageous to also optimise side chain geometry. Therefore, the "trend" of decreasing quality with increasing percentage identity may be artefactual. It is clear, however, that no single method outperforms any other and that achieving the correct alignment, even at low percentage identities, can lead to highly successful models with over 85% of the native contacts broadly correct. Similarly, even in this small subset of structures, it is clear that some target-template pairs are "easier" to model no matter which alignment method is selected. The obvious example here is the 1stfE-1cs8A pair.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
BLAST	1.48	1.89	1.22	0.53	2.45	1.26
COACH	1.58	1.95	1.30	1.05	0.76	1.01
MUSCLE	1.62	1.87	1.12	0.68	1.60	1.44
Profile-Profile	1.83	1.9	1.41	0.72	0.89	1.41
PSI-BLAST	1.48	1.8	0.74	0.70	1.08	1.53
Sequence-Profile	1.90	1.95	0.98	0.61	1.02	1.24
TM-align	1.86	1.76	0.86	0.63	0.57	1.04

Table 5.12. The Average Difference in Distances of Correct Contacts. The average difference in distance, in Angstroms, (between the method and the PDB structure file) for the contacts made below 5Å distance. The lowest (best) score for each pair (column) is shown in green, the worst in red. The results are for side chain – side chain contacts.

Table 5.12 shows the average difference in the distances of the correct contacts. Here the results are varied but do show some large differences in the distances of the contacts correctly predicted, revealing no one alignment method to be consistently the best at modelling the contacts. For the alignments sharing the lowest sequence identity (2kaiA+1cvwH and 2sicE+1r64A) COACH always achieves the lowest difference in distances by a significant amount (over 0.2of an Angstrom and 0.13 of an Angstrom). For the pairs sharing modest sequence similarity PSI-BLAST, BLAST and MUSCLE out-perform the other methods and the Sequence-Profile, Profile-Profile and COACH method seem perform the worst.

The differences in distances for these correct contacts (contacts within +/- 1Å of the PDB file contacts) can be seen in table A3.6. Again, for the higher sequence identity pairs (above 20% sequence identity) the best method (Table A3.6) at predicting contacts closest to the actual contact varies between the pairs; MUSCLE and Sequence-Profile do achieve the best results for two out of four pairs and Profile-Profile predicts the worst contacts for two out of four pairs, bearing in mind, only up to around 55% of the contacts are correct. 2sicE+1f64A (13.81% sequence identity) implies BLAST predicts the most accurate contacts, at an average of only +/- 0.36Å difference in distances to the correct contacts, and Profile-Profile and PSI-BLAST the least accurate. It is however, worth noting that in this instance that BLAST may have an average of +/- 0.36Å, but this is only for 33.52% of the 176 contacts made in the PDB structure file. If the rest of the methods were investigated (all, including the gold standard TM-align predicting around 50% of the contacts correctly) COACH and the Sequence-Profile methods would be more accurate than BLAST. For 2kaiA+1cvwH the methods order (the most accurate first, having the lowest difference in distance) would be: Profile-Profile, COACH, MUSCLE, Sequence-Profile, BLAST, and finally PSI-BLAST. This does need to be put in context with the amount of correct contacts predicted though; BLAST only predicts 13.89% of the contacts correctly, and MUSCLE only 38.89%. The rest of the methods achieve above 50% correct with COACH and Profile-Profile reaching 75% correct.

The results for the side chain – side chain correct contacts that were obtained with an error of +/- 2Å and +/- 3Å can be seenin Appendix 3, Tables A3.4 and A3.5. It was important to assess the side chain results even if the side chains were not optimised as the side chains interact with the ligand upon binding. These results show similar trends to the results for an error distance of +-/ 1Å, with BLAST and PSI-BLAST predicting the contacts with the largest difference in distance from the PDB structure file contacts, and COACH and Profile-Profile predicting the smallest difference in distances for the pairs below 20% sequence identity. The results do show more contacts are predicted correctly within an error distance of +/- 2Å and +/- 3Å, but this is found to be a trade-off with specificity, since the increase in correct contacts leads to a decrease in the accuracy of the contacts (the difference in distances increases).

For the main chain – main chain interactions (which use the same distance cut-offs) the results can be seen in Appendix 3, Table A3.3 the total number of contacts found in the PDB structure was less compared to the side chain – side chain interactions. The percentage of correct contacts is better than the side chain – side chain results in the majority of instances, except for pair 2sicE+1r64A and alignment method MUSCLE (49.32%) and BLAST (39.19%). With the side chain – side chain interactions MUSCLE achieved 65.34% and BLAST 77.27% for the same pair. 2sicE+1r64A does share a small 13.81% identity and has a relatively high number of main chain – main chain contacts (148) compared to the number of main chain - main chain contacts of the other pairs. Again, TM-align outperforms the other methods when aiming to obtain as many correct contacts as possible.

The average differences in distances between the contacts modelled by the methods and the actual contacts in the PDB structure file for the main chain – main chain interactions were consistently smaller than for the side chain – side chain interactions for the TM-align method, except for the pair 2sicE+1r64A. This could be due to the relatively high number of main chain – main chain interactions of the pair 2sicE+1r64A compared to the number of side chain – side chain interactions of the other pairs. In the majority of instances the differences in distances (for all of the methods) for the main chain – main chain interactions are better (smaller distances) than for the side chain – side chain interactions (Appendix 3, Table A3.6).

For the percentage of correct contacts made by each method for the six pairs (these are the contacts which are only a distance of +/- 1Å from the correct contact in the PDB file) the main chain – main chain results are consistently better. For the pair 2kaiA+1cvwH (10.63% sequence identity) and methods COACH, Profile-Profile, Sequence-Profile and TM-align, 100% of the main chain – main chain contacts are correctly modelled to within +/-1Å of the actual PDB contacts, compared to 75% for COACH, 75% for Profile-Profile, 69% for Sequence – Profile and 89% for TM-align. The differences in distances improved as well, for example improving from 0.52Å to 0.34Å for TM-align, pair 2sicE+1r64A, which has 148 main chain – main chain contacts compared to 176 side chain – side chain contacts. See Appendix 3 Table A3.6 the main chain – main chain, +/- 1Å results.

Again, for the main chain – main chain contact results for the distances of +/-2Å and +/3Å were an improvement over the side chain – side chain results, especially in the case of TM-align, even more correct contacts were obtained with smaller differences in distances than the side chain – side chain results. See Appendix 3, Tables A3.7 and A3.8, for the main chain – main chain results.

The specificity and positive predicted value (formulas and description of calculations in section 5.3.9) have been calculated for the difference in distance ranges one to three, for the side chain – side chain correct contacts; figure 5.10. ROC plots could not be produced for this data as the value for true negatives (contacts not predicted which were correct) would be very high and result in insignificant values.

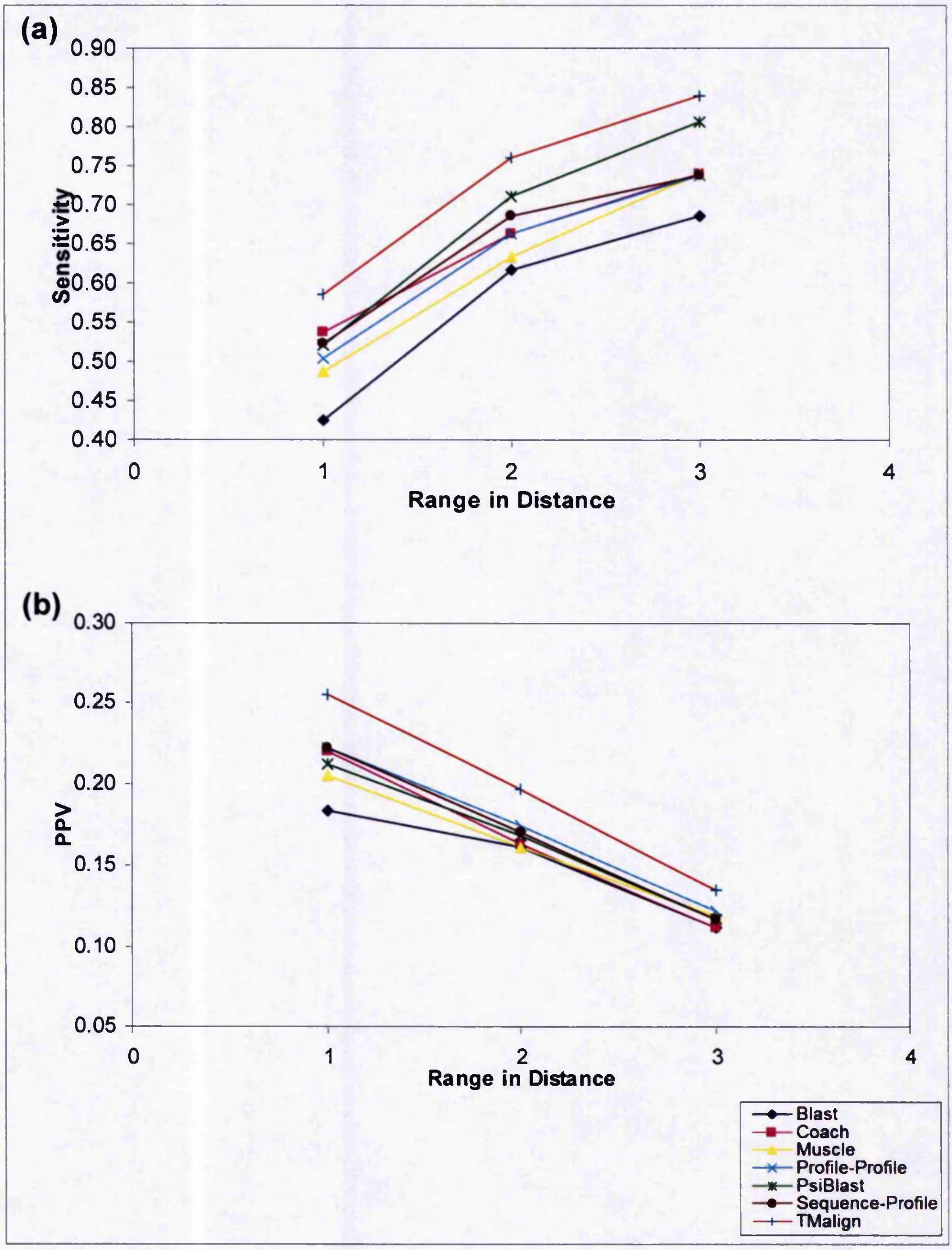


Figure 5.10. The Sensitivity and PPV of the Differences in Distances. The sensitivity and positive predicted value of the different alignment methods as an average over the six pairs, for the side chain – side chain correct contact results.

The main chain – main chain graphs for sensitivity can be found in Appendix 3, Figure A3.1. As the sensitivity of the methods increases (Figure

5.10), the positive predicted value decreases; more correct contacts are predicted but more incorrect contacts are made as well. For the +/- 1Å range, COACH outperforms all the other methods (except the gold standard TM-align) with Sequence-Profile and PSI-BLAST, Profile-Profile, MUSCLE and BLAST following in decreasing sensitivity (Figure 5.10a). This alters as the difference in distance increases; PSI-BLAST seems to increase in sensitivity above the other methods, with most of the methods becoming equally as effective at predicting the contacts. Again in the +/- 1Å range, BLAST is the poorest predictor, with little difference between the other methods, as the range increases, the PPV decreases and the difference in PPV between the methods becomes negligible. It is interesting to note that when the average performance is assessed in terms of contacts, the methods that use multiple sequence alignments and profiles are able to produce the most "useful" alignments, as measured by their ability to produce models with contacts closer to the true ones.

5.4.10 An Example

To provide a fuller understanding of the procedures used in this project, for chapters 4 and 5, it is useful to consider one example (chosen at random) in more detail for the methods BLAST, COACH and TM-align, and the 2sicE+1464A is described here. The alignments made by the three methods are shown in Figure 5.11. When assessing these alignments the positions had to be equivalent (a residue in the target for one method must be assessed against the same residue in the target in another method, on occasion they will be aligned differently if the method chops off some of the target sequence), and hence in Figure 5.11 the starts of the COACH and BLAST alignments for the target have been placed so they align with the starts of the TM-align target sequence. This means that in the BLAST alignment the target residues "DGSGQY" are aligned to the template residues "DITTEY", whereas in TM-align "DGSGQY" are aligned with "G-DITT", resulting in an incorrect alignment for these residues. In COACH, however "DGSGQY" in the target sequence, is aligned with "GD-ITT", the placement of the gap means that the "GS" residues are incorrectly aligned with "-D" instead of "D-", but the rest of the four residues are aligned correctly.

(a)

>2sicE

```
                                                  SQIKAPALHSQGYTGSNVK
VAVIDSGIDSSHPDLK---VAGGASMVPSETN---PFQDNNSHGTHVAGTVAAL--NNSIGVLGVAPSA
SLYAVKVLGADGSGQ-------YSWIINGI---EWAIANNMDVINMSLGGPSG---SAALKAAVDKAVA
SGVVVVAAAGNEGTSGSSSTV-GYPGKYPSVIAVGAVDSSNQRASFS----SVGPELDVMAPGVSIQST
-LPGNKYGAYNGTSMASPHVAGAAALILSKHPN--WTNTQVRSSL
```

>1r64A

```
                                                  SDINVLDLWYNNITGAGVV
AAIVDDGLDYENEDLKDNFCAEGSWDFNDNTNLPKPRLSDDYHGTRCAGEIAAKKGNNFCGV-GVGYNA
KISGIRILSGDITTEDEAASLIYGLDVNDIYSCSWGPADD----GRHLQGPSDLVKKALVKGVTEGRDS
KGAIYVFASGNGGTRGDNCNYDGYTNSIYS-ITIGAIDHKDLHPPYSEGCSAVMAVTYSSGSGEYIHSS
DINGRCSNSHGGTSAAAPLAAGVYTLLLEANPNLTWRDVQYLSIL
```

(b)

>2sicE

```
----------------------AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDL—KV-
--AGGASMVPSETN-----PFQDNNSHGTHVAGTVAAL-NNSIGVLGVAPSASLYAVKVLGADGSGQYS
WIINGIEWAIANNMDVINMSLGGPSGS----AALKAAVDKAVASG---------VVVVAAAGNEGTSGS
SSTV----GYPGKYPSVIAVGAVDSSNQRASFSSVGPELDVMAP----GVSIQSTLP-GNKYGAYNGTS
MASPHVAGAAALILSKHPNWTNTQVRSSLENTTTKLGD-------------------------------
---------------------------------------------------------------------
---------------------------------------------------------------------
---------SFYYGKGLINVQAAAQ
```

>1r64A

```
LLPVKEAEDKLSINDPLFERQWHLVNPSFPGSDINVLDLWYNNITGAGVVAAIVDDGLDYENEDLKDNF
CAEGS-WDFNDNTNLPKPRL---SDDYHGTRCAGEIAAKKGNNFCGVGVGYNAKISGIRILSGD-ITTE
DEAA-SLIYGLDV-NDIYSCSWG-PADDGRHLQGPSDLVKKALVKGVTEGRDSKGAIYVFASGNGGTR-
-GDNCNYDG-YT-NSIYSITIGAIDHKDLHPPYSEGCSAVMAVTYSSGSGEYIHSSDINGRCSNSHGGT
SAAAPLAAGVYTLLLEANPNLTWRDVQYLSILSAVGLEKNADGDWRDSAMGKKYSHRYGFGKIDAHKLI
EMSKTWENVNAQTWFYLPTLYVSQSTNSTEETLESVITISEKSLQDANFKRIEHVTVTVDIDTEIRGTT
TVDLISPAGIISNLGVVRPRDVSSEGFKDWTFMSVAHWGENGVGDWKIKVKTTENGHRIDFHSWRLKLF
GESIDSSKTE----------------
```

(c)

>2sicE

```
          -------------AQ-SV--PYGVS-------QIKAPALHSQGYTGSNVKVAVIDSGID
SSHPDL-K-VAG--GASMVPSETNPFQDN--NSHGTHVAGTVAALNN-SIGVLGVAPSASLYAVKVLGA
DGSGQYSWIINGIEWAIANNMDVINMSLGG------PSG-SAALKAAVDKAVAS-----GVVVVAAAGN
EGTSGSSSTV-GYPGKYPSVIAVGAVDSSNQRASFSSVGPELDVMAPGV----SIQSTLPGNKYG-AYN
GTSMASPHVAGAAALILSKHPNWTNTQVRSSLENTTTKL-----G-D---------SFYYGKGLINVQA
AAQ------------------------------------------------------------------
---------------------------------------------------------------------
------------
```

>1r64A

```
          LLPVKEAEDKLSINDPLFERQWHLVNPSFPGSDINVLDLWYNNITGAGVVAAIVDDGLD
YENEDLKDNFCAEGSWDFNDNTNLPKPRLSDDYHGTRCAGEIAAKKGNNFCGVGVGYNAKISGIRILS-
G-DITTEDEAASLIYGL-DVNDIYSCSWGPADDGRHLQGPSDLVKKALVKGVTEGRDSKGAIYVFASGN
GGTRG-DNCNYDGYTNSIYSITIGAIDHKDLHPPYSEGCSAVMAVTYSSGSGEYIHSSDINGRCSNSHG
GTSAAAPLAAGVYTLLLEANPNLTWRDVQYLSILSAVGLEKNADGDWRDSAMGKKYSHRYGFGKIDAHK
LIEMSKTWENVNAQTWFYLPTLYVSQSTNSTEETLESVITISEKSLQDANFKRIEHVTVTVDIDTEIRG
TTTVDLISPAGIISNLGVVRPRDVSSEGFKDWTFMSVAHWGENGVGDWKIKVKTTENGHRIDFHSWRLK
LFGESIDSSKTE
```

Figure 5.11. The Alignments of 2sicE+1r64A. The alignments made by (a) BLAST, (b) COACH and (c) TM-align for the pair 2sicE+1r64A. Equivalent target residues are in blue, template residues in red.

For this pair, sharing only 13.81% sequence similarity, BLAST retained 88% of its target and 55% of its template, whilst TM-align and COACH retained 100% of both sequences. However, 100% of the target sequence from the BLAST alignment was model-able compared to 93% from COACH and 69% from TM-align. COACH and TM-align both share around 45% gaps in their target sequence in the alignment whilst BLAST only has around 10%. BLAST also has 10 and 3 gapped instances in the target and template sequence, respectively, whilst COACH has 11 and 10 and TM-align has 19 and 3 gapped instances in their target and template sequences. This suggests (in this instance anyway) that retaining all of the alignment and having more gapped instances in the target sequence rather than the template sequence is important for a more accurate (or strictly, more useful) alignment (TM-align being the gold standard and COACH outperforming BLAST when looking at the contacts, not overall). It also implies that the amount of alignment that is model-able and the amount of gaps in the target aren't necessarily as important, since TM-align and COACH have less model-able alignment and more gaps introduced into the target than BLAST has. COACH seems to have difficulty in aligning the last part of the target sequence and introduces a large gap at the end of the target sequence. This could be due to the lack of optimisation of the gap penalties in the sequence alignment.

```
>2sicE
000000000000000000000000000000000000000000000000000000000000001210000
000000000000000000000000012112222221121000000000000000122222210000000
000000000000121222100000000000000000000000000000012100000000000000000
00000000122112100000000000000000000000000000000000000000000000000000
AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLKVAGGASMVPSETNPFQDNNSHGTHVA
GTVAALNNSIGVLGVAPSASLYAVKVLGADGSGQYSWIINGIEWAIANNMDVINMSLGGPSGSAALKAA
VDKAVASGVVVVAAAGNEGTSGSSSTVGYPGKYPSVIAVGAVDSSNQRASFSSVGPELDVMAPGVSIQS
TLPGNKYGAYNGTSMASPHVAGAAALILSKHPNWTNTQVRSSLENTTTKLGDSFYYGKGLINVQAAAQ
```

Figure 5.12. The Interface Assignments: 2sicE. The interface assignments used for the target 2sicE. 1r64A does not have any since it is the template in the I-vs-S set. Highlighted in blue are the residues used in the explanation.

The interface was assigned for 2sicE (+1r64A) as in Figure 5.12, and has been assigned four interface regions in total. This is not counted as six since the small interface regions (121) will not reach the threshold of three or more category "2"s in the seven residue sliding window. After assessing the alignments relative to the TM-align gold standard, it was found that BLAST correctly aligned 25% of the target correctly, with 43% of the interfacial positions being correct. By comparison COACH obtained 42% of the alignment correctly, with 65% of the interfacial positions aligned correctly. In Figure 5.11, the residues "DGSGQY" of the target 2sicE, described in the assessment of the alignments as an example, are shown in blue. These residues are assigned as being part of the interface, and BLAST fails to align any of these interface residues correctly whilst COACH aligns four out of six correctly. Although a Serine residue is present in this interface it is not the active site Serine (Serine 221).

Once the models were built with the different alignment protocols, they were assessed. Models built from BLAST alignments had an average global RMSD (over all of the standard refinement models) of 8.09Å, those from COACH alignments an average global RMSD of 6.93Å and from TM-align 1.97Å, with a global interfacial residues RMSDs of 7.12Å, 4.12Å, 2.32Å for BLAST, COACH and TM-align alignments respectively. However, the average local RMSDs were 1.89Å, 1.61Å, 0.86Å for the interface regions of the alignments and 1.57Å, 1.15Å, 0.67Å for the non-interface regions of the alignments for BLAST, COACH and TM-align respectively. In this case, for the pair 2sicE+1r64A, as with the overall trend, the interface appears to be modelled more accurately than the non-interface only on a global scale, but not when assessing the model's quality using local RMSDs. The structures of the models for BLAST, COACH and TM-align compared to the PDB target actual structures can be seen in figure 5.13. It is possible to see that overall the alignment method TM-align produces a model (red structures) closer in structure to the PDB structure (blue structures) than either BLAST or COACH, with COACH outperforming BLAST, paying particular attention to the helices at the front of the peptidase chain structures and to the lower right hand loop.

