Protein Model Construction and Evaluation

Daniel Peter Klose

University College London

Division of Mathematical Biology
National Institute for Medical Research
The Ridgeway
Mill Hill
NW7 1AA

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

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

Abstract

The prediction of protein secondary and tertiary structure is becoming increasingly important as the number of sequences available to the biological community far exceeds the number of unique native structures. The following chapters describe the conception, construction, evaluation and application of a series of algorithms for the prediction and evaluation of two and three-dimensional protein structure. In chapter 1 a brief overview of protein structure and the resources required to predict protein features is given. Chapter 2 describes the investigation of sequence identity and alignments on the prediction of two-dimensional protein structure in the form of long and short range protein contacts a feature which is known to correlate with solvent accessibility. It also describes the identification of a feature which is referred to as the 'Empty Quarter' which forms the basis of an evaluation function described in Chapter 3 and developed in Chapter 4. Chapter 3 introduces the Dynamic Domain Threading method used during round six of the CASP exercise. Phobic, a protein evaluation function based on predicted solvent accessibility is described in Chapter 4. The *de novo* prediction of α/β proteins is described in Chapter 5, the method introduces a new approach to the old problem of combinatorial modelling and breaks the size limit previously imposed on *de novo* prediction. The final experimental chapter describes the prediction of solvent accessibility and secondary structure using a novel combination of the fuzzy k-nearest neighbour and support vector machine. Chapter 7 closes this piece of work with a review of the field and suggests potential improvements to the way work is conducted.

Acknowledgements

I would like to thank Willie Taylor and Kuang Lin for their continual support and advice, Willie Taylor and members of the lab for reading numerous versions of this work in their spare time; Richard Goldstein, Ben Blackburne and Delmiro Fernandez-Reyes for their help with Bayesian statistics and the work described in chapter 6.

In addition, I would like to thank Jens Kleinjung, Ben Blackburne and Sebastian Wasilewski for the hours they spent solving computer problems which would otherwise have made this work impossible.

I dedicate this thesis to my family.

Contents

List Of Tables

List Of Figures

Abbreviations

1D	one dimensional
2D	two dimensional
3D	three dimensional
aa	amino acid
ANN	Artificial Neural Network
ASA	absolute solvent accessibility
BLAST	Basic Local Alignment Search Tool
CAFASP	critical analysis of fully automated structure prediction
CASP	critical analysis of structure prediction
CB	City Block Distance
CC	coiled-coil
cc	correlation coefficient
CDS	Coding sequence
CM	comparative modeling
CN	contact number
COOH	carboxyl group
cRMSD	carbon-alpha root mean squared deviation
DD	dynamic domain
DDT	dynamic domain threading
DNA	deoxyribonucleic acid
DSSP	Dictionary of Secondary Structure of Proteins
ED	Euclidean Distance
fkNN	fuzzy *k* nearest neighbour
FR	fold recognition
HIV	human immunodeficiency virus
HSI	hue saturation intensity

JTT	Jones, Taylor, Thornton
kNN	*k* nearest neighbour
LC	low complexity
LCS-GDT	longest continuous segment global distance test
MSA	Multiple Sequence Alignments
MST	multiple sequence threading
NF	new fold modeling
NH	Amino group
NMR	nuclear magnetic resonance spectroscopy
PDB	Brookhaven Protein Data Bank
POPS	parameter optimised surfaces
PSI-BLAST	Position Specific Iterative - Basic Local Alignment Search Tools
PSSM	position specific scoring matrix
PT	Periodic Table
RBF	Radial Basis Function
RGB	red green blue
RMSD	root mean squared deviation
RoG	radius of gyration
RSA	relative solvent accessibility
RWCO	residue wide contact number
SA	solvent accessibility
SASA	solvent accessible surface area
SAP	Structure Alignment Program
SD	Single domain
SS	secondary structure
SVC	Support Vector Classification
SVM	Support Vector Machines
SVR	Support Vector Regression

TCS	Taylor Colour Scheme
TM	trans-membrane
TUNE	Threading using Neural Networks
α	alpha
Å	Angstroms
β	beta
ε	epsilon (lower case)
γ	gamma (lower case)
ϕ	Phi
ψ	Psi
ξ	xi (lower case)

Chapter 1

An Introduction To Protein Structure

Introduction

Today there are in excess of 60 million linear amino acid sequences in the GenBank database (Benson et al., 2005). In comparison there are only 40,000 three dimensional (3D) structures available in the Worldwide Protein Data Bank (PDB) (Berman et al., 2000). With high throughput genome sequencing projects elucidating data at an astonishing rate it is likely that the sequence structure gap will continue to expand.

While biophysical techniques, such as X-ray crystallography (crystallography) and nuclear magnetic resonance spectroscopy (NMR), provide detailed information about the 3D coordinates of atoms within a protein they suffer from several, currently unavoidable, problems, including a size restraint of approximately 60kd for NMR and the ability to form crystals for X-ray crystallography. This means that structures such as trans-membrane proteins are difficult to solve because of their membrane bound location and large size. This is reflected in the PDB where there are only 234 structures of which 120 are unique[1]. Even without these problems a major hurdle remains: a vast amount of skill and time has to be invested in each structure to overcome a myriad of potential problems, making rapid, automated elucidation of structures very challenging.

With the advent of the Human Genome Project[1] the field of computational biology has taken on a new importance. Bioinformatics, as it is now often called, uses expertise from the fields of computer science and mathematics to record, analyse and predict biological features from sequences and structures. For over thirty years it has been generally accepted that the amino acid sequence provides enough information to specify

[1] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

the overall 3D shape of a protein. This concept is supported by the pioneering work of Anfinsen and co-workers (Anfinsen, 1972, Anfinsen, 1973) and, as a result of this, one of the grand challenges of bioinformatics has been to use the wealth of sequence data to predict the folded protein structure.

Protein Structure

Protein Primary Structure

All proteins (polypeptides) consist of a linear chain comprising a mix of twenty possible amino acids (monomers). Each monomer consists of an amino group (NH_2), a central carbon atom (C_α) and a carboxyl group (COOH). The polypeptide chain is synthesised by a condensation reaction which forms a peptide bond between the amino group of one monomer and the carboxyl group of another monomer (see figure 1.1). The order of the monomers is called the primary structure of a protein.

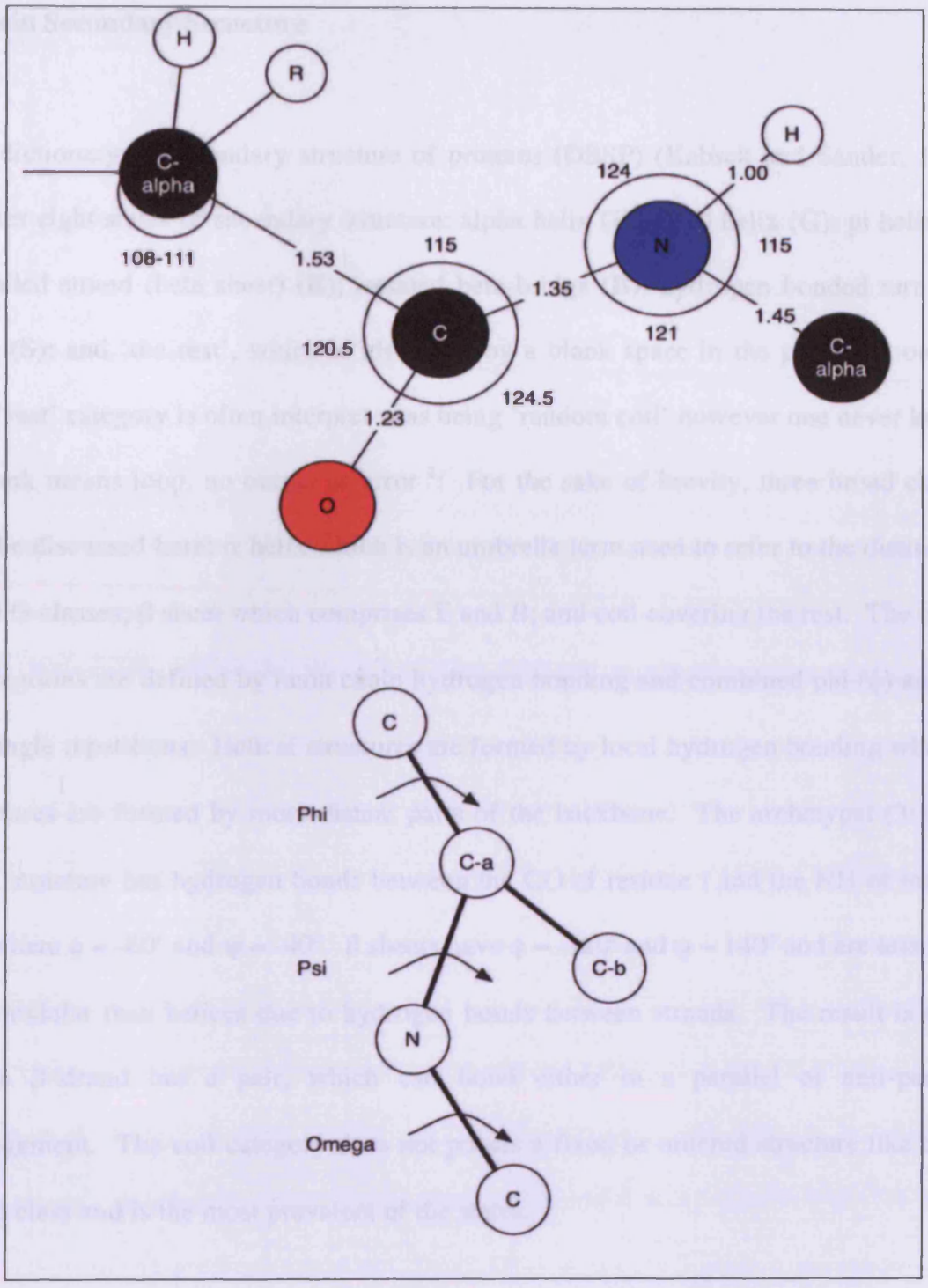


Figure 1.1: Amino acid structure: Individual amino acids are linked by a peptide bond synthesized during a condensation reaction. The angles between atoms intra and inter-residue are important to overall protein structure and especially the limited torsional freedom about the peptide bond. The diagram indicates the bond lengths, angles and names in proteins.

Protein Secondary Structure

The dictionary of secondary structure of proteins (DSSP) (Kabsch and Sander, 1983) defines eight states of secondary structure: alpha helix (**H**); 3/10 helix (**G**); pi helix (**I**); extended strand (beta sheet) (**E**); isolated beta-bridge (**B**); hydrogen bonded turn (**T**); bend (**S**); and 'the rest', which is identified by a blank space in the programs output. The 'rest' category is often interpreted as being 'random coil' however one never knows if blank means loop, no output or error [2]! For the sake of brevity, three broad classes will be discussed here: α helix which is an umbrella term used to refer to the distinct H, I and G classes; β sheet which comprises E and B; and coil covering the rest. The α and β categories are defined by main chain hydrogen bonding and combined phi (ϕ) and psi (ψ) angle repetitions. Helical structures are formed by local hydrogen bonding whilst β structures are formed by more distant parts of the backbone. The archetypal (3/10) α helix structure has hydrogen bonds between the CO of residue *i* and the NH of residue *i*+4 where $\phi \approx -60°$ and $\psi \approx -40°$. β sheets have $\phi \approx -120°$ and $\psi \approx 140°$ and are less local and modular than helices due to hydrogen bonds between strands. The result is not a single β-strand but a pair, which can bond either in a parallel or anti-parallel arrangement. The coil category does not posses a fixed or ordered structure like the α and β class and is the most prevalent of the states.

[2] http://swift.cmbi.ru.nl/gv/dssp/

Protein Tertiary Structure

The tertiary structure of a protein refers to its overall shape in 3D space. Tertiary structure also includes domains which are stable, compact units that fold autonomously and perform associated functions semi-independently.

Protein Folding

It has been said that proteins 'understand' how to fold but biochemists do not. Over the years many theories have tried to explain the mechanisms behind protein folding (see (Dill, 1990) for an in-depth review). Today the paradigm is that protein folding is driven by hydrophobic partitioning of amino acids based upon their physicochemical properties (Anfinsen, 1972, Anfinsen, 1973, Rose and Roy, 1980). Taylor showed that amino acids can be divided, on paper, into several overlapping classes based on these properties (Taylor, 1986, Taylor, 1997b). The two largest groups are the non-polar (hydrophobic) and the polar (hydrophilic) residues, it is the partitioning of the hydrophilic/phobic residues with respect to the solvent (essentially water) which is the overall driving force behind protein folding and stability. Hydrogen bonding is important in maintaining specific structural features but it is crucial to remember that for every hydrogen bond formed internally two bonds with water are lost and one water–water bond formed – a simple bond count reveals no net gain (Klose and Taylor, 2007). The resulting structure is a core of hydrophobic residues (from which water has been excluded), surrounded by a shell of hydrophilic residues which interface with the solvent making the protein soluble.

A consequence of these properties is that proteins typically fold so that the secondary structure elements are arranged into common topological patterns (Sternberg and Thornton, 1976, Sternberg and Thornton, 1977, Levitt and Chothia, 1976, Richardson, 1976). The fascinating arrangements of these elements shared by proteins with different functions (Orengo et al., 1993a) allows for them to be split into families (much like the kingdom of life) based upon similar tertiary structure (Overington et al., 1993, Orengo et al., 1993b, Yee and Dill, 1993). Today, such classifications are found in the CATH (Orengo et al., 1997) and SCOP (Murzin et al., 1995) databases which will be considered in more detail later.

Solvent Accessibility

Taylor's Venn diagram (Taylor, 1997b) has two main groups, the hydrophobic or non-polar residues and the hydrophilic or polar residues. When examining a folded protein it can be useful to look at the entire protein surface or single residues with this classification in mind (Manavalan and Ponnuswamy, 1978, Nozaki and Tanford, 1971). Examination of the protein surface may yield some insight into whether the protein is folded properly with the hydrophobic residues sequestered to the core and hydrophilic residues on the surface where they interact with the solvent (Rose and Roy, 1980). There are two values which can be used to describe the solvent accessibility of a residue, absolute solvent accessibility (ASA) and relative solvent accessibility (RSA) (Lee and Richards, 1971). The ASA is the total exposed surface area of a residue. while the RSA is a measure of exposure based on a residue being in a GLY-*x*-GLY (Rost and Sander, 1994) or ALA-*x*-ALA (Ahmad et al., 2004a, Ahmad et al., 2004b) tripeptide conformation depending on the scheme used. Both of measures can be calculated using

the ACCESS program (Lee and Richards, 1971), repacked as NACCESS (Hubbard, 1993) using the atomic radii of (Alden and Kim, 1979) which, along with DSSP (Kabsch and Sander, 1983), are regarded as the 'gold-standard'.

Contact Number

Contact number (CN) was first described and used for structure evaluation in the 1980s by Nishikawa and Ooi (Nishikawa and Ooi, 1980, Nishikawa and Ooi, 1986). They described a "simple and good measure" to show the relative location of a residue on the surface or interior of a protein (Nishikawa and Ooi, 1980). The Ooi number, as it was termed, was an estimate of the number of Cα atoms within an 8Å sphere centred on the Cα of a given residue. It was also proposed as a meaningful alternative to the prediction of secondary structure as, at the time, it could be predicted to similar accuracies (Nishikawa and Ooi, 1980). Today contact number is generally considered as an alternative measure to solvent accessibility, as they are strongly correlated and equally well predicted (Kinjo et al., 2005, Yuan, 2005, Hamelryck, 2005). While they are well correlated, it has been suggested that contact number is more conserved across a familial alignment (Hamelryck, 2005) and so should lend itself more towards prediction. The relationship between contact number is simple: a residue found in the core of a protein is likely to be surrounded by a number of other core residues, thus it is unlikely that much of its surface will be exposed to solvent, the result being a high contact number and a low ASA/RSA. A residue located in an exposed loop is unlikely to have many surrounding residues (neglecting its sequence neighbours) and thus has a low contact number and high ASA/RSA. The big difference between contact number

and solvent accessibility is that the former provides some clue as to the location a residue while the latter cannot (Hamelryck, 2005).

Today, contact number is typically defined as the number of residues within an *n* angstrom (Å) sphere of a central residue (*i*). The radius of the sphere is typically set to 10Å and is placed directly on the C_β of *i* (both the sphere location and size differ to that proposed by Nishikawa and Ooi). There are many variations on this method including altering the size of the sphere placed on *i*, the position of the sphere – on the $C_{\beta/\alpha}$ (Kinjo et al., 2005) for all residues except glycine or the use of half-spheres (Hamelryck, 2005).

Databases & Resources

For a reference to all of the databases in the following section please refer to (Galperin, 2007).

Sequence Databases

The Non-redundant database

The non-redundant database (nrdb) is compiled by the National Center for Biotechnology Information (NCBI) as a database for Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). The nrdb contains non-identical sequences from GenBank CoDing Sequence (CDS) translations (Benson et al., 2005), the Brookhaven Protein Data Bank (PDB) (Berman et al., 2000), SwissProt (Boeckmann et al., 2003),

Protein Information Resource (PIR) (Barker et al., 1999) and PRF (PRF, 2007). Sequence alignments are crucial to structure prediction (see chapter 2 for a detailed discussion), and the nrdb is a commonly used resource. Changes are made to the nrdb before being used in structural pursuits, this includes filtering to remove low complexity (LC), transmembrane (TM) and coiled coil (CC) regions. These regions are removed as they tend to produce spurious, insignificant matches with regions that do not share biological function.

UniProt

The Universal Protein Resource (UniProt) is a comprehensive resource or protein sequence and annotation data (Apweiler et al., 2004, Leinonen et al., 2004). The data consists of three projects: Knowledgebase (UniProtKB) (Boutet et al., 2007, Leinonen et al., 2006, Martin, 2005) which is a collection of functional information; Reference Clusters (UniRef) (Suzek et al., 2007, Apweiler, 2008) which contains clustered sets of sequences from knowledgebase as well as information from UniParc; and Archive (UniParc) (Apweiler, 2008) which is a comprehensive, non-redundant database that contains almost all publicly available sequence records. UniProt is a collaboration between the European Bioinformatics Institute, the Swiss Institute of Bioinformatics and the Protein Information Resource (mentioned above) and is designed as a replacement for the aforementioned databases.

The Conserved Domain Database (CDD)

The Conserved Domain Database or CDD (Marchler-Bauer et al., 2002) is a database of conserved domain alignments with links to 3D structures of domains. The alignments in CDD are based on publicly available data from Pfam (Bateman et al., 2000) and Simple Modular Architecture Research Tool (SMART) (Schultz et al., 1998) and are based primarily on sequence alignments.

Structure Databases

The Brookhaven/RCSB Protein Data Bank (PDB)

The Brookhaven/RCSB Protein Data Bank (Berman et al., 2000) was established in 1971 as a repository for biological macromolecular crystal structures. The number of submissions has grown year on year and it is now a requirement that structural data is submitted to the PDB prior to publication. The database is no longer limited to crystallographic structures and now includes Nuclear Magnetic Resonance (NMR) spectroscopy, cryoelectron microscopy and theoretical models. Alongside the 3D data the PDB also provides access to sequence, secondary structure, structural classification and function information via external databases such as the Gene Ontology (Ashburner et al., 2000).

Protein Structure Classification: SCOP and CATH

The Structural Classification of Proteins (SCOP) database (Murzin et al., 1995, Andreeva et al., 2004, Lo Conte et al., 2002, Lo Conte et al., 2000, Hubbard et al., 1999, Hubbard et al., 1998) provides a comprehensive, manually curated, description of the structural and evolutionary relationships of proteins with known 3D structure. The database works by classifying proteins on hierarchical levels; family; superfamily; common fold; class. The family level addresses the issue of common evolutionary origin, proteins that share 'significant' sequence similarity or extremely similar structure and function but dissimilar sequence, such as globins, are grouped. The superfamily group classifies proteins solely on the basis of function and *low* sequence identity, such a classification encompasses the immunoglobulins with their variable and constant domains. The third tier is based on the major secondary structure elements – all members of the class adhere to the same arrangement and topological connections. Proteins that follow these rules are said to share a common fold. The final level of classification comes at the secondary structure level where the proteins are grouped into 5 major classes:

1. All-α: proteins that consist predominantly of α-helices;

2. All-β: proteins that consist predominantly of β-sheets;

3. α/β: proteins containing intermixed α-helices and β-sheets;

4. α+β: proteins in which there are segregated α-helices and β-sheets;

5. Multi-domain: proteins that consist of domains belonging to different classes or for which there are no known homologues.

There are several other small classes that address peptides, small proteins, nucleic acids and carbohydrates, more details can be found in (Hubbard et al., 1999). The distinction between each category is important for theoretical work as the use of particular scoring functions can depend upon the class of the target (see chapter 4).

Based on an automatic comparison method (Taylor and Orengo, 1989b, Taylor and Orengo, 1989a) Orengo *et al.*, produced a method for semi-automated classification of proteins in response to the increasing number of protein structures (Orengo et al., 1997). The database groups proteins into four main categories: class (c), architecture (a), topology (t) and homologous superfamily (h) and was dubbed CATH. The first tier of CATH is *class*, as with SCOP, this division is based on the relative content of α-helices and β-sheets with the exception that there are only three groups - the α+β & α/β are merged into an α-β class. Also, in a similar fashion to SCOP, CATH has additional groups containing structures that have minimal secondary structure. The *A*-level, distinguishes structures which occupy the same class but differ in architecture. This does not include discrimination on a topological level, but along more general lines – such as the number of layers in an α-β sandwich i.e. it does not include the details of the connections between secondary structure elements. The provision of the architecture division is unique to the CATH database. The third tier addresses the issue of fold variation, proteins that share the same overall fold - arrangement of secondary structure elements and connections - are grouped together. At this level there is structural

similarity, yet at the same time there is not necessarily common function. The final (H) level groups proteins that share similar structure *and* function, which is strong evidence to suggest that they diverged from a common ancestor. Further details can be found in (Orengo et al., 1997). CATH classifications are also presented in the PDB alongside those of SCOP.

ASTRAL

The ASTRAL compendium (Brenner et al., 2000) provides a link between the structures in the PDB and SCOP protein domains. Sequence information is provided in the form of both the SEQRES and ATOM PDB records while structural information exists in the form of domains unique to ASTRAL. Tools and lists are also provided which give access to subsets of the data held, this includes the ability to extract sequences that share a maximum identity (pre-computed lists cover 40 and 95% identity) as well as the option to retrieve lists of structures. For a detailed description of ASTRAL see (Brenner et al., 2000, Galperin, 2007).

Family of Structurally Similar Proteins

The family of structurally similar proteins (FSSP) is a database of protein structure-structure alignments based on information from the PDB. Initiated in 1992 by Holm *et al.*, it was one of the original structure alignment databases (Holm et al., 1992), however as of November 2004 it was no longer maintained. FSSP consists of 'sets' of proteins. Each set represents proteins that share some structural similarity with a probe protein. In addition to the structural constraints there are also sequence restraints, each set has a

minimum sequence identity of approximately 30% and a maximum no greater than 70%. All structures that share greater than 70% identity are discarded because they have marked structural similarity with the probe. The alignments were computed using the DALI algorithm for optimal pairwise structure alignment (Holm and Sander, 1993).

Structural Alignment Database

The Structural Alignment Database (SAD (Marsden and Abagyan, 2004)) comprises a collection of structural alignments designed for derivation and optimisation of sequence-structure alignment algorithms. The alignments are sourced from HOMSTRAD (Mizuguchi et al., 1998b), BaliBase (Thompson et al., 1999) and SCOP-based Gerstein databases (Marsden and Abagyan, 2004). To maintain status as a high quality resource the creators of SAD define 6 criteria that have to be met for inclusion in the database:

1. Non-redundancy – sequences should be represented once.

2. Cover fold space – contain as many representatives as is required to represent fold space. At the same time the dataset must be normalised to avoid over-representation – IgG folds are abundant in the PDB.

3. High quality and quantity – must contain a number of alignments to be statistically viable. Alignments are in sufficient number to allow for derivation and optimisation of new algorithms.

4. Contain alignments derived from structures with good resolution (better than 2.5Å).

5. Contains alignments that are 'structurally significant'. Alignments with a small number of aligned pairs are not likely to be reliable.

6. Cover a wide range of sequence identities, allowing the effects of sequence identity to be studied.

One of the unique features of this dataset is that it does not use RMSD to evaluate structure alignments, instead a measure based on contact area distance (CAD) is used (Marsden and Abagyan, 2004). SAD contains 1927 high-resolution structures that cover a range of fold and sequence space.

Benchmark ALIngment dataBASE

The benchmark alignment database (BAliBASE (Thompson et al., 1999)) is a collection of manually refined multiple sequence alignments categorised by blocks of sequence conservation sequence length, similarity and the presence of N/C terminal extensions (Thompson et al., 1999). The constituent sequence information was gathered from FSSP, HOMSTRAD and manually constructed structural alignments from literature. Where there is insufficient structure data additional sequence information is gathered from the HSSP database (Sander and Schneider, 1993). Each alignment is manually checked so that conserved blocks and secondary structure elements are aligned. As of version 3 (October 2005) BaliBase contained 6255 alignments.

HOMologous STRucture Alignment Database (HOMSTRAD).

HOMSTRAD (Mizuguchi et al., 1998b) comprises a set of protein structure alignments for homologous families. Akin to SAD, HOMSTRAD enforces a resolution limit, a minimum number of structures per family and low sequence identity. The identity measure is not deliberate but is useful as it results in an average sequence identity greater than 20% for proteins that connect two subgroups. Information about local structural environments calculated by JOY (Mizuguchi et al., 1998a) are also made available. All information in HOMSTRAD is obtained using an automatic pipeline connected to the PDB and as such complements the SCOP and CATH databases. As of June 2007 HOMSTRAD contained 1032 families constructed from 3454 structures as well as 6412 singleton families.