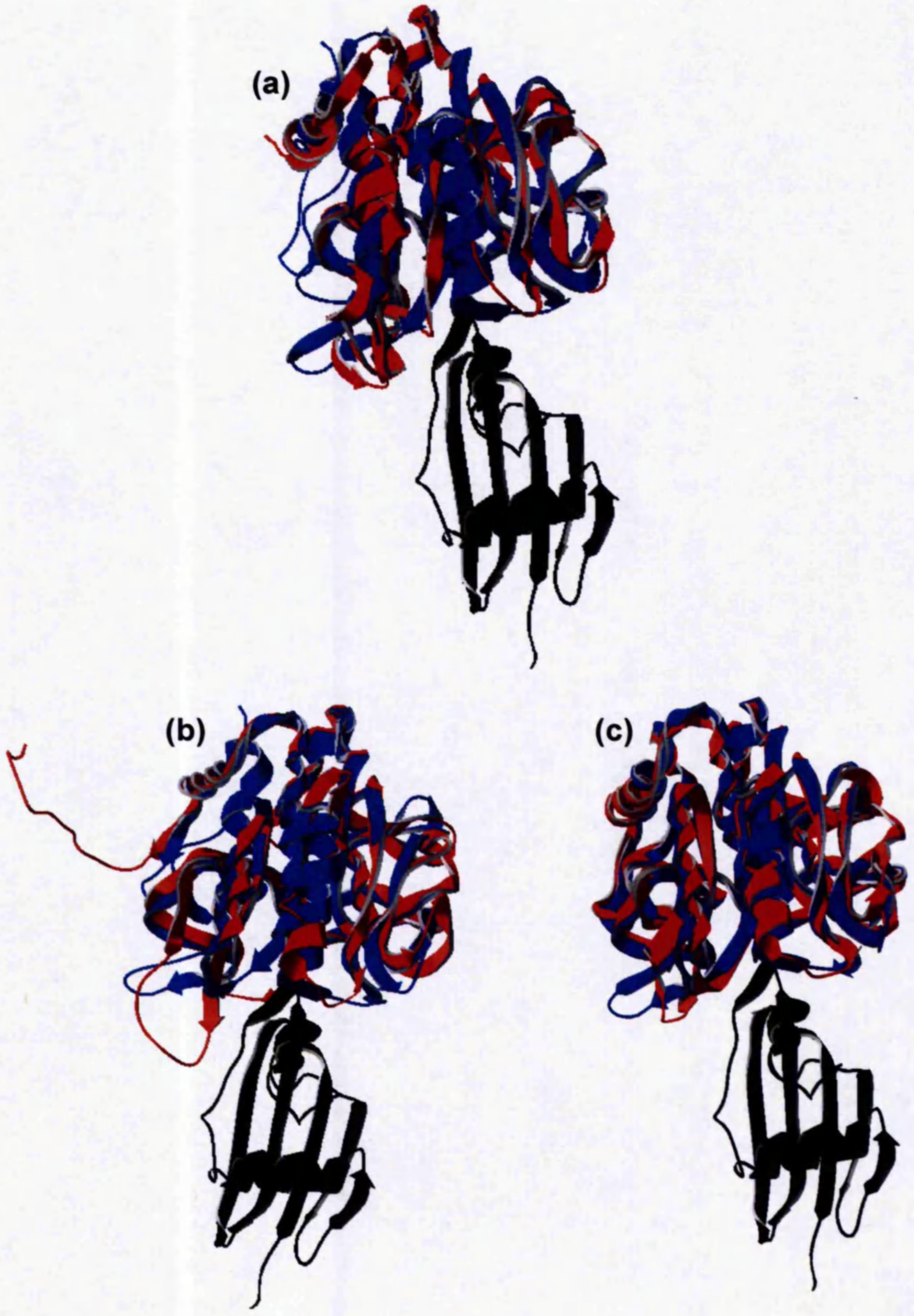


Figure 5.13. The Models of 2sicE. The blue structures are the PDB structure of the peptidase chain of 2sic, the red structures are the modelled structures of the peptidase chain of 2sic, and the black structure is the PDB structure of the inhibitor chain of 2sic. (a) compares the BLAST modelled structure with the PDB structure, (b) is COACH's modelled structure and (c) is TM-align's. SPDB Viewer was used to create these images (http://www.expasy.org/spdbv).

The contacts made between the peptidase chain and the inhibitor chain in the target PDB were compared to the modelled structure contacts. Overall (contacts correctly predicted within +/- 5Å error of the actual contacts) BLAST achieves an average accuracy of 65.34% contacts reproduced in the model, COACH an accuracy of 75.70% and TM-align 90.34% for the 2sic structure, with average difference in distances of 1.26Å, 1.01Å and 1.04Å for BLAST, COACH and TM-align respectively. Although COACH obtains a lower difference in distance than TM-align, TM-align gets around 15% more of the contacts correct. When the error is reduced to +/- 1Å, BLAST's accuracy falls to 33.52%, COACH's to 56.82% and TM-align's to 58. 52%. As the margin for error is increased to +/- 3Å, the accuracy also increases, as might be expected since the distance cut-off is now more generous. As the margin of error decreases, fewer of the correct contacts are modelled, but it is still possible for COACH to obtain 56.82% of the contacts correctly that exist within +/- 1 Å, having an average difference in distance of 0.52Å. Figure 5.14 displays the side-chains that participate in contacts made between the inhibitor chain and the peptidase chain of the target 2sic; for clarity, the inhibitor chain is not shown. There are a few instances where it appears that COACH models side-chains better than TM-align (the green and yellow structures of COACH are closer to one another than the green and yellow structures of TM-align), but overall BLAST models the contacts worse than COACH, and COACH models the contacts worse than TM-align.

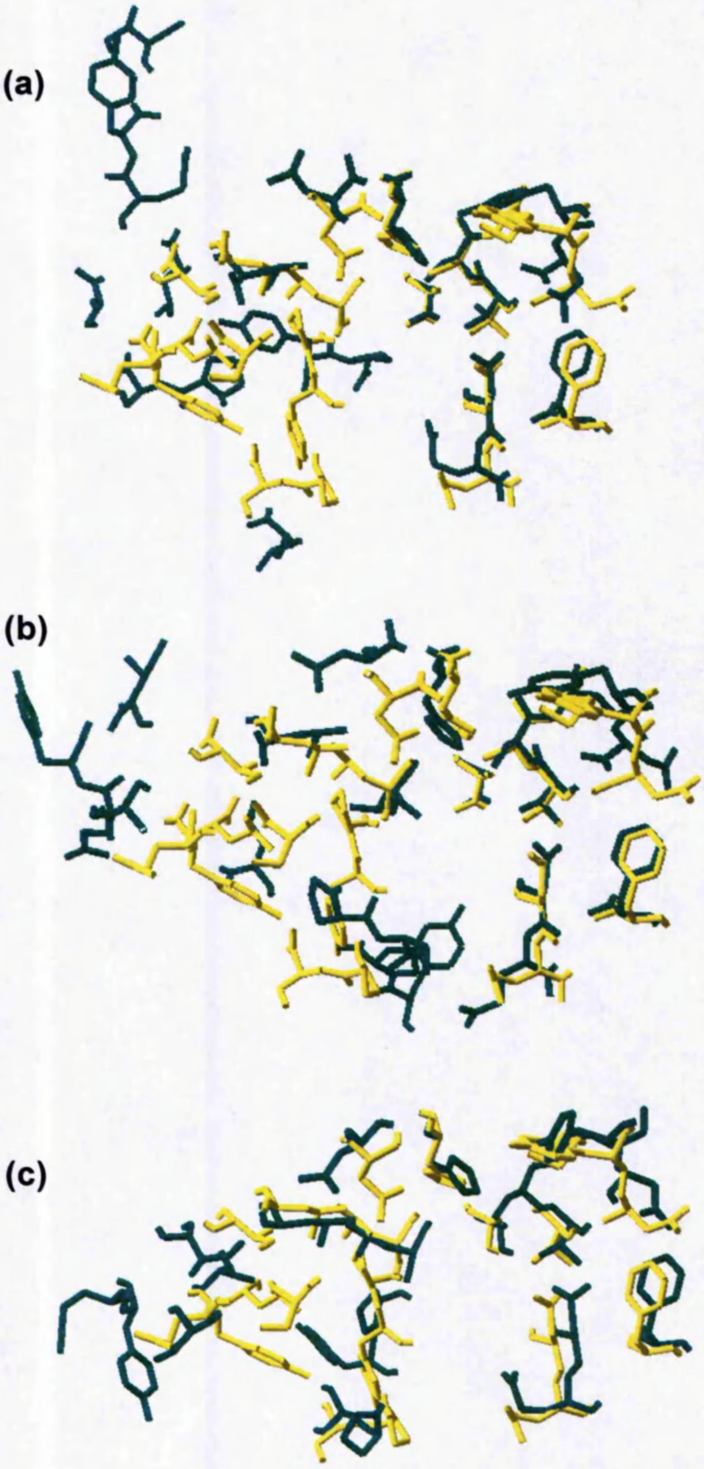


Figure 5.14. The Contacts of 2sicE. The yellow structures of the side-chains involved in contacts between the peptidase chain and inhibitor chain are from the PDB structure, the green from the modelled structure. (a) compares the BLAST side-chains involved in modelled contacts with the PDB structure contacts, (b) is COACH's modelled side-chains and (c) is TM-align's.

5.5 CONCLUSIONS AND FUTURE WORK

This chapter covers the investigation into the ability of different alignment protocols to produce models using the resulting alignment. The alignments were assessed in chapter 4. The results in this chapter reflect just the I-vs-S set since, as explained previously, not enough target-template alignment pairs existed sharing below 20% sequence in the I-vs-I set. For each target-template alignment generated by the various alignment protocols, a total of seventy models were built, representing numerous refinement levels of the models and the loop structures within the models. These models were then tested globally and locally with respect to the actual structure of the target in terms of the RMSDs of the models overall and residue segments. The contacts made between the inhibitor chain and the peptidase chain of the target were also validated by determining the number of correctly predicted contacts, as well as the difference distance these contacts were predicted in relation to the actual contacts. Six alignment target-template pairs were chosen from the I-vs-S dataset to assess in more detail, eventually leading to one pair which was examined in more detail. The aim was to find out how accurate the interface residues could be modelled in and around the traditional "twilight zone", compared to the non-interfacial residues, and which alignment method would lead to models with the most precision in these recognition regions. This potentially could lead to an optimal protocol for modelling the interface regions to an acceptable level of accuracy into the twilight zone, even though normally modelling is not expected to be generally high quality with such distance homologues as templates.

After the completion of this project, it has been concluded that modelling into the twilight zone is by no means a trivial task, with results being complex to decipher. There are however, a few concepts that present themselves with some consistency.

Only MUSCLE provided alignments which resulted in all of the models being built for every pair. It was not possible to obtain all of the models using the other alignment protocols. Fewer models were built in the loop set, these proved more difficult to build than the standard models. This was probably due to more challenging modelling tasks presented to MODELLER, whereby the solution of spatial restraints could not be satisfied and the modelling protocol failed.

As the percentage identity between the target and template decreases, as expected, the quality of the models produced from this alignment also decreases. Models produced from alignments sharing below 20% sequence similarity had RMSDs nearly double of those sharing below 80% sequence identity, however these RMSDs are still considered useable.

On average, and based on the global RMSDs of the models, it appears that BLAST is consistently the best alignment method to use across all of the percentage identities and MUSCLE the worst. However, this somewhat surprising result must be examined in context, and considered properly in the context of the problem at hand in this study. This result is likely a consequence of BLAST retaining the least of its alignment, which in turn leads to a lower RMSD. Fewer residue positions are therefore evaluated in the RMSD calculation, and moreover, they will probably be the easiest to align and most conserved as a consequence of BLAST being a local alignment algorithm. Hence, this will lead to a lower global RMSD for BLAST alignments. In comparison MUSCLE retains all of its target sequence in the alignment and MODELLER therefore builds even the most difficult models/segments. Additionally, more of the target sequences are "lost" as the percentage sequence identity decreases. As the percentage identity of the pairs decrease, the profile and HMM methods (containing more evolutionary information) achieve lower RMSDs than the sequence based methods. It is also evident that more gaps are introduced into the alignment as the percentage sequence identity decreases and the model becomes less reliable. This is because gaps in the template mean that the equivalent portion of the target has no equivalent structure onto which to build the model. Fewer gaps tend to occur in the local alignment methods, offering some advantage over the global methods, which have to retain both of the sequences in the alignment.

Usually the percentage of alignment that is model-able is not an issue, unless a particularly difficult alignment is attempted, with sequences of greatly differing length. TM-align, when faced with this prefers to reduce the number of gapped instances and increases the gap penalty, resulting in a small percentage of the original alignment being model-able. MODELLER can not build the start and/or end of the models if there are overhangs of gaps at the start and end of the alignment. However, as TM-align prefers, it is more productive to have an alignment which is less model-able and has fewer gapped instances in the alignment breaking up the target sequence.

Using loop refinement seems to increase the average RMSD values of the models, and increases the variation of the resulting RMSDs of the different models. If there are alignment errors at the stem residues, loop modelling is not likely to result in an accurate model. This means loop modelling is most useful for target sequences that share more than 30% sequence identity with the template structures (Fiser *et al.*, 2000).

The importance of obtaining highly accurate alignments in terms of comparative modelling is demonstrated by the result that TM-align (the gold standard) always outperforms the other methods in every way. This is not only a reassuring result, but highlights the optimal theoretical performance that could be achieved by any comparative modelling performance based on this target-template pairing. The fact that all methods do not achieve the model accuracy of the gold standard TM-align suggests that there is still room for improvement, even in the most accurate of alignment protocols. Although in most cases, even TM-align does not build models with 100% accuracy, allowing for the errors being introduced in comparative modelling in the other steps o f the model building process.

When the global RMSD of the interface was compared to the global RMSD of the whole (interface and non-interface residues) modelled structure, the RMSD of the interface was always lower than the whole structure and BLAST provided the best models and MUSCLE the worst. The increase in accuracy for the interface may, in part, be a result of fewer residues being assessed. However, the recognition region is made up from discontinuous segments of the protein in all cases, and therefore it still remains challenging to build these with high quality just like the protein fold as a whole. In other words, it is still challenging for MODELLER to locate the different segments that make up the recognition surface with the correct relative disposition to each other. This suggests that this result is not really artefactual, but genuinely represents superior performance in structure prediction at the protease recognition interface. Nevertheless, the local, average, RMSDs were also calculated; this consistently resulted in the non-interface region having lower RMSDs. Th is result too, needs careful consideration. Firstly, the non-interface regions cover much of the protein, including both accessible surface and buried core. As shown in Figure 5.9, highly accessible regions of the protein are not generally modelled as well as those in the core. Hence, given that recognition regions are by definition accessible, the comparison between interface and non-interface might be biased against the interface regions. Another factor to consider is that peptidase active sites contain interface regions with both highly variable positions as well as highly conserved regions. The key residues in the mechanistic catalytic part of the active site will be conserved, but those in the specificity pockets which give each protease its primary cleavage rules must be more variable across a protein family, to provide the diversity in function which is observed in biological systems. The variable conservation of interface results in low and high accuracy regions, thus, higher accessible regions of the interface can also contribute to lower precision in the modelling of the interface. This is a major challenge for predicting protease function, since arguably the most interesting residues which provide specificity to the enzyme are the hardest to model.

The contacts between the peptidase chain and the inhibitor chain of the models could be predicted with relatively high accuracy, even if the rest of the model was not so precise at low percentage identities. Indeed, in one instance, the COACH alignment resulted in a model with 56.82% of the contacts correctly within +/-1Å (averaging a distance of 0.52Å) of error for a pair sharing only around 13% sequence identity. COACH appears to be the most accurate and BLAST the worst method at modelling these contacts. As the percentage sequence identity increases, the method to choose for accurate contact modelling becomes less clear, with all of the methods modelling contacts to similar accuracies. There is a trade-off between the amount of correctly modelled contact residues and the accuracy of these contacts; the more contacts that are correctly modelled, the less accurate the contacts become.

Overall, trying to predict an accurate comparative model is a non-trivial task, involving many different steps. When trying to model accurate interface regions of an alignment pair with a sequence identity in the twilight zone, the important aspects include: keeping gapped instances in the alignment to a minimum, some refinement being applied to the whole model, choosing the model with the lowest global RMSD in this case as the structure is known (taking into consideration the local RMSDs may not always reflect this choice), and choosing an alignment method which uses profiles or HMMs. This is by no means a protocol to be used in every modelling task, but hopes to provide a better understanding of the important concepts to consider when building a comparative model with an accurate interface using an alignment sharing a sequence identity falling into the twilight zone.

Building comparative models accurately has important applicability, amongst others, drug design is one area. It has been said that most proteins to be modelled fall into the twilight zone and one of the downfalls of comparative modelling is the difficulty in obtaining an accurate model when the sequence identity of the alignment pair reaches into the twilight zone. Most models are deemed useless when their alignment shares less than 20% sequence identity. If the interactions between a protein and its inhibitor could be modelled to modest accuracy, regardless of the accuracy of the global structure, specificity predictions could be useful, even with pairs sharing low sequence similarity. This chapter has hopefully revealed that this is plausible.

In order to extend the progress in this chapter, it may be beneficial to investigate loop modelling and side-chain modelling further. If the results here suggest the interface can still be modelled fairly accurately in the twilight zone, with minimal loop refinement and no loop modelling or side-chain modelling, it is plausible that the interface accuracy could be improved. Since the interface contains regions of highly similar and variable regions, one suggestion to improve loop modelling would be to attempt loop modelling on the conserved regions first, then complete fragment assembly on the variable regions, finally applying a loop closure method to model the loops. More work could be done on finding better ways to assess the models, for example, finding a way to normalise the RMSD results. Improvements could be made in the modelling stage of the project, a more formal loop modelling method would be beneficial as well as incorporating multiple templates into the model building step. There is also the possibility of identifying regions which are likely to result in poor models (or high quality ones) from the alignments and of assessing the model quality for low quality regions.

6. CONCLUSIONS AND Outlook

The prediction of protein structure is an important step to help bridge the sequence-structure gap that exists at present in structural biology, and is also an essential asset in obtaining structures for understanding protein function. The three-dimensional structure of a protein is important for precisely predicting the binding mode between a protein and its ligand. This knowledge also plays a key part in accurately designing drugs to bind in a desired manner (DeWeese-Scott & Moult, 2004).

Comparative modelling is seen as one of the most accurate methods for protein structure prediction (Venclovas & Margelevicius, 2003) and is most useful when target-template pairs share above 30% sequence identity. This would be appropriate if most potential modelling pairs existed above 30% sequence identity but most actually exist below 30% identity (section 3.5.1).

One of the major limiting factors of comparative modelling is producing good quality models within the twilight zone. Errors accumulate in the alignment step and frequently, when building non-trivial models, in the loop regions. Even if the resulting models aren't perfect, it is still possible to extract useful information in most cases. Modelling loop regions accurately, which tend to correspond to interface regions, is important for ascertaining the function of the protein.

Thus, this project was focused on improving protein structure prediction using comparative modelling in the twilight zone. It also aimed at improving the prediction of loop regions (more specifically the N-termini of alpha-helices) and determining me thods for accurately aligning target -template pairs w ithin th e twilight zone. The distinction between methods for predicting the interface more accurately than the rest of the protein was made, resulting in useful implications in comparative modelling.

Secondary structure prediction was used to improve the modelling of proteins sharing below 30% sequence similarity. Using the improved secondary structure prediction program ELEPHANT (Wilson *et al.*, 2004) to provide restraints in the model building process, loop regions (specifically the N-termini of alpha-helices) were modelled more accurately. ELEPHANT was found to improve the fringes of alpha-helices whilst maintaining the overall accuracy of the secondary structure. The ELEPHANT algorithm worked on the basis that secondary structure is based on multiple sequence alignments, and secondary structure prediction methods frequently identify the correct core regions but not the fringes of the secondary structures due to the difference in lengths in protein families. A comparison of the target secondary structure predicted by JPRED and ELEPHANT and the secondary structure of the template assigned by DSSP was compared to the actual (assigned by DSSP) secondary structure of the target, revealing the predicted secondary structure of the target (combined with the actual secondary structure of the template) being closer in structure to the actual structure of the target than the actual secondary structure of the template (the secondary structure of the template alone). This was completed across a range of percentage sequence identities. More improvement was found in and around the twilight zone where most potential models exist. It was found that using the ELEPHANT secondary structure restraints to model the N-termini of alpha-helices produced more accurate models than using the JPred secondary structure restraints. This offered the potential for improved comparative modelling, particularly of loop fringes, stretching into the twilight zone with implications in drug design.

To help benefit from accurately predicted models, a study on various alignment protocols was completed, including sequence based and profile based methods. The alignments were assessed as a whole, as well as local regions containing interface residues. This was completed in the hope that the recognition regions would be modelled more accurately than expected in and around the twilight zone. On average, as the percentage identity decreased, the alignment accuracy decreased, with more gaps (and frequently of longer lengths) being inserted into the alignments. Local alignment methods did not retain all of the sequence submitted, however, only CE and BLAST were found to discard parts of interface. Lower percentage identity pairs increased the alignment accuracy variation between the methods, the more sophisticated profile ones outperforming the sequence based. COACH performed the best in the 20% and under sequence identity category, with 50% of the residues being correctly aligned with reference to TM-align.