Protein Structure Prediction

Before starting this section, there are two terms that require definition as they will be used extensively throughout this work. The first is target; this refers to the protein sequence that we are trying to assign structure to. The second is template, which refers to a complete, or section of, protein chain that has a known structure and can or has been used to infer structure on the target. A generic approach to structure prediction is shown in figure 1.2 - it should be noted that not all steps are used in all methods.

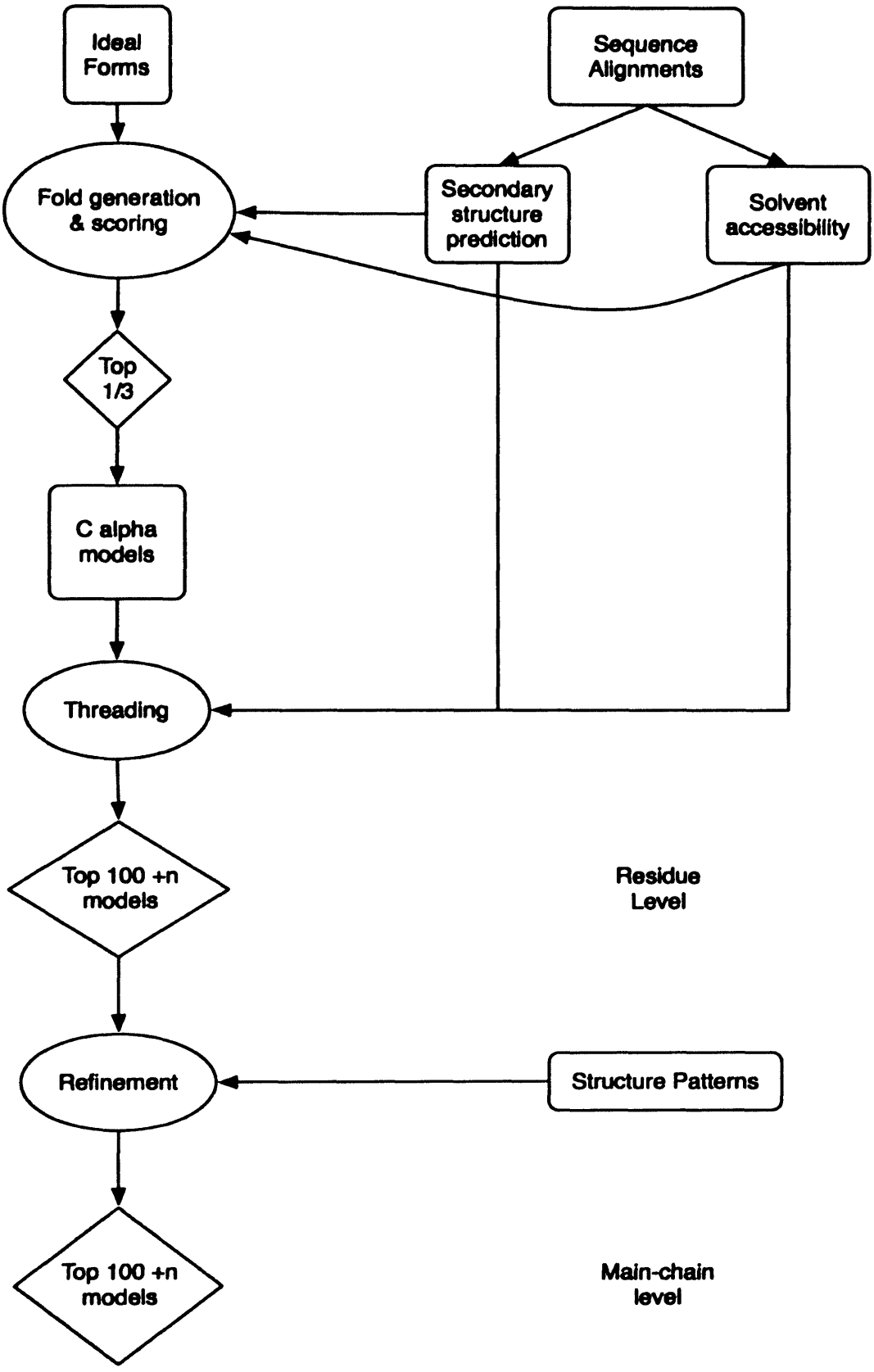


Figure 1.2 Generic Protein Structure Prediction Pipeline: The flow chart presents a generic approach to prediction of three-dimensional structure. It should be noted that a large number of groups do not look for functional annotations and predict their own domain boundaries. The three shaded boxes represent the main approaches for predicting protein structure.

As shown in figure 1.2 there are many aspects to protein structure prediction, ranging from the prediction of 2D features all the way to 3D structure and substrate docking (see CAPRI (Henrick, 2006) for details).

Two Dimensional Protein Structure Prediction

Two dimensional structure prediction thrives as a field distinct from three-dimensional structure. It has it roots in the 1970s when Chou and Fasman developed a technique for assigning secondary structure by hand (Chou and Fasman, 1978). Their method was simple – using observed frequencies of amino acids in known structures probabilities were assigned to residues being in certain structures. The structure classifications are not those used today but rather strong helix formers (**H**), weak helix formers (**h**), indifferent (**I**), weak helix breakers (**b**) and strong helix breakers (**B**). This method was quickly developed by Garnier, Osguthorpe and Robson (GOR) (Garnier et al., 1978) which aside from being ideal for computational use, was more sophisticated. Since the 1980s the number of methods for prediction of secondary structure has grown rapidly including the use of Bayesian Theory (Thompson and Goldstein, 1997), fuzzy *k* nearest neighbours (chapter 6), neural networks (Rost, 1996, Jones, 1999b) and support vector machines (Ward et al., 2003) and combinations of each (chapter 6).

Three Dimensional Protein Structure Prediction

Protein structure prediction methods can be roughly split into three categories (shaded grey in figure 1.2), however it should be noted that the boundaries between fold recognition and *de novo* prediction are becoming increasingly less well defined.

Comparative / Homology Modelling (CM)

Where there is clear sequence similarity, the problem of constructing a 3D model is largely how to substitute the existing side-chains with the new side-chains to which they have been matched (aligned) (Marti-Renom et al., 2000, Guex and Peitsch, 1997, Schwede et al., 2003). Since there is clear overall similarity, many of these will be the same and most will involve only minor substitution of groups (for example, ASP → ASN). In addition, most of the substitutions will occur on the surface of the protein leaving much of the hydrophobic core intact, this is due to evolutionary constraints which mean that it is easier to accept a mutation when the residue is not buried. In this situation, the simple axiom: 'if it ain't broke, don't fix it' is the best advice to follow. Indeed, even better advice is to let it all be done automatically as there are now several programs that can construct good models providing the sequences are clearly related. Many of these are commercial but the Swiss-model (Schwede et al., 2003) program can be used freely over the internet for non-commercial purposes.

Where sequence similarity decreases, the problem of indels (relative insertions and deletions) becomes important as they imply that the protein backbone will need to be remodelled (to close the gaps after deletion, or add new chain for an insertion).

Fortunately, most of these changes tend to be found on the protein surface where there is usually scope to make larger changes. If changes apparently do not occur on the surface, then this is a strong indication that the alignment between sequence and structure may be incorrect. While such problems can be attempted using programs like Swiss-model, the limiting factor becomes being able to specify the correct alignment on which to base the model.

Fold Recognition (FR)

Jones *et al.*, coined the term threading in 1992 (Jones et al., 1992a) to describe a method for predicting 3D structure where sequence similarity drops into the twilight zone (Doolittle, 1986). Since then, threading has been used more widely to describe any fold recognition (FR) method. Following common usage, ''threading'' will be used to refer to all FR techniques.

The strategy of aligning protein sequences using typical alignment methods and then building a model is no longer the standard approach for distantly related sequences . Interaction is needed between the emerging model and the alignment. As mentioned above, if an insertion is found in the core, then the answer is usually to change the alignment - not the structure. Historically, this was carried out in a series of iterations with manual realignment at each stage; as for example, in the construction of the HIV protease model (Pearl and Taylor, 1987). Eventually it became apparent that this progress could become more automated with the alignment and model being calculated simultaneously.

True threading

To thread a sequence over a structure two components are necessary: a packing measure for the substituted amino acid and an alignment method that can optimise the sum of the packing scores. The former is available in the 'rough' empirical potentials of Sippl (Sippl, 1990), referred to as 'potentials of mean force'. These are 'rough' in that they do not directly consider side-chain interactions (to do so would be impossible until the full model was constructed) but capture the preference of an amino acid to be in a particular environment and, indirectly, secondary structure state. The second component, the alignment method, is readily available from the sequence alignment field, but cannot be used directly. One solution is to apply the alignment in a series of iterations, gradually substituting new residues into the existing structure. Another solution is to take the double dynamic programming method developed for structure comparison (Taylor and Orengo, 1989b) and apply it to the threading problem (Jones et al., 1992a).

3D/1D Alignment

The sequence/structure matching problem was also approached from the sequence alignment side. Beginning with a pure sequence alignment, structural features are predicted (such as secondary structure state and degree of burial) which are then matched to features of protein structures along with its sequence (Bowie et al., 1991, Luthy et al., 1991, Rice and Eisenberg, 1997). Unlike the 'true' threading methods described above, this approach does not take account of 3D interactions in the

calculation of the alignment and so can use the dynamic programming algorithm without complication.

In theory, the 3D/1D approach is less powerful than the 'true' threading methods, however, when applied to very distant relationships, there is little perceptible difference in the methods. This probably results from the common incorporation of multiple sequence data into the 3D/1D methods and from accurate prediction of secondary structure. The 'true' threading methods also make the assumption that the basic core structure of the model protein will be the same as the structure on which it was built: for distant relationships this is seldom completely true.

Ab initio / *De novo* / New Fold Modelling (NF)

The last resort – although arguably the most exciting – is *ab initio* prediction. *Ab initio* (Latin: "from the beginning") prediction relies on the assumption that natively folded proteins exist in a state of low free energy. To obtain the structure of a protein one simply has to compute all possible interactions between all residues in a sequence until the lowest free energy conformation is found! In reality this problem is far from trivial – in fact it has only been done for short polypeptides up to 30 residues (Duan and Kollman, 1998). Frustrated with a lack of progress, *ab initio* methods have begun to use structural information, often in the form of fragment packing, allowing for proteins up to 100 residues to be predicted – although no longer from first principles (Rohl and Baker, 2002) (Bradley et al., 2005). In order to retain correct nomenclature – as well as to keep physicists happy – the name of this approach has been changed to *de novo* or new fold modelling.

Model Evaluation

Regardless of the method used to generate the protein structures many thousands of models can be produced by a single prediction attempt. While some models resemble real proteins, with well-formed secondary structure and a hydrophobic core, the majority tend to be poorly formed with little or no secondary structure and 'unnatural' packing (exposed hydrophobics and buried hydrophilics). To build and identify models that resemble real proteins it is crucial to use reliable evaluation functions, such as CAO (Lin et al., 2003) and Phobic (Klose, in preparation) an updated version of burial/hydrophobic matching described in (Taylor et al., 2006), for hydrophobic core evaluation. Scoring functions are diverse but can be categorised as physical or knowledge based.

Physical scoring functions are based around force fields such as CHARMm, which aim to describe the physical interactions that occur within the protein between residues and atoms (Brooks et al., 1983). Such functions include bonded and unbonded energies, dihedral (torsion) angles and Van der Waals terms. As such they are mathematically complex and require all-atom models to be constructed.

Knowledge based scoring functions rely on the identification of characteristics that are common among native protein structures. Additionally, some methods use characteristics that are common to decoy (non-native) structures to differentiate between the two sets (native and non-native). The characteristics are then used to design a function which empirically captures these features from a limited amount of information.

Sequence Alignment – Pairwise and Multiple.

The first step in predicting protein structure is to scour resources looking for information on the target, this typically involves sequence alignments. Historically sequence alignment began with the alignment of two sequences, which is referred to as a 'pairwise alignment' (Needleman and Wunsch, 1970, Smith and Waterman, 1981) only later, with the development of sophisticated algorithms, were methods expanded to deal with rapid alignment of numerous sequences (Altschul et al., 1997, Edgar, 2004, Higgins and Sharp, 1988, Taylor, 1987).

Both methods work by examining sequences for a series of elements or patterns that occur in the same order. By hand, short sequences can be aligned by writing them down in two columns. Characters that match are placed in the same column while dissimilar characters are aligned as a mismatch. In more advanced methods another 'character', a gap, can be introduced. Mismatched positions and gaps are placed so as to maximise the number of identical characters in register, as pioneered by Saul Needleman and Christian Wunsch (Needleman and Wunsch, 1970).

Pairwise sequence alignments can be categorised as either local or global. In global alignment an attempt is made to maximise the register over the entire sequence. Local alignment (Smith and Waterman, 1981) aims to maximise vertical register where sequence similarity is greatest, resulting in the formation of islands of aligned positions. The nature of local alignment makes the technique suitable for the comparison of a small segments to large expanses of sequence – as found in modern genomics.

Machine Learning

Machine learning is a broad sub-field of the artificial intelligence which covers the development of algorithms and techniques that allow computers to 'learn' (Cristianini and Shawe-Taylor, 2000). The methods can be split into one of several types: transductive, where observed, specific training cases are used to solve a specific problem (as opposed to inductive learning); inductive, where general rules and patterns are elucidated from a training set; deductive, where a conclusion is necessitated by previously know premises (Vapnik, 1998, Cristianini and Shawe-Taylor, 2000). The methods in this area include decision trees, *k* nearest neighbour (*k*NN (Wilson, 1972), genetic algorithms (Holland, 1975), artificial neural networks (Minksy and Edmonds, 1954) and support vector machines (Vapnik, 1998). It is not possible, or relevant, to cover all methods here, instead the reader should refer to (Klose and Taylor, 2007) for an overview. The methods that are relevant to this thesis are those based on nearest neighbour and support vector machine learning. Support vector machines play an important role in chapters 2 and 6, while the *k*NN is crucial for chapter 6, the basics of each method is described briefly below, more detail is given in the respective chapters.

The *k* Nearest Neighbour Algorithm

The *k* nearest neighbour method is one of the most simple methods for inferring class to an unknown 'object' given prior knowledge (Wilson, 1972). It is simple in as much as there is only one parameter to optimise – the number of neighbours (*k*) required to optimally infer class. Figure 1.3 shows a basic example using a two-dimensional feature vector to distinguish between squares and circles. Using a set of known

examples, the optimal value of k is determined to be equal to 3. By applying the same measure of distance as used to define k, the three closest known samples are determined, the class of these determines the class of the unknown point x. In this instance x is determined to be a square.

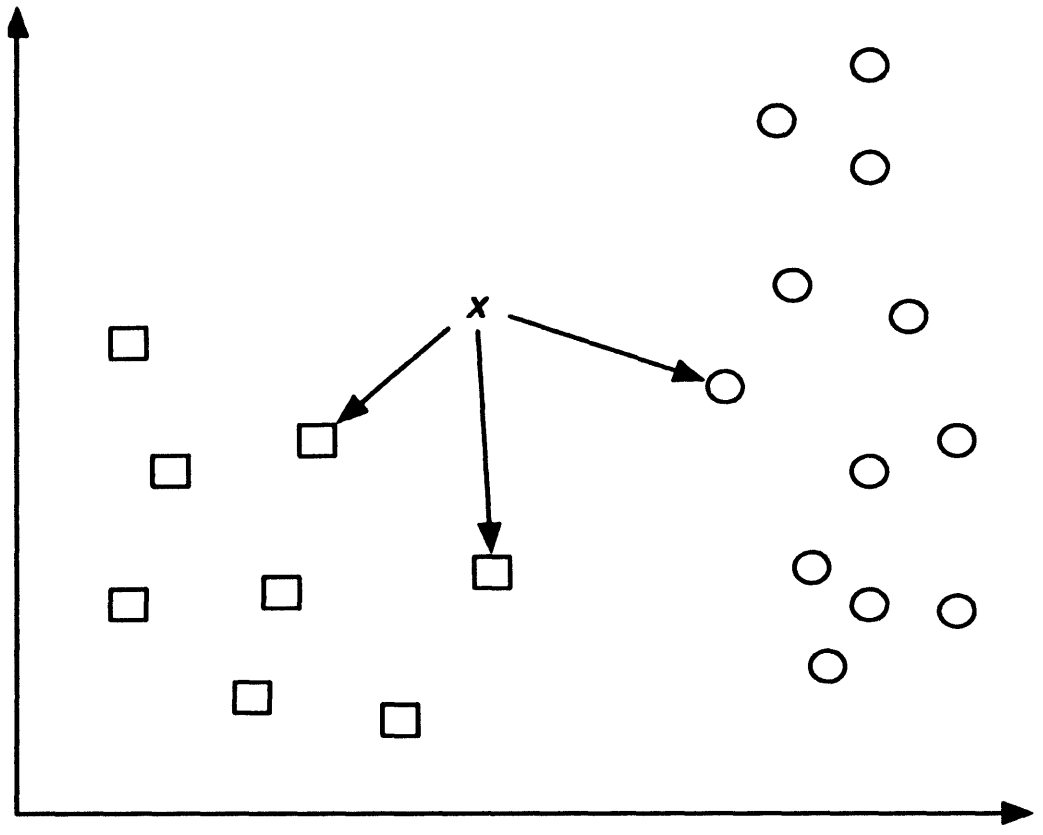


Figure 1.3: An idealised *k* nearest neighbour model: The *k*NN approach is simple, requiring only one parameter, the number of *k* used to infer class, to be identified. With *k* established as three the unknown vector (*x*) is compared to each known example in the dataset. The class inferred on *x* is the most represented of the *k* nearest neighbours, in this example a square.

Although effective in some situations the *k*NN can be improved. In a *k*NN the contribution to class membership is treated equally for each of the *k* in figure 1.3, to solve this problem a 'fuzziness' parameter can be added. This parameter controls the contribution to class membership of each of the *k* neighbours by using a function to weight the class contribution by distance. The result of this is a probability of the unknown sample belonging to the circle and square classes. Although basic, the *k* nearest neighbour based approaches have been used with success in the filed of protein structure prediction (Sim et al., 2005, Bondugula and Xu, 2007).

Support Vector Machines

Support vector machines (SVMs)(Boser et al., 1992, Vapnik and Lerner, 1963, Scholkopf and Smola, 2002) are more complex than the *k*NN-like approaches. SVMs belong to a class of machine learning methods called maximum margin classifiers because they were initially constructed to optimally separate linearly separable data. Figure 1.4 shows how a linearly separable problem could be solved by adopting any one of the possible dividing hyperplanes on which to make classifications. The overall effectiveness of these hyperplanes is unlikely to be optimal when classifying unknown examples.

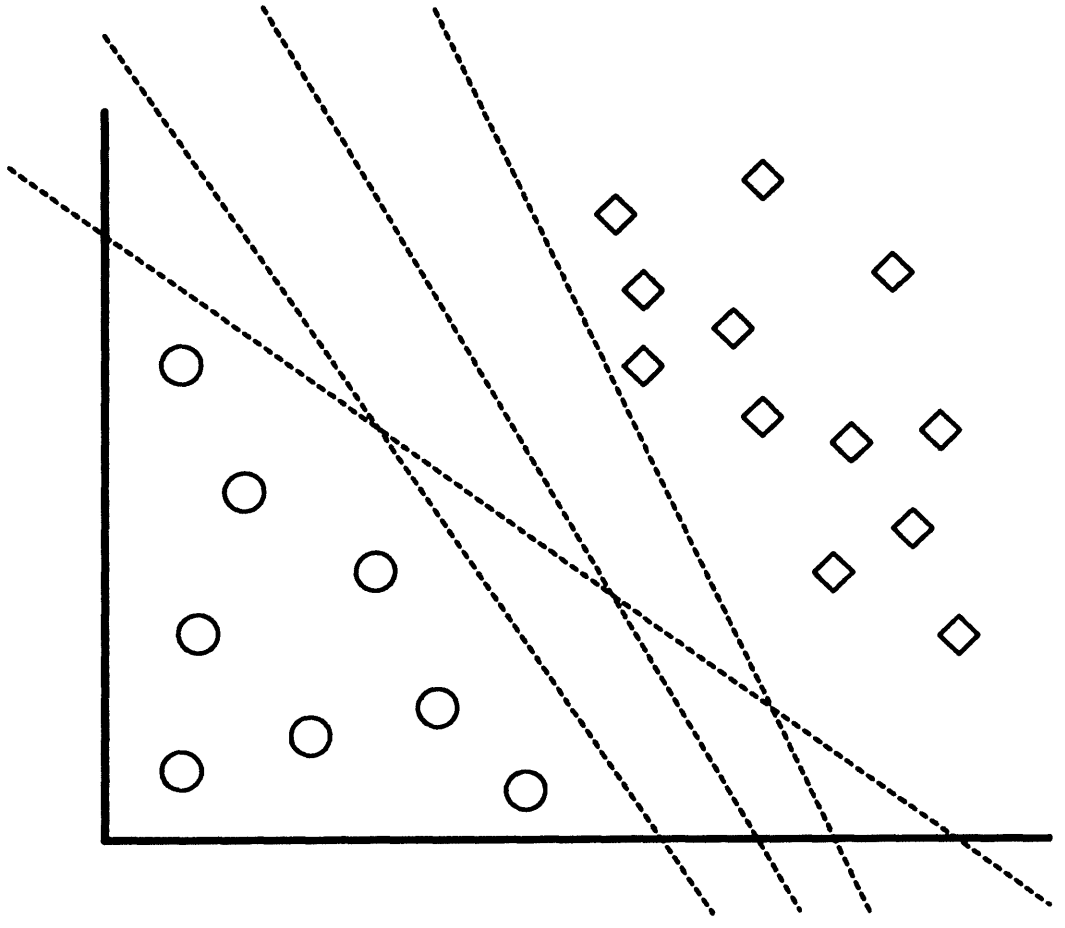


Figure 1.4: A linearly separable problem: The two classes (circles and diamonds) can be divided by a number of lines, some of which are shown. The margin hyperplanes (shown) are non-optimal but would allow for classification on unknown samples. Because the hyperplanes are non-optimal there is scope for misclassification which would be minimised if the hyperplane was placed such that it was an equal distance from both classes.

This is where the support vector machines stand out, by calculating the hyperplane which is equidistant from both classes (maximum margin hyperplane). This is achieved by calculating two additional hyperplanes which define the boundary from each class to the dividing hyperplane as shown in figure 1.5.

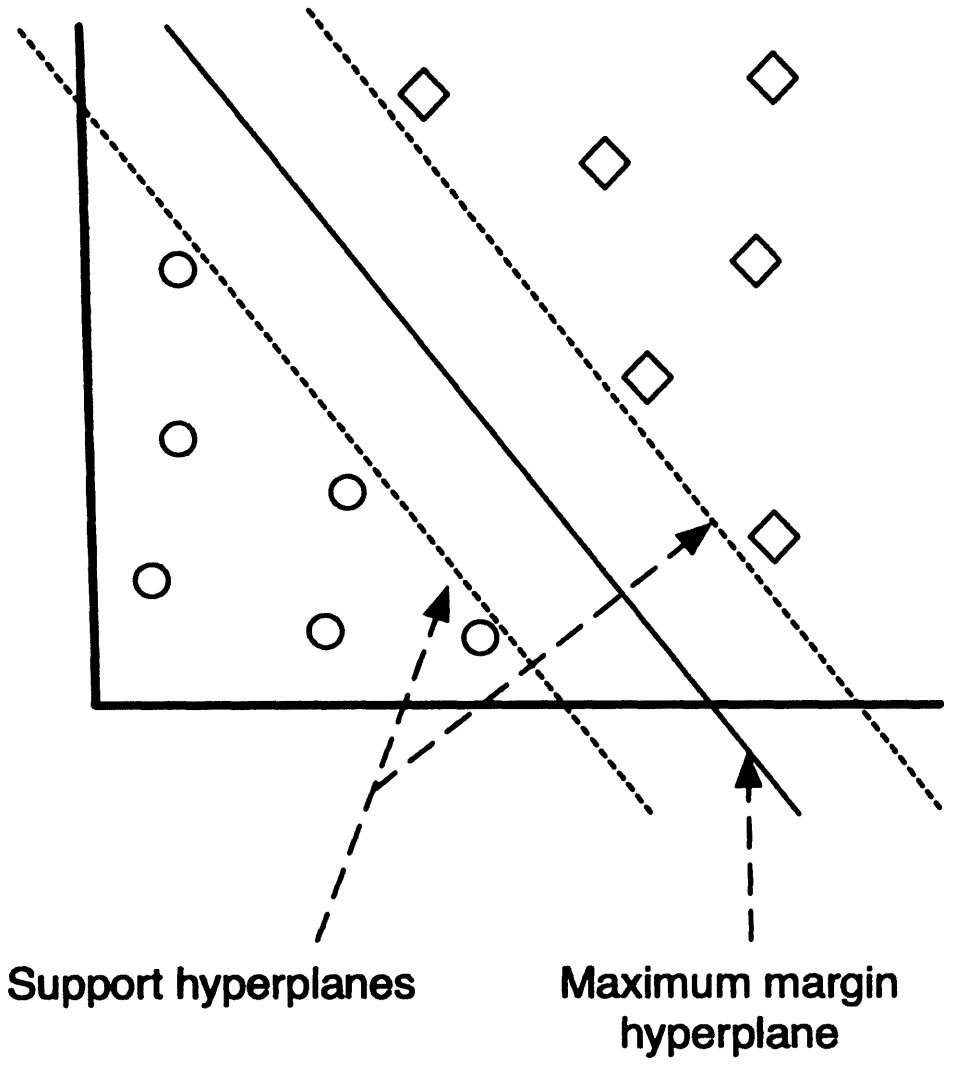


Figure 1.5: The maximum margin hyperplane: To define the maximum margin hyperplane, two further planes, support hyperplanes, have to be identified. The support hyperplanes allow for the definition of the maximum margin hyperplane, which is equidistant from both classes, these hyperplanes are defined by support vectors.