The interface regions of the alignment were more accurately aligned than the rest of the alignment, even below 20% identity. Again, profile methods and methods based on hidden Markov models were better at aligning interface regions than the other methods. This increase in accuracy may be due to parts of the interface being more conserved than the rest of the peptidase. This is a result of the additional evolutionary pressures exerted on the recognition regions, and the profile/HMM methods are better at distinguishing evolutionary related positions from non-related positions. For example, using the COACH method and a target-template pair having below 20% sequence identity, the interface positions were aligned up to 80% correct. The corresponding non-interface residues were aligned with around 50% accuracy.

The alignments were then used to build comparative models. The ability of the different alignment protocols to produce accurate comparative models was assessed. Also their ability to model the interface regions more accurately than the non-interface regions at low percentage sequence identity was determined. Different refinement levels were applied to models and to the loops in the models in the hope that the interface regions may be modelled more precisely.

As the percentage sequence identity stretched into the twilight zone, the average RMSDs of the models nearly doubled in comparison to those pairs sharing above 20% sequence identity. Increasing numbers of gaps in the alignment resulted in higher global RMSDs. The more conserved the multiple sequence alignment was in the building of the alignment used in the modelling process, the more accurate the model was. A correlation also existed between the accessibility of the proteins and the RMSD; as the accessibility increased, the RMSD also increased. These trends were more pronounced for TM-align.

Model refinement and loop refinement reduced the global RMSDs in a small percentage of pairs, broadening the minimum and maximum RMSD of the multiple models built. Even TM-align did not achieve models with 100% accuracy; errors in the other steps in the modelling process were introduced.

It appeared that the local alignment method BLAST was the best at producing models with the lowest global RMSDs in all of the percentage identity areas. This may be due to the more difficult aspects of the alignment being discarded, as well as some of the important recognition regions, which is not advantageous. Since this project was focused on obtaining high quality interface regions rather than globally correct models, the local RMSDs of the recognition regions were also obtained. The RMSDs of the local interface regions seemed to be higher than the RMSDs of the local non-interface region, this may be due to highly variable and also surface regions of the peptidase .

To assess the modelled interfaces in more detail, the number of correctly modelled contacts and the contact difference in distances was obtained for six pairs; two having below 20% sequence identity. As percentage sequence identity increased, the most reliable method to model the contacts became less clear. Well into the twilight zone, trends emerged; when modelling the contacts, the profile methods performed better than the sequence based methods, and COACH outperformed all of the other methods, achieving more correct contacts and smaller average differences in distances. In one instance, the COACH alignment resulted in a model with 56.82% of the contacts correctly within +/-1Å (averaging a distance of 0.52Å) of error for a pair sharing only around 13% sequence identity. The more contacts modelled (increasing the distance range) the greater the difference in distances between the actual contacts and the modelled contacts; a trade-off between specificity and sensitivity occurred.

It may be advantageous to assess the area of template selection in comparative modelling; this is another step where errors are present. Using multiple templates may improve the building of models. One suggestion would be to try and combine multiple templates after investigating which ones, and which parts, would be the best to use. It is thought that these hybrids could include secondary structure predictions of the target and the actual secondary structure of the template, s o experiments into how these hybrids could be combined would be needed. The use of more profile based methods in the construction of the alignments and more optimisation for the individual methods would probably result in more accurate models, as would the use of loop and side-chain modelling with more emphasis on the refinement procedure. Investigations into the definition and assignment of interface residues may result in more accurate model building. This may be aided by some form of loop modelling being performed on the conserved regions of the interface, followed by fragment assembly on the variable regions, using a loop closure method to complete the loop. Assessing the contacts of more pairs and pairs where both of the target and template contained an inhibitor would also be interesting.

In structural genomics, a major goal is to obtain a set of useful models for all protein domains. One definition of "useful" is that the functional information that can be deduced from the model approaches that which could be obtained from an experimental structure (DeWeese-Scott & Moult, 2005). It is clear from the analysis presented in this Thesis that modelling proteins below and within the twilight zone is a non-trivial task, with many variables to consider. The success of aiding the prediction process of loop regions using secondary structure prediction, and determining methods more suited to the prediction and modelling of interface regions within the twilight zone may provide the basis for an enhanced approach to this prediction process.

7. REFERENCES

Al-Lazikani, B., Jung, J., Xiang, Z., & Honig, B. (2001). Protein structure prediction. **Curr Opin Chem Biol, 5**(1), 51-56.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. **J Mol Biol, 215**(3), 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res, 25**(17), 3389-3402.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. **Science, 181**(96), 223-230.

Armougom, F., Moretti, S., Keduas, V., & Notredame, C. (2006). The iRMSD: a local measure of sequence alignment accuracy using structural information. **Bioinformatics, 22**(14), e35-39.

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. **Science, 294**(5540), 93-96.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. **Bioinformatics, 15**(11), 937-946.

Barrett, A. J., Rawlings, N. D., & O'Brien, E. A. (2001). The MEROPS database as a protease information system. **J Struct Biol, 134**(2-3), 95-102.

Barton, G. J., & Sternberg, M. J. (1987). A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. **J Mol Biol, 198**(2), 327-337.

Bates, P. A., Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. **Proteins, Suppl 5**, 39-46.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J., & Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. **Nature, 326**(6111), 347-352.

Bolognesi, M., & Smith, J. (2006). Proteins: Stay Tuned! **Current Opinion in Structural Biology, 16**, 710-713.

Bower M. J., Cohen F.E., and Dunbrack R.L., Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. **J Mol Biol, 267,** 1268-1282.

Bowie, J. U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. **Science, 253**(5016), 164-170.

Brandon, C., & Tooze, J. (1999). Introduction to Protein Structure. **Garland.**
Bystroff, C., Thorsson, V., & Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. **J Mol Biol, 301**(1), 173-190.

Bruccoleri, R, E., Karplus, M. (1987). Prediction of the folding of short polypeptide segments in proteins by systematic search. **Biopolymers 26**,137-168.

Burke, D., Deane, C., Blundell, T. (2000). Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. **Bioinformatics 16**: 513-519.

Chakrabarti, S., Bhardwaj, N., Anand, P. A., & Sowdhamini, R. (2004). Improvement of alignment accuracy utilizing sequentially conserved motifs. **BMC Bioinformatics, 5**, 167.

Chandonia, J. M., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., & Brenner, S. E. (2002). ASTRAL compendium enhancements. **Nucleic Acids Res, 30**(1), 260-263.

Chen, C. C., Singh, J. P., & Altman, R. B. (1999). Using imperfect secondary structure predictions to improve molecular structure computations. **Bioinformatics, 15**(1), 53-65.

Chen, L., Zhou, T., & Tang, Y. (2005). Protein structure alignment by deterministic annealing. **Bioinformatics, 21**(1), 51-62.

Chen, Y., & Crippen, G. M. (2005). A novel approach to structural alignment using realistic structural and environmental information. **Protein Sci, 14**(12), 2935-2946.

Chou, P. Y., & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. **Biochemistry, 13**(2), 211-222.

Chou, P. Y., & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. **Adv Enzymol Relat Areas Mol Biol, 47**, 45-148.

Contreras-Moreira, B., Fitzjohn, P. W., & Bates, P. A. (2003). In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. **J Mol Biol, 328**(3), 593-608.

Cozzetto, D., & Tramontano, A. (2005). Relationship between multiple sequence alignments and quality of protein comparative models. **Proteins, 58**(1), 151-157.

Crooks, G. E., & Brenner, S. E. (2004). Protein secondary structure: entropy, correlations and prediction. **Bioinformatics, 20**(10), 1603-1611.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. **Bioinformatics, 14**(10), 892-893.

Dayhoff, M. O., & Eck, R. V. (1968). Atlas of Protein Sequence and Structure. **National Biomedical Research Foundation, Silver Spring, Md**.

DeWeese-Scott, C., & Moult, J. (2004). Molecular modeling of protein function regions. **Proteins, 55**(4), 942-961.

Dunbrack, R. L., Jr. (2006). Sequence comparison and protein structure prediction. **Curr Opin Struct Biol, 16**(3), 374-384.

Dunbrack & Karplus (1993). R.L. Dunbrack, Jr and M. Karplus, Backbone-dependent rotamer library for proteins: Application to side-chain prediction. **J. Mol. Biol. 230**, 543–574.

Eddy, S. R. (1998). Profile hidden Markov models. **Bioinformatics, 14**(9), 755-763.

Eddy, S.R. (2001). HMMER: Profile hidden Markov models for biological sequence analysis. http://hmmer.wustl.edu.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Res, 32**(5), 1792-1797.

Edgar, R. C., & Batzoglou, S. (2006). Multiple Sequence Alignment. **Current Opnion in Structural Biology, 16**, 368-373.

Edgar, R. C., & Sjolander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. **Bioinformatics, 20**(8), 1309-1318.

Feldman, D. E., & Frydman, J. (2000). Protein folding in vivo: The importance of molecular chaperones. *Curr. Opin. Struct. Biol.* **10:** 26–33.

Fernandez-Fuentes, N., Zhai, J., & Fiser, A. (2006). ArchPRED: a template based loop structure prediction server. **Nucleic Acids Res, 34**(Web Server issue), W173-176.

Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. **Proteins, 51**(3), 434-441.

Fiser, A., Do, R. K., & Sali, A. (2000). Modeling of loops in protein structures. **Protein Sci, 9**(9), 1753-1773.

Fiser, A., Feig, M., Brooks, C. L., 3rd, & Sali, A. (2002). Evolution and physics in comparative protein structure modeling. **Acc Chem Res, 35**(6), 413-421.

Fiser, A., & Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. **Methods Enzymol, 374**, 461-491.

Fiser, A., & Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. **Bioinformatics, 19**(18), 2500-2501.

Fiser, A., Sanchez, R., Melo, F., & Sali, A. (2001). Comparative Protein Structure Modeling. **Computational Biochemistry and Biophysics**, 275-312.

Floudas, C., Fung, H., McAllister, S., Monnigmann, M., & Rajgaria, R. (2006). Advances in Protein Structure Prediction and De Novo Protein Design: A Review. **Chemical Engineering Science, 61**, 966-988.

Friedberg, I., Jambon, M., & Godzik, A. (2006). New avenues in protein function prediction. **Protein Sci, 15**(6), 1527-1529.

Friedberg, I., Kaplan, T., & Margalit, H. (2000). Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. **Protein Sci, 9**(11), 2278-2284.

Frishman, D., & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. **Protein Eng, 9**(2), 133-142.

Fuentes-Prior, P., Noeske-Jungblut, C., Donner, P., Schleuning, W.D., Huber, R., Bode, W. (1997) Structure of the thrombin complex with triabin, a lipocalin-like exosite-binding inhibitor derived from a triatomine bug. Proc.Natl.Acad.Sci.USA **94:** 11845-11850

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. **J Mol Biol, 120**(1), 97-120.

Godzik, A. (1996). The structural alignment between two proteins: is there a unique answer? **Protein Sci, 5**(7), 1325-1338.

Godzik, A., & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. **Proc Natl Acad Sci U S A, 89**(24), 12098-12102.

Goldsmith-Fischman, S., & Honig, B. (2003). Structural genomics: computational methods for structure analysis. **Protein Sci, 12**(9), 1813-1821.

Greer, J. (1990). Comparative modeling methods: application to the family of the mammalian serine proteases. **Proteins, 7**(4), 317-334.

Gribskov, M. (1994). Profile analysis. **Methods Mol Biol, 25**, 247-266.

Gribskov, M., Luthy, R., & Eisenberg, D. (1990). Profile analysis. **Methods Enzymol, 183**, 146-159.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. **Proc Natl Acad Sci U S A, 84**(13), 4355-4358.

Grishin, N. V. (2001). Fold change in evolution of protein structures. **J Struct Biol, 134**(2-3), 167-185.

Gto, N. (1976). Statistical mechanics of protein folding, unfolding and fluctuation. **Adv Biophys.** 65-113.

Guzzo, A., V., (1965). The influence of amino acid sequence on protein structure. ***Biophys, J.*** **5**:809-822.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. **Proc Natl Acad Sci U S A, 89**(22), 10915-10919.

Higgins, D. G., & Taylor, W. R. (2001). Bioinformatics - Sequence, Structure and Databanks. **Oxford Press**.

Hogeweb, P., & Hesper, B. (1984). The Alignment of Sets of Sequences and the Construction of Phylogenetic Trees. An Integrated Method. **Journal of Molecular Evolution, 20**, 175-186.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., & Vriend, G. (1992). A database of protein structure families with common folding motifs. **Protein Sci, 1**(12), 1691-1698.

Holm, L., & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. **J Mol Biol, 233**(1), 123-138.

Hooft, R. W., Sander, C., & Vriend, G. (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. **Proteins, 26**(4), 363-376.

Hubbard, S. J., Campbell, S. F., & Thornton, J. M. (1991). Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. **J Mol Biol, 220**(2), 507-530.

Hughey, R., Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. **Comput Appl Biosci. 12**:95–107

Jacobson, A., & Sali, A. (2004). Comparative Protein Structure Modeling and its Applications to Drug Discovery. **Annual Reports in Mdeical Chemistry, 39**, 260-276.

Jaroszewski, L., Li, W., & Godzik, A. (2002). In search for more accurate alignments in the twilight zone. **Protein Sci, 11**(7), 1702-1713.

John, B., & Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. **Nucleic Acids Res, 31**(14), 3982-3992.

Johnston, C. R., & Shields, D. C. (2005). A sequence sub-sampling algorithm increases the power to detect distant homologues. **Nucleic Acids Res, 33**(12), 3772-3778.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. **J Mol Biol, 292**(2), 195-202.

Jones, D. T., & McGuffin, L. J. (2003). Assembling novel protein folds from super-secondary structural fragments. **Proteins, 53 Suppl 6**, 480-485.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. **Nature, 358**(6381), 86-89.

Jones, T. A., & Thirup, S. (1986). Using known substructures in protein model building and crystallography. **Embo J, 5**(4), 819-822.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers, 22**(12), 2577-2637.

Kabsch, W., & Sander, C. (1983). How good are predictions of protein secondary structure? **FEBS Lett, 155**(2), 179-182.

Karchin, R., Cline, M., Mandel-Gutfreund, Y., & Karplus, K. (2003). Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. **Proteins, 51**(4), 504-514.

Karypis, G. (2006). YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. **Proteins, 64**(3), 575-586.

Kazemian, M., Moshiri, B., Nikbakht, H., & Lucas, C. (2007). A new expertness index for assessment of secondary structure prediction engines. **Comput Biol Chem, 31**(1), 44-47.

Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. **J Mol Biol, 299**(2), 499-520.

Kimura, R., Brower, R., Vajda, S., & Camacho, C. J. (2001). Dynamical View of the Positions of Key Side Chains in Protein-Protein Recognition. **Journal of Biophysics, 80**, 635-642.

Kneller, D. G., Cohen, F. E., & Langridge, R., (1990). "Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network" **J. Mol. Biol. 214**, 171-182.

Koehl, P. and Delarue, M. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. **J. Mol. Biol. 239,** 249–275.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. **Proteins, 64**(3), 559-574.

Krissinel, E. (2007). On the relationship between sequence and structure similarities in proteomics. **Bioinformatics, 23**(6), 717-723.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. **J Mol Biol, 235**(5), 1501-1531.

Laskowski, R. A., MacArthur, M. W., & Thornton, J. M. (1998). Validation of protein models derived from experiment. **Curr Opin Struct Biol, 8**(5), 631-639.

Lee, J. (2006). Measures for the assessment of fuzzy predictions of protein secondary structure. **Proteins, 65**(2), 453-462.

Lesk, A. M. (2001). Introduction to Protein Architecture. **Oxford Press.**

Lesk, A. M. (2002). Introduction to Bioinformatics. **Oxford Press.**

Levin,J.M. and Garnier,J. (1988) Improvements in a secondary structure prediction method based on a search for local sequence homologies. **Biochim. Biophys. Acta, 955**, 283–295.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. **J Mol Biol, 226**(2), 507-533.

Lim, V. I. (1974). Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. **J Mol Biol, 88**(4), 873-894.

Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. **Nucleic Acids Res, 30**(1), 264-267.

Lopez-Otin, C., & Overall, C. M. (2002). Protease degradomics: a new challenge for proteomics. **Nat Rev Mol Cell Biol, 3**(7), 509-519.
Madhusudhan, M. S., Marti-Renom, M. A., Sanchez, R., & Sali, A. (2006). Variable gap penalty for protein sequence-structure alignment. **Protein Eng Des Sel, 19**(3), 129-133.

Martin, J., Gibrat, J. F., & Rodolphe, F. (2006). Analysis of an optimal hidden Markov model for secondary structure prediction. **BMC Struct Biol, 6**, 25.

Martin, C., Cheetham, C. & Rees, A. R. (1989). Modelling antibody hypervariable loops: a combined algorithm. **Proc. Natl Acad. 86**. 9268-9272.

Martin, A., MacArthur, M., & Thornton, J. (1997). Assessment of comparative modelling in CASP2. ***Proteins Suppl 1***:14–28.

Marti-Renom, M. A., Fiser, A., Madhusudhan, M. S., John, B., Stuart, A. C., Eswar, N., et al. (2003). Modeling Protein Structure From Its Sequence. **Current Protocols In Bioinformatics, Supplement 3**(5.1.1-5.1.32).

Marti-Renom, M. A., Madhusudhan, M. S., Fiser, A., Rost, B., & Sali, A. (2002). Reliability of assessment of protein structure prediction methods. **Structure, 10**(3), 435-440.

Marti-Renom, M. A., Madhusudhan, M. S., & Sali, A. (2004). Alignment of protein sequences by their profiles. **Protein Sci, 13**(4), 1071-1087.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. **Annu Rev Biophys Biomol Struct, 29**, 291-325.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2001). What are the baselines for protein fold recognition? **Bioinformatics, 17**(1), 63-72.

Mittl, P. R., & Grutter, M. G. (2006). Opportunities for structure-based design of protease-directed drugs. **Curr Opin Struct Biol, 16**(6), 769-775.

Montgomerie, S., Sundararaj, S., Gallin, W. J. & Wishart, D. S. (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. **BMC Bioinformatics, 7**, 301.

Moult, J. (1999). Predicting protein three-dimensional structure. **Curr Opin Biotechnol, 10**(6), 583-588.

Myers, E. W., & Miller, W. (1988). Optimal alignments in linear space. **Comput Appl Biosci, 4**(1), 11-17.

Nature. (2007). Looking Ahead with Structural Genomics. **Nature Structural and Molecular Biology, 14**, 1.

NC-IUBMB. (1992). Enzyme Nomenclature. **Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.**

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J Mol Biol, 48**(3), 443-453.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. **Journal of Molecular Biology, 302**(1), 205-217.

Notredame, C., Holm, L., & Higgins, D. G. (1998). COFFEE: an objective function for multiple sequence alignments. **Bioinformatics, 14**(5), 407-422.

Nuin, P. A., Wang, Z., & Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. **BMC Bioinformatics, 7**, 471.

Ohlson, T., Wallner, B., & Elofsson, A. (2004). Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. **Proteins, 57**(1), 188-197.

Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., & Sillitoe, I. (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. **Proteins, Suppl 3**, 149-170.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. **Structure, 5**(8), 1093-1108.

Orengo, C. A., & Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. **Methods Enzymol, 266**, 617-635.

Ortiz, A. R., Strauss, C. E., & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. **Protein Sci, 11**(11), 2606-2621.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., & Notredame, C. (2004). 3DCoffee: combining protein sequences and structures within multiple sequence alignments. **J Mol Biol, 340**(2), 385-395.

Ouali, M., & King, R. D. (2000). Cascaded multiple classifiers for secondary structure prediction. **Protein Sci, 9**(6), 1162-1176.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. **Phil. Trans. R. Soc. A, 187**, 253-318.

Peitsch, M. C. (2002). About the use of protein models. **Bioinformatics, 18**(7), 934-938.

Peitsch M. C., Jongeneel C.V. (1993). A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. **Int Immunol 5**, 233-238.

Petrey, D., & Honig, B. (2005). Protein structure prediction: inroads to biology. **Mol Cell, 20**(6), 811-819.

Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., et al. 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. **Proteins, 53** (6), 430–435.

Pollastri, G., Martin, A. J., Mooney, C., & Vullo, A. (2007). Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. **BMC Bioinformatics, 8**(1), 201.

Pollastri, G., Przybylski, D., Rost, B., & Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. **Proteins, 47**(2), 228-235.

Prasad, J. C., Comeau, S. R., Vajda, S., & Camacho, C. J. (2003). Consensus alignment for reliable framework prediction in homology modeling. **Bioinformatics, 19**(13), 1682-1691.

Rai, B. K., & Fiser, A. (2006). Multiple mapping method: a novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. **Proteins, 63**(3), 644-661.

Rawlings, N. D., & Barrett, A. J. (1993). Evolutionary families of peptidases. **Biochem J, 290 (Pt 1)**, 205-218.

Rawlings, N. D., & Barrett, A. J. (1994). Families of cysteine peptidases. **Methods Enzymol, 244**, 461-486.

Rawlings, N. D., & Barrett, A. J. (1995). Evolutionary families of metallopeptidases. **Methods Enzymol, 248**, 183-228.