The support hyperplanes are defined by the feature vectors which identify the boundary of each class (Poggio and Girosi, 1990), these vectors are termed the support vectors and form the basis of the model. All training examples which are not identified as support vectors are discarded, resulting in a reduction of the data used to define the model which is later tested against.

So far the focus has been on linearly separable problems. In real-world applications linearly separable problems appear to be in the minority and, as such, a method needs to be able to function on non-linearly separable data. This is achieved through the introduction of slack variables (Smith, 1968, Bennet and Mangasarin, 1992) and the kernel trick (Aronszajn, 1950) (see chapter 2 for mathematical description). Slack variables (ξ) measure the degree of misclassification for each point (figure 1.6), all non-zero slack variables are then penalised such that the definition of the model becomes a trade off between the margin size and the error penalty.

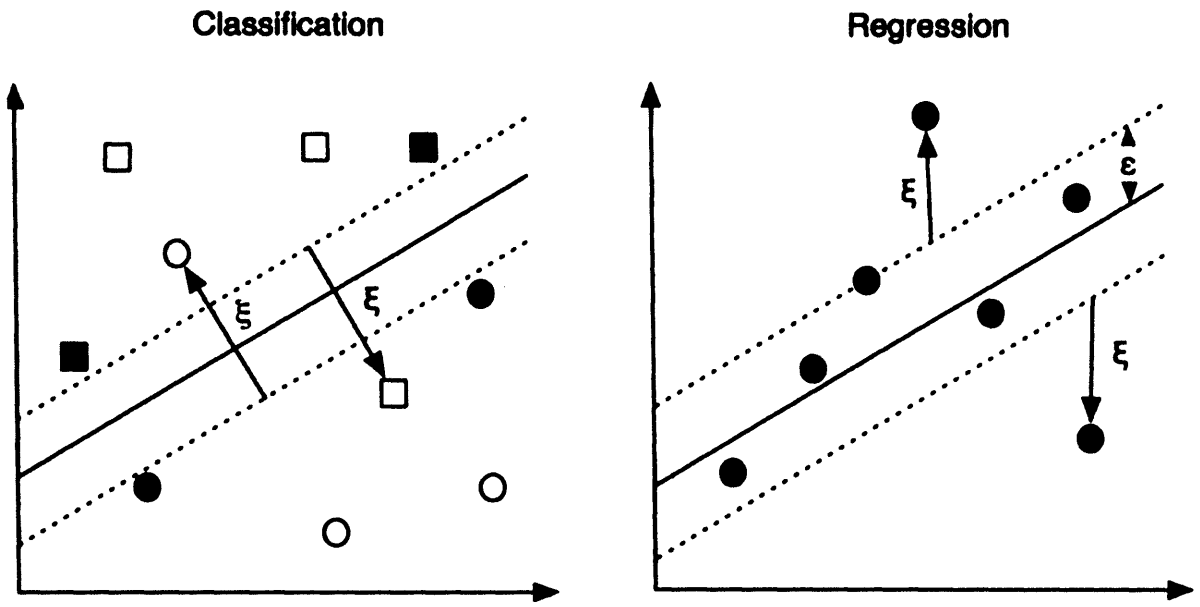


Figure 1.6: Slack variables in Classification and Regression: The left hand side of the diagram shows the use of slack variables in a linearly separable classification problem, the shaded squares and circles represent the support vectors. The right hand side of the diagram shows the use of slack variables in a linear regression problem. The mathematical explanation is covered in chapters 2 and 6 for regression and classification respectively.

The kernel trick (Aizerman et al., 1964) is the feature which allows for classification and regression solutions to be found for non-linearly separable problems. In the original solution for linear problems, proposed by Vapnik, dot products were used; the kernel trick replaces each dot product with a kernel function (see chapters 2 & 6 for details). The function allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space, such transformations can be non-linear and the transformed space can be multidimensional, in fact if a Gaussian radial basis function is applied, the feature space is a Hilbert space[3] of infinite dimensions. There are numerous kernels which can be applied and many of the existing architectures, such as Libsvm (Chih-Chung and Chih-Jen, 2001) and SVM^{light} (Joachims, 1999), include upwards of four methods as well as the option for a user defined functions. Typical kernel functions include the polynomial, sigmoid and radial basis functions as shown in table 1.1.

Table 1.1: Standard Kernel Options

Kernel Function	Formula
Polynomial	$(x \cdot x')^d$
Radial Basis	$\exp\left(-\gamma\|x - x'\|^2\right)$ for γ greater than 0
Gaussian Radial Basis	$\exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$
Sigmoid	$\tanh(kx \cdot x' + c)$ for values where k>0 & c < 0.

So far, only classification problems have been addressed, however SVMs are also applicable for regression problems. As illustrated above, classification relies on a set of training data, discarding all points that do not aid in the identification of the support hyperplanes. Support vector regression is analogous with the exception that the model

[3] a mathematical concept which generalises the notion of Euclidean space in a way that extends methods of vector algebra from 2D/3D space to infinite-dimensional spaces, allowing for distances and angles to be measured.

'ignores' data points that are within a threshold ε (diagram 1.6) of the margin. These concepts are important for the work described in chapters 2 and 6.

Aims

The aim of this work was to contribute to the field of protein structure prediction by utilising existing tools in a novel fashion to predict and asses two and three dimensional protein structure. The work presented here describes the construction of two model evaluation functions; one based on the prediction of contact number from idealised sequence alignments; the second based on a feature referred to as the 'empty quarter' which attempts to exploit the theory of protein folding through hydrophobic collapse. Two novel methods for the prediction of three dimensional structures are then described and evaluated; the first method, based on existing fold recognition methodology, aims to provide a solution to the problem of domain definition through the application of Ising-like models as described by Taylor (Taylor, 1999a). The second method describes an approach for the *De novo* prediction of large (greater than 100 amino acid) α/β proteins that share a simple α/β sandwich architecture. This method, also designed on existing methods, uses Taylor's periodic table as a start point for the identification of a suitable architecture on which 'threading' templates are based. The final section describes the novel combination and application of f*k*NN and SVM to predict secondary structure and solvent accessibility to improve the performance of both the 3D modelling procedures as well as two dimensional structure prediction.

Chapter 2

Assessment of Sequence Structure Alignments and the Hydrophobic Quarter

Introduction

One of the first reports of a sequence being solved was that of the phenylalanyl chain of insulin by Sanger and Tuppy in 1951 (Sanger and Tuppy, 1951a, Sanger and Tuppy, 1951b). The development of their method led to the sequences of several proteins, representative of protein families, being elucidated. In the early 1960s Margaret Dayhoff and her colleagues began to assemble the first sequence databases which eventually became the Protein Information Resource (PIR) (George et al., 1986). The development of this resource resulted in construction of the Dayhoff substitution matrices which play a role in protein sequence alignment today.

Deoxyribonucleic acid (DNA) sequences followed later as a result of work by Sanger (Sanger and Coulson, 1975, Sanger et al., 1977) and Maxam & Gilbert (Maxam and Gilbert, 1977) which eventually resulted in a Noble prize in chemistry for Sanger and Gilbert in 1980. As more DNA and protein sequences became available so to did demand for computer algorithms to analyse them. Gibbs and McIntyre (Gibbs and McIntyre, 1970) had already described a method for comparing two amino acid sequences which, despite its simplicity, remains in use today. The method requires one sequence to be placed along the x-axis the other along the y-axis, at every position along each sequence where two positions match a dot is placed. This graph is scanned, by eye, for diagonal lines which reveal sequence similarities, insertions and deletions. Despite these features the so called DOT-plot does not lend itself to the automatic identification of regions that are similar but interrupted by regions of low sequence similarity. This problem was largely solved by Needleman and Wunsch (Needleman and Wunsch, 1970) and redefined by Smith and Waterman (Smith and Waterman,

1981). The techniques, although designed for different purposes, global and local alignment respectively, relied upon the same approach, called dynamic programming. In turn these methods formed the basis for multiple sequence alignment, which today is fundamental to biological research from the creation of PCR primers to phylogenetic analysis (Higgins and Sharp, 1988, Taylor et al., 1994).

One of the most widely used alignment tools is PSI-BLAST (Altschul et al., 1997). Probably an artefact of its age, PSI-BLAST forms the base of a number of structure prediction tools through the creation of position specific scoring matrices (PSSMs) (Gribskov et al., 1987, Staden, 1988). A PSSM is a common method for representation of biological sequences, more specifically it is a matrix of scores that gives a weighted match to any given substring of fixed length (N). Each row represents a symbol in the starting protein sequence, while columns are used for observations of constituents of the full alphabet (in the case of proteins, 20 amino acids). The log odds values in the columns are used to evaluate matches with target sequences, for each column the log odds scores are summed to obtain a new log odds score for the alignment to that sequence position, the higher the logs odds score the more significant the match.

Multiple sequence alignments and PSSMs play a crucial role in the prediction of protein structure, from the estimation of solvent accessibility to the identification of templates used to predict 3D structure in comparative modelling, as such, the axiom "rubbish in, rubbish out" holds true – meaning that if the alignment is poor then any predictions that depend on it will also be poor – it is, therefore, beneficial to have a way of identifying if one alignment is better than another for predicting 2D or 3D structure.

In this chapter I will present an unsuccessful approach to identify ideal sequence alignments for protein structure prediction. The first part of this work led to the identification of an interesting feature which will be referred to as the 'empty quarter' and the development of a structure fragment library used in the development of a novel protein refinement method (Jonassen et al., 2006) and reconstruction of the SPREK scoring function library (Taylor and Jonassen, 2004). The second section describes the use of alignments to assess the effect of sequence diversity on the prediction of protein contact number using support vector regression. This approach was taken as an alternative means of examining the effect of ideal sequence alignments.

Methods

Evaluation of Sequence Alignments part 1: Can Ideal Alignments be identified?

In the initial phase of this work the approach was to examine hydrophobic positions in multiple sequence alignments. In an 'ideal' protein, hydrophobic positions would be buried within the hydrophobic core. These positions play a crucial role in stabilising the protein structure (Taylor, 1986), as such it is expected that they are well conserved – i.e. not so prone to mutation as exposed hydrophilic residues. By creating and examining sequence-structure alignments, positions of conserved hydrophobicity can be identified. Making the assumption that good alignments should preserve a number of crucial core blocks, a search was conducted as described below.

Generating Sequence Structure Alignments.

Sequence-structure alignments were extracted from the following high quality alignment databases: the Structural Alignment Database (SAD) (Marsden and Abagyan, 2004); HOMologous STRucture Alignment Database (HOMSTRAD) (Stebbings and Mizuguchi, 2004); Families of Structurally Similar Proteins (FSSP) (Holm and Sander, 1994). The sequence structure alignments, where necessary were augmented using alignments from the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2002) and the Benchmark Alignment dataBase (BAliBase) (Thompson et al., 1999). One of the criteria used in the construction of SAD was that alignments must sample a number of sequences to be structurally statistically significant. A similar criteria was applied in this work, with any alignment containing less than four sequences aligned to the probe being discarded.

An additional set of sequence alignments were generated from the SCOP dataset. Using the ASTRAL database a set of protein domains sharing a maximum pairwise identify of 40% were selected from the PDB. Alignments for each of these domains was generated using a method based on the combination of MULTAL and MULSEL (Higgins and Taylor, 2000), described below.

MULTAL is designed to deal with a large number of sequences that are typical of family analysis or database wide sequence searching. It uses single-linked clustering over a number of user-defined cycles to filter sequences. At the start of each cycle only sequences that have a pairwise identity above the cycle threshold are kept. Calculation of similarity is controlled by three parameters: span, window and peptide pre-sort.

Rather than optimising the positions of sequences within the alignment, for example ranking sequences from left to right based on sequence identity, MULTAL considers pairwise identity over a number of adjacent sequences – the number of sequences to be considered is called the span, by default the span starts small and expands from one cycle to the next. As cycles progress the number of sequences (including sub-alignments) decreases relative to the span so that by the final cycles the number of sub alignments and sequences (called 'blocks') is less than the span resulting in an all against all comparison. The window parameter controls the width of the path the alignment back-tracks through the scoring matrix and is increased in each cycle. The final parameter, the peptide sort, is a method for sorting the initial starting sequences. It uses a dynamic radix tree to store and compare sets of strings (tripeptides) which are used to define a rough order of starting sequences.

To control alignment quality there are two adjustable parameters, the substitution matrices and the gap penalty. Two matrices are used by MULTAL, an identity matrix (where amino acid identities score 10, 0 otherwise) and the PAM matrix, however these are not exclusive and can be replaced by BLOSUM, PAM250 or the JTT matrices. The gap penalty for MULTAL is very soft, in as much as there is a single gap penalty which is paid only once when gap is opened. The rational is that locations at which insertions can occur in the protein are generally on the surface and that if a small insertion can be made then there are probably few constraints on the formation of a linker between domains to an even larger insertion. As with all other MULTAL parameters the gap penalty can be changed but was kept in the range of 20-30 over the entire run. Establishing where to stop is a problem for all alignment methods so in this work MULTAL was run with default parameters.

Measuring Conserved Hydrophobicity

In 1997 Taylor proposed a scheme to 'illuminate' multiple sequence alignments with colour (Taylor, 1997b) (see figure 2.1). The scheme forms the foundation of a measure of conserved hydrophobicity which was implemented as follows. The original scheme relied on each amino acid being assigned a pure (spectral) colour and that as a collection they could be prescribed a cyclic ordering. The order of the amino acids is dominated by two crucial chemical and physical properties – hydrophobicity and size both playing a critical role in protein stability (Taylor, 1986). As in the original method, cysteine is yellow, the negatively charged acidic residues red, positively charged basic groups are blue. The inclusion of yellow as a primary colour allows for four equidistant points to be placed around a circle. Using these four points the remaining amino acids are placed at equal points around the circle. Hydrophobic residues are green, aromatics are green-blue, amino acids commonly found in loops are red-orange and large polar amino acids are purple-blue as shown in figure 2.1.

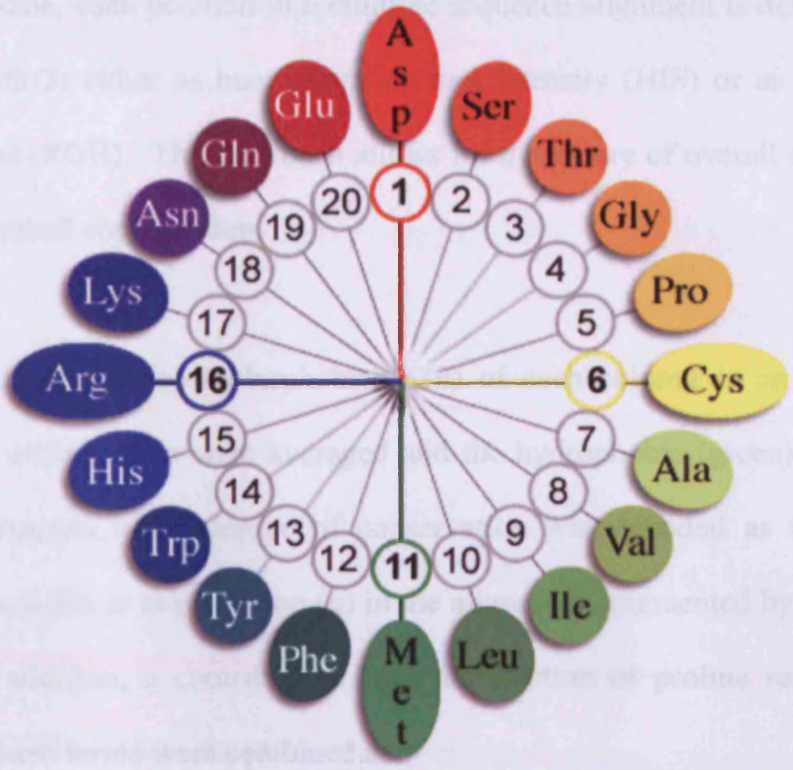


Figure 2.1: Colours of the Amino Acids: Each amino acid is described by a three element vector: either as hue, saturation and intensity (HSI) or as a mix of red, green and blue (RGB). For simplicity yellow is included as a primary colour allowing for equidistant points to be placed on the circle (Asp, Cys, Met, Arg). These four residues are then used to place the remaining amino acids around the circle at equal distances.

Using this scheme, each position in a multiple sequence alignment is described using a vector of length 3: either as hue, saturation and intensity (HIS) or as a mix of Red, Green and Blue (RGB). This approach allows for a measure of overall conservation as well as biochemical conservation.

To establish the conserved hydrophobicity (h) of each column in an alignment the colours of the amino acids were averaged and the hydrophobic (green) component of the vector extracted. The degree of conservation was encoded as the number of different amino acids at any position (a) in the alignment, augmented by the fraction of gaps (g). In addition, a contribution from the fraction of proline residues (p) was introduced. These terms were combined as:

$$c = (h+1)\cdot\varsigma(a,A)\cdot\varsigma(g,G)\cdot\varsigma(p,P)-1, \qquad (2.1)$$

where the function $\varsigma(x,y)$ is the Gaussian transform: $\exp(-2x^2 10^{-y})$, A = 2, G = 1 and P = 0, the result is that all values of c fall into the range -1 ... +1.

Calculating Solvent Accessibility from Structure

To evaluate the solvent accessibility (SA) of models that contain only Cα atoms it was not possible to use DSSP (Kabsch and Sander, 1983) or NACCESS (NACS) (Hubbard, 1993) as they rely on full atomic structures. Instead a heuristic program, *P*arameter *OP*timised *S*urfaces (POPS) (Cavallo et al., 2003) was used, the version of POPS used in this work was a C++ reimplementation of POPS-Residue (POPS-R). The POPS

algorithm is based on the technique proposed by Still et al., (Still, 1990) and the probabilistic method of Wodak and Janin (Wodak, 1980).

The Solvent Accessible Surface Area (SASA) of the protein (A) is defined as:

$$SASA(A) = \sum_{i=1}^{n} a_i \qquad (2.2)$$

where the protein A has n residues.

The Wodak and Janin formula is defined as:

$$A_i(r_n) = S_i \prod_{i=1}^{n} 1 - \frac{p_i p_j b_{ij}(r_{ij})}{S_i}, \qquad (2.3)$$

where $S_i = 4\pi(R_i + R_{solv})^2$ which is the SASA of the i^{th} atom (or residue in POPS-R) with radius R_i and a solvent probe with a radius R_{solv}. b_{ij} is the SASA of S_i covered by the overlap of atoms i and j at a set distance $r_{ij} = |r_i - r_j|$. If $r_{ij} > R_i + R_j + R_{solv}$, $b_{ij}(r_{ij}) = 0$, otherwise $b_{ij}(r_{ij}) = \pi(R_i + R_{solv})(R_i + R_j + 2R_{solv} - r_{ij})[1 + (R_j - R_i)r_{ij}^{-1}]$. In order to work, residue level spheres were centred over the C_α. p_i depends upon the atom type in the original definition while p_{ij} serves as an additional reduction factor that distinguishes between the first and next neighbours. These factors were optimized by Hasel et al., (Hasel, 1988). In order to evaluate the packing of amino acids within the structure the overall SASA is ignored in favour of the residue level SASA.

The residue exposure levels were mapped into the range -1 ... +1 (exposed ... buried) using the Gaussian transform:

$$e = 2\exp(-ca^2) - 1 \qquad (2.4)$$

where a is the surface area estimated by POPS and c (the inverse variance) has the value 0.0003, which was found over a number of native structures to give an approximate mean of zero for the transformed value e.

The use of a measure of conserved hydrophobicity as a measure of sequence alignment quality is a defensible approach to the problem, based of the concept that protein folding is driven by the aversion for water on the nonpolar residues (Dill, 1990). Despite the role that the hydrophobic effect plays in protein folding, the observations made did not yield a measure which could be used to definitively assign one alignment 'better' for prediction or modelling purposes than any other. While some interesting observations were made, as will be described in the results, the decision was reached to change the focus of the investigation to a feature which can be observed, measured and controlled, namely sequence identity.

Evaluation of Sequence Alignments part 2: Testing the effect of sequence identity on Contact Prediction.

In the second phase of this work the original question 'can ideal alignments be identified?' was rephrased. The objective became to examine if altering the level of sequence similarity across multiple sequence alignments, less error prone predictions of

contact number could be made for application in model ranking and fold recognition. To achieve this, the Representative Protein Databank (Noguchi et al., 1997) was used to select structures from the PDB which matched the following criteria: minimum resolution of 3Å by X-ray crystallography only; R-factor less than 0.3; no chains breaks or non-standard residues and a minimum length of 60 residues. In addition to these features, all membrane proteins, mutant and complex structures were discarded. Of the initial 914 proteins only 815 passed the criteria. From this set of structures only those which were identified as α/β proteins were selected, leaving 172 proteins. The choice of α/β proteins was based upon ongoing work in the laboratory which is described in chapter 5 and (Taylor et al., 2008). Sequence information was extracted from the ATOM records. The advantage of a small dataset is the speed at which the support vector machines can be trained and tested.

Generating 'Ideal' Sequence Alignments.

Where sequences with high similarity and structure can be found, the comparative modelling approach should be applied, however this approach is not possible for every target. For such cases it becomes necessary to create an alignment which includes information from close to distance homologs giving some insight into the evolution of the target. In this work, alignments were generated such that a specific range of sequence identities were sampled – these alignments are referred to as 'ideal alignments'. The sequence alignments were generated using the procedure outlined in figure 2.2. The method used is an extension of the MULTAL-MULSEL method, described above, to generate a series of sequence alignments which conform to specific

sequence identity cut-offs using a combination of the MULTAL-MULSEL scores (figure 2.3).

Initial multiple sequence alignments were generated using a standard PSI-BLAST search performed against a filtered copy of the non-redundant database (nr). The nr database was pre-filtered to remove all low complexity and coiled-coil regions using Pfilt, this was done to avoid spurious hits. Instead of creating a position specific scoring matrix (PSSM), PSI-BLAST was run with the –m 6 switch to return the default PSI-BLAST multiple sequence alignment.

Sequences from the PSI-BLAST search were used to form a local 'hits' list which was used as input to the MULTAL-MULSEL pipeline. Prior to execution of the pipeline, each sequence in the PSI-BLAST list was extended by 10 residues at both the N and C terminal to aid in realignment during the later stages of the pipeline. If more than 1000 sequences were identified during the initial PSI-BLAST run, then sequences were discarded at random until there were only 1000 left in the set.

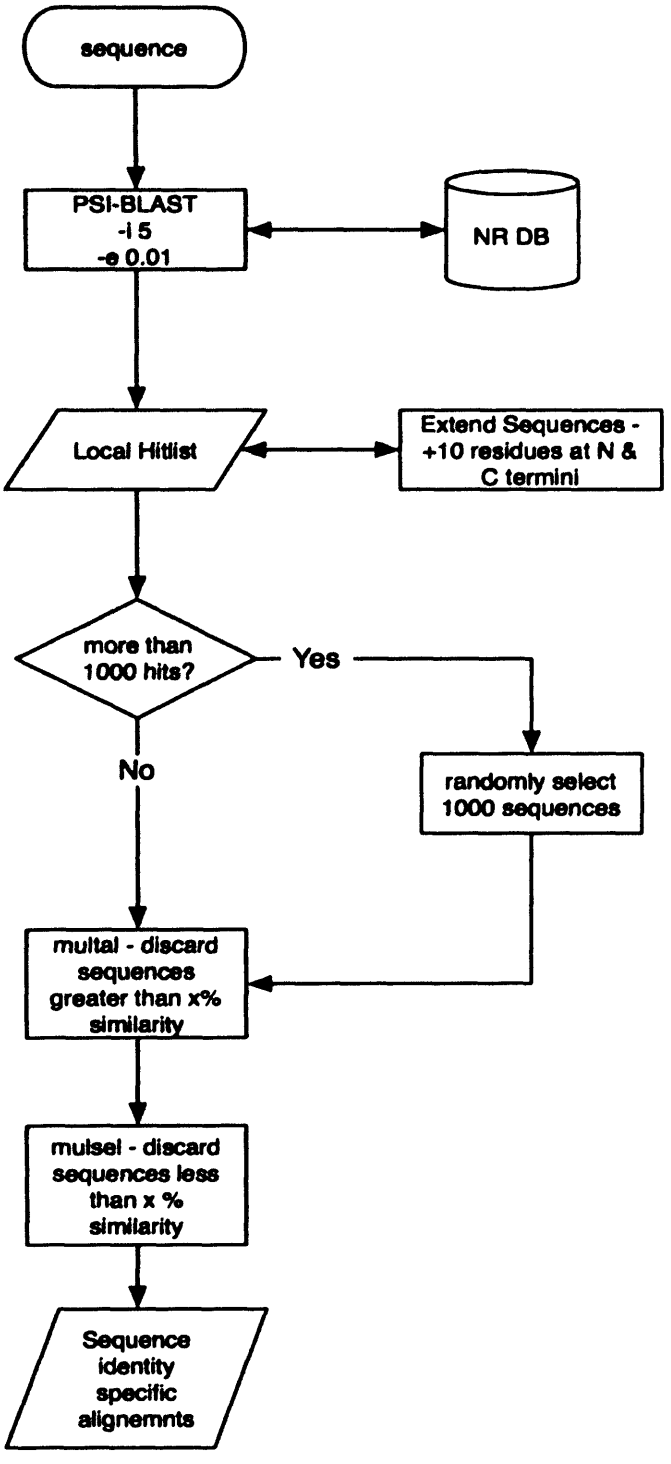


Figure 2.2: Generation of Similarity Specific Sequence Alignments: The input sequence is scanned against the non-redundant database (nrdb) using five iterations of PSI-BLAST and a sequence threshold of 0.01. The output of the scan is parsed into a list of hits – sequence identity numbers and associated amino acid sequences. Each sequence is extended at the N and C termini to help reduce clustering errors that could potentially be introduced in later steps. If there are more than 1000 hits returned then 1000 proteins are selected at random from the hit list. The final MULTAL and MULSEL steps filter the remaining sequences on identity. Two brackets are defined at the start of the process, one for low and the other for upper tolerance of sequence identity. Sequences that fall outside of these brackets are discarded.