Rawlings, N. D., Morton, F. R., & Barrett, A. J. (2006). MEROPS: the peptidase database. **Nucleic Acids Res, 34**(Database issue), D270-272.

Rawlings, N. D., Tolle, D. P., & Barrett, A. J. (2004). Evolutionary families of peptidase inhibitors. **Biochem J, 378**(Pt 3), 705-716.

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. **Methods Enzymol, 266**, 525-539.

Rost, B. (2001). Review: protein secondary structure prediction continues to rise. **J Struct Biol, 134**(2-3), 204-218.

Rost, B., & O'Donoghue, S. (1997). Sisyphus and prediction of protein structure. **Comput Appl Biosci, 13**(4), 345-356.

Rost, B., Schneider, R., & Sander, C. (1997). Protein fold recognition by prediction-based threading. **J Mol Biol, 270**(3), 471-480.

Russell, R. B., & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. **Proteins, 14**(2), 309-323.

Sadreyev, R., and Grishin, N. (2003). COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. **J. Mol. Biol. 326:** 317–336.

Salamov, A., and Solovyev, V. (1995). Prediction of protein secondary structure by combining nearest-neighbr algorithms and multiple sequence alignment. **J. Mol. Biol. 247**: 11-15.

Sali, A. (2001). MODELLER: A Program for Protein Structure Modelling Release 6.

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. **J Mol Biol, 234**(3), 779-815.

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., et al. (2000). Protein structure modeling for structural genomics. **Nat Struct Biol, 7 Suppl**, 986-990.

Sanchez, R., & Sali, A. (1997). Advances in comparative protein-structure modelling. **Curr Opin Struct Biol, 7**(2), 206-214.

Sanchez, R., & Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. **Proteins, Suppl 1**, 50-58.

Sanchez, R., & Sali, A. (1998). Large Scale Structure Modelling of the *Saccharomyces Cerevisiae* Genome. **Proc Natl Acad Sci, 95**, 13597-13602.

Sanchez, R., & Sali, A. (2000). Comparative protein structure modeling. Introduction and practical examples with modeller. **Methods Mol Biol, 143**, 97-129.

Saqi, M. A., Russell, R. B., & Sternberg, M. J. (1998). Misleading local sequence alignments: implications for comparative protein modelling. **Protein Eng, 11**(8), 627-630.

Schulz, G. E., & Schrimer, R. H. (1978). Principles of Protein Structure. **Berlin:Springer-Verlag**.

Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. **Nucleic Acids Res, 31**(13), 3381-3385.

Sen, T. Z., Cheng, H., Kloczkowski, A., & Jernigan, R. L. (2006). A Consensus Data Mining secondary structure prediction by combining GOR V and Fragment Database Mining. Protein Sci, 15(11), 2499-2506.

Shannon, C. E. (1948). A Mathematical Theory of Communication. **Bell Sys. Tech. J., 27**, 379-423, 623-656.

Shi, J., Blundell, T. L., & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. **J Mol Biol, 310**(1), 243-257.

Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. **Protein Eng, 11**(9), 739-747.

Siew, N., Elofsson, A., Rychlewski, L., & Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. **Bioinformatics, 16**(9), 776-785.

Simons, K. T., Strauss, C., & Baker, D. (2001). Prospects for ab initio protein structural genomics. **J Mol Biol, 306**(5), 1191-1199.
Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. **Proteins, 17**(4), 355-362.

Skolnick, J., Kihara, D., & Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. **Proteins, 56**(3), 502-518.

Smith, D. K., & Thornton, J. M., (1989). SSTRUC Computer Program. Department of Biochemistry and Molecular Biology, University College, London.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. **J Mol Biol, 147**(1), 195-197.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. **Bioinformatics, 21**(7), 951-960.

Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. **Proteins, 28**(3), 405-420.

Sternberg, M. J. (1996). Protein Structure Prediction. **IRL Press.**

Suhrer, S. J., Wiederstein, M., & Sippl, M. J. (2007). QSCOP--SCOP quantified by structural relationships. **Bioinformatics, 23**(4), 513-514.

Summers,N.L. and Karplus,M. (1989) Construction of side-chains in homology modelling. Application to the C-terminal lobe of rhizopuspepsin. **J. Mol. Biol., 210,** 785–811.

Sutcliffe M. J., Haneef I., Carney D., and Blundell T.L. 1987. Knowledge based modelling of homologous proteins, Part I: Three- dimensional frameworks derived from the simultaneous superposition of multiple structures. **Protein Eng 1**, 377-384.

Tan, Y. H., Huang, H., & Kihara, D. (2006). Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. **Proteins, 64**(3), 587-600.

Taylor, W. R., Flores, T. P., & Orengo, C. A. (1994). Multiple protein structure alignment. **Protein Sci, 3**(10), 1858-1870.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res, 22**(22), 4673-4680.

Tramontano, A. (1998). Homology modeling with low sequence identity. **Methods, 14**(3), 293-300.

Tramontano, A. (2003). Comparative Modelling Techniques: Where Are We? **Comp. Funct. Genom., 4**, 402-405.

Tramontano, A., Leplae, R., & Morea, V. (2001). Analysis and assessment of comparative modeling predictions in CASP4. **Proteins, Suppl 5**, 22-38.

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: Structure, Function, and Genetics.* **43(2):** 227-241.

Venclovas, C. (2003). Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. **Proteins, 53 Suppl 6**, 380-388.

Venclovas, C., & Margelevicius, M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. **Proteins, 61 Suppl 7**, 99-105.

Vogt, G., Etzold, T., & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. **J Mol Biol, 249**(4), 816-831.

Wang, G., & Dunbrack, R, L. Jr. (2003) PISCES: a protein sequence culling server. **Bioinformatics, 19,** 1589-1591.

Wang, G., & Dunbrack, R. L., Jr. (2004). Scoring profile-to-profile sequence alignments. **Protein Sci, 13**(6), 1612-1626.

Watson, J. D., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., et al. (2007). Towards fully automated structure-based function prediction in structural genomics: a case study. **J Mol Biol, 367**(5), 1511-1522.

Westhead, D. R., & Thornton, J. M. (1998). Protein structure prediction. **Curr Opin Biotechnol, 9**(4), 383-389.

Wilson, C. L., Boardman, P. E., Doig, A. J., & Hubbard, S. J. (2004). Improved prediction for N-termini of alpha-helices using empirical information. **Proteins, 57**(2), 322-330.

Wilson, C. L., Hubbard, S. J., & Doig, A. J. (2002). A critical assessment of the secondary structure alpha-helices and their termini in proteins. **Protein Eng, 15**(7), 545-554.

Wistrand, M., & Sonnhammer, E. L. (2005). Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. **BMC Bioinformatics, 6**, 99.

Wlodawer, A., & Vondrasek, J. (1998). Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. **Annu Rev Biophys Biomol Struct, 27**, 249-284.

Xiong, J. (2006). Essential Bioinformatics. **Cambridge University Press.**

Yang, A. S. (2002). Structure-dependent sequence alignment for remotely related proteins. **Bioinformatics, 18**(12), 1658-1665.

Yi, T. M., & Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. **J Mol Biol, 232**(4), 1117-1129.

Yona, G., & Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. **J Mol Biol, 315**(5), 1257-1275.

Zhou, H., & Zhou, Y. (2005). SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. **Bioinformatics, 21**(18), 3615-3621.

Percentage Identity	SST1	JPred	ELEPHANT	SST2	SST3
<=100	0.88	1.26	1.25	0.73	2.30
<=80	0.89	1.27	1.26	0.79	2.32
<=60	0.90	1.28	1.28	0.88	2.35
<=40	0.94	1.33	1.33	1.02	2.41
<=30	0.97	1.32	1.33	1.16	2.43
<=20	0.98	1.36	1.35	1.47	2.59

Table A1.1. Average Helical RMSDs for the Explicit Method. The average alpha-carbon helical RMSD over all of the models built using the different secondary structure restraints is displayed.

Percentage Identity	SST1	JPred	ELEPHANT	SST2	SST3
<=100	1.61	1.71	1.71	0.72	1.93
<=80	1.66	1.74	1.73	0.79	1.96
<=60	1.69	1.76	1.75	0.87	1.98
<=40	1.73	1.79	1.80	1.00	2.03
<=30	1.77	1.84	1.85	1.18	2.09
<=20	1.95	2.02	2.02	1.52	2.26

Table A1.2. Average Sheet RMSDs for the Explicit Method. The average alpha-carbon sheet RMSD over all of the models built using the different secondary structure restraints is displayed.

```
INCLUDE
SET ALNFILE = '1a28_A.1e3g_A.ali'
SET KNOWNS = '1e3g_A'
SET SEQUENCE = '1a28_A'
OUTPUT_CONTROL = 11112
READ_MODEL = '1e3g_A'
SEQUENCE_TO_ALI
WRITE_ALIGNMENT FILE = '1e3g_A'
SET PDB_EXT = '.1e3g_A_ELE_AB.modpdb'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
SET DEVIATION = 4.0
SET LIBRARY_SCHEDULE = 1
SET MAX_VAR_ITERATIONS = 300
SET MD_LEVEL = 'refine_3'
SET REPEAT_OPTIMIZATION = 3, MAX_MOLPDF = 1E6         #Repeat the
whole cycle 3-times and do not stop unless obj.func. > 1E6
SET FINAL_MALIGN3D = 1

SET RAND_SEED = -12312        # to have different models from another
TOP file
CALL ROUTINE = 'model'
CALL ROUTINE = 'special_restraints'
SUBROUTINE ROUTINE = 'special_restraints'
  SET ADD_RESTRAINTS = on
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '9' '16'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '37' '56'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '67' '90'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '117' '134'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '138' '149'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '161' '181'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '190' '220'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '229' '237'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'alpha', RESIDUE_IDS = '242' '246'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'strand', RESIDUE_IDS = '98' '100'
  MAKE_RESTRAINTS RESTRAINT_TYPE = 'strand', RESIDUE_IDS = '251' '252'
  RETURN
END_SUBROUTINE
```

Figure A1.1. Example MODELLER Input File. This file shows how the secondary structure restraints are applied in the model building process.

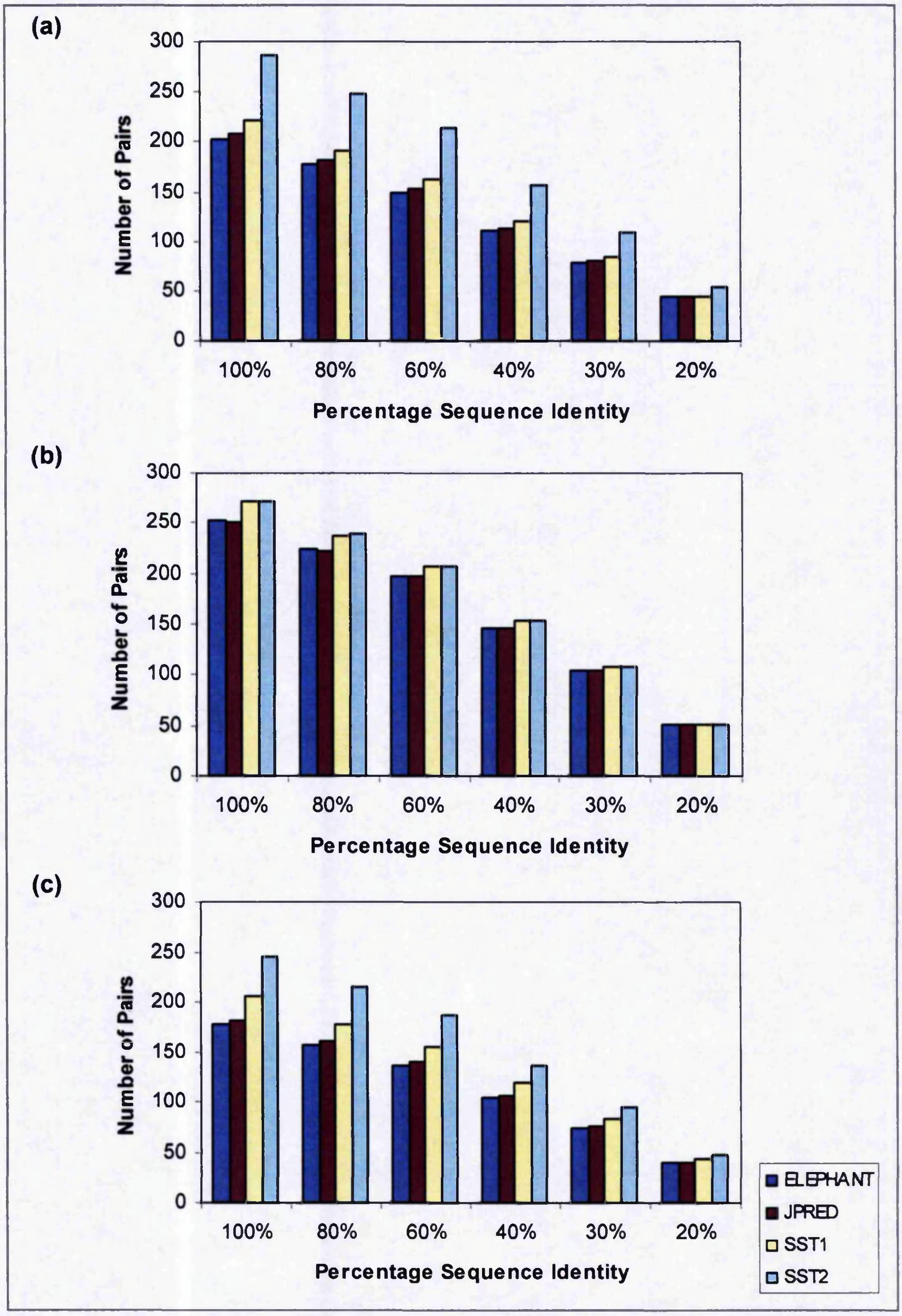


Figure A1.2. Number of Pairs with Lower RMSDs. (a) shows the number of pairs which have lower RMSDs than SST3 (the template with explicit secondary structure restraints), (b) displays the results for the regions which are in helical conformation and the graph in (c) shows the results for the regions in beta strand conformation. All RMSDs are between the alpha-carbons of the target and template.

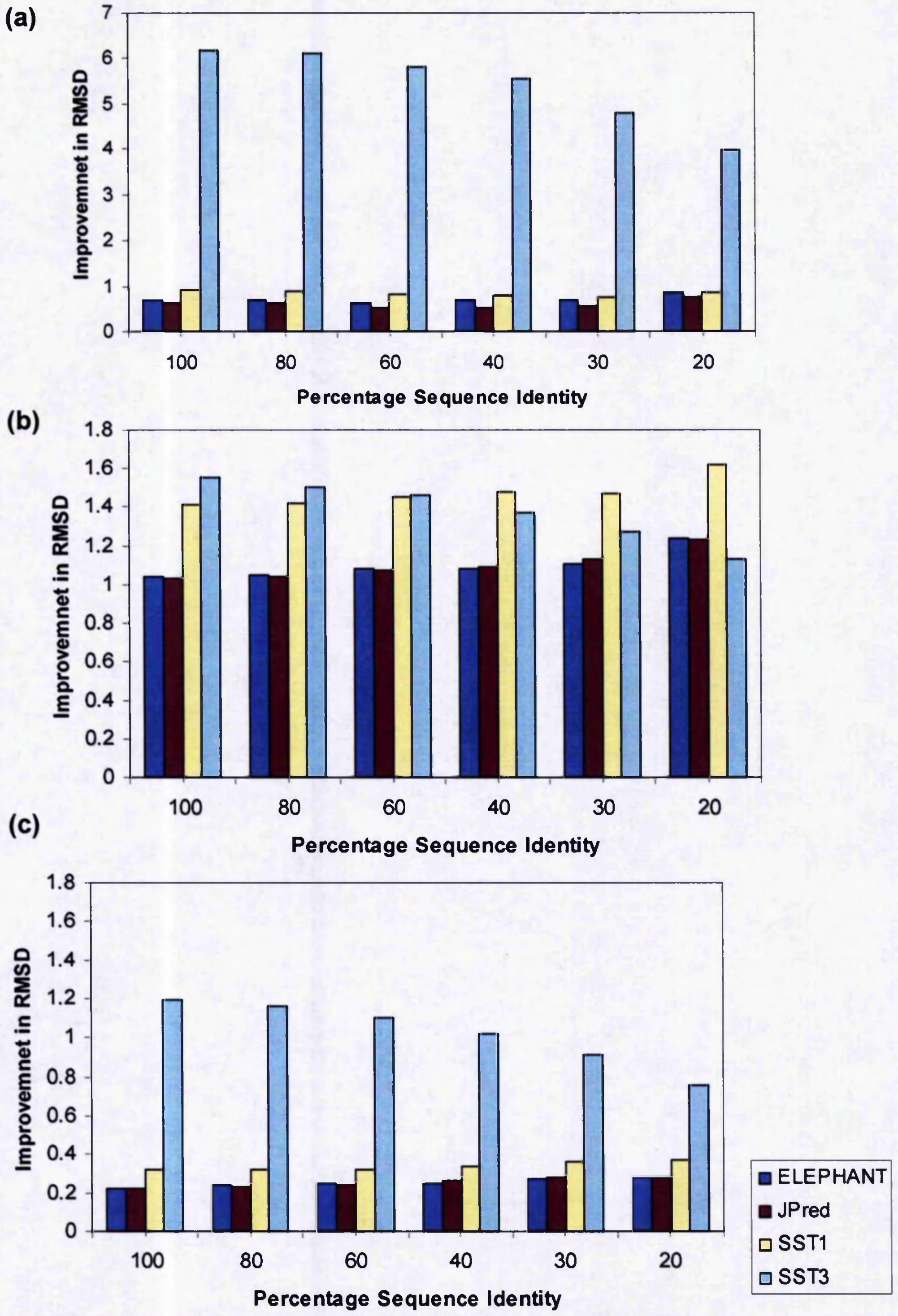


Figure A1.3. Average Improvement in RMSD. The graphs show the average improvement in RMSD when comparing the resulting models built using the different secondary structure restraints to those models built using the restraints explicitly from the template (SST2). (a) shows the average improvement in RMSD between the alpha-carbons of the target and template, (b) displays the RMSD results for the helical regions and (c) shows the average RMSD results for the beta-strand regions.

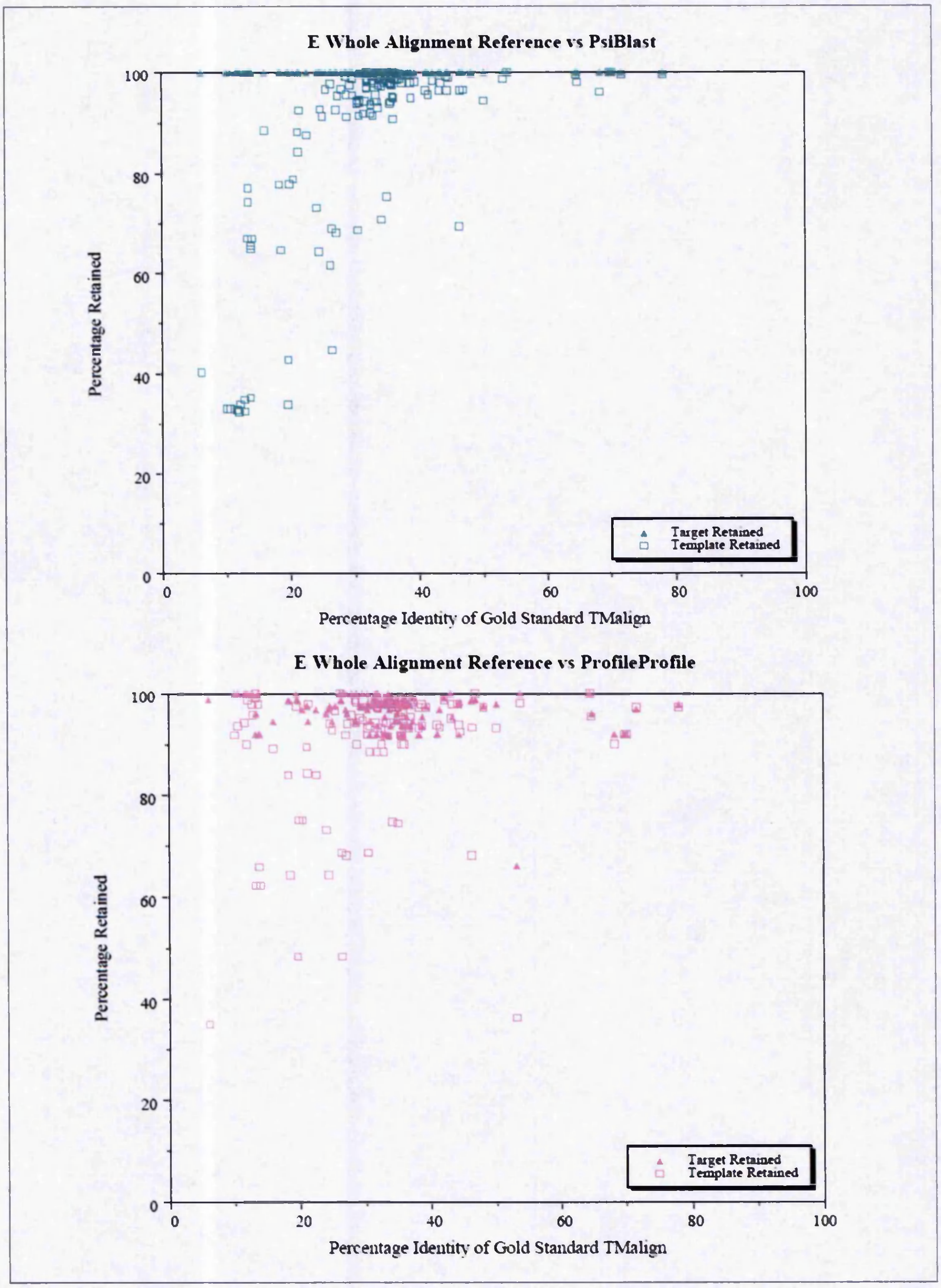


Figure A2.1. The Percentage of Sequence Retained at Different Percentage Identities. The percentage retained of the target and template sequences *versus* the percentage identity of the aligned pair (from structural alignment). Data from the PSI-BLAST and Profile-Profile alignment methods are displayed only from the I-vs-S set.