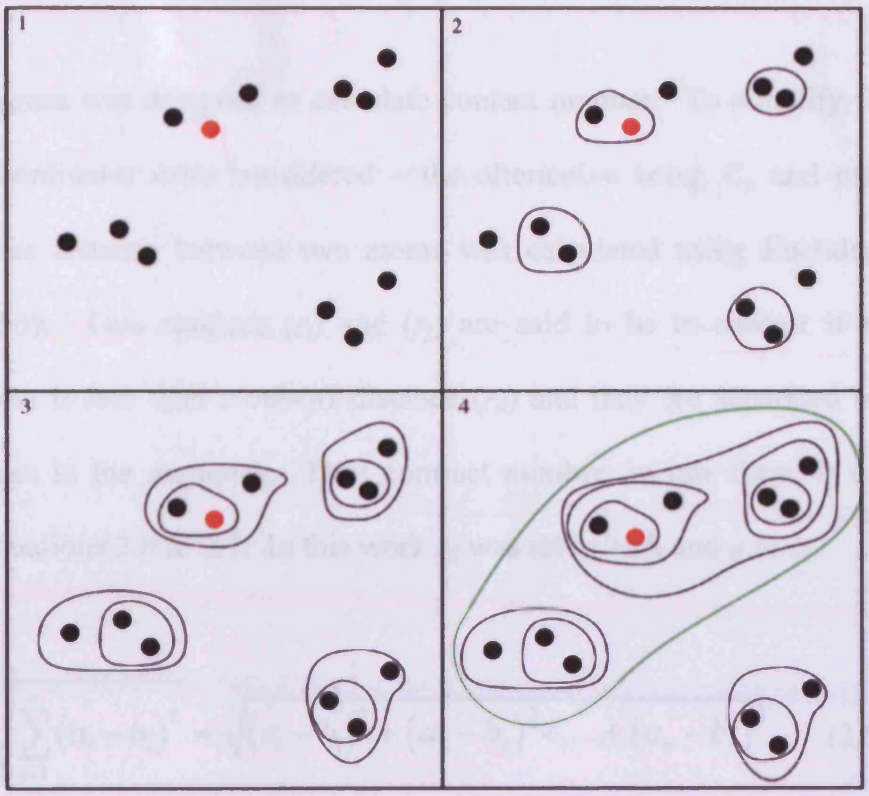


Figure 2.3 Clustering Similar Sequences Using MULTAL & MULSEL: 1. The red circle represents the seed sequence. Each of the black circles represents a single sequence identified by the initial PSI-BLAST search. Through stages 2-4 sequences are clustered by their MULTAL-MULSEL scores. Clustering leads to the identification of 4 clusters, 3 of which fall below a predefined MULTAL limit. From each of these clusters a representative sequence is chosen (rather than a consensus sequence), if a sequence has an associated structure in the PDB it is short listed as the representative of the family. In this example, after the MULTAL and MULSEL steps three sequences are left that fit the predefined criteria.

Calculation of Contact Number.

A C++ program was designed to calculate contact number. To simplify the problem, only Cα coordinates were considered – the alternative being C_β and pseudo-C_β for glycine. The distance between two atoms was calculated using Euclidean Distance (equation 2.5). Two residues (r_i) and (r_j) are said to be in contact if the distance between them is less than a cut-off distance (r_d) and they are separated by at least q other residues in the sequence. Thus contract number, in raw form, is calculated as shown in equations 2.6 & 2.7. In this work r_d was set to 12Å and q to 3.

$$\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ,\ldots, + (a_n - b_n)^2} \qquad (2.5)$$

where a and b are vectors of length n.

$$Cn_i = \sum_{j:|j-i|>q} f(r_{ij}) \qquad (2.6)$$

Where Cn_i is the contact number of the i^{th} residue and r_{ij} is defined as:

$$1 \Big/ \left(1 + \exp[\omega(r_{ij} - rd)]\right) \qquad (2.7)$$

To smooth the contact number equation 2.7 is applied. This sigmoid function blurs the cut-off boundary such that the contact numbers are continuous rather than discrete (Kinjo et al., 2005).

Prediction of Contact Number using Support Vector Regression

To predict contact number, a function is required to learn a relationship between an input, the information contained in the multiple sequence alignments, and the output, the number of contacts a particular residue has. There were several methods which could have been applied to this problem however, due to the small size of the dataset, support vector regression was opted for.

The dataset is defined as (X, Y) where $X = (x_1, x_2, \ldots, x_i)$ and $Y = (y_1, y_2, \ldots, y_i)$ where x_i is an n-dimensional vector of length l and y_i is the associated label. The aim of epsilon (ε) insensitive SVR (ε-SVR) is to map a set of features to an output space.

$$f : x \rightarrow y \text{ is defined by } f(x_i) := \langle \omega, \Phi(x_i) \rangle + b \qquad (2.9)$$

where ω is the weight vector and b is the bias. The function $\langle x, y \rangle$ is the inner product of the weight vector ω and $\Phi(x_i)$. Where $\Phi(x_i)$ is a non-linear function that maps a data point from the inner dimension space to 'feature space' allowing for non-linear separation to be performed (a kernel function).

In order to obtain the best solution to the regression problem, the values of ω and b are found using an optimisation criteria – in this case the pr_LOQO method (Smola, 1997). The aim of this step is to minimise the following function:

$$\frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \qquad (2.10)$$

which is subject to:

$$f(x_i) - y_i \leq \varepsilon + \xi_i$$
$$y_i - f(x_i) \leq \varepsilon + \xi_i^* \qquad (2.11)$$
$$\xi_i, \xi_i^* \geq 0 \text{ for } i = 1 \ldots l$$

The parameter c controls the trade off between the slack variables (ξ_i, ξ_i^*) and the margin. c is also referred to as the softness/hardness parameter as increasing it makes the margin greater. The parameters ξ_i, ξ_i^* measure the deviation of x_i outside of the ε-insensitive tube, the solution is then given by:

$$f(x) = \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) \langle \Phi(x_i), \Phi(x) \rangle + b \qquad (2.12)$$

The two α parameters are Lagrangian multipliers and those that assume non-zero values are the support vectors, those with zero values are discarded. The inner product $\langle \Phi(x_i), \Phi(x) \rangle$ can then be replaced by a kernel function such that:

$$\langle \Phi(x_i), \Phi(x) \rangle = K(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \qquad (2.13)$$

The third term in equation 2.13 is called the Radial Basis Function (RBF). The gamma (γ) parameter defines the width of the Gaussian and was optimised using five fold cross validation. SVMs are not limited to the RBF kernel, indeed there are several other

options including polynomial and sigmoid functions, however the RBF kernel has been shown to deal well with linear and non-linear problems.

For this work, the input vectors encoded the local environment using a window of seven residues either side of the residue being predicted. Each position in the window was represented by a 21 element vector, the first twenty elements corresponding to the transition values extracted from the PSSMs and the 21st element representing observations outside the target sequence. Where observations were made outside the sequence (at the N & C-terminals) the first twenty elements were set to zero and the twenty-first element to 0.5. The final input vectors had 315 dimensions. Three fold cross validation was completed using 100 randomly selected proteins. The remaining 72 proteins comprised the hold-out test set.

Evaluation of Protein Models using predicted contacts

To assess the potential application of the predictions, in the later stages of development the predicted contact numbers were used to create model evaluation functions, the output of which is shown in figures 2.5-2.10 and described further in the results and discussion. The first function used the difference between the predicted contact number and the observed contact number calculated from models generated using a method similar to that described in (Taylor et al., 2008). The second approach applied a Bayesian rule outlined below:

$$fitness = \sum_{i=1}^{l} P\left(C^{i}_{native} \middle| C^{i}_{prediction}\right) \qquad (2.8)$$

Equation 2.8 translates as: the probability of residue *i* (in the native structure) having a contact number C_{native} given that the predicted contact number of *i* is $C_{prediction}$. The concept was that a Bayesian rule would apportion different penalties to under and over prediction of contact number resulting in better discriminatory power. Using the training dataset a lookup table was generated, allowing for probability assignments to be made (the probability of a residue having 3 contacts when 1 .. *n* contacts are predicted *etc*). When a new target is attempted, either from the hold out test set or new target, the fitness is the sum of the probabilities, theoretically the greater the score the more native-like the protein should be. The problem experienced here was that the distribution of predicted and actual contact numbers were marked, i.e. the bins were quite distinct, while this sounds good, in fact on such a small dataset it means that sampling was not sufficient and that there is no generalisation. Several smoothing functions were used to overcome this limitation (caused by the small dataset) however none produced satisfactory results and so a simple difference between the predicted contact number and that seen in models was adopted.

Results and Discussion

The aim of this work was to find one or more characteristics which could be used to identify multiple sequence alignments as good or bad for structure prediction. This started as an examination of the conserved hydrophobic positions across sequence/structure alignments. The alignments were extracted from several existing datasets as well as the construction of a custom set of structures from the PDB40 and ASTRAL database. The starting hypothesis was that hydrophobic positions, which are conserved across alignments play a crucial role in defining structures – being situated in the core and conserved suggest, from an evolutionary perspective, that this is the case. A simple count of the number of conserved hydrophobic positions did not yield any pattern. Additionally it becomes very difficult to examine the alignment quality without something to compare to – random sequences do not fill this role. The advantage of using the conserved hydrophobic measure was that is could easily be compared to solvent accessibility derived from structural information. This comparison showed an interesting pattern, as shown in figure 2.4, which was seen across all structure alignment datasets listed in the methods section.

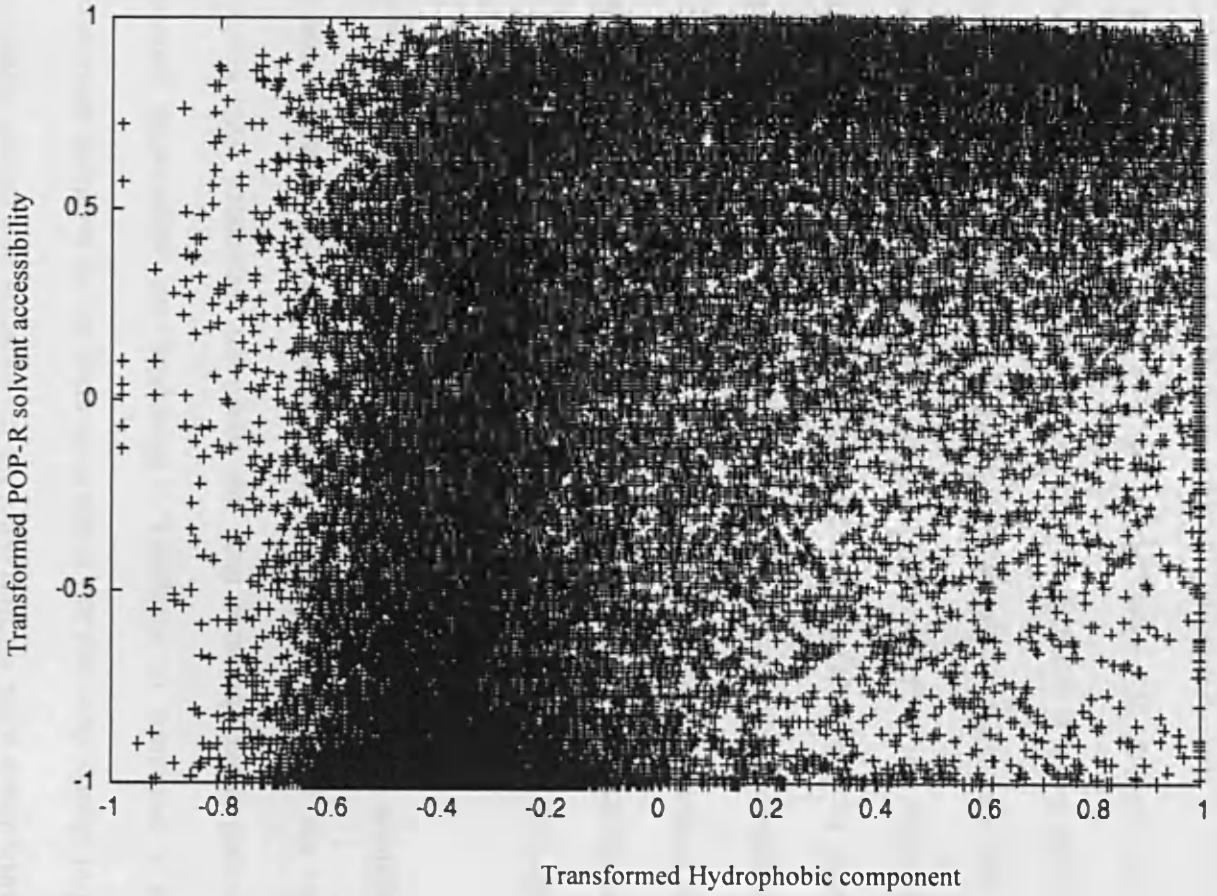


Figure 2.4: The Empty Quarter: The x-axis is the hydrophobic component extracted from the Taylor colour scheme, -1 corresponds to totally exposed while +1 is totally buried. The y-axis is the Gaussian transformed POPS-R value where -1 represents total exposure and +1 total burial. Each data point represent a single amino acid from a structure within the data. The bottom right hand corner is sparsely populated compared to the rest of the plot and as such was coined the 'Empty Quarter'. This pattern is visible in all the datasets mentioned in the method section (data not shown).

The pattern shown in figure 2.4 was coined the 'Empty Quarter' and shows that, in native structures, it is rare for conserved hydrophobic positions identified in the alignments to correlate with exposed residues in native structures. By generating a set of non-native/decoy structures is was clear that this property was not well conserved - the empty quarter becomes heavily populated. The combination of the structure and sequence provides a solid starting point for the creation of a protein structure evaluation function, but clearly does not aid in the evaluation of the multiple sequence alignments, given that a structure is necessary. However, the alignments generated for the aforementioned work, because they are based on sequences with known structures, allowed for the generation of a protein structure library for use in the SPREK and Furball fragment tessellation programs (Taylor et al., 2006, Jonassen et al., 2006).

Because the first phase of this work did not yield a result directly relevant to the initial line of investigation an alternative approach was adopted. The new approach was to explore the effect of varying sequence identity on the prediction of contact number to see if a specific range of sequence variation was better for prediction than a 'default' MSA. Rather than attempting to reverse engineer the problem, the focus is changed to examine the effect of sequence identity on the prediction of a particular protein feature. While solvent accessibility can be predicted from multiple sequence alignments, it has been suggested that it is not a feature which is well conserved across a family sequence alignment (Przybylski and Rost, 2002). Instead, contact number, a value that is well correlated with solvent accessibility, was calculated as it has been suggested that it is better conserved across family alignments (Hamelryck, 2005).

Prediction of contact number, like solvent accessibility, can be approached as either a classification or regression problem. As a predicted feature, contact number is a relatively new area but has been explored, most notably, by Akira Kinjo who used linear regression to predict contact number to a similar degree of accuracy as solvent accessibility (Kinjo et al., 2005). Considering the previous work and that contact number is supposed to be more conserved across a familial alignment it seemed a good alternative feature to predict, in addition accurate prediction of contact number has potential applications in threading and model evaluation.

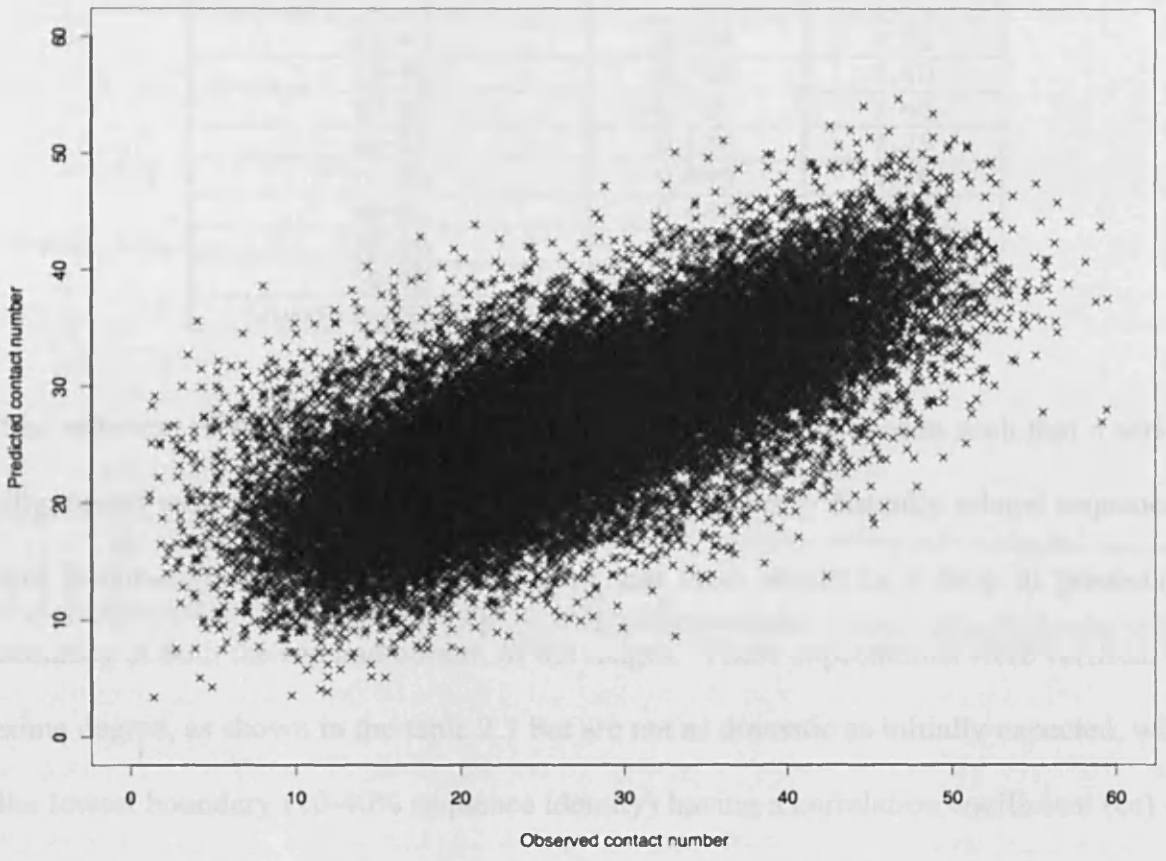


Figure 2.5 Alpha/Beta protein contact predictions: The observed Cα contact numbers run along the x-axis. The predicted Cα contact numbers run along the y-axis. The overall correlation is reasonable and the means square error is 6.8.

Table 2.1: The effect of sequence identity on the prediction of contact number

Sequence Identity Range (%)	correlation coefficient	MSE
10-40	0.684	7.75
10-50	0.704	7.40
10-60	0.700	7.40
20-50	0.688	7.75
10-90	0.699	7.70
30-60	0.660	8.00
30-90	0.640	8.00
50-90	0.570	8.50
Standard PSI-BLAST	*0.730*	*6.80*

The selection of the boundaries for sequence alignment were chosen such that a series alignments were made ranging from close homologs to very distantly related sequences and in-between. Initial expectations were that there would be a drop in prediction accuracy at both the top and bottom of the ranges. These expectations were verified, to some degree, as shown in the table 2.1 but are not as dramatic as initially expected, with the lowest boundary (10-40% sequence identity) having a correlation coefficient (cc) of 0.68, insignificantly less than the best predictions across the 10-60 and 20-50 ranges. The worst performance was obtained in the 50-90 range, which addresses close homologs (50% ... 90% similarity). The correlation coefficient was 0.57 and the mean error was 8.5, although this is not much worse than other thresholds, it does reflect an increased error in what is a sensitive measure. The decrease is probably a result of the lack of variation in the sequences from which the PSSM is derived. When compared to a standard PSI-BLAST generated PSSM – i.e. one that 'naturally' samples the same sequence database – the overall performance is worse still, as the PSI-BLAST PSSM obtains a cc of 0.73 and an error of 6.8 (figure 2.5) which, as mentioned above, is comparable to that of (Yuan, 2005). This suggests that attempting to generate 'artificial' alignments, i.e. those that cover specific sequence identities, does nothing to improve contact number prediction, in fact it does the opposite. Repeating the results

for long and sort range contacts shows the same pattern. It is possible that the effect of sequence similarity may have a more profound effect on feature prediction where there is a stronger conservation of structure – such as secondary structure.

To test the potential use of these predictions, now based on the PSI-BLAST derived PSSMs, two simple scoring functions were designed, the first based on simple difference and the second based on a Bayesian approach as described in the methods. The Bayesian approach suffered from the outset because the overall size of the datasets and inconsistency of prediction. The simple difference did however work with the limited dataset but does not punish under and over prediction equally, something that the Bayes-like approach would have. The overall results are interesting, in as much as it is obvious that the predictions are not robust enough to be used in final model evaluation – with an MSE of 6.84 which is comparable to that of (Yuan, 2005) and (Kinjo et al., 2005). When applied to five *de novo* predictions (figures 2.6-10) there is some indication that this type of function may be use, as outlined below.

As will be detailed in chapters 3 & 5, it is often useful to evaluate models post construction and prior to final refinement more often than not to save compute time, as demonstrated in (Taylor et al., 2008, Taylor et al., 2006). It is this niche were the contact predictions could be of future use. Figures 2.6-.10 show, despite the function being simple, that there is some discriminatory power. The worst performance is shown in figure 2.6 for glycerol-3-phosphate cytidylyltransferase (pdb 1coz chain A) a 129 amino acid protein. For this target no 'good' models were produced (those under 5Å) and this is reflected in the performance of the evaluation function which fails to discriminate the lower RMSD models (6Å) from the rest. Using the evaluation function

solely on this performance would be unwise based on the fact the even by selecting the top 200 models no low score models make it into the final ensemble. Figure 2.7 shows the polar opposite performance of the evaluation function on 1DI0, for this target a number of good models were produced. Selecting the top 200 constructs would result in almost all of models being taken forward to the refinement stage being within 6Å of the native structure. For the remaining three structures, shown in figures 2.8-.10, the results fall between the two prior examples. For each model there is no clear distinction between the ensembles of structures, however by selecting the top 200 models, a mix of good models (those below 6Å) and less desirable models (greater than 6Å) would be taken forward to the refinement stage. This means, that by using the techniques described in chapters 3 and 5, the number of 'good' models should increase relative to the 'bad'.

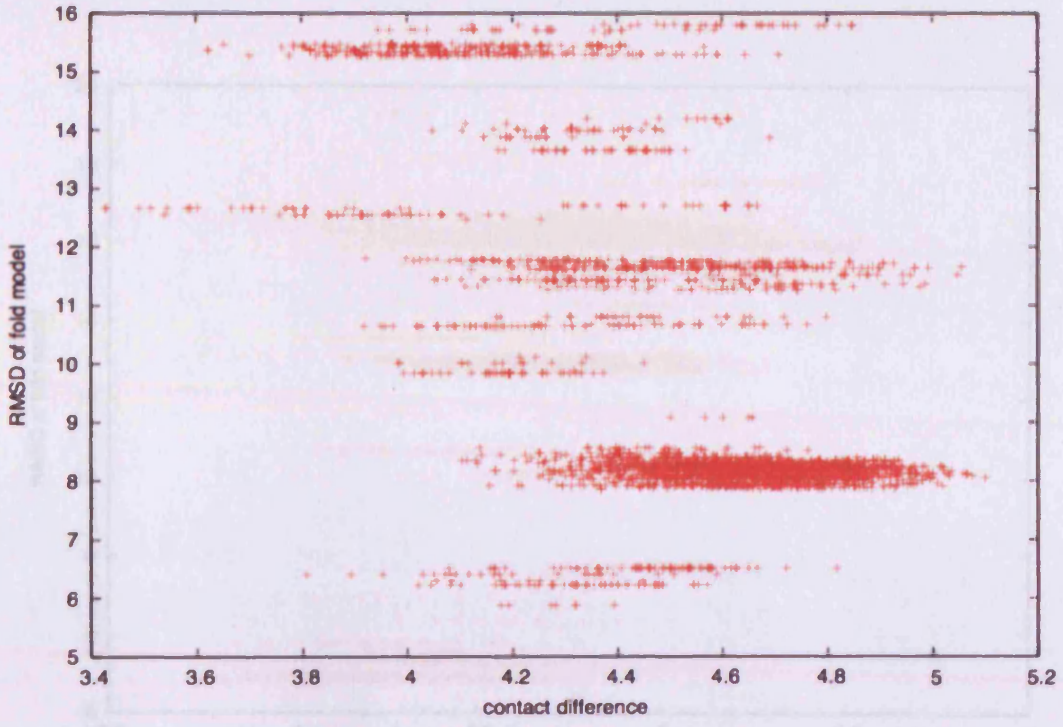


Figure 2.6 1COZA contact score: 1COZA produced the worst results for contact prediction. The top scoring native fold structure has a score of ~3.8, the rest scoring the same as, or less than, the non-native structures. Although some native-fold structures do make it into the top 500 they are in minority.

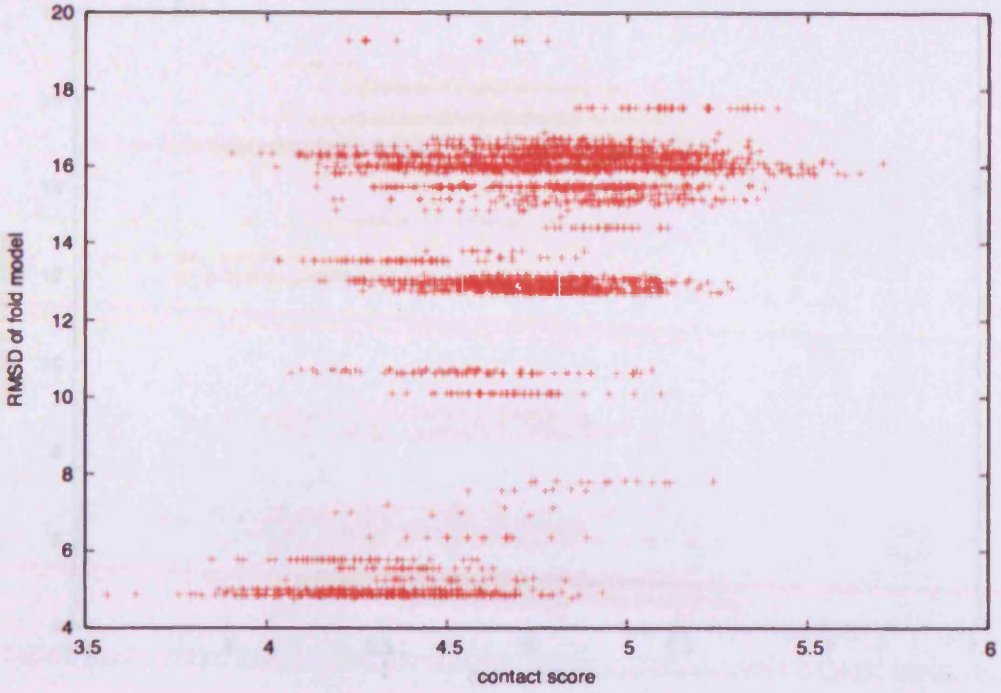


Figure 2.7 1DI0 contact score: The prediction of contact number for 1DI0 was good, this is shown by native-fold structures ranking better than the non-native structures. Such a prediction suggests that homology was present in the training set, however 1DI0 was identified in the *HOT* set not the training set. All structures in both the training and testing sets shared a maximum of 25% sequence identity so memorisation can be ruled out.

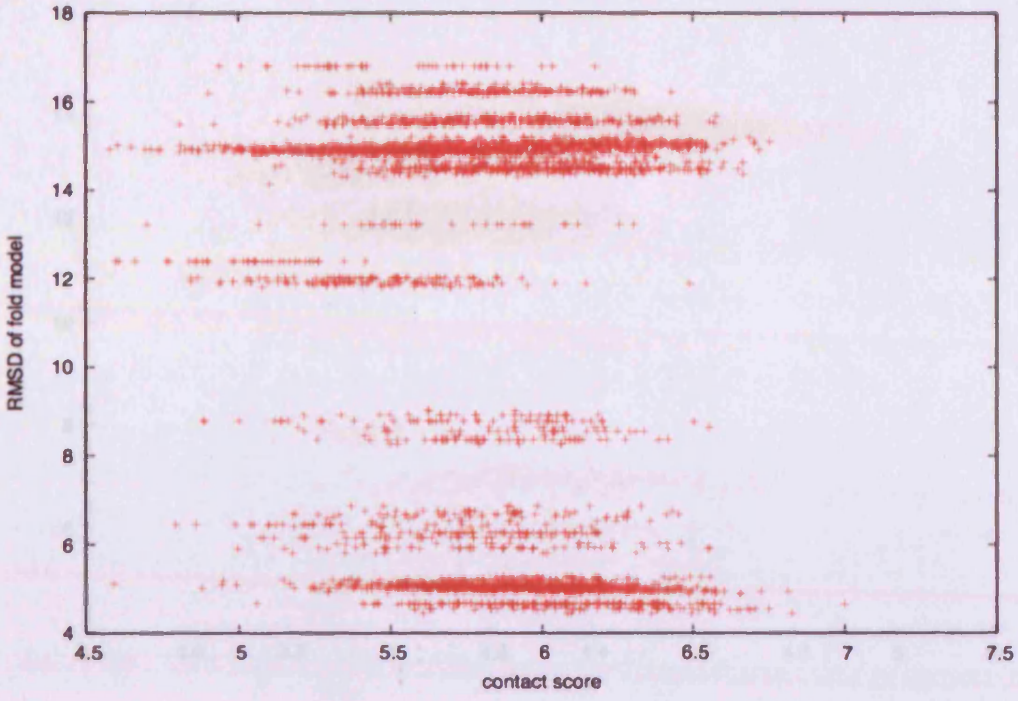


Figure 2.8 1F4P contact score: Lower scores represent structures which exhibit contact patterns closer to the predicted pattern. Although there is not a clear separation of the ensembles this score would be useful in the prediction pipeline where the top 500 models are taken. For 1F4P some native folds would proceed into the next round of refinement as well as incorrect folds which may be identified by more sophisticated functions such as Phobic, Sprek and TUNE.

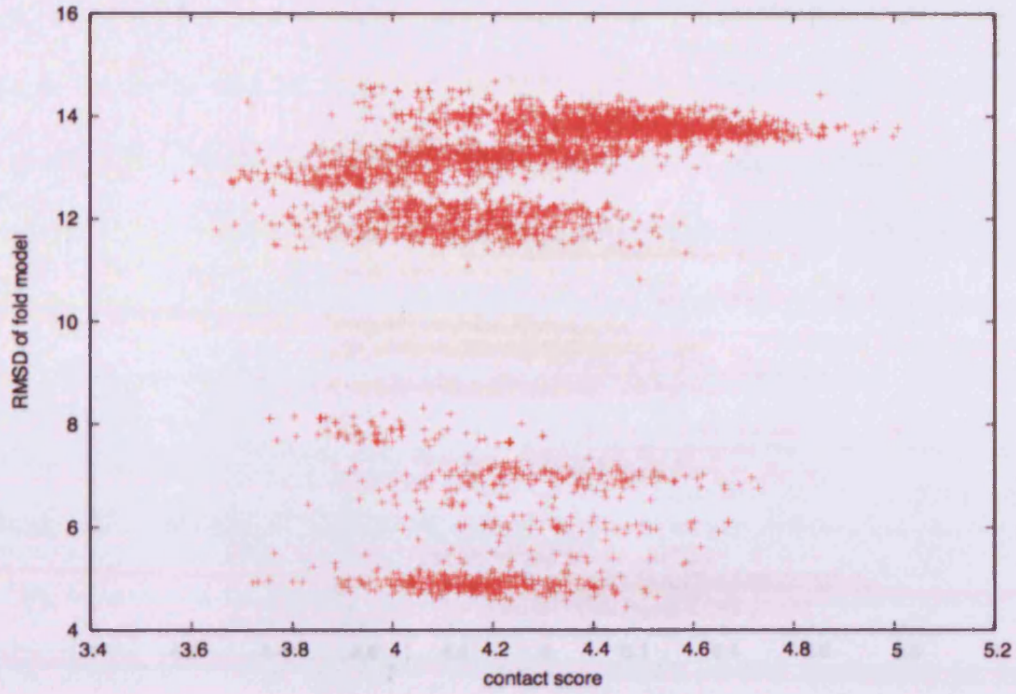


Figure 2.9 2trxA contact score: As with 1F4p (figure 2.8) native folds are identified in the top scoring models but there is a clear overlap with structures which are dissimilar from the native target structure. As with the previous structures, selection of the top 200 models would result in a number of good structures being taken forward to refinement.

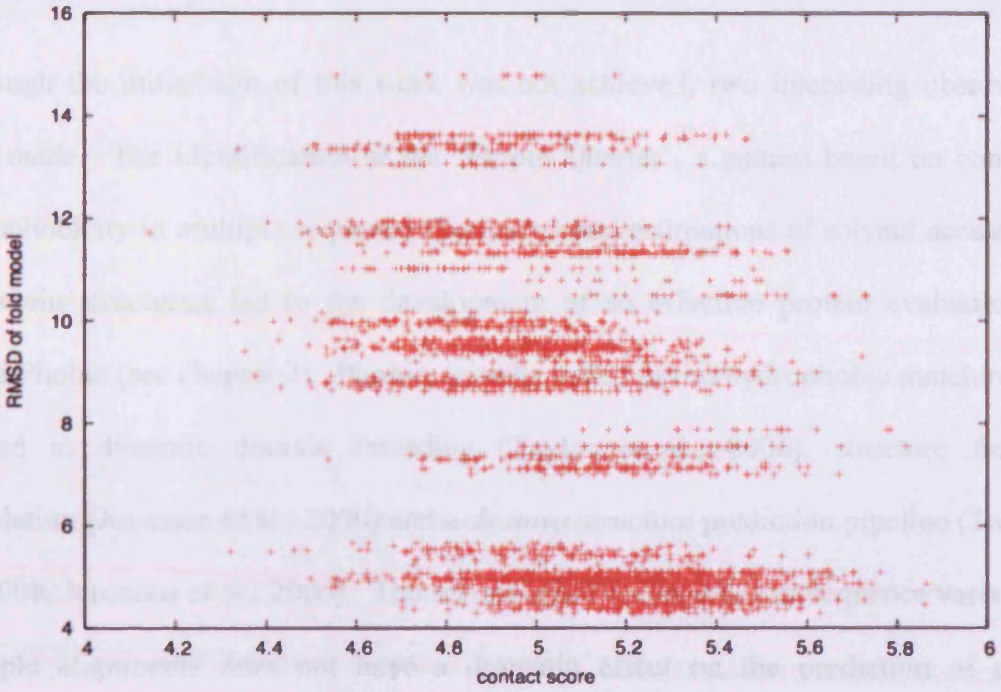


Figure 2.10 3CHY contact score: Structures which adopt the native 3CHY fold are identified in the top 500 structures, but like the previous examples structures which do not adopt the correct fold also score well.

Conclusion

Although the initial aim of this work was not achieved, two interesting observations were made. The identification of the 'Empty Quarter', a pattern based on conserved hydrophobicity in multiple sequence alignments and estimations of solvent accessibility in protein structures, led to the development of an effective protein evaluation tool called Phobic (see chapter 2). Phobic, initially called burial/hydrophobic matching, was applied in dynamic domain threading (Taylor et al., 2006), structure fragment tessellation (Jonassen et al., 2006) and a *de novo* structure prediction pipeline (Taylor et al., 2008, Jonassen et al., 2006). The second observation was that sequence variation in multiple alignments does not have a dramatic effect on the prediction of contact number, showing that only marginally better predictions can be made in the presence of remote sequence homology than close. Furthermore it is clear that better predictions can be made using a PSI-BLAST derived PSSM than using the routines currently applied in our prediction pipelines.

It has been suggested that all one needs to accurately model the three dimensional structure of a protein are accurate predictions of secondary structure, contact number and residue wise contact number (RWCO (Kinjo et al., 2005)). This work shows that, despite having a correlation coefficient of 0.74, the margin of error is sufficient to prevent the use of predicted contact number in a final stage model evaluation function. However, it does show that the predictions may be of some use in post-construction / pre-refinement evaluation functions (as will be described in chapter 5). While the simple difference function, examined here, shows some, albeit limited, potential for the selection of native-like folds from the ensembles of models, it is not unreasonable to

assume that the performance could be improved. This improvement may be achieved through the application of a larger training dataset and implementing an algorithm that is capable of punishing under and over prediction independently, such as the Bayesian method described previously.

In summary, despite failing to achieve the aims described at the start of this chapter, the work conducted as part of this investigation played an important role in the initial development of two protein structure prediction pipelines, one model evaluation function and a novel method for the prediction of secondary structure and solvent accessibility which will be described in the following chapters.

Chapter 3

The Critical Analysis of Structure Prediction:

Round 6 and Dynamic Domain Threading

Introduction

The aim of protein structure prediction is to map a sequence with unknown structure, referred to as the target, to structural space using any information available. This problem is one of the grand challenges in computational biology and is assessed biennially by the *C*ritical *A*nalysis of *S*tructure *P*rediction (CASP) and *C*ritical *A*nalysis of *F*ully *A*utomated *S*tructure *P*rediction (CAFASP) exercises.

CASP was initiated in 1994 by John Moult with the aim of becoming an internal control mechanism to direct the future of the field (Moult et al., 1995). The procedure is simple, CASP runs for approximately seven months of which four are devoted to structure prediction and three to analysis. During the four months sequences with privately known structure are distributed to participants, none of whom know the native structure of the target. Each of the targets is assigned a deadline dependent on the use of fully automated servers or 'expert' guidance. The deadlines range from 24 hours (for server groups) to several weeks depending on when the native structure is released to the scientific community.

At the end of the prediction period, all submitted models are evaluated by a panel of experts using the longest-continuous-segment-global-distance-test (LCS-GDT) (Zemla, 2003) as well as any tools of their choice. The analysis is composed of several categories each with a number of sub-categories – the major division of CASP can be drawn between two dimensional (2D) and three dimensional (3D) predictions. In the past, the 2D category has covered secondary structure, solvent accessibility, disorder

and contact number, however it is generally regarded as 'second fiddle' to three dimensional prediction as 2D features are used to improve overall structure prediction.

Overall performance is assessed by the ability of each method to consistently predict the three dimensional coordinates of all C_α atoms of the protein. Assessment at this level is dived into three categories based on the approach deemed most suitable by the assessors. These categories are identical to those introduced in chapter 1: Comparative modelling (CM), which is divided into easy and hard targets based on the ease of which a sequence with known structure (template) can be identified using tools such as PSI-BLAST (Altschul et al., 1997). Hard targets fall into an area close to the twilight zone (Doolittle, 1986) where diminishing sequence similarity makes template identification difficult; Fold recognition (FR) is also split into easy and hard targets. The boundary between hard CM and easy FR is not clearly defined but could potentially be measured by sequence similarity. Hard FR targets are typically attempted by complicated fragment packing methods that borrow sections of structure from a number of templates and 'glue' them back together, an effective example of this approach is the SAM series by Karplus (Karplus et al., 2003); New fold modelling (NF), as with CM and FR, overlaps with FR. NF is applied in cases where there is very little or no information available on a target. Many methods aim to solve this problem using prior knowledge by incorporating fragment packing – the idea that reducing the size of the probe sequence increases the probability of getting a hit from the databases but at the expense of the noise-signal ratio. NF modelling is beset with difficulties including the introduction of knowledge based scoring potentials which suffer from massively reduced capacity in the absence of detectable homology. Ultimately the NF category is

'the' problem to solve and arguably over recent years more progress has been made here than in CM and FR modelling fields.

Post assessment the success of each groups attempts on all structures is made publicly available and an unofficial 'winner' of the current round is identified – over the last three rounds there has been no change at the top! Assessment at this resolution allows widespread changes to be made to methods and effectively forces participants to adopt new methods. At the group level CASP allows for identification of weak spots in their methods such as consistent failure to identify the best template or the absence of a suitable model evaluation function(s).

In this chapter I will introduce the method used in the 6th round of the critical analysis of structure prediction exercise (CASP6). The method, dynamic domain threading, is a new threading approach that attempts to side-step the problem of domain definition in templates. I will present a review of the overall performance at CASP6 and describe how the method could be improved.

Methods

Construction of Protein Models Using Dynamic Domain Threading.

A full description of the model construction process used in CASP6 can be found in (Taylor et al., 2006). Before giving a brief overview of the techniques used it is important for the reader to be aware of the meaning of two terms: target, the name used

to refer to sequences whose structure is to be modelled; template, the term used to refer to sequences with known structure which are used to construct models of the target.

The feature which distinguishes DDT from other threading techniques is the use of an Ising-like model to define a series of sub domains within each template (as described below). Each sub domain is based around different positions within the template and for each variation the sequence, secondary structure and degree of residue burial are compared against that of the target (Taylor, 1997a). The resultant alignment is used to generate Cα models for the target. Because the sub-domains are borne from different regions of the template variation is introduced. The result of this variation is that a number of structurally different models are produced from a single template. Each of these models is then evaluated using a number of scoring functions described below.

Domain Definition using an Ising-like Model

Domain definition was based on an Ising-like method described in (Taylor, 1999a). This method uses multiple seed points within the template structure from which sub-domains grow and form a combined sub-domain or remain separate. This means that the domains defined are not guaranteed to be similar in size to the target. To avoid the problem of small or giant domains the growth of the domain model was biased to select the same number of residues as the target. The bias was not perfect and occasionally fragments of structure were generated. To suppress this problem a smoothing stage was introduced to unite or remove fragments.

Alignment of the Target and Domains

The DDT program was developed from the MST program (Taylor, 1997a) which uses a comparison of secondary structure state and predicted residue exposure along with a measure of sequence similarity to align the target and the template. Unlike the original MST method, pairwise spatial restraints are ignored in DDT. This approach is like the 3D/1D methods described in Chapter 1.

Secondary structure matching

Template secondary structures were defined by the STICK program (Taylor, 2001) that uses only Cα coordinate data, giving a consistent definition across both native structures and models.

Secondary structures were predicted using the PSI-Pred program (Jones, 1999b). To obtain variation in the secondary structure predictions, a separate prediction was made for each sequence in the family-based alignments. The result, on average, was that at least one prediction was a close approximation of the native structure.

Burial/hydrophobic matching

Residue exposure was calculated using a version of the POPS program (Fraternali and Cavallo, 2002, Cavallo et al., 2003) which had been optimised to work with Cα models, which like STICK, gives a consistent measure across native structures and models. The

output of POPS was mapped into the range -1 ... +1 (exposed – buried) using the Gaussian transform:

$$e = 2\exp(-ca^2) - 1$$

where a is the surface area estimated by POPS and c is the inverse variance which has the value 0.003, which was found to give a zero mean for the transformed value e. Predicted burial was calculated as a function of conserved hydrophobicity, which was calculated using the colour scheme described in chapter 2. The degree of conservation was encoded as the number of different amino acids at a particular position in an MSA (a) augmented by the fraction of gaps (g). In addition, a contribution from the prolines (p) present was introduced. The terms were combined as follows:

$$c = (h+1)\cdot\zeta(a,A)\cdot\zeta(g,G)\cdot\zeta(p,P) - 1$$

where the function $\zeta(x,y)$ is the Gaussian transform:

where A=2, G=1, P=0. As the hydrophobic component (h) falls into the range -1 ... +1 and all other values are transformed using equation (<NUMBER>), c always falls into the range -1 ... +1. The value of c was shifted slightly to give a zero mean over the positions in the alignment to correspond with the exposure measure (e).

Sequence/Structure alignment

The alignment of the sequence and structure properties used a conventional dynamic programming algorithm. The score matrix was built by the addition of a sequence component derived from a Dayhoff-like matrix of amino acid similarity (Jones et al., 1992b) of the template/probe residue match scaled into the range 0 ... 1. The score matrix was then supplemented by secondary structure matching where a β match was evaluated to +4, α matching +2 and loop matching +1. The burial hydrophobic match was then included.

The matrix generated by this scoring scheme was then modified by the current domain definition by reducing the score by an order of magnitude for positions excluded from the domain. The matched residues form the framework over which models were constructed using a modified version of the RAMBLE program (Taylor et al., 2002). The new program (called TRACK) imposes a constraint to 'encourage' selected residues to lie close to given probe points. To be a 'core' residue (*m*) the following conditions had to be met:

- *m* is buried;
- *m* is in, and not on the end, of a secondary structure fragment;
- the percentage match in secondary structure must be greater than 70%;
- the template residue (*n*) aligned with *m* must also be within a predicted secondary structure fragment;
- the number of matched residues must be more than half the length of the template.

To avoid directional bias during construction, RAMBLE starts at residue *n* (aligned to residue *m* in the template) and expands towards both termini. Where the template and domain residues are paired, a large number of semi-random trial positions were tested until a new position was found within 1.5Å of the template for residue *n* – this step generates a small amount of variation between the template and target structures. If this was successful the new position was shifted towards the template position by a random factor between 0 and 1. If the new position was within 1Å of the template it was accepted, otherwise a default torsion based random walk position was used, with torsional angles appropriate for the local secondary structure.

Each model was then refined using routines similar to those from the DRAGON and GADGET programs (Petersen and Taylor, 2003). These procedures adjust the Cα-Cα bond length between residues and also the distance between the residue and residue *i*+2. Non-bonded contacts were repelled if the distance between them was less that 5.0Å in a β-sheet, 5.5Å in α-helix and 6.5Å otherwise. Pairs of residues predicted as β were refined towards 5.0Å if they met the criteria defined in (Taylor, 1999a).

Model Evaluation

Before application of sophisticated scoring functions each model was filtered using some 'basic' geometric checks described below.

Radius of Gyration.

The predicted radius of gyration (how much the protein spreads from its center (RoG)) was calculated from the probe sequence length (N) as:

$$R_c = \frac{\sqrt{10}}{5}\left(\frac{3N}{3\pi\rho}\right)^{1/3} \quad (3.1)$$

where rho (ρ) is the density of the protein estimated as $3*10^{-3}$ and $\left(\frac{3N}{3\pi\rho}\right)$ is the estimated radius of a spherical protein while $\frac{\sqrt{10}}{5}$ converts this to the RoG. The RoG of the model was also calculated from the template sub-domain over which the model was constructed (R_d). The value can take a wide range of values dependant upon the shape of the domain, to moderate the values an average was taken between the structures. To allow for variation 2Å was added:

$$R = 2 + (R_c + R_d)/2 \quad (3.2)$$

where $R_d = 1/N \sum x_i^2$ for a set of coordinates (x) with zero centroid. Models with a RoG score greater than R were discarded.

Hydrogen Bonded β-sheets, Tangles and Distortions

The maximum number of hydrogen bonded pairs in a β sheet was estimated from the number and length of predicted β-strands. This is achieved using the sum of the lengths

of all the β-strands with the exception of the longest. Models with approximately 20 bonded pairs typically contained a sheet of five well connected strands.

Although tangles and knotted structures occur only rarely in native structures, they pose a problem for modelling and can be justifiably excluded due to their rarity. The current modelling strategy employed a smoothing algorithm in which the number of steps (s) required to reduce the protein to linear form and the number of chain 'bumps' (r) experienced in the process were used as indicators. The logarithm of the product of s and r was used to give a score:

$$t = \log(1 + r \cdot s) \qquad (3.3)$$

Geometric quality of the final model were measured as the number of non-adjacent residues within 4Å (b in equation 3.4) and the number of adjacent residues with a Cα-Cα RMSD of 4Å (a in equation 3.4). Again the logarithmic value of the product of a and b was taken:

$$d = \log((1 + a)(1 + b)) \qquad (3.4)$$

As the scores are related they were combined (G) as:

$$G = t + d \qquad (3.5)$$

Models with a G-score greater than 20 were rejected. The value of 20 was a result of coarse grain optimisation over a small set of proteins.

Detailed Evaluation

While the previous scores are capable of removing some structures that violate basic constraints they are too coarse to discriminate good models from the ensemble of predicted structures. While any (or all possible) evaluation functions can be applied to the models, in the current work only three methods were applied, each derived from a distinct physical basis to avoid redundancy. The scores used were based on the POPS program (as described in chapter 2 and above), TUNE-3D (Lin et al., 2002) and SPREK (Taylor and Jonassen, 2004), because each of these scores was applied to models of the same length it was unnecessary to normalise for length or structural type.

Burial and secondary structure matching

Using POPS, the solvent accessible surface area of residues in the model were estimated and mapped into the range -1 ... +1 as described above for the template structure. The value was then compared to the measure of conserved hydrophobicity (*c*, mentioned above) by taking the sum of their product over all residues. Positive values of this sum indicated a better than random correspondence.

The segment based method (STICK) was then applied to each of the models. Effects of the modelling process meant that each model may have different secondary structure and also differ from those in the template protein. As in the template/probe matching, a simple count was taken of the number of residues at which the observed and predicted structure matched.

Residue Packing using TUNE and SPREK

TUNE encodes the propensity of pairwise residue interactions in an artificial neural network and uses a reduced representation of protein structure based on the Cα and the residue centroid (Lin et al., 2002). For this work TUNE was modified so that the centroid was constructed 2Å beyond the bisector of the Cα bond. The network was retrained using the new representation and evaluated on the 4-state decoys of Park and Levitt (Park and Levitt, 1996) prior to the construction of the pipeline.

The SPREK method scans local fragments of structure against a non-redundant database of known protein structure. The number of patterns matched against the database is used as an indication of how protein-like the construct is. In this work the database of known structures was extended using a set of alignments based on the SCOP domain database (Murzin et al., 1995) (Klose & Taylor, unpublished data) described in chapter 2.

Comparison of Protein Structures

Two methods were used to compare the predicted structures to the native structures. Each method is described below.

Structure Alignment (SAP) Program

Structure alignments were made with the SAP program which is based on the sequence structure alignment algorithm of Needleman and Wunsch (N&W) (Needleman and

Wunsch, 1970). The application of the N&W algorithm allows for incorporation of insertions and deletions, a necessary requirement when dealing with distantly related proteins. The difference between the sequence and structure alignment is the situation of the amino acid in the structure – a buried core residue is different to an exposed loop residue, whereas two residues in a linear sequence are fundamentally the same. The difference in location is used to advantage in SAP, with a measure of local structural environment of each residue forming the basis of similarity score. In addition to this basic measure, a representation of the 3D structure is required. In SAP this is achieved by aligning sections of each protein and then maximising their local environment scores. SAP is suited to the current work as it requires only Cα coordinates and can be forced to give one-to-one alignments as well as optimal alignments. The output of SAP is the number of equivalenced carbon alphas, the number of selected residues that lay on the alignment and the root mean squared deviation (RMSD). If either of the first two values falls far below 50% caution must be exercised in the interpretation of the results. An additional file is created by SAP that contains a rigid body superposition of the two structures which can be further analysed in molecular graphics software. There is also a plot of the RMSD with increasingly large subsets of residues – similar to the LCS-GDT measure used to assess CASP.

The DALI Method

DALI is an algorithm for pairwise structure alignment using Cα coordinates (Holm and Sander, 1993). As stated by the authors the method has two steps, each consisting of several sub-steps and several scores. The first score measures how similar two structures are on a sub-structure level. In this measure only residues that are matched

contribute to the score and as such the larger the value the more optimal the set of equivalenced residues are. The remaining two scores address the problems of searching for predefined patterns in structures and the search for the largest common sub-structure between two structures. The first step is a pairwise comparison of all elementary contact patterns in the two distance matrices. DALI stores equivalenced hexapeptide-hexapeptide contact patterns between two proteins (A & B) in a non-exclusive list of pairs. The second step assembles pairs of contact patterns into sets of pairs in order to maximise the similarity score. After this step a Monte Carlo algorithm is used to build alignments from contact patterns, this step is then followed by a final refinement stage. DALI is available as a pairwise and multiple structure alignment tool, however it was only used for multiple alignment in this work.