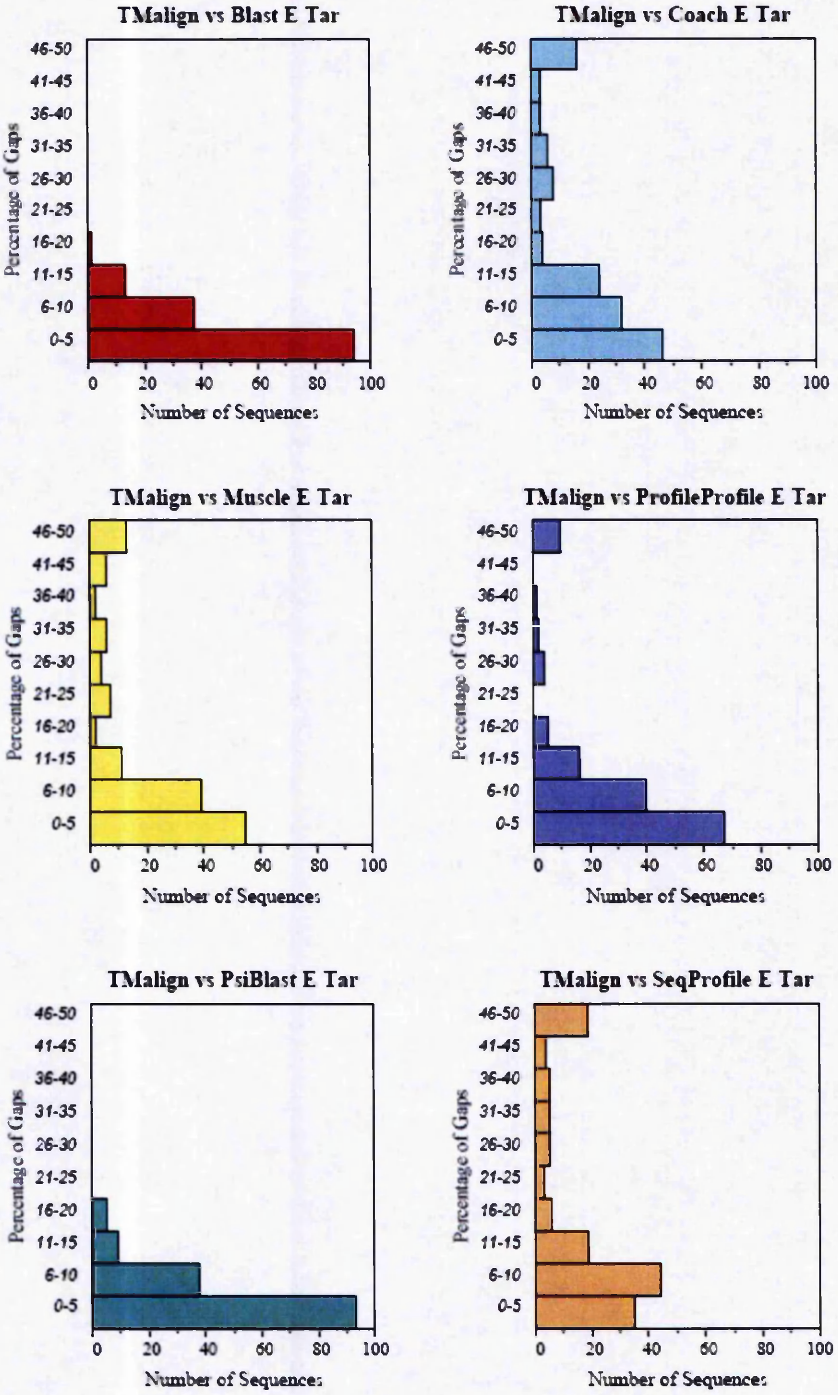


Figure A2.2. The Percentage of Gaps in the Target Sequences. The percentage of gaps introduced into the target sequences by the different alignment methods. This is for the I vs S set.

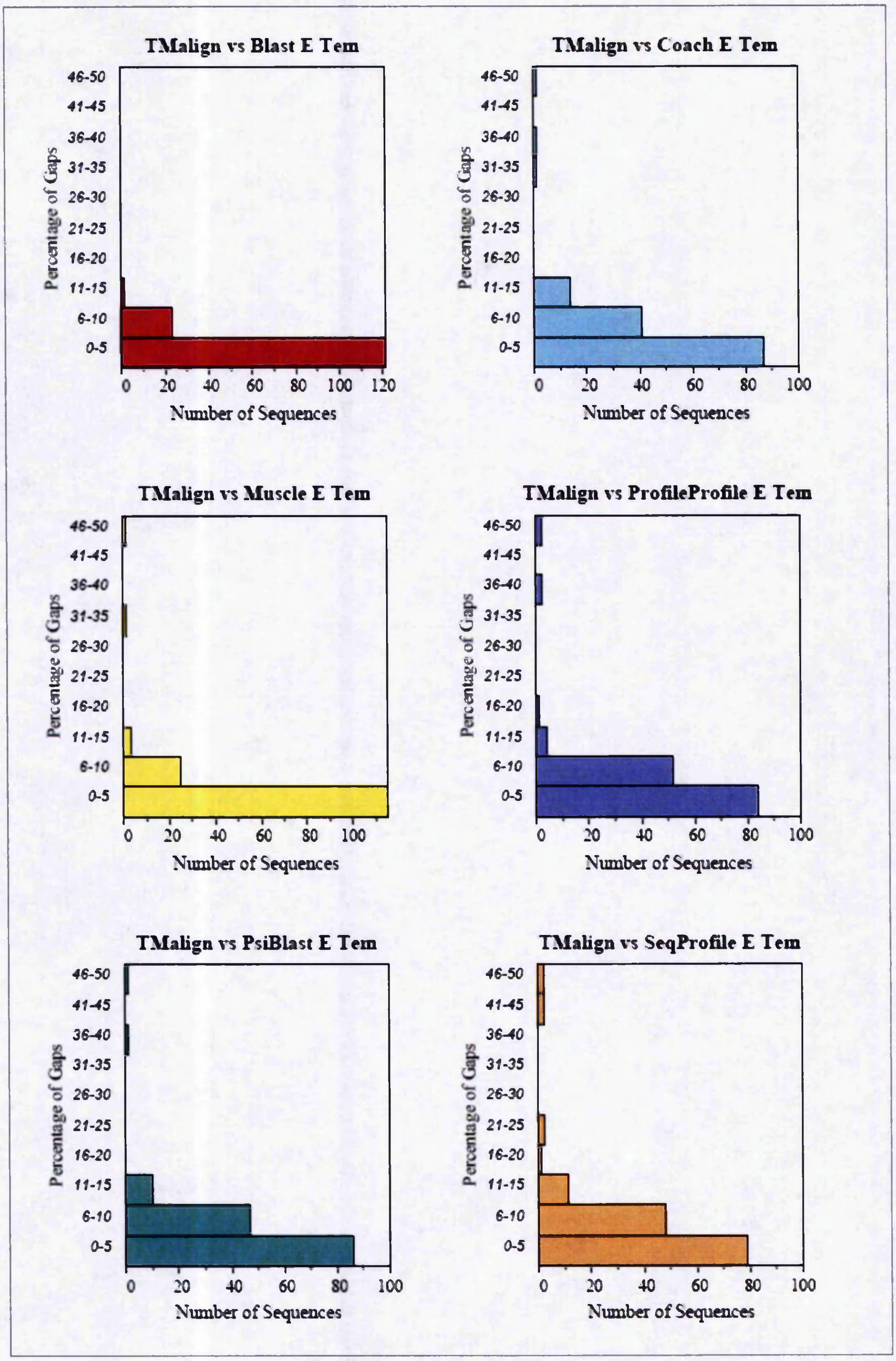


Figure A2.3. The Percentage of Gaps in the Template Sequences. The percentage of gaps introduced into the template sequences by the different alignment methods. This is for the I vs S set.

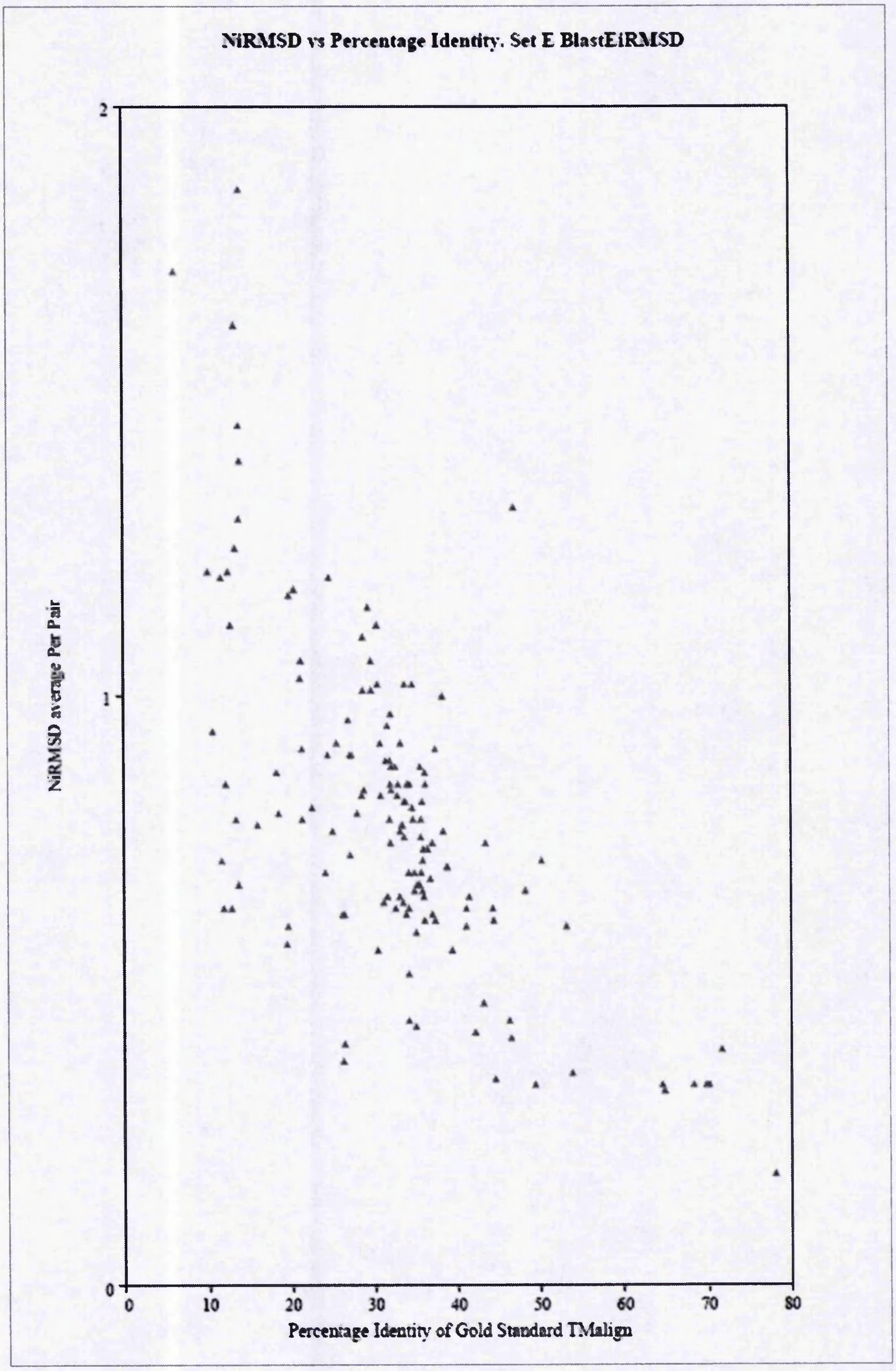


Figure A2.4. The Percentage Identity and the NiRMSD for the BLAST method. For the I-vs-S set, the average NiRMSD per pair was plotted for the BLAST method against the percentage identity of that alignment pair.

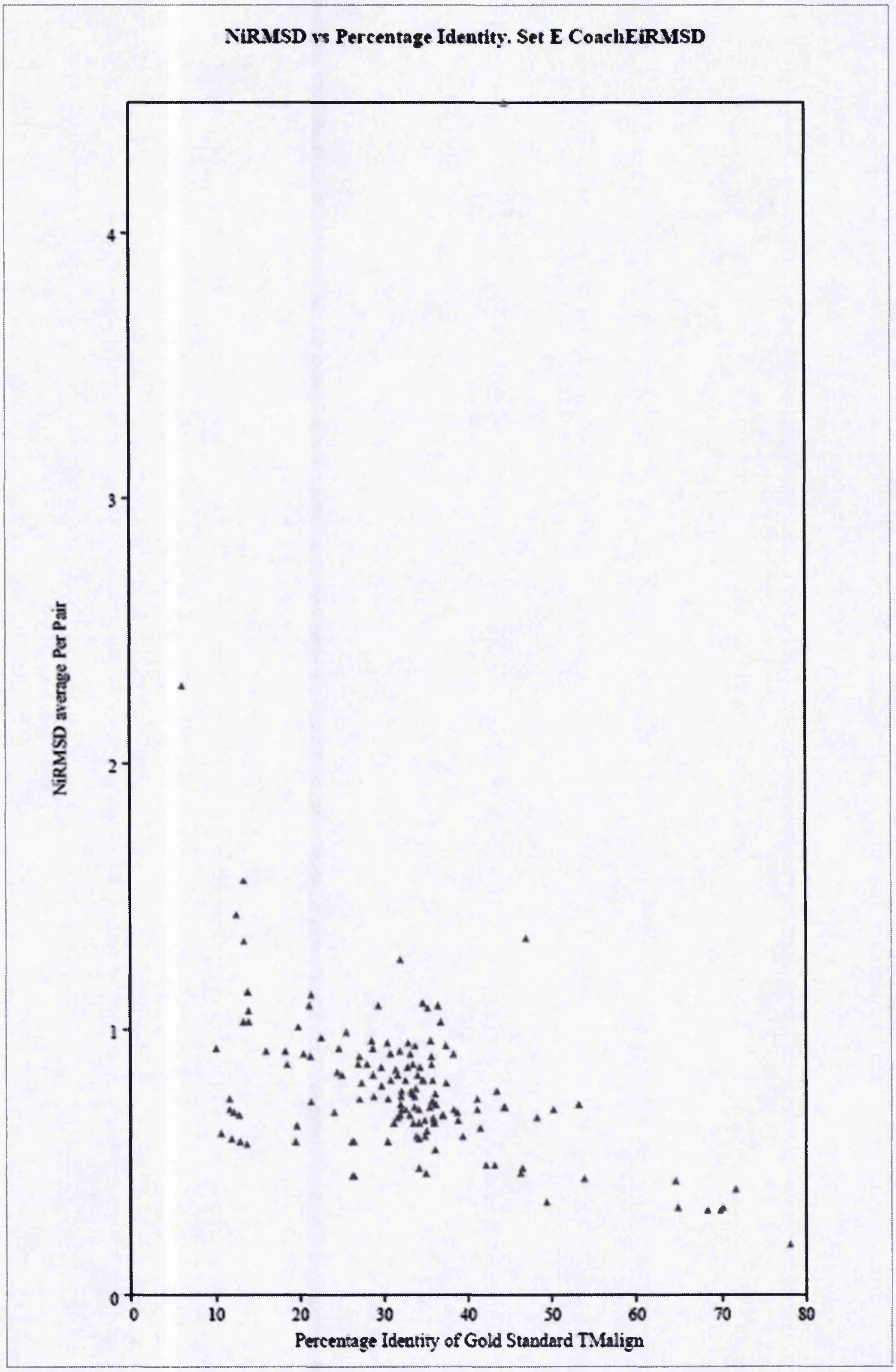


Figure A2.5. The Percentage Identity and the NiRMSD for the COACH method. For the I-vs-S set, the average NiRMSD per pair was plotted for the COACH method against the percentage identity of that alignment pair.

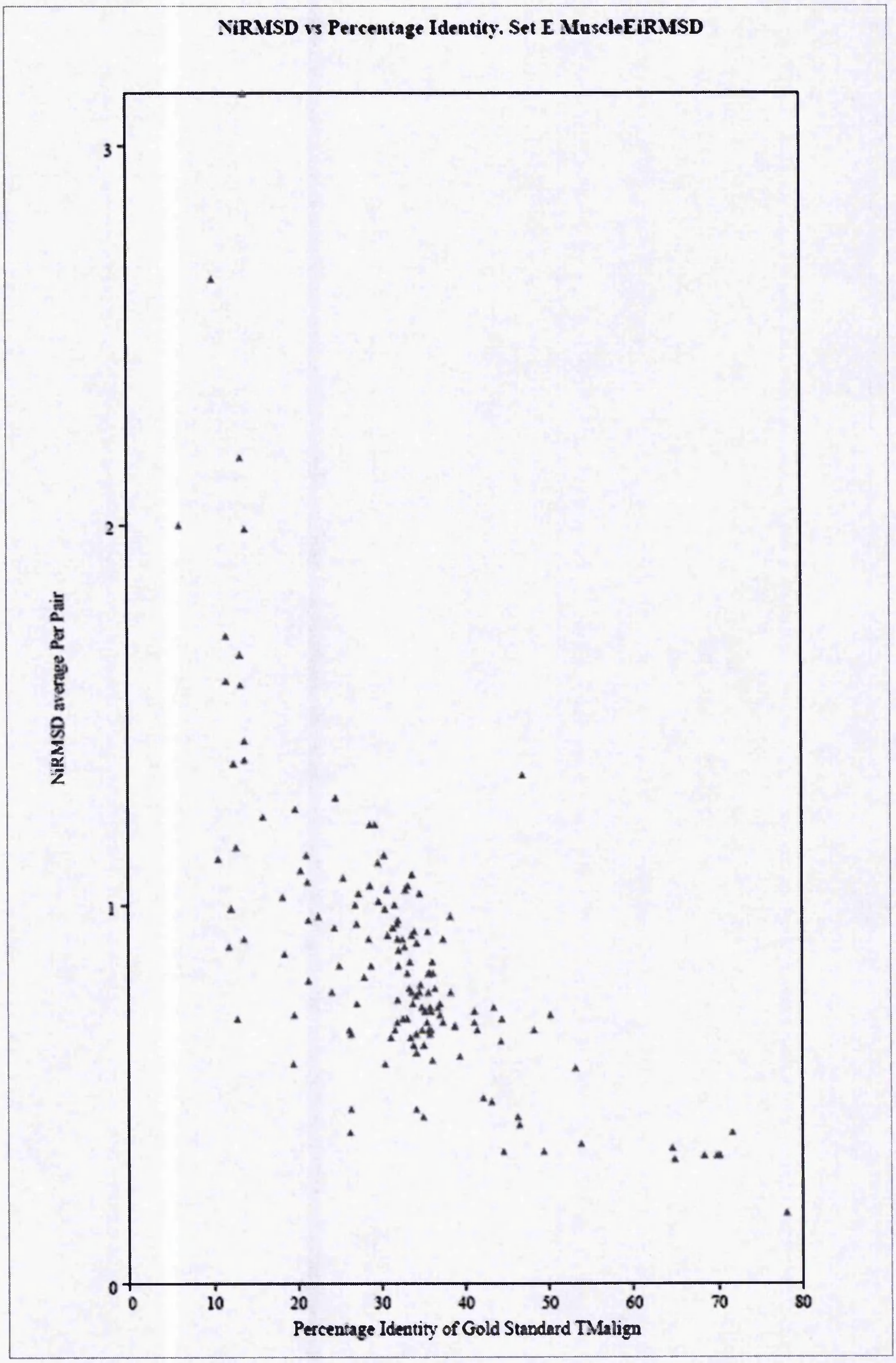


Figure A2.6. The Percentage Identity and the NiRMSD for the MUSCLE method. For the I-vs-S set, the average NiRMSD per pair was plotted for the MUSCLE method against the percentage identity of that alignment pair.