The output from the web-based DALI algorithm contains a multiple sequence alignment plus five measures of similarity: raw-score, the value initially computed by the DALI algorithm; the Z-score – the number of standard deviations from the mean, the larger this value the more similar two structures are considered to be; the id, a simple measure of percentage sequence identity; the length of the structure alignment (similar to SAP); RMSD, as returned by most structure comparison algorithms however this value varies between methods (SAP and DALI) as the protocol for calculation differs.

Results and Discussion

Conserved Hydrophobicity, TUNE and SPREK

From preliminary investigation it was observed that the scatter plot of conserved hydrophobicity against solvent accessible surface area (first introduced in chapter 2 as the 'empty quarter') did not show a linear trend, but tended to be well populated everywhere except for the quadrant corresponding to conserved hydrophobic residues that were exposed in native structures ($c > 0$ and $s < 0$) (see chapter 2, figure 2.4). A simple evaluation function was developed from a count of the number of points (residues) found within the 'empty quarter', the greater the count the less ideal the model.

To evaluate the effectiveness of the three model evaluation functions, a set of decoy models were generated from all SCOP domains starting with the code "d1a". The decoy structures were constructed by reversing the native structure and using this as a template on which to construct models (Taylor and Jonassen, 2004, Jonassen et al., 2000). For each protein in the decoy set two classes of models were generated: those based on the native structure which deviated by approximately 1.4Å on average (over 500 structures) from the native structure; those based on the reversed structure which deviated by approximately 10Å on average. The same geometry parameters, described previously, were used to select these models, resulting in many models being rejected. The result of this was a set of decoy models that were compact and protein-like with reasonable secondary structure and a minimum of ten forward and reverse models per structure. The final dataset consisted of 36 proteins with 20,000 models in all (11,500

native and 8,000 reverse). The ability of each method to discriminate the native (true) from the reversed (false) was quantified using a simple measure: for each method, a value of its score (x) was found where the number of false models scoring over x (false positions, FP) equalled the number of True models scoring less than x (false negatives, FN). The number of false negatives is a measure of how well the two populations have been discriminated.

Over the sample of 543 structures 16 were misclassified by TUNE, 18 by SPREK and 46 by The Empty Quarter. In contrast to TUNE and SPREK, the POPS based score is basic, having no facility to check for homology and lacking any form of optimisation, this resulted in a limited ability to discriminate between the true and false models. For this reason it was removed from the fine grain evaluation and the SPREK and TUNE scores were combined as follows:

$$S = (s(t+10))^{1/2} \qquad (3.6)$$

where t is the TUNE score with 10 added to it to ensure it is positive and s is the SPREK score. The square root was taken to avoid large numbers and has no effect on final rankings. The combination of scores in this fashion resulted in 8.3 errors per protein, a drop by half when compared to the individual techniques.

DDT Compared to Standard Threading

To evaluate the quality of models and their scores in a more realistic environment, the bacterial chemotaxis protein family (CHEY) was modelled on a variety of templates.

Using an alignment of 8 sequences based on the Che-Y protein (pdb 3CHY), a diverse collection of template structures was selected, ranging from the structure itself, through homologs included in the alignment to analogous folds identified using the DALI structure comparison server (Holm and Sander, 1995). For the structures in the set, a large number of models were generated with and without the dynamic domain threading turned on. The method without the DDT is identical to the DDT method with the exception that the domain definition was set to include all residues in the template structure. This minor change means that the same number of modelling attempts was made for each template structure, allowing the numbers and quality of the models resulting from each group to be compared directly. In the following paragraphs, the dynamic domain method will be referred to as DD and the single domain SD.

Assessment of Structural Quality

The quality of the models was assessed by structural superposition of the model onto the known structure of the probe protein (3CHY) using the SAPit program (Taylor, 1999b). The models were then ranked on their RMSD fit and plotted from smallest to largest value for both the DD and SD methods as shown in figure (3.1).

For the homologous single domain templates, there was essentially no difference between the plots of the ranked RMSD values, except sometimes where a small number of very poor structures had been generated, probably as a result of unpredicted secondary structure elements allowing the chain path to deviate from the template too far to get back on the 'right' track. Many of these models were avoided by discarding the worst 20% of the models, typically those over 460 in figure 3.1.

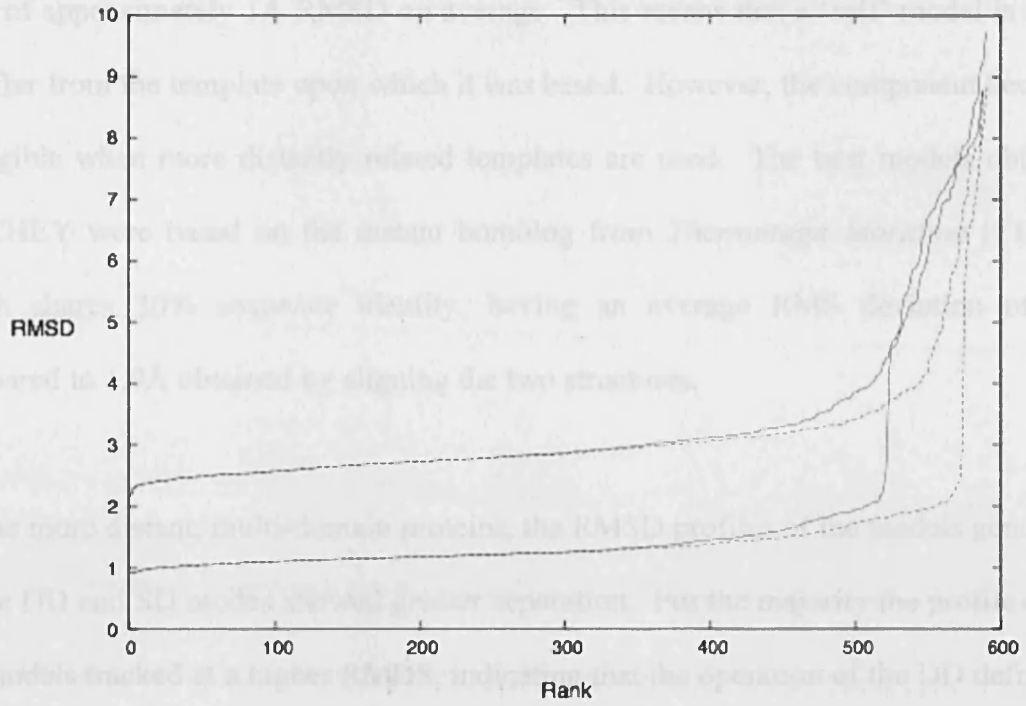


Figure 3.1 Ranked RMSD deviations for closely related template structures: The RMSD deviations for the models constructed using the dynamic-domain approach (solid lines) are plotted along with the data obtained without domain definition (dashed lines). Data are plotted for two proteins: the lower curves are from modeling the CHEY sequence on its own structure (3chy) while the upper is from a homologous CHEY protein (1tmy) with 30% sequence identity. The sharp rise in RMSD values to the right is discussed in the text body. The RMSD values (Y-axis in Å) for each model are plotted in rank order along the X-axis.

It should be noted that in this method, the stoichastic component contributes a 'noise' level of approximately 1Å RMSD on average. This means that a 'self' model is likely to differ from the template upon which it was based. However, the component becomes negligible when more distantly related templates are used. The best models obtained for CHEY were based on the distant homolog from *Thermotoga Maritima* (1TMY), which shares 30% sequence identity, having an average RMS deviation of 2Å, compared to 1.9Å obtained by aligning the two structures.

On the more distant, multi-domain proteins, the RMSD profiles of the models generated by the DD and SD modes showed greater separation. For the majority the profile of the SD models tracked at a higher RMDS, indicating that the operation of the DD definition had resulted in an increase in lower RMSD models. This is shown for the double domain protein NARL (1A04, 205aa) (figure 3.2) and methylmalonyl-voA mutase (1REQA, 727aa) (figure 3.3). The sequence identity with CHEY is low, 25% for 1A04 and 11% for 1REQA (as measured by structure comparison).

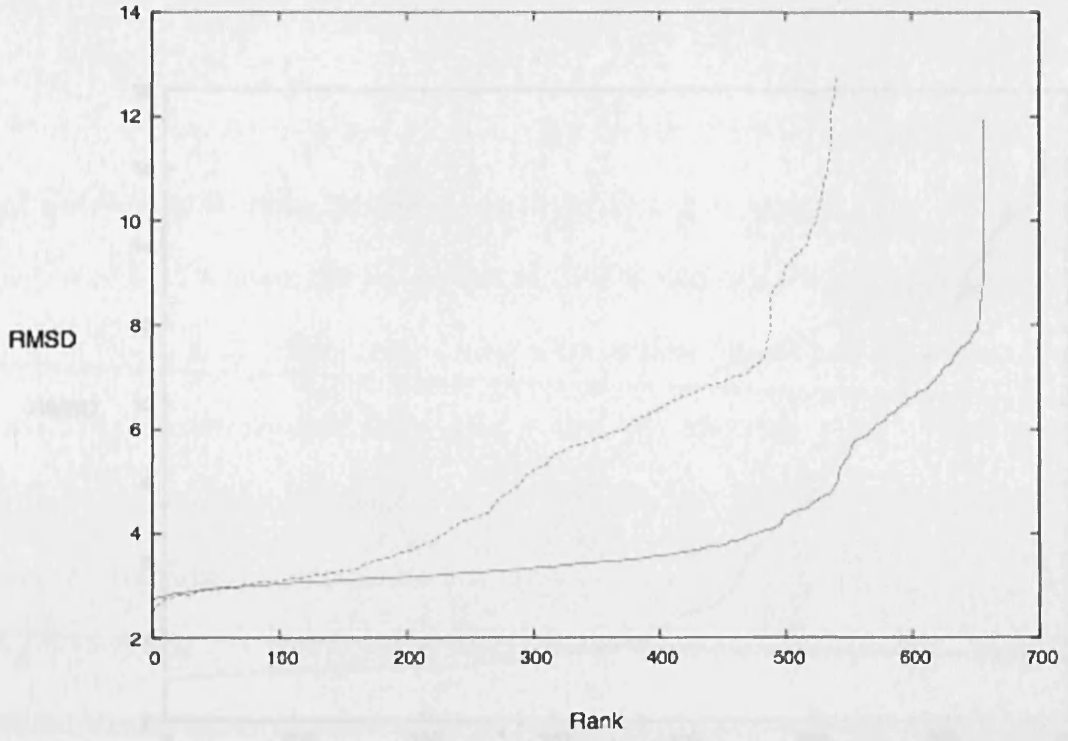


Figure 3.2 Ranked RMSD for double domain protein 1A04: The RMSD deviations for the models constructed using the dynamic-domain approach (solid lines) are plotted along with the data obtained without domain definition (dashed lines). The RMSD values (Y-axis) for each model re plotted in rank order along the X-axis.

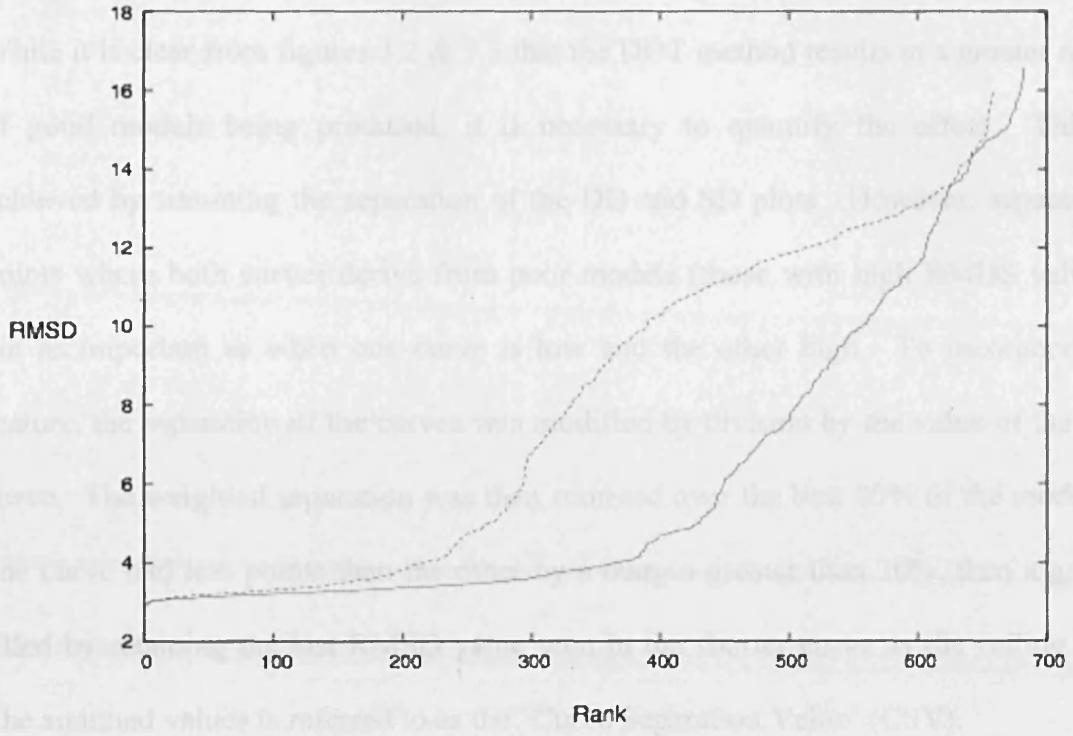


Figure 3.3 Ranked RMSD for double domain protein 1REQ: The RMSD deviations for the models constructed using the dynamic-domain approach (solid lines) are plotted along with the data obtained without domain definition (dashed lines). The RMSD values (Y-axis) for each model are plotted in rank order along the X-axis.

Quantifying the DDT improvement

While it is clear from figures 3.2 & 3.3 that the DDT method results in a greater number of good models being produced, it is necessary to quantify the effect. This was achieved by summing the separation of the DD and SD plots. However, separation at points where both curves derive from poor models (those with high RMDS values) is not as important as when one curve is low and the other high. To incorporate this feature, the separation of the curves was modified by division by the value of the lower curve. The weighted separation was then summed over the best 80% of the models. If one curve had less points than the other by a margin greater than 20%, then a gap was filled by retaining the last RMSD value seen in the shorter curve as the ceiling value. The summed values is referred to as the 'Curve Separation Value' (CSV).

The CSV was calculated for the proteins identified by the DALI method as having a Z-score of 6. Although this value appears high, proteins in this family which have a Z-score less than 10 cannot be considered as homologous, furthermore none under Z=15 have more than 21% sequence identity. To avoid complications with domain definition and modelling across gaps, structures that contained chain breaks greater than 8Å were omitted. The CSV for the remaining 43 proteins that fulfilled the above criteria were plotted against their chain length (figure 3.4).

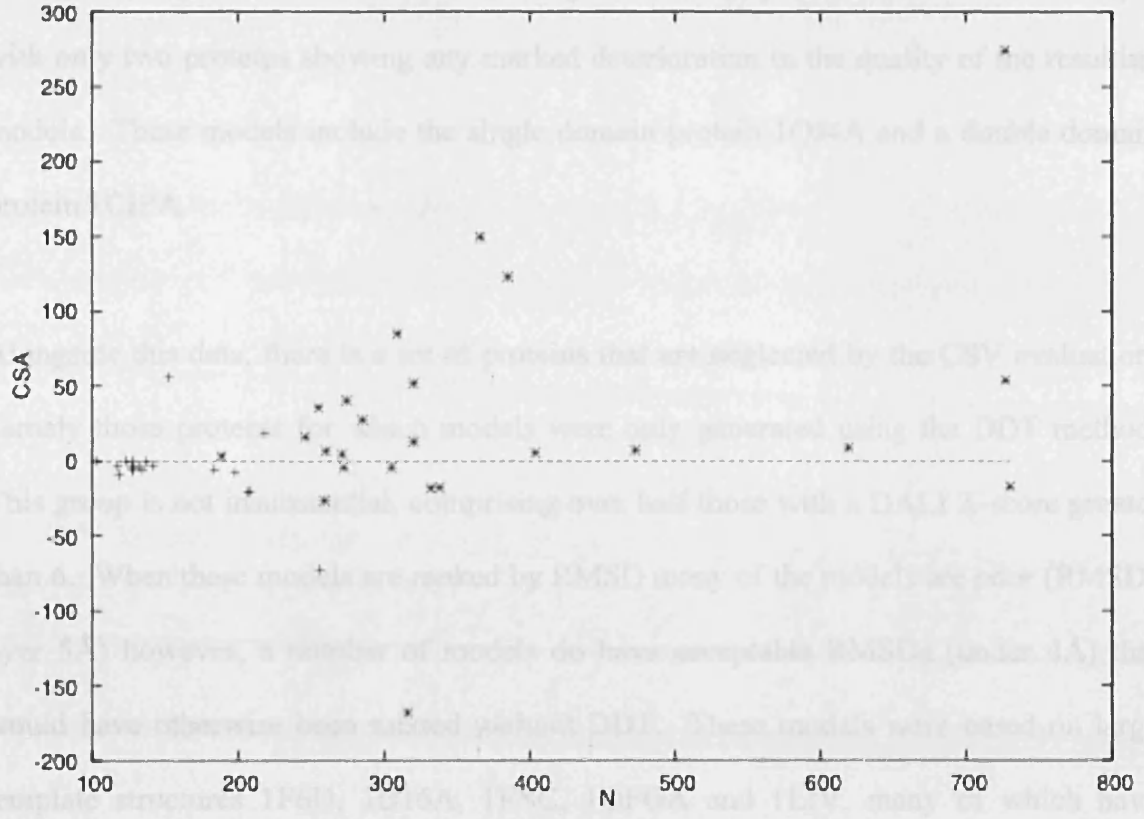


Figure 3.4 Quantifying DDT Improvements: The effect of the DDT algorithm on the number of good models produced was quantified using the CSV measure relative to the SD models. The CSV was plotted for each protein against the chain length (N).

The data in figure 3.4 shows an increasing trend for improvement with chain length, with only two proteins showing any marked deterioration in the quality of the resulting models. These models include the single domain protein 1QJ4A and a double domain protein 1CIPA.

Alongside this data, there is a set of proteins that are neglected by the CSV evaluation, namely those proteins for which models were only generated using the DDT method. This group is not insubstantial, comprising over half those with a DALI Z-score greater than 6. When these models are ranked by RMSD many of the models are poor (RMSDs over 5Å) however, a number of models do have acceptable RMSDs (under 4Å) that would have otherwise been missed without DDT. These models were based on large template structures 1F6D, 1B16A, 1ESC, 1OFGA and 1LIV, many of which have complex multi-linked domain structures.

Evaluation of CASP 6

During CASP6 attention was focused on comparative modelling and fold recognition targets at the expense of new fold modelling. For each target a sequence alignment was generated using the method described previously. Templates were identified using the TUNE and GenThreader tools and then run through the DDT method with dynamic domain threading switching turned on and off. All models were pooled and assessed as described above, the top three structures where submitted to the CASP evaluators for assessment.

To establish if the best template structure was identified the target structure was scanned against the PDB using the DALI algorithm. For all the targets where a template should have been identified there are several outcomes: the best template is identified and used to construct models; the best template is identified but discarded; the best template is not identified and therefore not used. To be considered as an appropriate match, the template used to construct models had to be in the top 5 DALI structures.

Table 3.1 Target Template Identification at CASP6.

No match	Match - discarded	Match - used
197	213	203
198	228	222
199	237	223
202	243	224
206	248	230
209	249	235
212	251	249
214	280	263
262		
272		
272		
281		

Each number in the above table is an identifier from the sixth round of CASP. There were 87 targets released during CASP6, not all targets were predicted and some were cancelled because of early release or failure to solve the structure on time. The above table shows for the targets attempted, with the exception of *de novo*, whether the best possible templates were identified and used. Match means the template used to construct models was found in the DALI top five hits.

Twenty-eight targets were identified as suitable for comparative modelling or fold recognition approaches (Table 3.1). Table 3.1 identifies the attempted CASP targets for which suitable template should have been identified. For 12 of these targets the best possible templates were not identified and, as a result, predictions were poor. For a further 8 targets the best template was identified but discarded by the evaluation functions. For the remaining 8 targets the template used to construct models was identified (afterwards) by DALI as being an ideal template. For targets T0230 & T0249, where good templates were identified, the method produced some of the better submitted models from all groups (see figure 3.5). Nevertheless, table 3.1 shows that, for the majority of targets, the best possible template was either not identified or discarded.

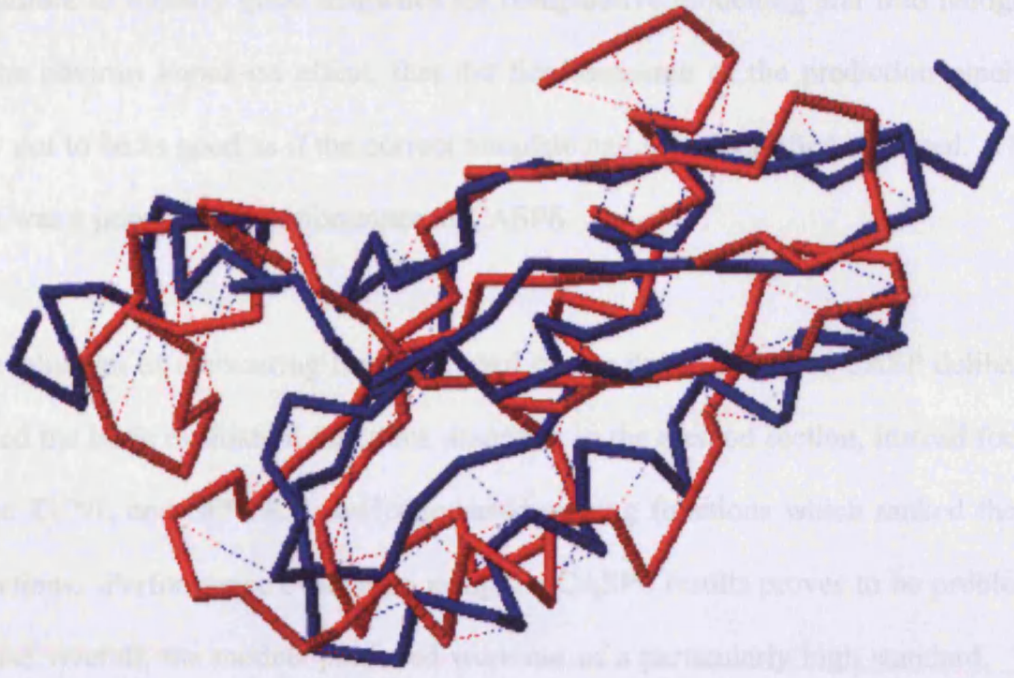


Figure 3.5 SAP structure superposition of T0230 and 1WCJ: The native structure is coloured blue is an α/β protein 102 residue in length. The red structure is first model submitted to CASP for evaluation. The overall RMSD is 5.61Å with much of the variation occurring in the long interconnecting loops between secondary structure elements.

The failure to identify good templates for comparative modelling and fold recognition has the obvious knock-on effect, that the final outcome of the prediction pipeline is likely not to be as good as if the correct template had been identified and used. The net result was a poor overall performance at CASP6.

The evaluation of the scoring functions used during the 6th round of CASP deliberately ignored the basic evaluation functions described in the method section, instead focusing on the TUNE and SPREK knowledge-based scoring functions which ranked the final predictions. Performance evaluation using the CASP6 results proves to be problematic because, overall, the models produced were not of a particularly high standard. There are two further extenuating circumstances, visual assessment of the top scoring models often conflicted with the evaluation function score, resulting in the submitted models not being the ones with the highest rank. Manual refinement of models also took place, this included changing connections, loop lengths and secondary structure. Indeed, for all targets, the model with the lowest RMSD produced by the prediction pipeline was not identified and, for some targets, the best models were in excess of 10Å from the native structure. Even for targets T0230 (figure 3.6) and T0249 (figure 3.7) the evaluation functions did not identify the model with the lowest RMSD, but a member of the closest ensemble of structures. Figure 3.6 shows the RMSD plotted against the evaluation score for each predicted structure; each point on the plot represents one of these models. In addition the two models submitted for assessment are highlighted rank 1 (green), rank 2 (red). By contrast figure 3.7 shows a threading target, where the template was correctly identified (table 3.1) and used. The figure clearly shows that the combination of the scoring function fails to identify the 'best' models which fall within 3Å of the native structure, the figure also shows the three models submitted to CASP

for assessment (green, red and blue). One thing that figures 3.6 and 3.7 clearly show is that the performance of the scoring functions is not consistent.

In summary, the CASP6 results were not unsurprising, with the DDT method able to generate good approximations for some comparative modelling targets, where suitable templates were identified. For fold recognition, the overall performance was 'hit and miss', depending on identification of a good template and performance of the evaluation function.