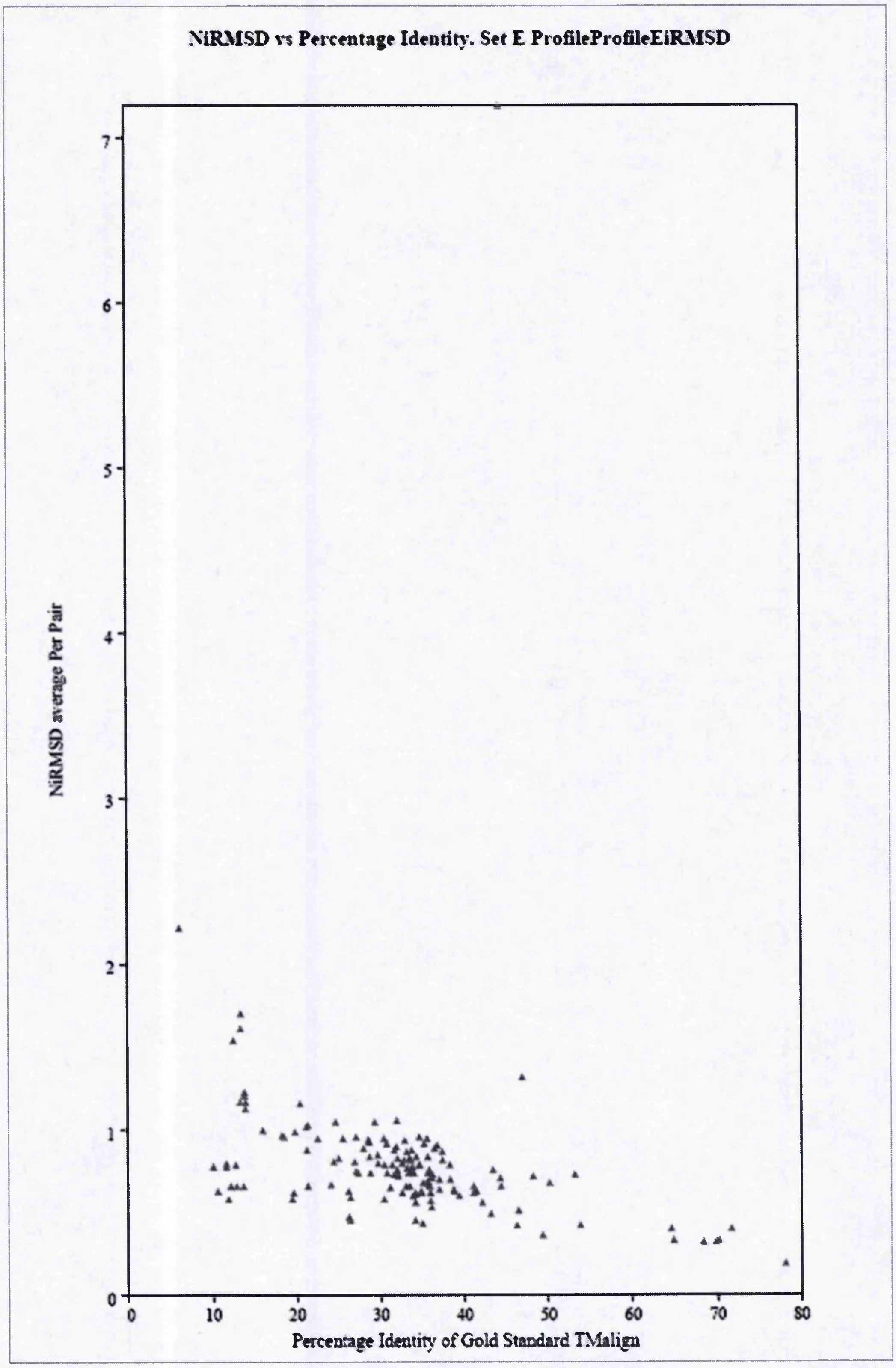


Figure A2.7. The Percentage Identity and the NiRMSD for the Profile-Profile method. For the I-vs-S set, the average NiRMSD per pair was plotted for the Profile-Profile method against the percentage identity of that alignment pair.

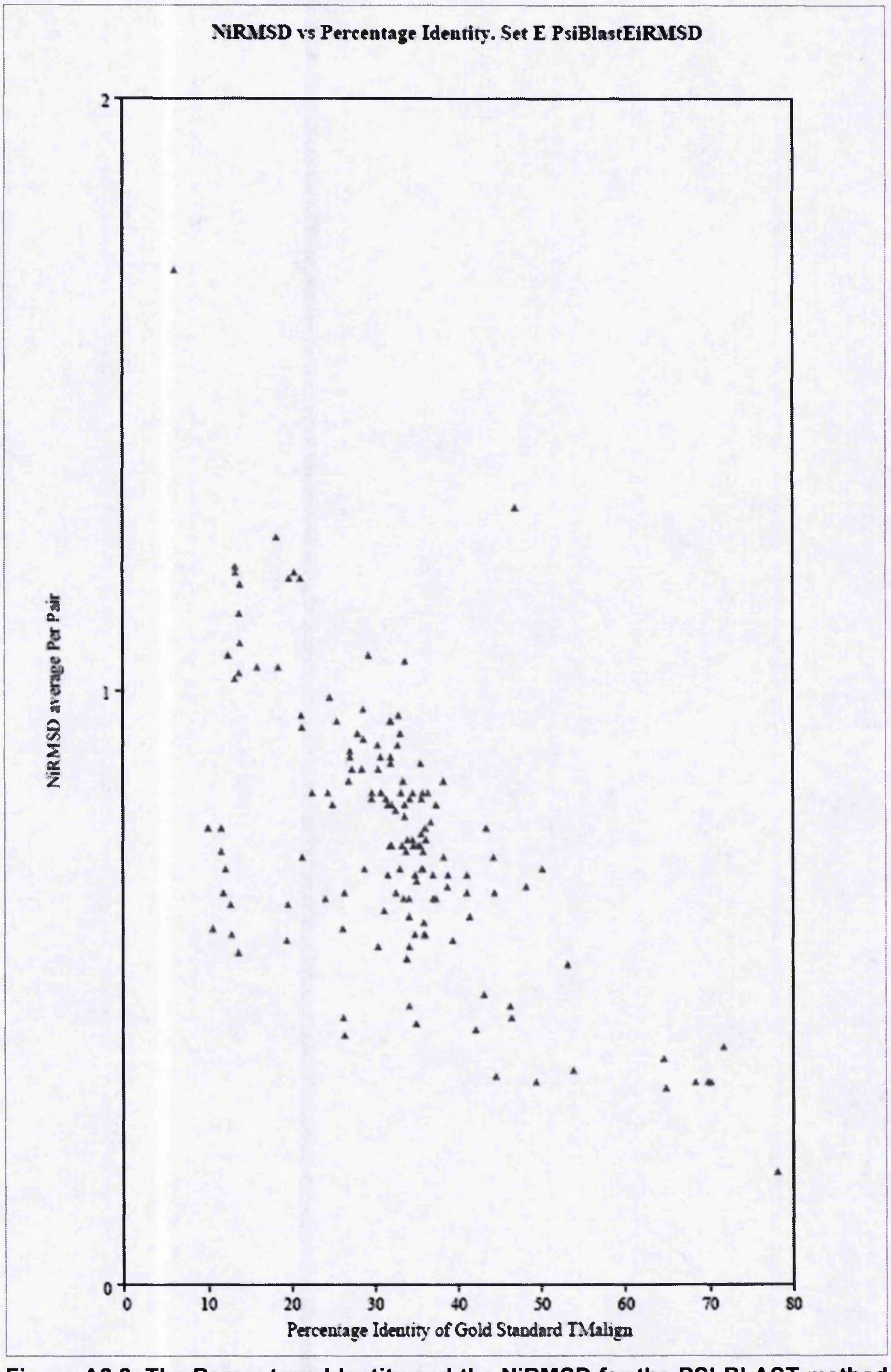


Figure A2.8. The Percentage Identity and the NiRMSD for the PSI-BLAST method. For the I-vs-S set, the average NiRMSD per pair was plotted for the PSI-BLAST method against the percentage identity of that alignment pair.

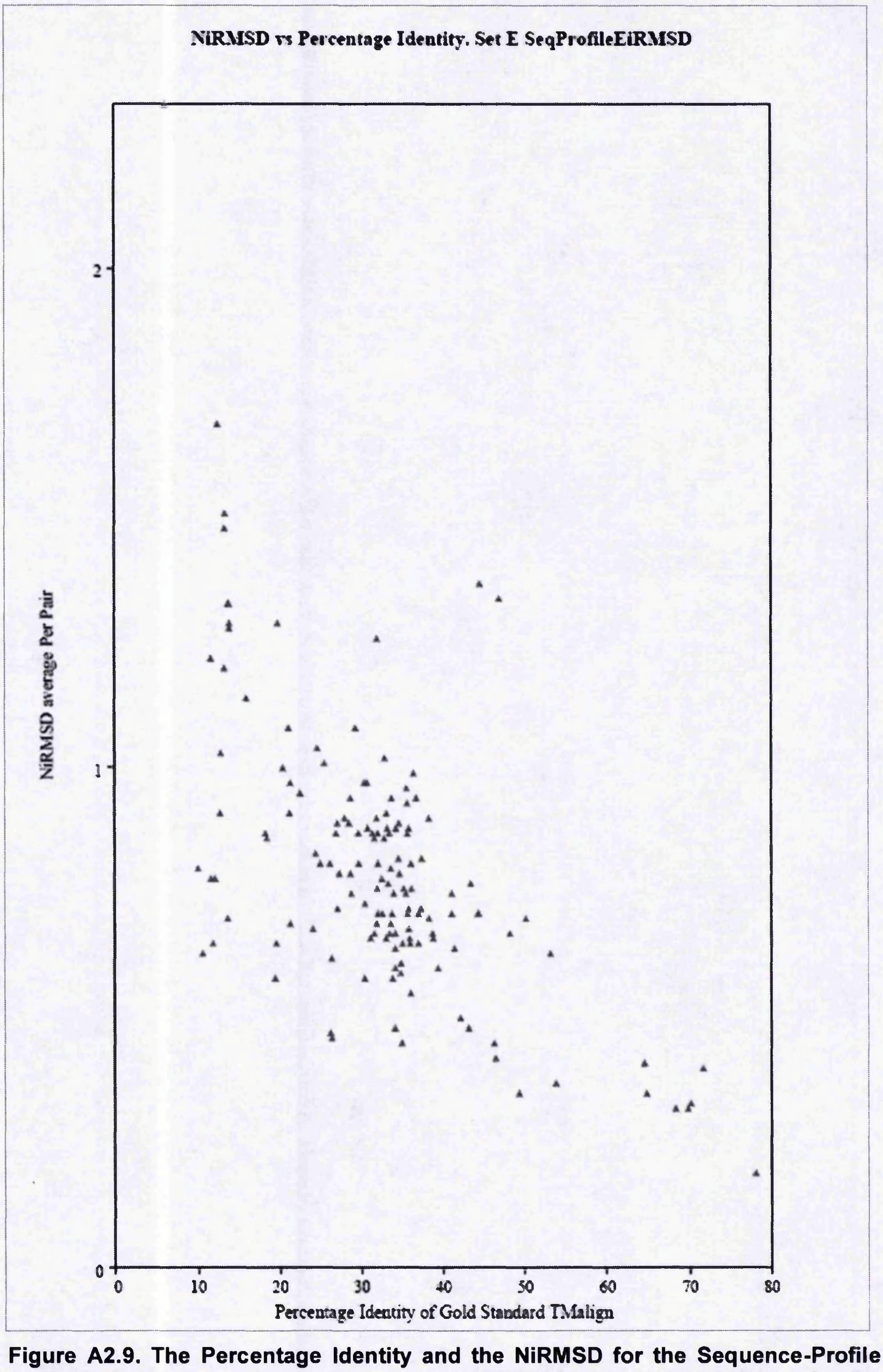


Figure A2.9. The Percentage Identity and the NiRMSD for the Sequence-Profile method. For the I-vs-S set, the average NiRMSD per pair was plotted for the Sequence-Profile method against the percentage identity of that alignment pair.

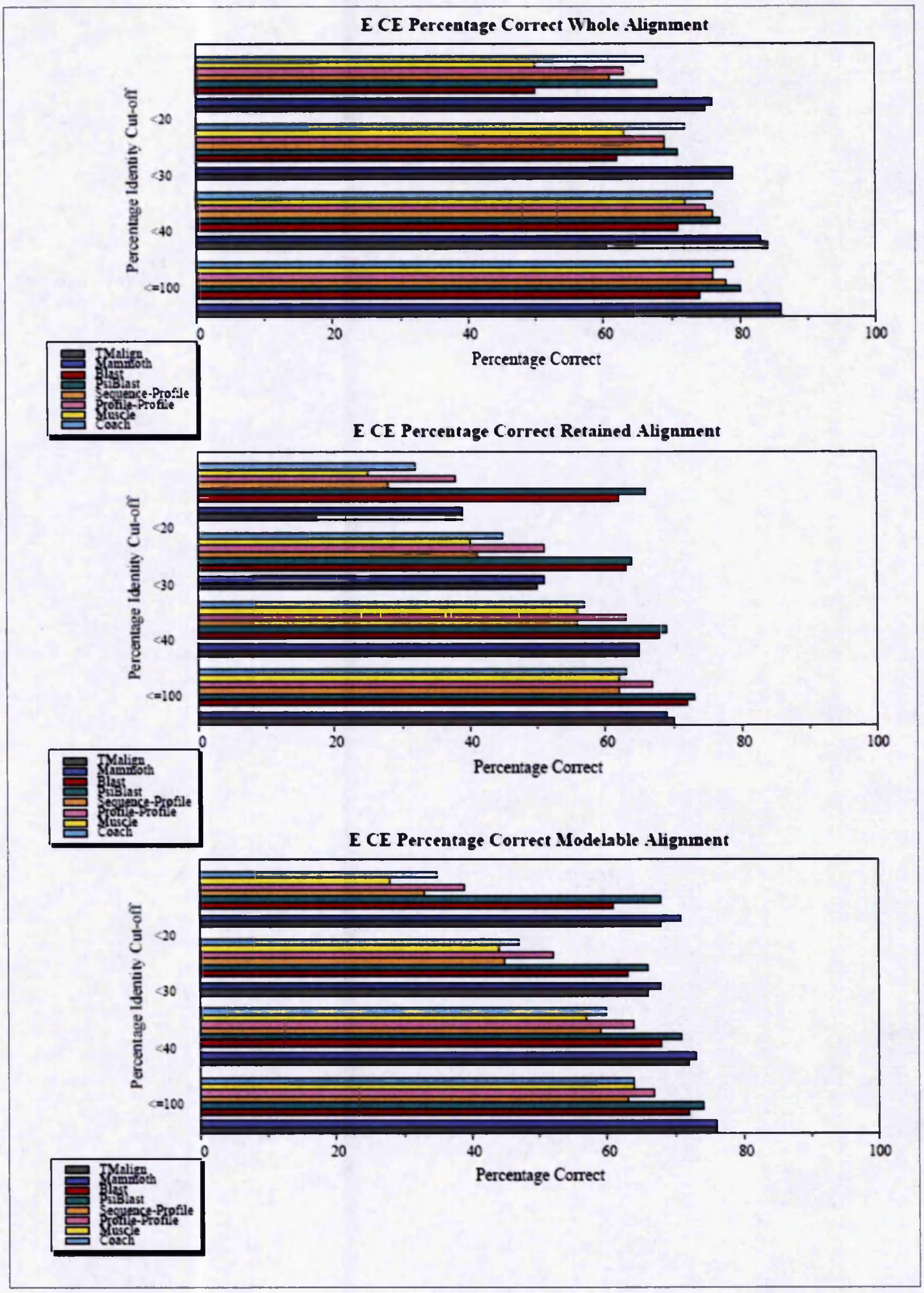
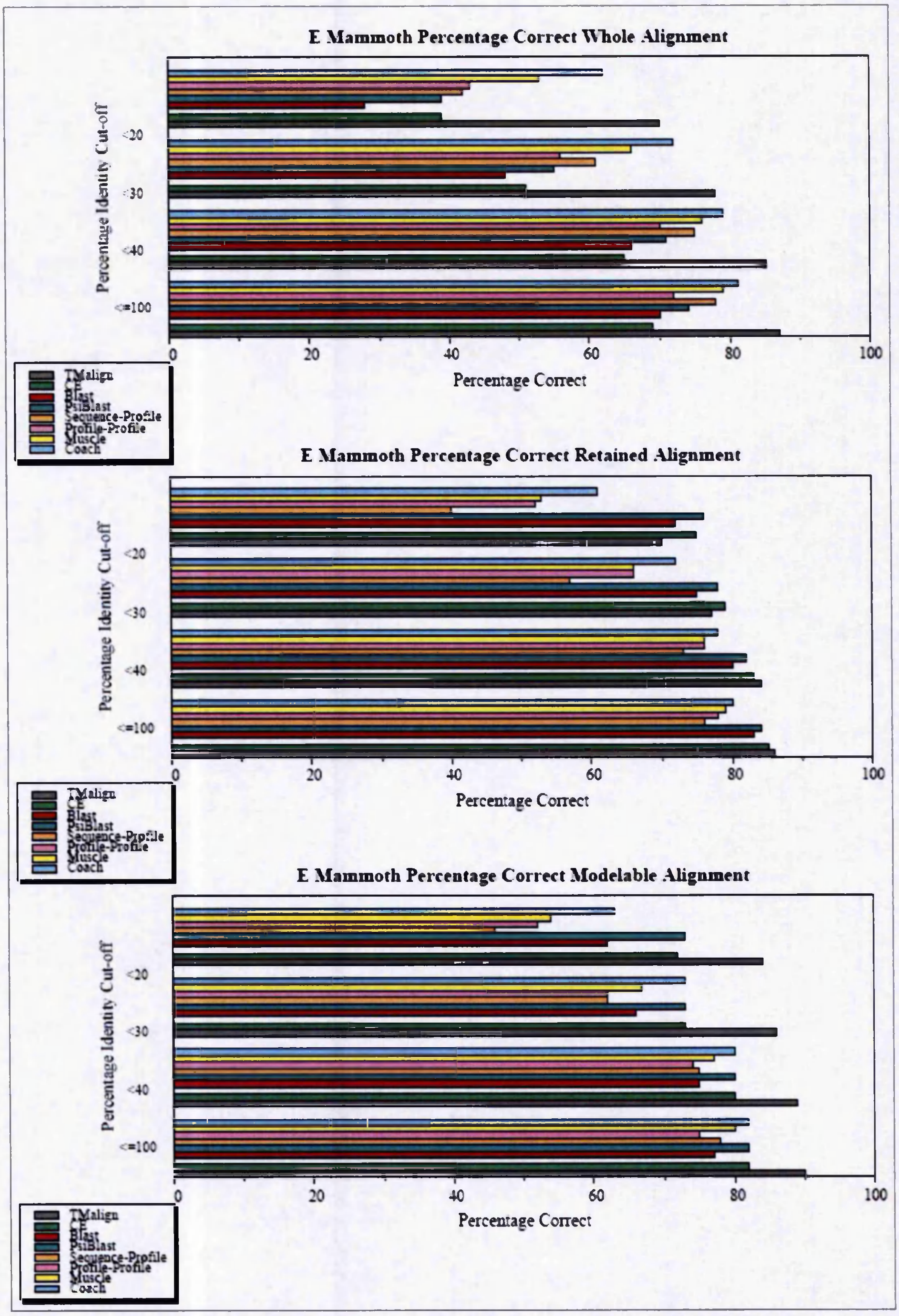


Figure A2.10. Percentage of Correctly Predicted Residues for CE. Results for the percentage of correctly aligned residues by each alignment method, as an average across all pairs in the I-vs-S set, assessed against the gold standard CE are shown. The percentage identity bins are inclusive as indicated by the "less than" signs.



E Mammoth Percentage Correct Whole Alignment

E Mammoth Percentage Correct Retained Alignment

E Mammoth Percentage Correct Modelable Alignment

Figure A2.11. Percentage of Correctly Predicted Residues for MAMMOTH. Results for the percentage of correctly aligned residues by each alignment method, as an average across all pairs in the I-vs-S set, assessed against the gold standard MAMMOTH are shown. The percentage identity bins are inclusive as indicated by the “less than” signs.

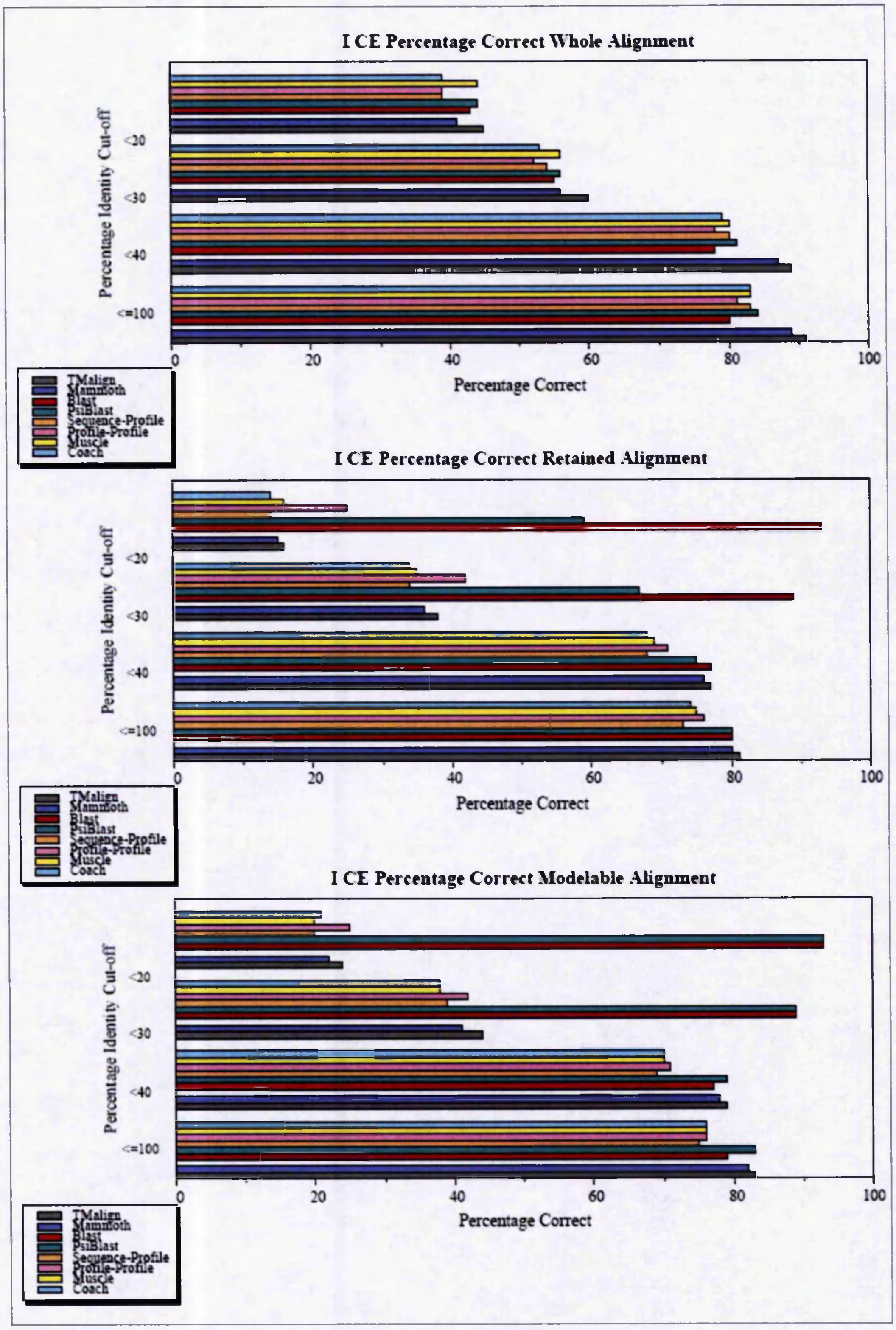


Figure A2.12. Percentage of Correctly Predicted Residues for CE, I-vs-I. Results for the percentage of correctly aligned residues by each alignment method, as an average across all pairs in the I-vs-I set, assessed against the gold standard CE are shown. The percentage identity bins are inclusive as indicated by the “less than” signs.

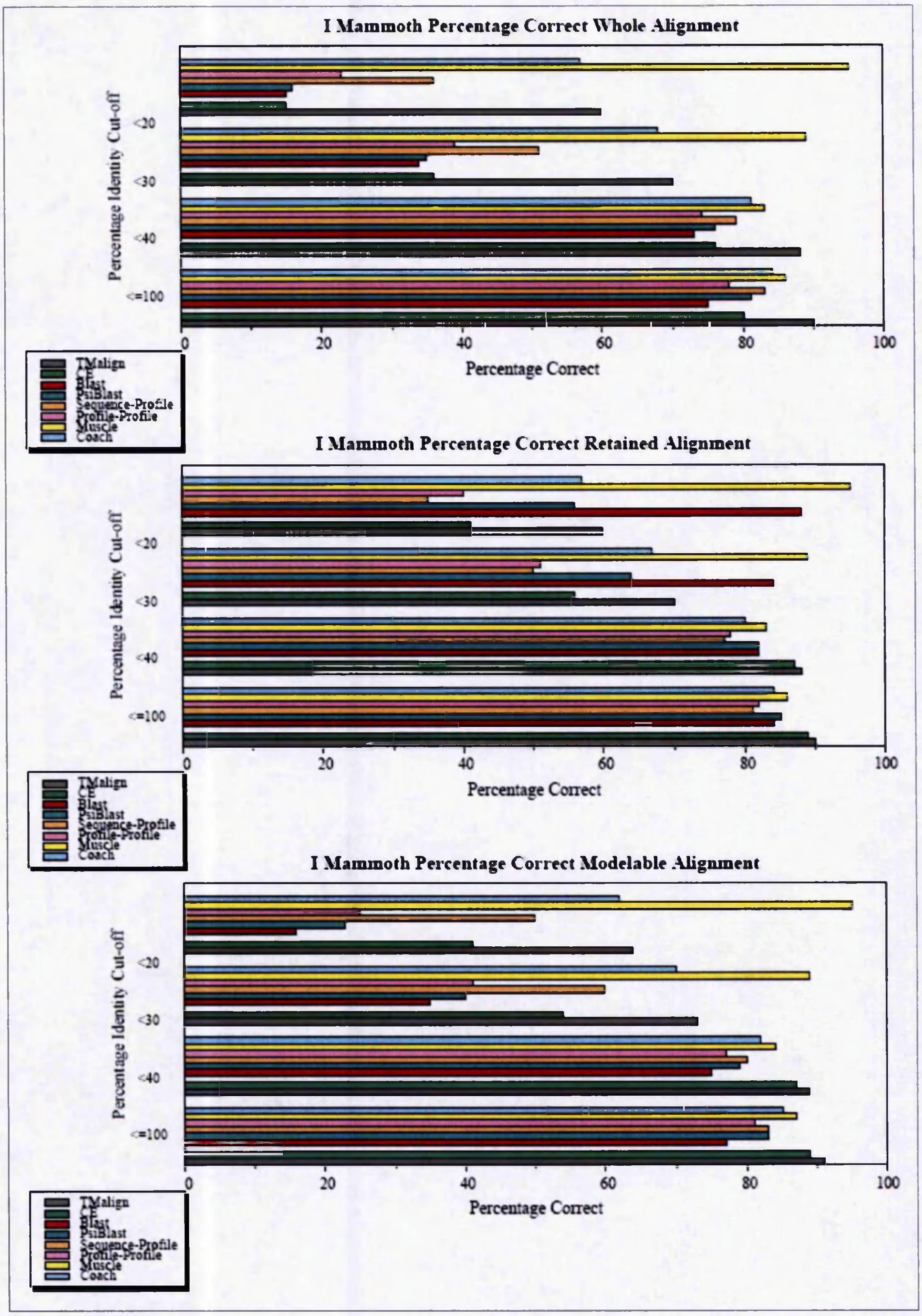


Figure A2.13. Percentage of Correctly Predicted Residues for MAMMOTH, I-vs-I. Results for the percentage of correctly aligned residues by each alignment method, as an average across all pairs in the I-vs-I set, assessed against the gold standard MAMMOTH are shown. The percentage identity bins are inclusive as indicated by the "less than" signs.

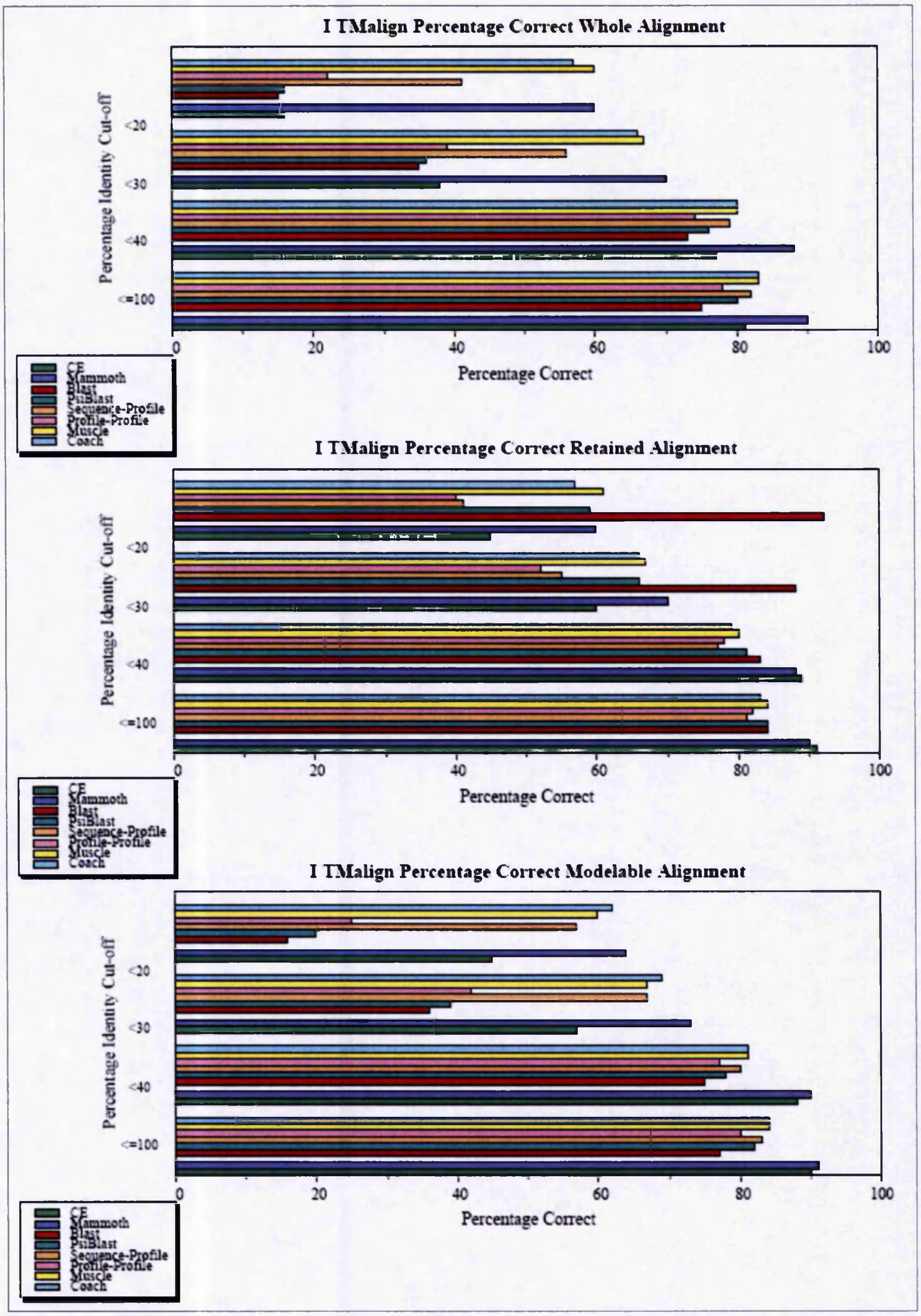


Figure A2.14. Percentage of Correctly Predicted Residues for TM-align, I-vs-I. Results for the percentage of correctly aligned residues by each alignment method, as an average across all pairs in the I-vs-I set, assessed against the gold standard TM-align are shown. The percentage identity bins are inclusive as indicated by the "less than" signs.

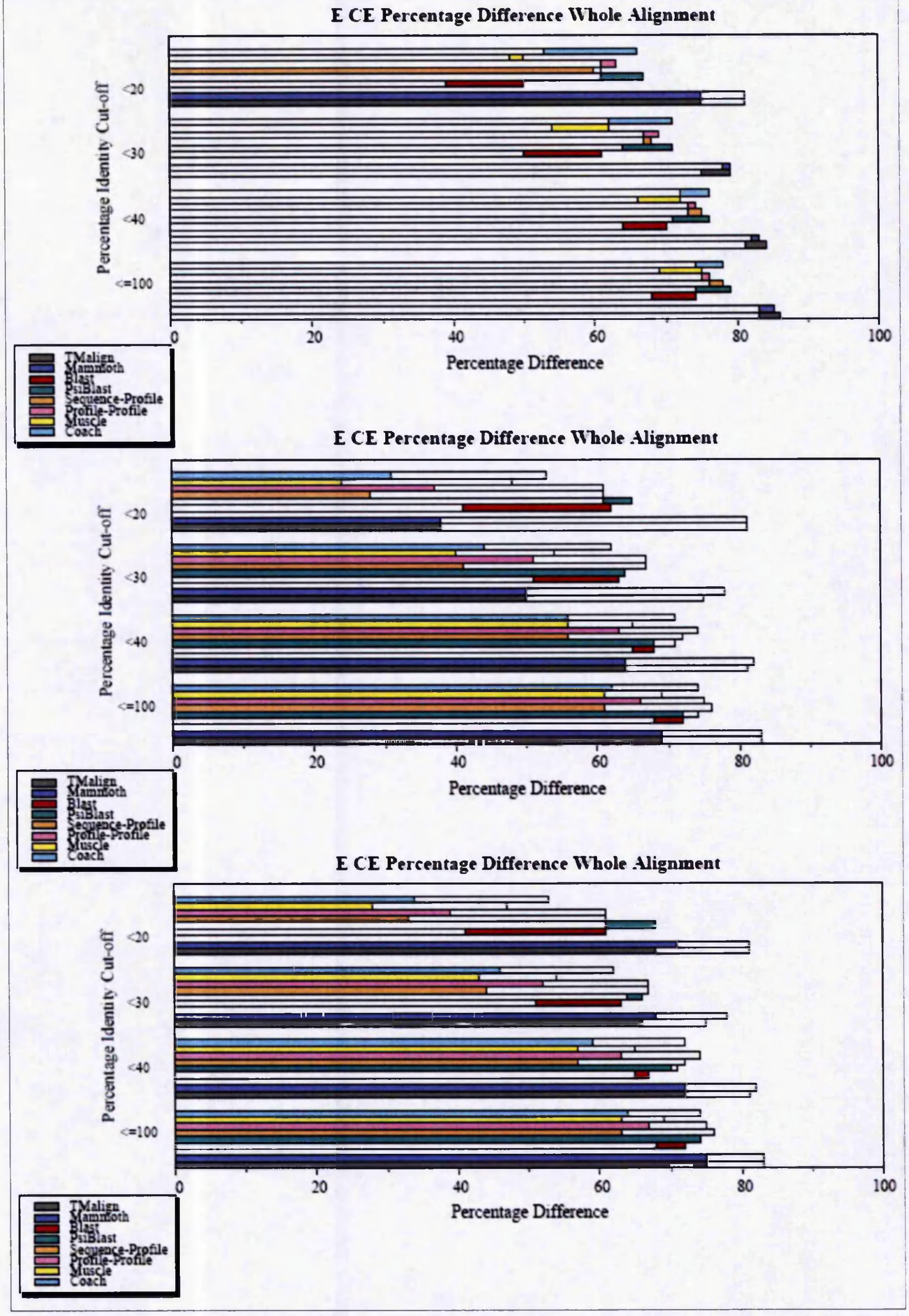
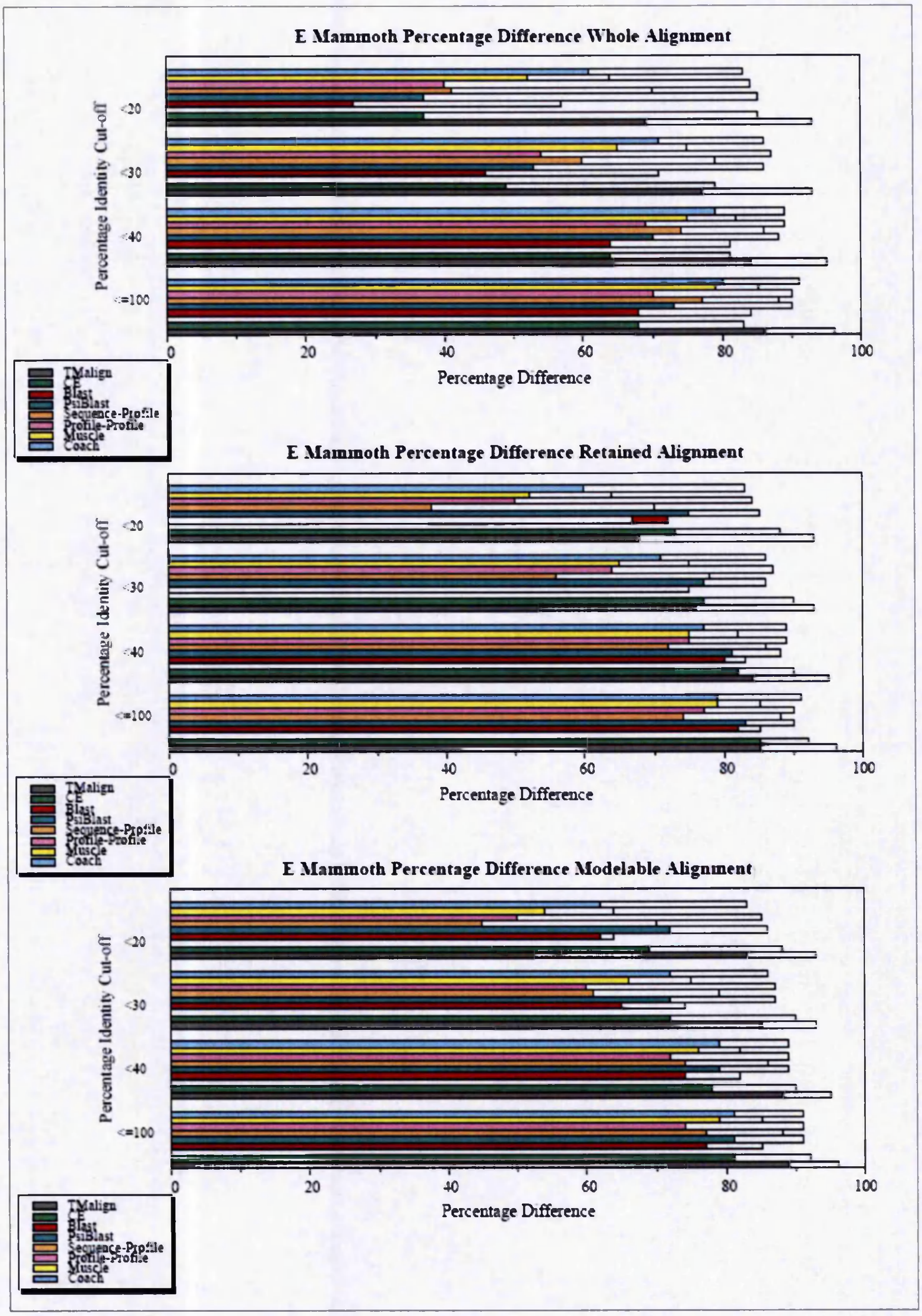


Figure A2.15. The Accuracy of the Non-interface Residues and Interface Residues for the CE Method. The different methods (using CE as the gold standard) were displayed on the graph with the percentage identity bins for the I-vs-S set, plotted against the percentage of correctly predicted residues of the alignment (non-interface residues; the solid coloured bars) with the percentage of correctly predicted interface residues (the white, empty bars).

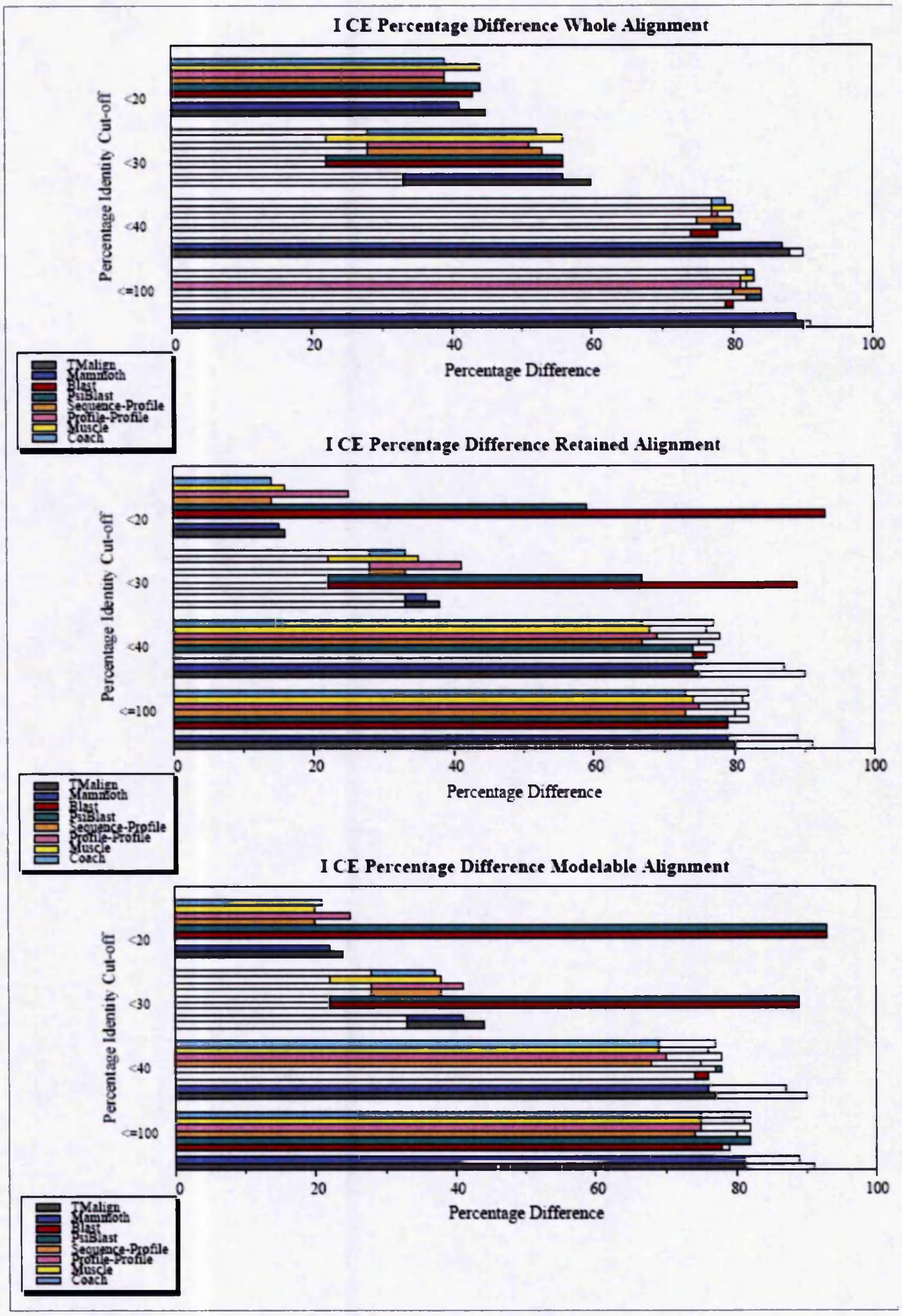


E Mammoth Percentage Difference Whole Alignment

E Mammoth Percentage Difference Retained Alignment

E Mammoth Percentage Difference Modelable Alignment

Figure A2.16. The Accuracy of the Non-interface Residues and Interface Residues for the MAMMOTH Method. The different methods (using MAMMOTH as the gold standard) were displayed on the graph with the percentage identity bins for the I-vs-S set, plotted against the percentage of correctly predicted residues of the alignment (non-interface residues; the solid coloured bars) with the percentage of correctly predicted interface residues (the white, empty bars).