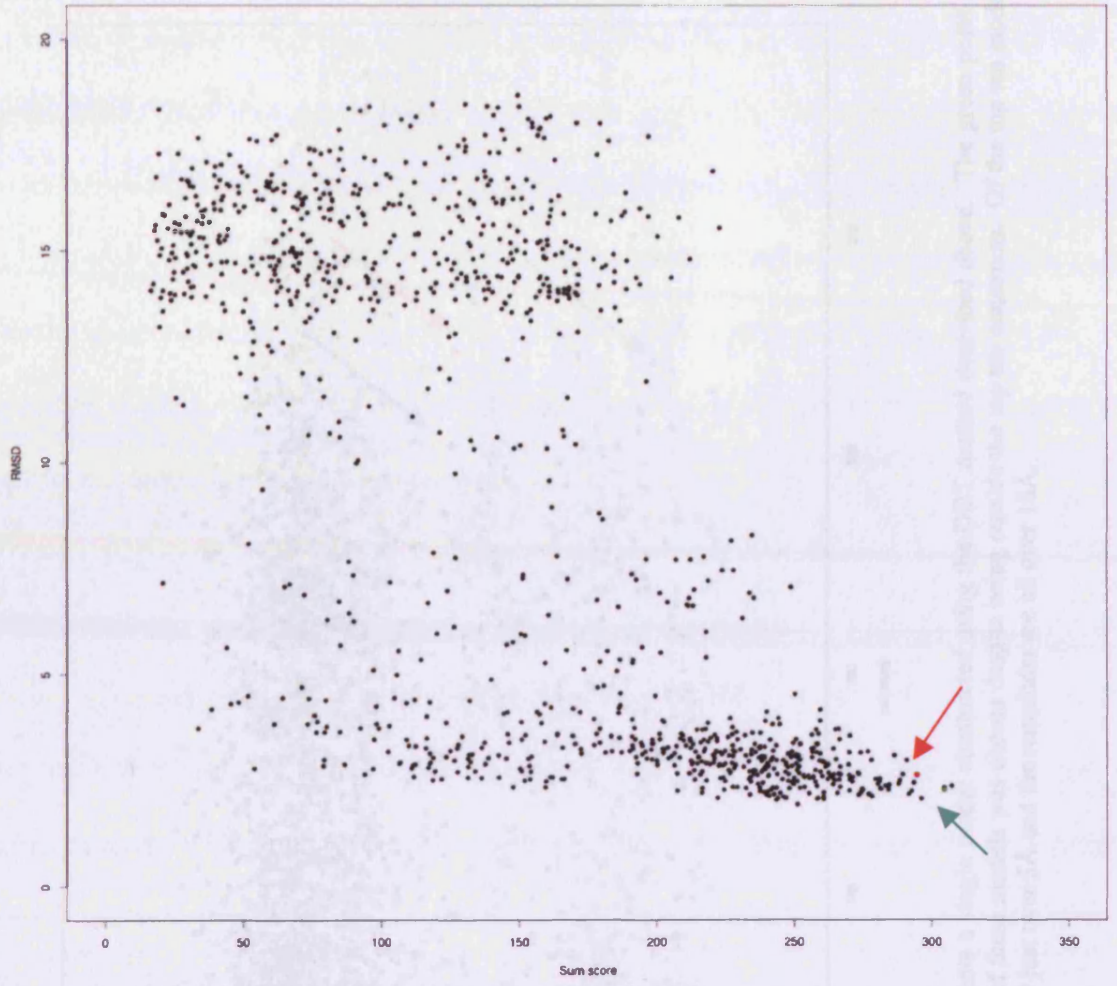


Figure 3.6 CASP Sum Scores for Target T0231: Target T0231 was a comparative modelling target, 142 residues in length. Each dot represents one of the models generated by the DDT pipeline. The red dot represents the 'best' structure submitted to CASP6, while the green dot represents the second 'best' structure submitted to CASP6. Both models were approximately 2.5Å from the native structure.

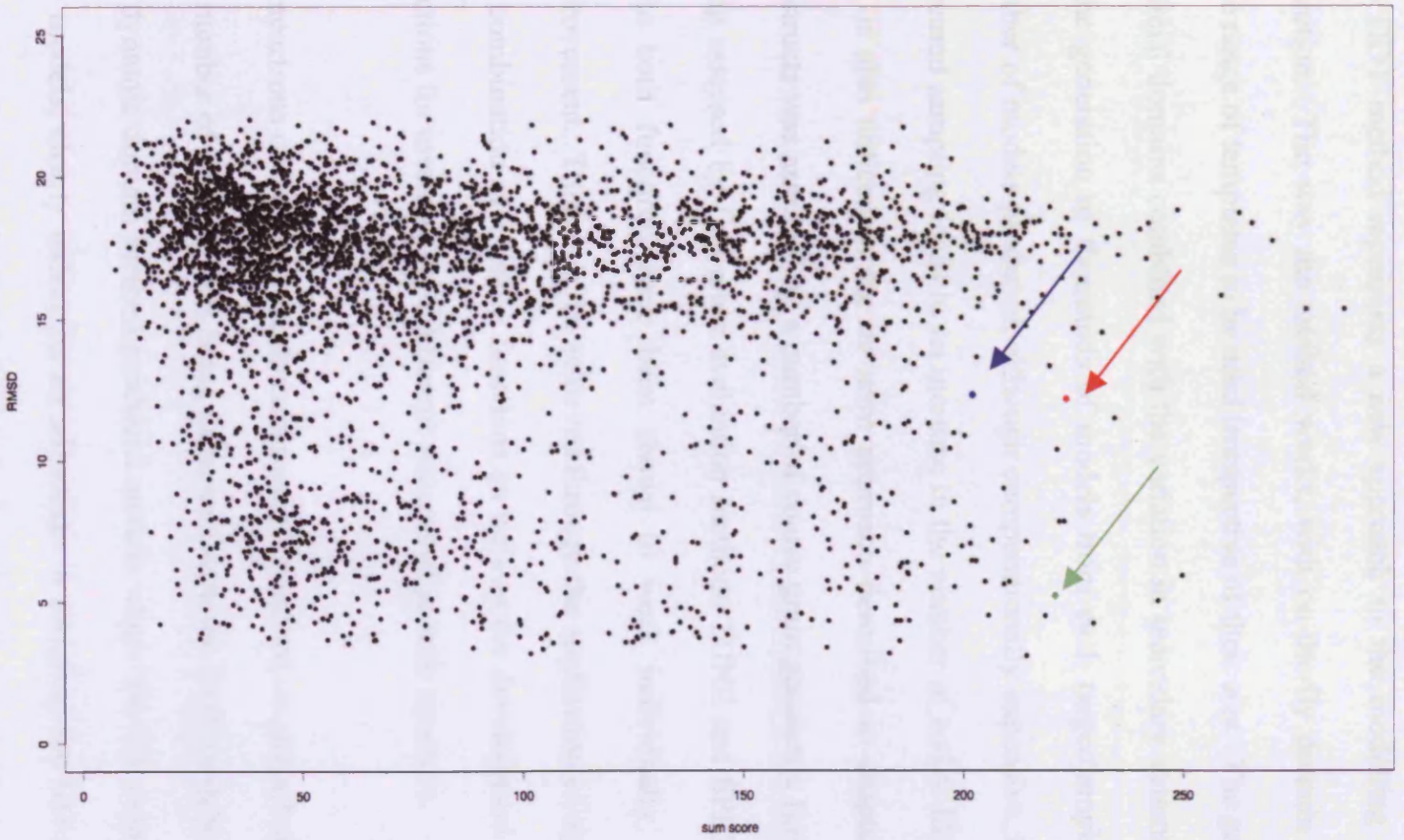


Figure 3.7 Sum Scores for T0223: Each point represents a single model constructed using the DDT method described above. The green point is the first model submitted to CASP, red is second and blue third. Each of these models was chosen despite being outside the top ten constructs. Of the top ten models the first six all have RMSDs over 10Å, the seventh model has an RMSD just over 5Å and the remainder are all over 10Å.

Conclusion

The DDT method represents a new approach to the modelling problem of domain definition. The way the method works, with on-the-fly domain definition, allows a wide range of templates to be used irrespective of their size. The generation of so many artificial domains combined with the variation in secondary structure prediction results in the generation of thousands of models from each target/template alignment. The number of models produced, although computationally expensive, has an advantage, as increased sampling leads to an increase in the number of native-like models – a feature that is also noticed in the *de novo* approach described in chapter 5. Each of these constructs was assessed by a number of coarse grain geometric functions before finally being assessed by fine-grain evaluation methods TUNE and SPREK. It is clear that, while both functions have been shown to work individually, there is scope for improvement. This could be achieved through the application of algorithms to optimise the combination of existing functions as well as the development of novel evaluation functions for assessment of different features of protein structure.

Comparisons of the dynamic domain method switched on and off showed an increase in the number of good models being generated with the DDT switched on. In some cases the dynamic domain method produced models where the SD method failed to produce any models, clearly indicating an advantage in adopting the DD method over the SD method. One drawback of the method is that it is not ideal for modelling close relationships as the random component, designed to introduce variation, draws the model away from the template, which helps to explain the performance in the comparative modelling area of CASP 6.

While the DDT method is an improvement over the SD method, it is also abundantly clear from this work that two areas needed to be improved: template identification and model evaluation. Having already attempted to utilise a measure of solvent accessibility to evaluate multiple sequence alignments and identifying the 'empty quarter' (see methods and chapter 2), the decision was made to examine the possibility of improving the DDT method through the design and application of a novel model evaluation function.

Chapter 4

The Construction and Evaluation of Protein Models: Phobic

Introduction

Where *Ab initio* structure prediction pipelines allow single structures to change over time, an alternative approach is to generate many static "snap-shots" of possible structures for the protein and then and pick the best one. The later approach is referred to as combinatorial modelling and was demonstrated in chapter 3. The combinatorial approach requires a vast amount of compute time and, unlike *ab initio* modelling, a number of evaluation functions for the selection of good models from the ensemble. There are two types of evaluation function: physical and knowledge based, of course, combinations of the two are common as demonstrated by Simons and Baker (Simons et al., 1999b).

Physical scoring functions (also called potentials of mean force and effective energy functions) are based on the thermodynamic hypothesis. This theory postulates that the native state of a protein is the state of lowest free energy under physiological conditions. The aim of physical scoring functions (PSFs) is to capture the relevant free energy components that contribute to the overall stability of a protein in a native state compared to misfolded or unfolded conformations. The most important contributions in PSFs comes from intra-molecular bonded and non-bonded energy terms as well as the free energy of solvation in the aqueous solution. One of the most popular approaches is the Molecular-Mechanics-Poisson Boltzmann/Surface area (MMPB/SA) method developed by Srinivasan et al (Srinivasan et al., 1998):

$$\Delta G = \Delta G_{mm} + \Delta G_{PB} + \Delta G_{sa} - T\Delta S \quad (4.1)$$

where ΔG_{mm} is the internal protein energy derived from a molecular mechanics force field; ΔG_{PB} is the polar contribution to free energy of solvation obtained as a solution to the Poisson or Poisson-Boltzmann equation; ΔG_{sa} is the hydrophobic contribution to free energy of solvation from the solvent accessible surface area obtained by equation 4.1; $T\Delta S$ is the relative protein entropy.

$$\gamma \cdot SASA + b \quad (4.2)$$

where $\gamma = 5.42$ cal mol^{-1} x Å2 and $b = 920$ cal mol^{-1}.

The $T\Delta S$ term is expensive to calculate and has been found to vary little for similarly compact proteins, and as such is regularly ignored (Feig and Brooks, 2002). Further to this, it is common practice to estimate free energies of solvation by more empirical implicit solvent models that produce relative free energies:

$$\Delta G = \Delta G_{mm} + \Delta G_{solvation} \quad (4.3)$$

These models are parameterised to fit experimental data (Wesson and Eisenberg, 1992). PSFs have proven to be successful at discrimination of native and non-native-like folds on standard decoy sets (Lazaridis and Karplus, 1999) as well during CASP exercises (Feig and Brooks, 2002). Despite this misleading simplicity, PSFs require full atomic models and considerable compute time. A poor historical record, through misinterpretation of data, has led to a decline in their use in favour of knowledge based potentials (Novotny et al., 1984, Lazaridis and Karplus, 2000). Knowledge based potentials (also called pseudo-potentials) are an alternative to PSFs. They are derived

from two sources: observed pairing frequencies of amino acids in databases of protein structures and approximations & assumptions about the physical processes that these quantities measure (Thomas and Dill, 1996). The idea was proposed by Tanaka and Scherga in 1975 (Tanaka and Scheraga, 1975) and later developed by Miyazawa & Jernigan in 1985 (Miyazawa and Jernigan, 1985) before taking a step forward in 1990 when Manfred Sippl developed his Potentials of Mean Force (PMFs) (Sippl, 1990). Since then methods have expanded to include various terms, from hydrogen bonding to contact number and solvent accessibility.

This chapter introduces a novel knowledge based statistical scoring function that assess the hydrophobic packing of protein models based on predicted and observed patterns of hydrophobicity. The scoring function, called Phobic, is shown to outperform methods used in current prediction pipelines and effectively discriminate low RMSD models from non-native models as well as native structures from ensembles of models (Taylor et al., 2006, Jonassen et al., 2006).

Methods.

Structure Data

Sequence and structure information was obtained from the PDB25 (Hobohm et al., 1992) and the Astral domain database. The PDB25 is a list of proteins that share a maximum of 25% sequence identity between any two sequences in the list, it can be obtained from the Imperial Cancer Research Fund[4]. The Astral database (Brenner et al., 2000) is a compendium of SCOP allowing direct access to sequence information from SEQRES and ATOM records as well as full three dimensional coordinates. From this set all structures that were less than sixty residues in length or were not identified as belonging to SCOP classes α, β, α+β or α/β were removed, resulting in a dataset of 1852 proteins.

Measuring Solvent Accessibility from All Atom Structures

The solvent accessible surface area (SASA) of an amino acid indicates how buried or exposed it is. There are two ways of expressing this value - absolute solvent accessible surface area (ASA) and relative solvent accessible surface area (RSA). RSA is the most convenient method as it defines the ratio of the surface exposed to the solvent ($SASA_i$) and the maximum solvent accessible surface for a particular amino acid (Max_i) as shown in equation 4.4.

$$RSA_i = \left(\frac{SASA_i}{Max_i}\right) \cdot 100 \qquad (4.4)$$

[4] www.bmm.icnet.uk/loop

The RSA was calculated using two tools, DSSP (Kabsch and Sander, 1983) and NACCESS (Hubbard, 1993) both of which require full atomic models. NACCESS calculates the RSA and ASA for each residue by default while DSSP only calculates the ASA. To obtain RSA values for DSSP the maximum solvent accessibility values of Ahmad (Ahmad et al., 2004a), established using an extended Ala-*x*-Ala tri-peptide conformation, were used. Comparison of the NACCESS and DSSP values yielded a correlation coefficient of 0.98.

Estimating Solvent Accessibility from Cα Chains

The prediction pipelines produce Cα-only models which are later completed using a program like SCWRL (Canutescu et al., 2003). To obtain a measure of solvent accessibility at the Cα level heuristic tools are required. These tools provide an accurate and rapid estimation of solvent accessibility which would otherwise not be possible.

In this work two heuristic methods were applied: POPS-R (Cavallo et al., 2003) as described in chapter two and SACAO (*S*olvent *A*ccessibility from *C*ont*A*ct *O*rder), a new method (described below) that relies on contact number to estimate solvent accessibility from pseudo-Cβ models (Lin, K and Klose, D, unpublished).

SACAO

SACAO is based on the fact that contact number and solvent accessibility are well correlated (Hamelryck, 2005). SACAO uses two spheres to represent each residue in Cα model. One sphere represents the Cα atom and the other represents a pseudo-Cβ. The sphere placed over the Cα has a fixed radius while the sphere placed over the pseudo-Cβ varies with residue size – smaller residues have a smaller radius (Alanine CH3) and larger have bigger radii (Lysine $(CH_2)_4NH_3^+$). A third 8Å sphere is then placed over the center of the Cβ as illustrated in figure 4.1. Any residues that fall within this sphere are flagged as forming potential contacts. Within this area a true contact is identified where the distances between two pseudo-Cβ (R_d) is less than the sum of their radii ($RC\beta_{i-j}$ plus two times the solvent radius (R_{solv} – typically set to 1.4Å) (equation 4.5). Residues also have to be in 'line of sight' and no further apart that the size of the solvent (R_{solv}). The number of contacts a residue has (Cn_i) is then compared to the maximum possible number of contacts.

$$Cn_i = \sum_{i=1}^{j} RC_{\beta_i} + RC_{\beta_j} + 2R_{solv} \qquad (4.5)$$

The value returned is scaled into the range 0 ... 1, where 0 infers total burial and 1 total exposure.

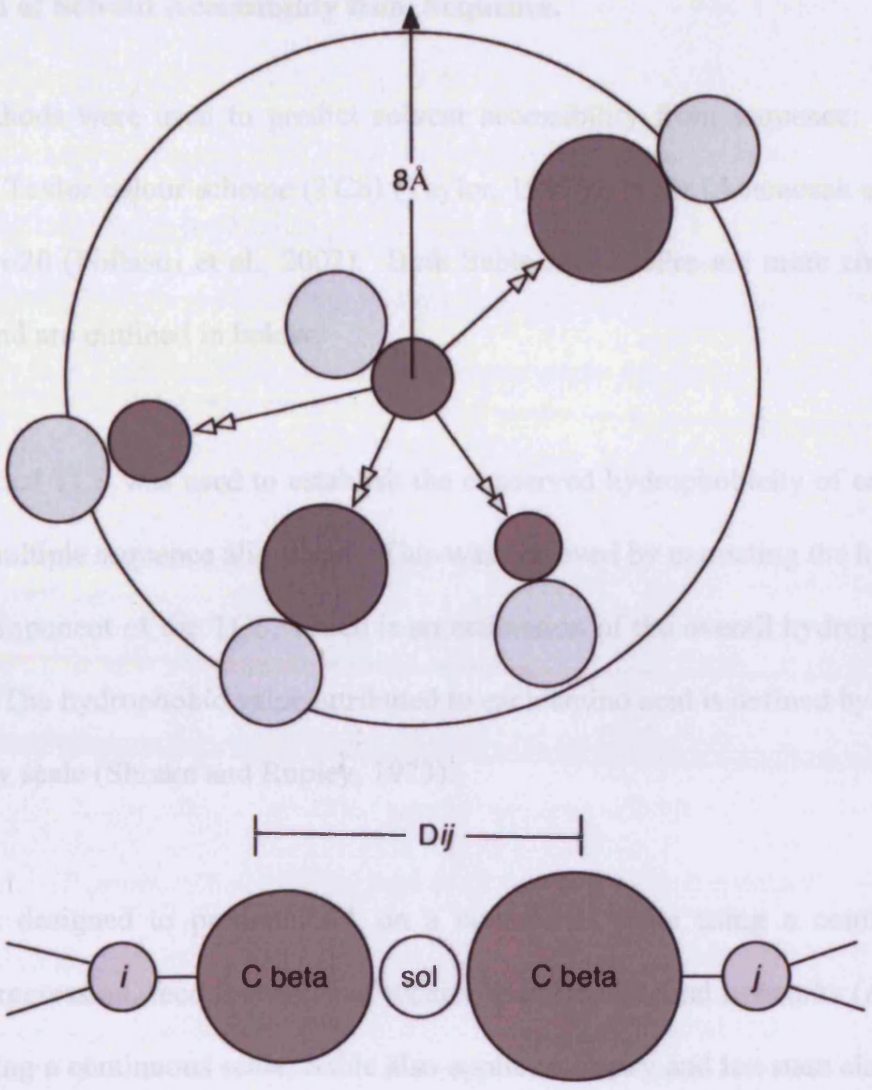


Figure 4.1 SACAO: A schematic representation Each amino acid is represented by two spheres, one placed on the C_α the other place on the C_β. The sphere placed over the C_α is of fixed size while the radius of the C_β reflects the size of the amino acid side. A third sphere of 8Å is then placed over the center of the C_β. All amino acids within this sphere are assigned as potential contacts. True contacts are identified where the distances between two pseudo-C_β is less than the sum of their radii plus two times the solvent radius.

Prediction of Solvent Accessibility from Sequence.

Three methods were used to predict solvent accessibility from sequence: a modified version of Taylor colour scheme (TCS) (Taylor, 1997b); Sable (Adamczak et al., 2004) and AccPro20 (Pollastri et al., 2002). Both Sable and AccPro are more complex than the TCS and are outlined in below.

The modified TCS was used to establish the conserved hydrophobicity of each column within a multiple sequence alignment. This was achieved by extracting the hydrophobic (green) component of the TCS, which is an estimation of the overall hydrophobicity of a column. The hydrophobic value attributed to each amino acid is defined by the Sharke and Rupley scale (Shrake and Rupley, 1973).

Sable was designed to predict RSA on a continuous scale using a combination of nonlinear regression, feed forward and recurrent artificial neural networks (ANNs). As well as using a continuous scale, Sable also applies a binary and ten state classification. For the purpose of this work the ten state classification was used. Each class corresponds to a bin such that class 0 encompasses all residues within 0...10% solvent accessibility while class 9 represents all residues with 90...100% accessibility. The artificial neural networks were trained on a dataset of 860 protein structures derived from the PFAM database and tested on a set of 603 non-homologous protein structures from the PDB. In the 860 protein dataset the optimal prediction threshold, the point where 50% of residues are buried or exposed, was 17%. The original publication reports an accuracy of 77% at the 25% threshold which represents a slight bias toward buried residues, theoretically making prediction easier. The most important part of the system is the input, like many other machine learning methods, Sable, uses information

extracted from position specific scoring matrices (PSSMs). The PSSMs are derived from a PSI-BLAST search against unfiltered versions of the SWISS-PROT and non-redundant database. The profiles form the base of the feature vectors, using a sliding window of 11 residues, the residue of interest is located at position 6. To the basic 11 residue window, measures of average hydrophobicity, volume and entropy are added as well as a binary vector which describes the secondary structure propensities for the central residues and its two immediate neighbours.

AccPro is one half of a web server for prediction of contact number and solvent accessibility. In this work only solvent accessibility was used. AccPro addresses solvent accessibility solely as a binary classification problem (buried or exposed) and is achieved using bi-directional artificial neural networks (BRNN). The theory is that BRNNs are less prone to over fitting than feed forward neural networks. AccPro was trained on a larger set of protein structures than Sable, however this set represented a slightly lower sequence identity threshold of 22%. The 1008 proteins in the set were split using a three fold cross-validation protocol and twenty different classification schemes were applied. Each scheme represents a 5% increment in solvent accessibility and covers thresholds from 0...95% exposure. The input to the system is again important, having reaching implications for the Phobic function. Like Sable, AccPro uses a PSI-BLAST PSSM as the base on which feature vectors are constructed. Unlike Sable, a COILS and SEG filtered version of the nrdb was used with a *E*-value threshold of 10^{-10} per iteration and a final threshold $E < 10^{-3}$ for inclusion into the multiple sequence alignments, the number of iterations are fixed to three. The profiles generated in this step are then used to scan SWISS-PROT, TrEMBL and PDB before each sequence is weighted using the following scheme:

$$W(s) = -\sum_{c} \log P[s(c)]$$

where $P[s(c)]$ is the probability of letter s in profile column c. Unlike Sable, no additional features, such as sequence entropy, are added to the input vectors.

Method Combinations – Corners & Phobic.

When examining native structures, structures present in the PDB, certain patterns are expected. One such pattern is the exposure of hydrophilic residues on the surface of the protein and burial of hydrophobic residues in the core (chapter 1). This is a gross-oversimplification as, in an Orwellian sense, all residues are hydrophobic but some residues are more hydrophobic than others.

One problem with prediction of structure, from a theoretical perspective, is the lack of information available about the target. For instance, the location of each residue within a 3D structure, while it is possible to identify residues as hydrophilic (see chapter 2) this gives no idea of how buried, or exposed, it may be in a native structure. The solution to this problem is to predict solvent accessibility – here this is achieved using Sable and AccPro. The prediction forms the foundation of what is expected to be 'true' and deviation from it is assumed to be 'bad'. The following sections describe the development of the Phobic scoring function (Taylor et al., 2006, Jonassen et al., 2006) through a method, Corners, derived from the "Empty Quarter" concept introduced in chapter 2.

Corners

Corners is based on the 'Empty Quarter' (chapter 2) however, instead of using the Taylor Colour Scheme, it leverages the power of the Sable neural network to estimate the relative solvent accessibility for each residue in a sequence. The predicted solvent accessibility was compared to structurally observed solvent accessibility using the POPS-R tool for a set of native and non-native (decoy) structures. Decoy structures were generated using the methods described in (Taylor et al., 2006, Taylor, 2006). Comparison of the native and non-native (decoy) structures reveals two points of interest: first, residues which are predicted as being exposed are more often found to be buried in decoy than in native structures; second, residues which are predicted as exposed are more prone to be buried in decoy than in native structures.

The aim of Corners was to exploit the observed differences by defining two planes which optimally separate native and non-native structures based on this pattern. In practice this is a function that takes the sum of the positions in each zone penalising structures which have more residues falling into either area. This means that Corners completed a similar function to the Burial/Hydrophobic matching described in Chapter 2 and (Taylor et al., 2006) with the addition of a contribution from residues predicted buried but exposed in Cα models. Corners was a stepping stone towards Phobic and as such will not be discussed in further detail.

Phobic

The Phobic scoring function (Taylor et al., 2006, Jonassen et al., 2006) is based on the same dataset as Corners but exploits two different tools to achieve a similar function. In place of POPS-R the SACAO tool (described above) was used to estimate the solvent accessibility of individual amino acids. To predict solvent accessibility AccPro took the place of SABLE as it is able to predict binary state exposure at 5% increments (SABLE predicts at 10%).

The original AccPro method is described above, however to be used in this work modifications had to be made to the code. AccPro is a 'black-box', training and testing was performed outside the laboratory and so and all potential influences of homology had to be removed prior to application. These changes occurred in the generation of the input profiles passed to the ANN. The original method uses PSI-BLAST to scan the target against the nrdb which has been filtered for low complexity, trans-membrane and coiled-coil regions. This scan produces a sequence profile which is then used to scan the TrEMBL, Swiss-Prot and PDB databases before generating a weighted profile. In the modified version, a standard PSI-BLAST search is performed against a filtered version of the nr database. Scanning of other databases, such as SWISSPROT, is prevented so that the overall profile passed to the ANN would not be the same for proteins used in the training set. In addition to this AccPro employs a correction facility. After making an initial prediction, a scan is completed against a local database of sequences and associated structure, if a suitable hit is found the prediction is altered to match information gathered from the database search. With this method switched on, predictions reach 100% accuracy. By removing these two steps the accuracy of AccPro

at the 25% threshold falls to 77% - a similar result to the initial test dataset, indicating that input sequences had not been memorised.