I CE Percentage Difference Whole Alignment

I CE Percentage Difference Retained Alignment

I CE Percentage Difference Modelable Alignment

Figure A2.17. The Accuracy of the Non-interface Residues and Interface Residues for the CE Method for the I-vs-I set. The different methods (using CE as the gold standard) were displayed on the graph with the percentage identity bins for the I-vs-I set, plotted against the percentage of correctly predicted residues of the alignment (non-interface residues; the solid coloured bars) with the percentage of correctly predicted interface residues (the white, empty bars).

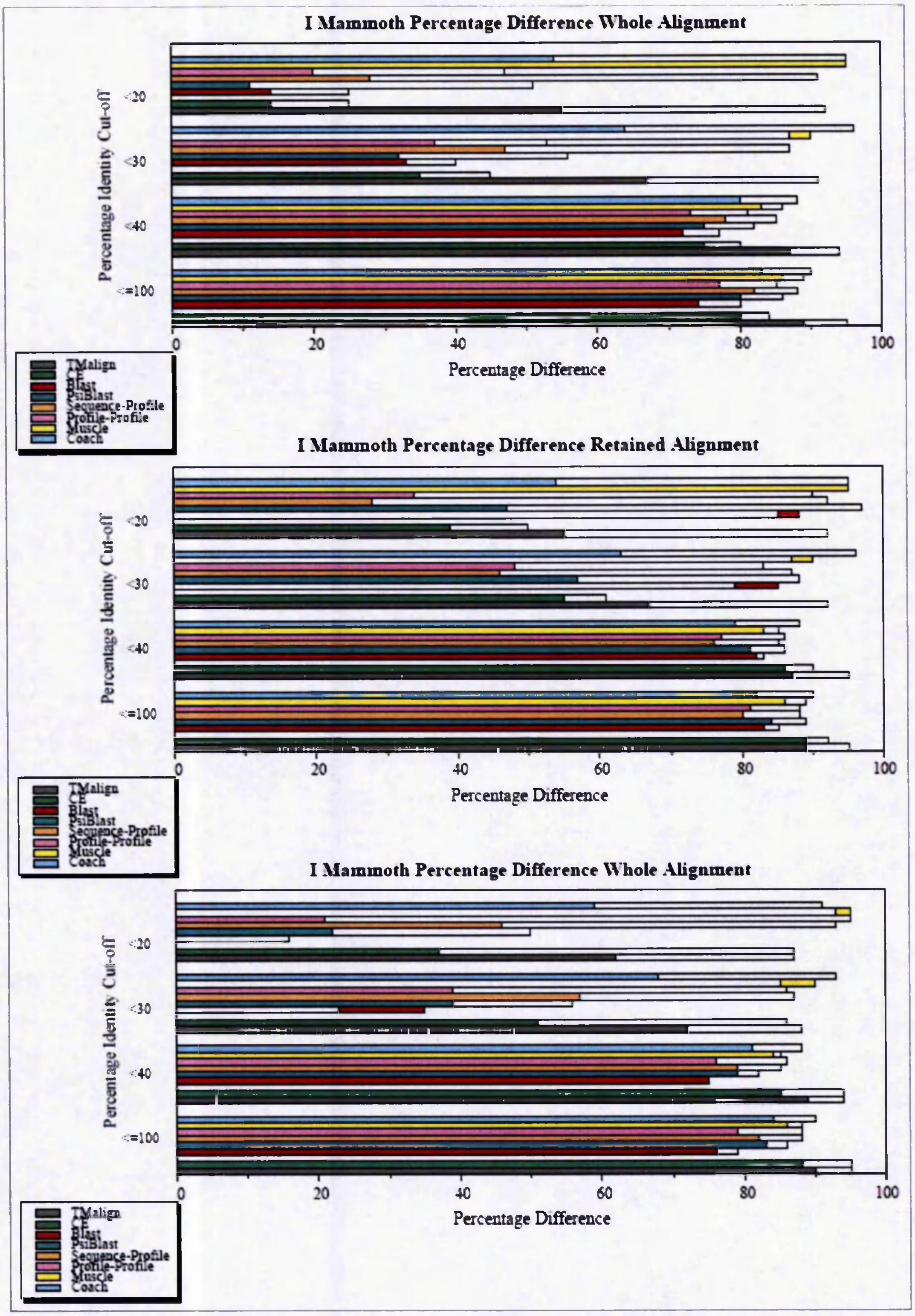


Figure A2.18. The Accuracy of the Non-interface Residues and Interface Residues for the MAMMOTH Method for the I-vs-I set. The different methods (using MAMMOTH as the gold standard) were displayed on the graph with the percentage identity bins for the I-vs-I set, plotted against the percentage of correctly predicted residues of the alignment (non-interface residues; the solid coloured bars) with the percentage of correctly predicted interface residues (the white, empty bars).

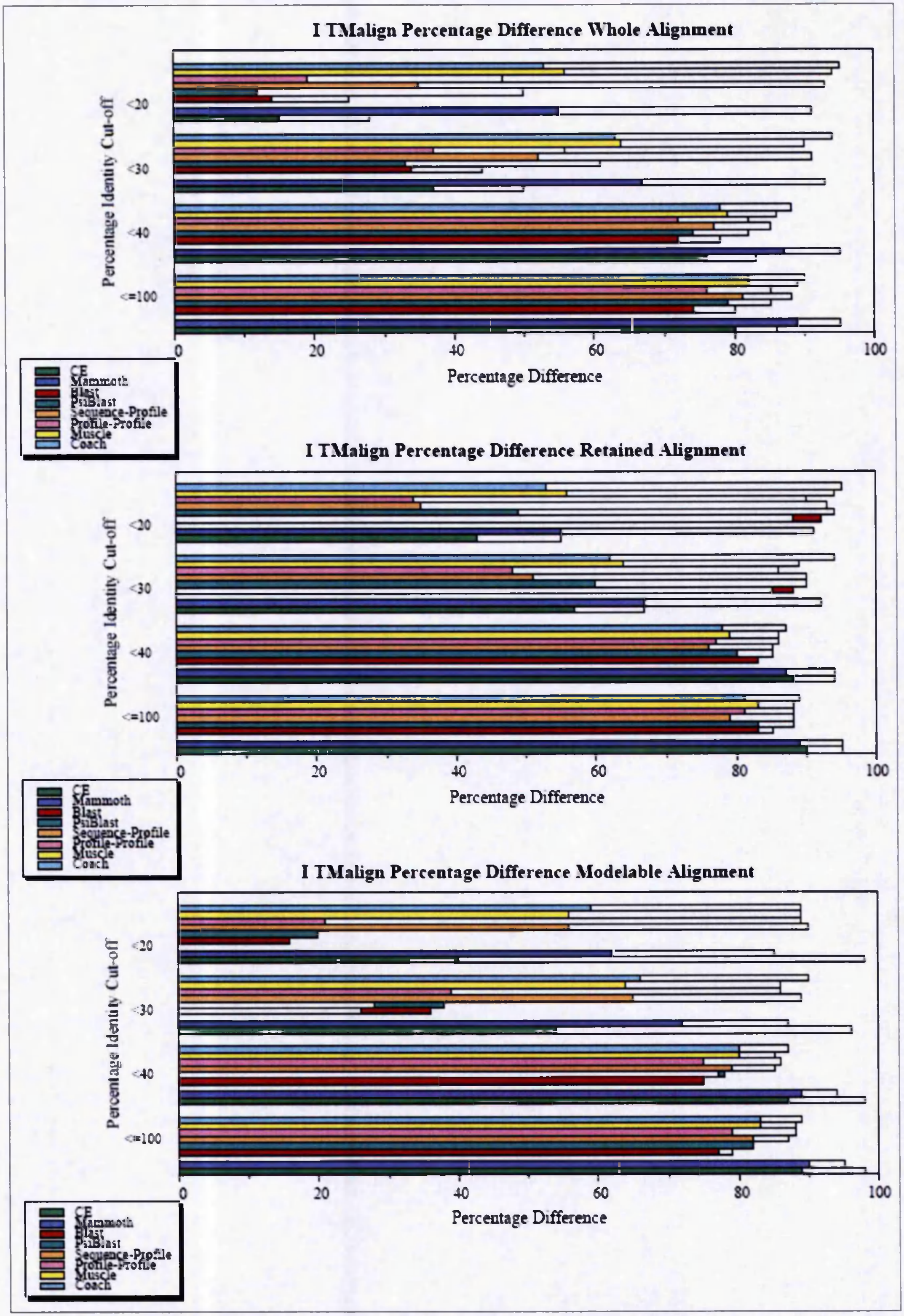


Figure A2.19. The Accuracy of the Non-interface Residues and Interface Residues for the TM-align Method for the I-vs-I set. The different methods (using TM-align as the gold standard) were displayed on the graph with the percentage identity bins for the I-vs-I set, plotted against the percentage of correctly predicted residues of the alignment (non-interface residues; the solid coloured bars) with the percentage of correctly predicted interface residues (the white, empty bars).

(a)

Methods	Standard		Loop	
	All	I	All	I
TM-align	2.65	2.31	2.89	2.64
PSI-BLAST	3.91	3.28	4.01	3.35
BLAST	3.05	2.93	3.14	3.03
MUSCLE	4.47	3.67	4.56	3.77
Sequence-Profile	3.81	3.31	3.88	3.39
Profile-Profile	3.89	3.34	4.03	3.52
COACH	3.88	3.23	4.02	3.36

(b)

Methods	Standard		Loop	
	NI	I	NI	I
TM-align	0.74	0.96	0.78	1.04
PSI-BLAST0.87	0.87	1.13	0.89	1.15
BLAST	0.86	1.05	0.87	1.07
MUSCLE	0.97	1.20	0.98	1.22
Sequence-Profile	0.90	1.15	0.91	1.05
Profile-Profile	0.88	1.01	0.91	1.05
COACH	0.89	1.14	0.91	1.16

Table A3.1. The Global and Local RMSD Results for the Main chain – Main chain. For the I-vs-S set, the global (table a) and the local (table b) main chain – main chain RMSDs are shown. "Standard" and "loop" refer to the ten or fifty models, respectively, and the refinement level these results were averaged over. "All" indicates the global RMSD value of all of the residues and "I" indicates the interface global RMSD result. "NI" indicates the non-interface RMSD value of all of the residues that were not part of the interface and "I" indicates the interface local RMSD result. The green boxes highlight the best (lowest) RMSD result, the red, the worst – ignoring the benchmarking TM-align which is a structural alignment method.

(a)

Methods	Standard		Loop	
	All	I	All	I
TM-align	3.42	3.23	3.66	3.57
PSI-BLAST	4.65	4.19	4.74	4.27
BLAST	3.80	3.83	3.90	3.95
MUSCLE	5.19	4.56	5.29	4.66
Sequence-Profile	4.53	4.17	4.60	4.28
Profile-Profile	4.24	4.07	4.39	4.26
COACH	4.58	4.11	4.72	4.25

(b)

Methods	Standard		Loop	
	NI	I	NI	I
TM-align	1.71	1.95	1.78	2.07
PSI-BLAST	4.65	4.19	1.91	2.21
BLAST	3.80	3.83	1.75	2.00
MUSCLE	5.19	4.56	2.01	2.30
Sequence-Profile	4.53	4.17	1.93	2.23
Profile-Profile	1.69	1.84	1.75	1.92
COACH	1.89	2.17	1.94	2.23

Table A3.2. The Global and Local RMSD Results for All Atoms. For the I-vs-S set, the global (table a) and the local (table b) all atoms RMSDs are shown. “Standard” and “loop” refer to the ten or fifty models, respectively, and the refinement level these results were averaged over. “All” indicates the global RMSD value of all of the residues and “I” indicates the interface global RMSD result. “NI” indicates the non-interface RMSD value of all of the residues that were not part of the interface and “I” indicates the interface local RMSD result. The green boxes highlight the best (lowest) RMSD result, the red, the worst – ignoring the benchmarking TM-align which is a structural alignment method.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
PDB	14	175	119	94	30	148
BLAST	100	87.43	78.99	100	96.67	39.19
COACH	100	86.29	78.99	100	100	74.32
MUSCLE	100	87.43	80.67	100	100	49.32
Profile-Profile	100	87.43	78.99	100	100	66.22
PSI-BLAST	100	87.43	78.99	100	100	95.95
Sequence-Profile	100	86.29	79.83	100	100	67.57
TM-align	100	86.86	100	100	100	97.30

Table A3.3. The Percentage of Correct Contact s. The percentage of correct contacts made by each method for the six pairs. The total number of contacts made in the PDB structure file below 5Å distance is shown in the top row. The results are for main chain – main chain contacts.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
BLAST	0.85	0.76	0.71	0.49	1.11	0.80
COACH	0.80	0.80	0.51	0.57	0.52	0.52
MUSCLE	0.74	0.80	0.52	0.48	0.73	0.59
Profile-Profile	0.88	0.84	0.68	0.54	0.47	0.68
PSI-BLAST	0.66	0.81	0.54	0.55	0.93	0.70
Sequence-Profile	0.82	0.80	0.57	0.53	0.64	0.58
TM-align	0.99	0.81	0.57	0.54	0.52	0.73

Table A3.4. The Average Difference in Distances of Correct Contacts Within +/- 2Å. The difference in distances of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 2Å from the correct contact in the PDB file. The results are for side chain – side chain contacts.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
BLAST	1.06	1.08	0.93	0.53	1.31	0.88
COACH	0.94	1.14	0.62	0.82	0.68	0.65
MUSCLE	1.09	1.17	0.73	0.55	1.15	0.76
Profile-Profile	1.21	1.15	0.94	0.62	0.53	0.82
PSI-BLAST	0.92	1.12	0.71	0.63	1.08	1.01
Sequence-Profile	1.03	1.14	0.79	0.57	0.64	0.60
TM-align	1.26	1.12	0.76	0.63	0.57	0.89

Table A3.5. The Average Difference in Distances of Correct Contacts Within +/- 3Å. The difference in distances of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 3Å from the correct contact in the PDB file. The results are for side chain – side chain contacts.

Methods	1f34A +1htrB	1avgH +1autC	1ppfE +1pytD	1stfE +1cs8A	2sicE +1r64A	2kaiA +1cvwH
BLAST	0.48	0.47	0.39	0.39	0.36	0.45
COACH	0.47	0.50	0.38	0.39	0.41	0.38
MUSCLE	0.48	0.43	0.29	0.37	0.42	0.40
Profile-Profile	0.51	0.46	0.41	0.34	0.45	0.32
PSI-BLAST	0.44	0.45	0.35	0.34	0.45	0.55
Sequence-Profile	0.47	0.43	0.36	0.33	0.41	0.41
TM-align	0.46	0.52	0.37	0.37	0.52	0.41

Table A3.6. The Average Difference in Distances of Correct Contacts Within +/- 1Å. The difference in distances of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 1Å from the correct contact in the PDB file. The results are for side chain – side chain contacts.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
BLAST	0.19	0.46	0.28	0.28	0.53	0.22
COACH	0.29	0.47	0.25	0.27	0.23	0.29
MUSCLE	0.81	0.45	0.26	0.26	0.52	0.26
Profile-Profile	0.17	0.46	0.25	0.31	0.34	0.37
PSI-BLAST	0.47	0.45	0.17	0.27	0.16	0.26
Sequence-Profile	0.92	0.47	0.24	0.28	0.38	0.28
TM-align	0.28	0.40	0.25	0.28	0.26	0.34

Table A3.7. The Average Difference in Distances of Correct Contacts Within +/- 1Å. The difference in distances of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 1Å from the correct contact in the PDB file. The results are for main chain – main chain contacts.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
BLAST	0.65	0.67	0.33	0.30	1.04	0.45
COACH	0.82	0.75	0.34	0.28	0.23	0.60
MUSCLE	1.03	0.67	0.31	0.27	0.71	0.5
Profile-Profile	0.68	0.65	0.34	0.33	0.34	0.64
PSI-BLAST	1.12	0.64	0.24	0.28	0.20	0.46
Sequence-Profile	1.38	0.75	0.28	0.28	0.38	0.46
TM-align	0.58	0.64	0.39	0.28	0.26	0.58

Table A3.8. The Average Difference in Distances of Correct Contacts Within +/- 2Å. The difference in distances of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 2Å from the correct contact in the PDB file. The results are for main chain – main chain contacts.

Methods	1avgH +1autC	1f34A +1htrB	1ppfE +1pytD	1stfE +1cs8A	2kaiA +1cvwH	2sicE +1r64A
BLAST	1.35	1.15	0.44	0.30	1.28	0.62
COACH	0.82	1.22	0.55	0.28	0.23	0.78
MUSCLE	1.73	1.21	0.41	0.27	0.71	0.69
Profile-Profile	0.68	1.16	0.53	0.33	0.34	0.71
PSI-BLAST	1.78	1.15	0.35	0.28	0.20	0.60
Sequence-Profile	1.67	1.22	0.48	0.28	0.38	0.50
TM-align	0.84	1.08	0.66	0.28	0.26	0.70

Table A3.9. The Average Difference in Distances of Correct Contacts Within +/- 3Å. The difference in distances of correct contacts made by each method for the six pairs, these are the contacts which are only a distance of +/- 3Å from the correct contact in the PDB file. The results are for main chain – main chain contacts.

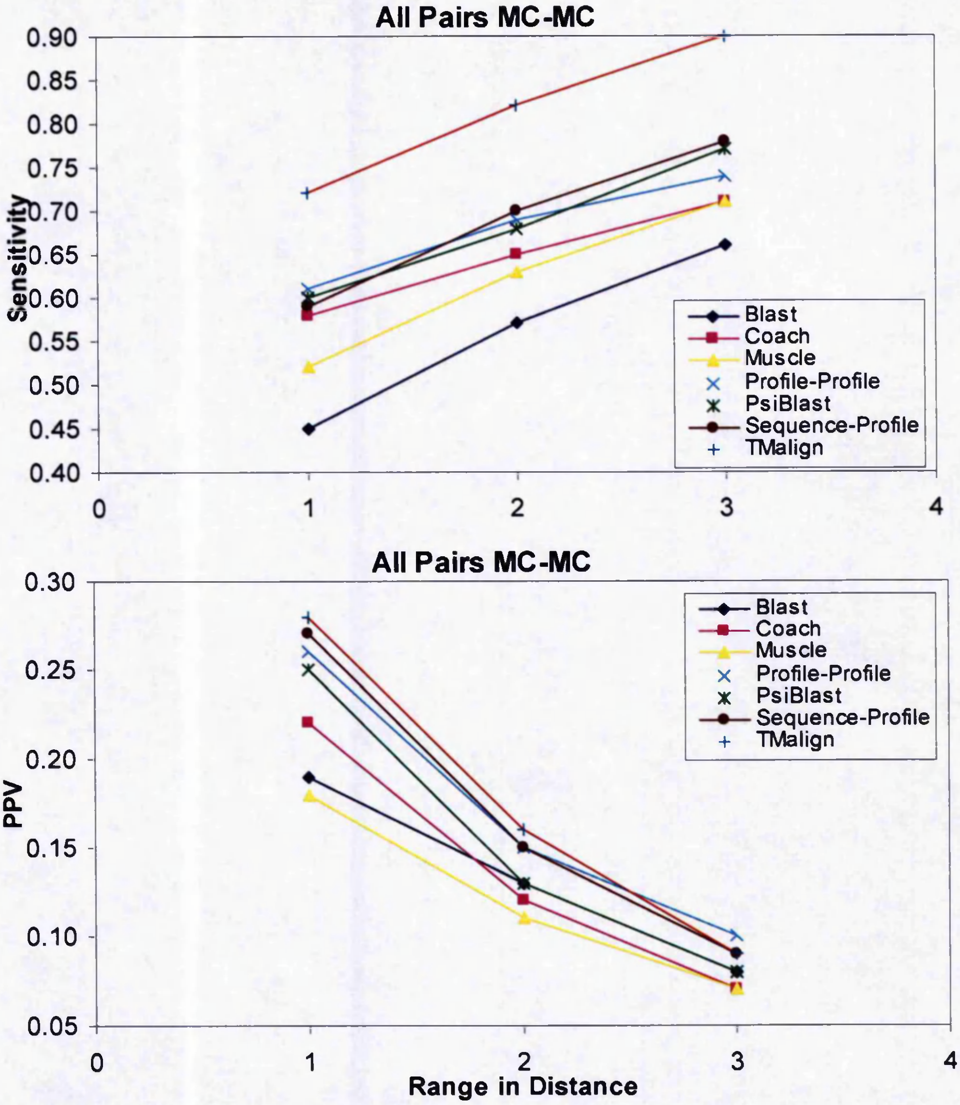


Figure A3.1. The Sensitivity and PPV of the Differences in Distances for the Main chain – Main chain. The sensitivity and positive predicted value of the different alignment methods as an average over the six pairs, for the main chain – main chain correct contact results.