To obtain a prediction of the relative solvent accessibility of each residue AccPro had to be further modified so that a single PSI-BLAST run was completed (instead of one per threshold). At each threshold a binary output, exposed or buried, is presented, when all predictions are combined, they form an output similar to that illustrated in figure 4.2.

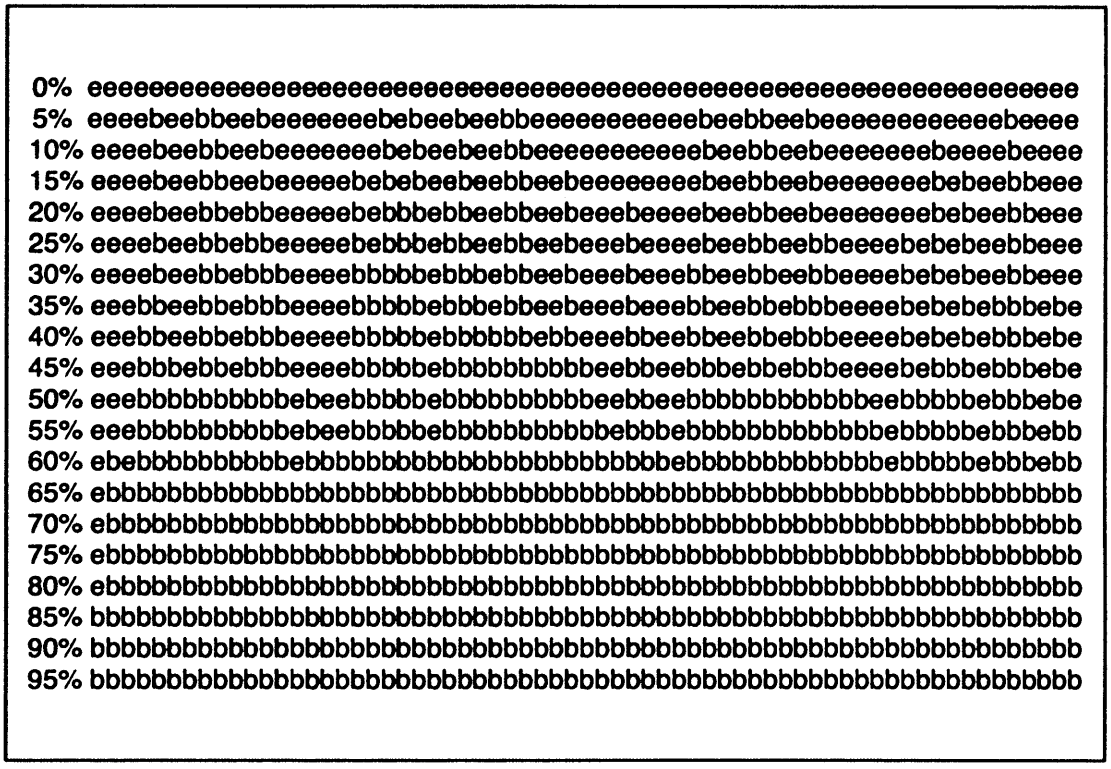


Figure 4.2 Binary Prediction of SA using AccPro: At each threshold a prediction of each residue being buried or exposed is made. When combined across all thresholds a profile can be constructed from which the relative solvent accessibility (RSA) can be estimated. The RSA is taken as the final threshold in which the residue is seen as exposed. For the first residue this yields an RSA of 80% while for the last residue the RSA is 50%. This example was created for illustrative purposes only and does not deliberately relate to any protein in the dataset.

For each residue, the maximum threshold at which the predicted state is exposed is assigned as the RSA. For the fist residue in the residue in figure 4.2 this yields an RSA of 80%, while for the last residue the RSA is estimated as 50%. This process was completed for each of the proteins in the dataset. The same information was extracted from native protein structures using the SACAO tool.

Using the decoy dataset described above, the SACAO solvent accessibility values are calculated and compared to the AccPro output. The combination of AccPro and SACAO for both the native and non-native datasets gives two samples for which there are twenty underlying distributions (one per threshold prediction), these distributions are used to construct the Phobic scoring matrix. For each state the native and non-native distributions are normalised and split into ten discrete bins according to the SACAO value. The division of the SACAO value was based on steps of 0.1 resulting in 10 bins per threshold and a final scoring matrix of 20 * 10. The scores in the matrix are simply the difference between the two distributions:

$$Matrix_{ij} = \sum Native_{ij} - \sum Random_{ij} \qquad (4.7)$$

where $Matrix_{ij}$ is the recall value used in the final matrix, i is the index of the AccPro prediction and j is the index of the SACAO bin. When the matrix is complete it is used as a "look-up" table. When presented with a new target the solvent accessibility is predicted from the sequence, this gives a value for each amino acid which is constant for every predicted model – given that the sequence is constant. Then for every model (~10,000 per target) the solvent accessibility is estimated using SACAO. These values are then used to extract the corresponding 'score' from the look-up table based on the

bin values. The sum of the scores across all residues is used as a measure of fitness, the more positive the score the more native-live the model is.

At this point there had been no effort to distinguish between the four major SCOP classes (α, β, α+β, α/β). It has been shown that all α proteins cause problems for model evaluation functions due to the sheer number of packing possibilities of the helices (Berglund et al., 2004). In an attempt to circumvent this problem four class-specific matrices were generated using the method described above however no overall improvement was observed.

Results & Discussion

There are several components to the Phobic scoring function. This section explains why specific tools were used and how effective each component and combination was.

POPS-R and SACAO

To assess the accuracy of SACAO and compare it to POPS-R, 2000 proteins were selected at random from the PDB40 dataset (all proteins share a maximum of 40% pairwise identity). RSA was obtained from all-atom structures using both DSSP and NACCESS as described previously. The correlation coefficients (cc) for DSSP and POPS was 0.68 while for DSSP and SACAO it was -0.74.

Sable and AccPro

It is difficult to compare two structure prediction tools irrespective of function – secondary structure, solvent accessibility, contact number and so on. There are several problems, each of which are outlined below. First is the dataset – it is rare to find two papers describing prediction tools that are built using the same dataset. By changing the dataset and leaving the method essentially untouched, performance jumps can be gained (Rost, 1996, Jones, 1999b). This is tied to a common method of comparing two unique methods (developed on different datasets) using a set of commonly used sequences such as the Rost and Sander secondary structure set or the Manesh set for solvent accessibility which will be discussed later.

Corners

As mentioned previously, Corners was a stepping stone between the Empty Quarter and the Phobic function, improving on the Empty Quarter, such that it compared favourably against TUNE and SPREK on the forward and reverse models described in chapter 3 (see table 4.1 for results and chapter 3 for a description of model construction and the evaluation procedure). The performance of each method is summarised in table 4.1 which shows the target and the score for each of the method: TUNE; SPREK; Empty Quarter; Corners; Phobic. The smaller the values the better the function performs at distinguishing the native-like models from the non-native.

Table 4.1 Performance of Evaluation Functions on Taylor Derived Decoy Sets

Target	TUNE	SPREK	Empty Quarter	Corners	Phobic
1A0A_A	3	0	0	8	7
1A12_A	73	85	0	0	0
1A1W_	22	8	47	16	8
1A2P_A	3	20	13	0	0
1A6M_	15	9	124	15	9
1A7S_	22	38	28	11	16
1AC5_	0	1	0	0	0
1ACF_	25	3	52	1	0
1AEP_	21	0	11	20	10
1AFJ_	12	12	28	10	6
1AGR_E	32	31	1	1	1
1A11_A	78	68	91	1	0
1A1U_	107	130	4	3	2
1AMX	5	34	7	1	1
1AOE_A	2	0	10	0	0
1AQZ_A	0	0	3	4	2
1ASS_	3	0	2	0	0
1ATZ_A	1	0	19	0	0
1AU1_A	27	18	183	1	0
1AUI_A	2	0	61	0	0
1AUY_A	3	33	46	0	0

The above table shows how the original Empty Quarter method, described in chapter 3, did not perform as well as TUNE and SPREK. When comparing the Corners function to SPREK and TUNE the performance is similar, with Corners being equal to and better than both TUNE and SPREK across most proteins. The Phobic function improves on the Corners scores, although not dramatically, with only SPREK outperforming it on three proteins (1A0A_A, 1AEP_ & 1AQZ_A).

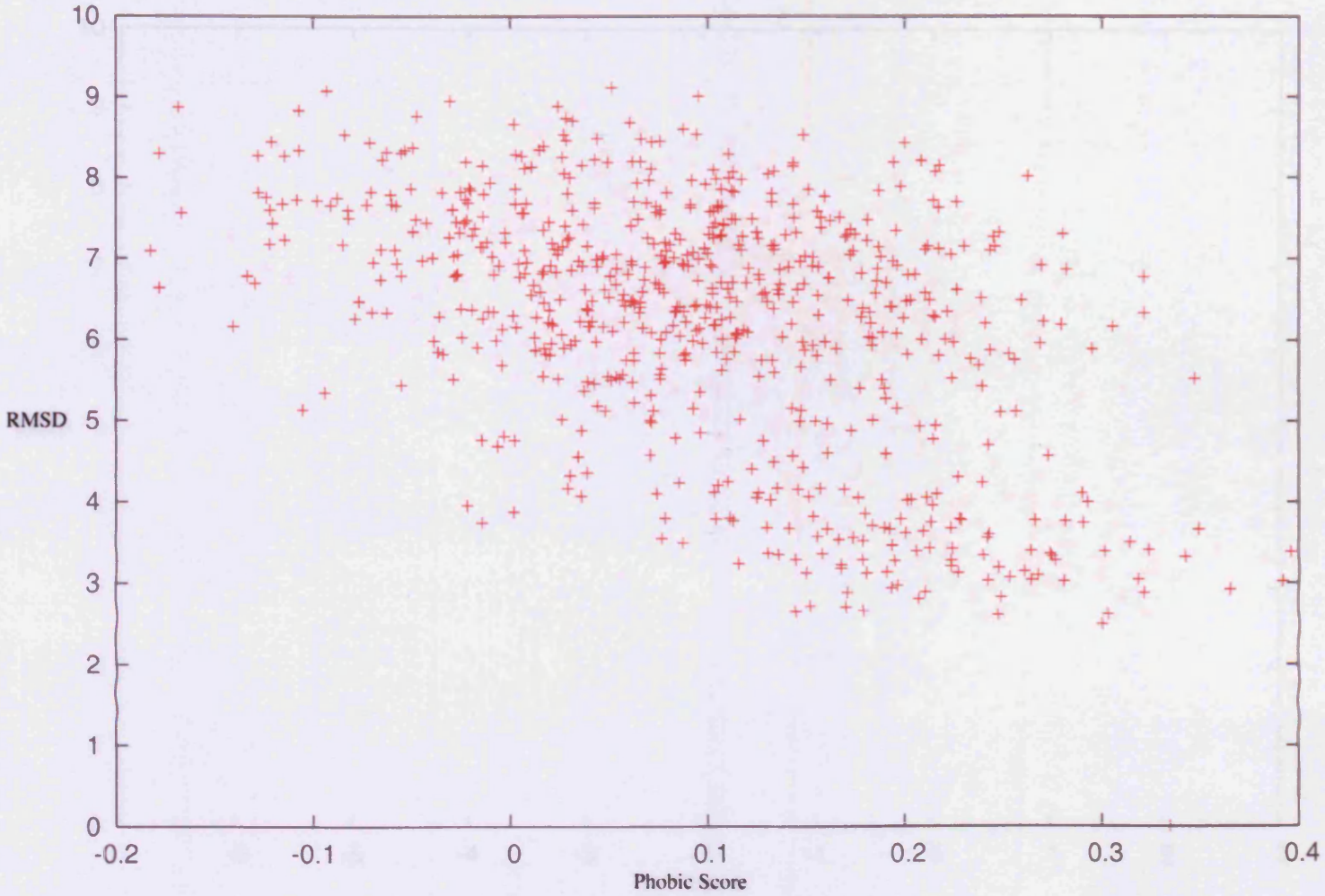
Performance of the Empty Quarter, first described in Chapter 3, is worse than the TUNE and SPREK functions, failing to differentiate between the forward and reverse structures (having large values in table 4.1). Table 4.1 shows that the Corners function improves on the Empty Quarter, probably by taking better account of the buried hydrophilic residues, for all but one of the structures (1AEP). 1AEP is a helix bundle protein, more specifically an apolipoprotein, whose ampipathic nature may help explain why there was a minor decrease in performance. The overall performance is better than that of TUNE and is comparable to SPREK. Lastly is the Phobic function, which performs better than both its predecessors as well as the TUNE and SPREK functions, producing lower scores for the majority of the decoys in the test set.

Phobic

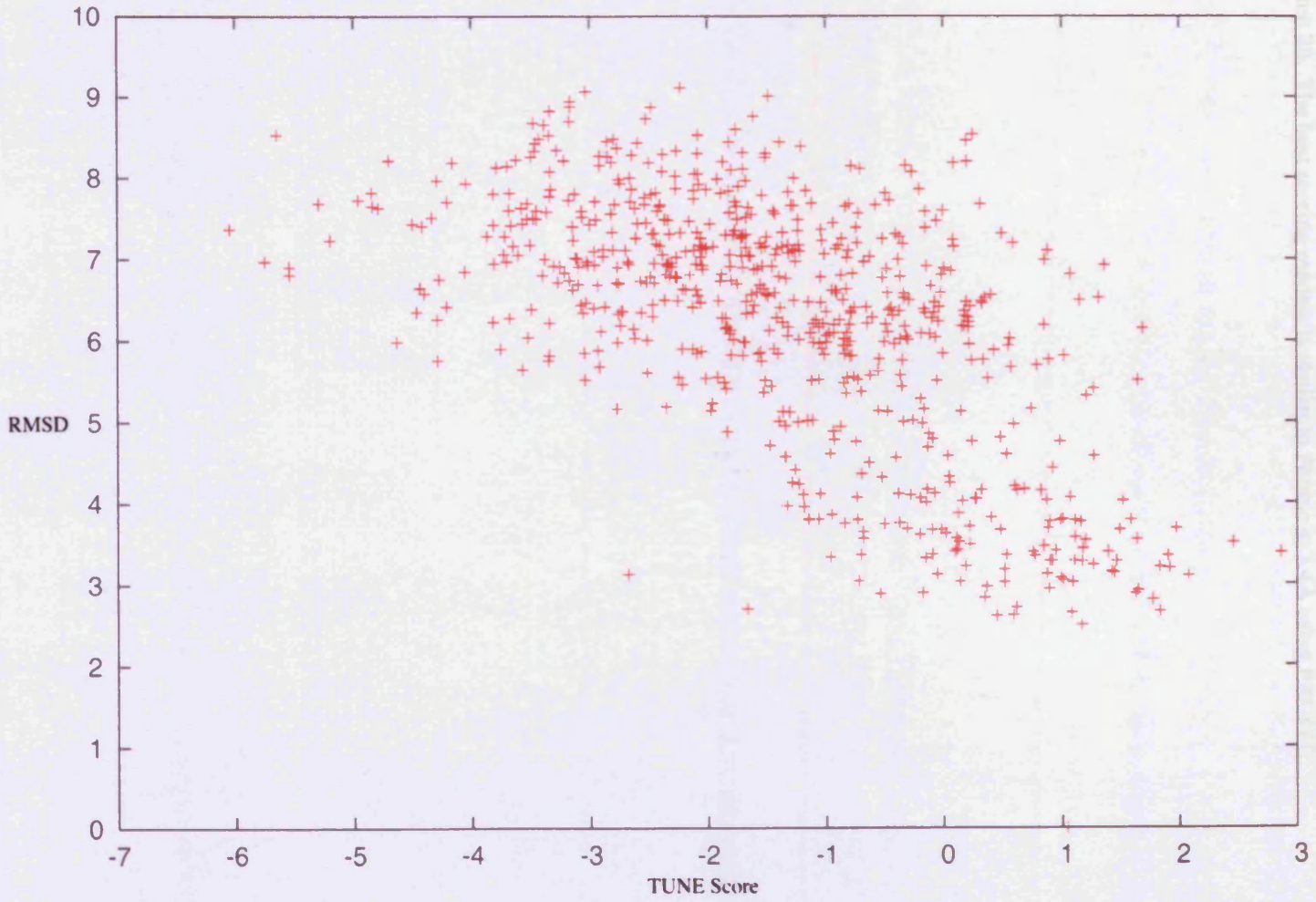
In addition to the above test, Phobic underwent testing on the 4State decoy set of Park and Levitt (Park and Levitt, 1996) and four random proteins from the Rosetta decoy set (Tsai et al., 2003). For reasons that will be explained below, a large number of real threading attempts were also used to analyse the performance of Phobic. So that an accurate analysis could be completed, Phobic was also compared to two evaluation tools currently used in all Taylor group prediction pipelines.

The 4State Decoy Set

The 4State decoy set consists of seven small proteins which cover a number of different folds and classifications. The proteins are all 'small' ranging in size from 54 to 75 amino acids and, as such, are not really large enough to form an extensive hydrophobic core, something that Phobic is designed to look for. Despite this size disadvantage Phobic performs comparably to the re-trained TUNE function described in chapter 3. Figures 4.3-9 show the TUNE and Phobic scores plotted against the cRMSD. Table 4.2 summarises the information, showing the cRMSD of the top three ranking models and if the native structure were identified in the top 10 models.



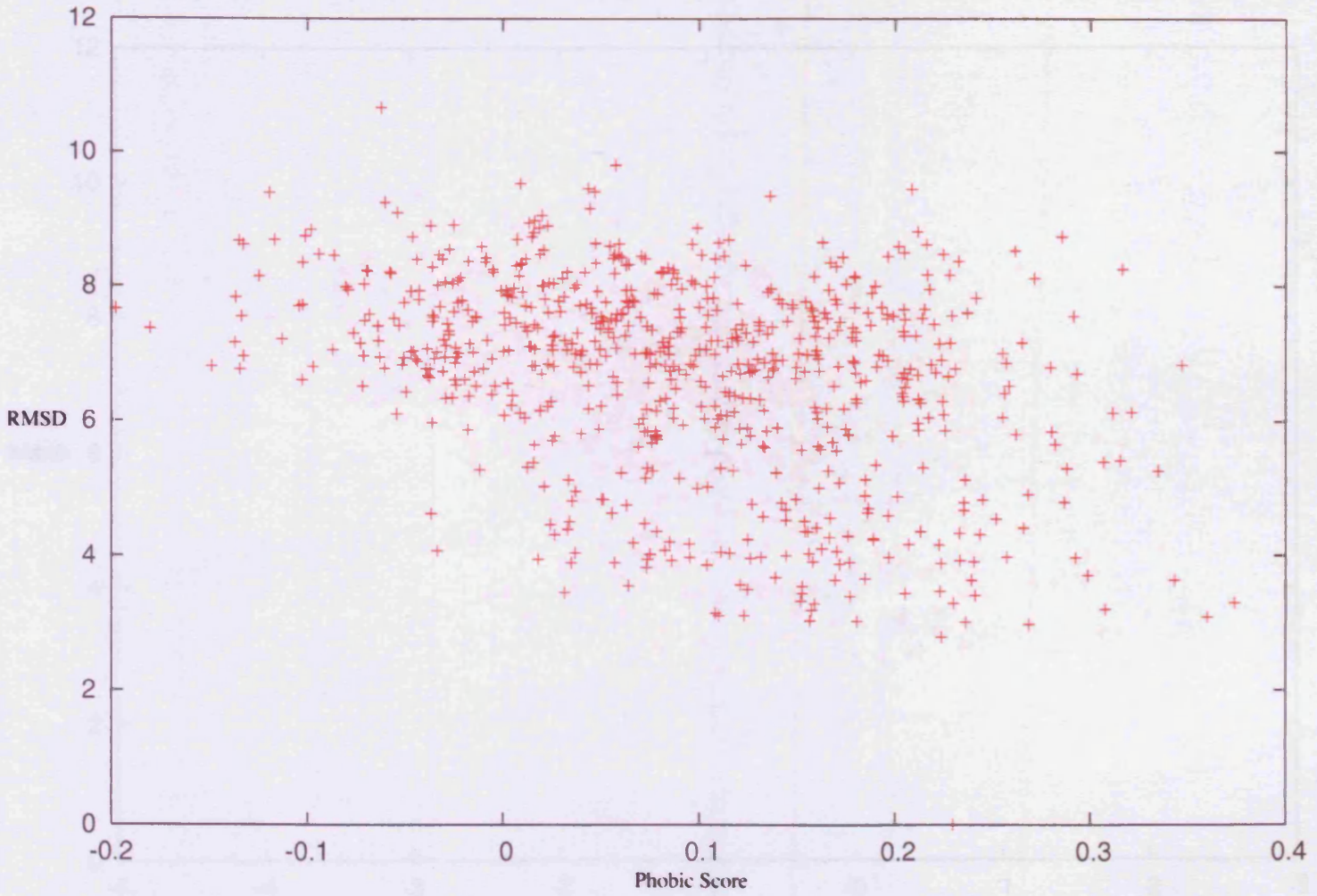
a)



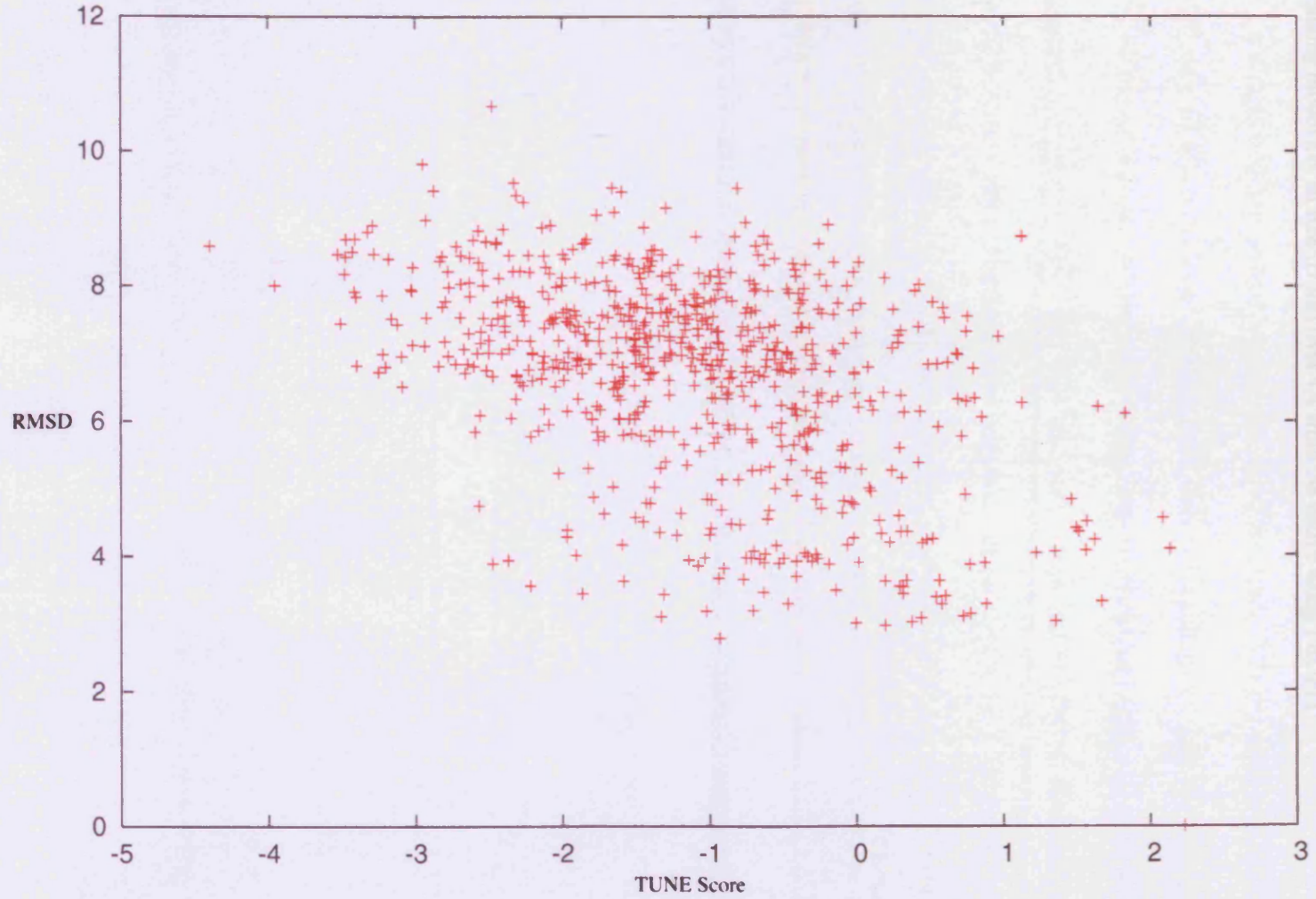
b)

Figure 4.3 Target 4RXN evaluation using Phobic and TUNE: 4RXN, classified as a small protein by SCOP, is 54 residues long. **a)** Phobic does not identify the native structure as the best model, however it is ranked 6th overall. The Top scoring model is 3.4Å from the native structure using PROFIT for a 1:1 structure alignment. **b)** Tune fails to identify the native structure from the ensemble ranking it outside the top 20. The best scoring models is similar to Phobic at 3.4Å using PROFIT.

Rubredoxin, PDB code 4RXN, is classified as a small protein by SCOP. It is 54 residues in length, of which 9 residues are incorporated into α-helical structures and 12 residues into β-strands. For both TUNE and Phobic the native structure is not identified as the 'best' structure in the set, Phobic ranks it at 6th while it is falls outside the top 20 for TUNE. Both functions identify the best model at 3.4Å from the native structure (see figure 4.3).



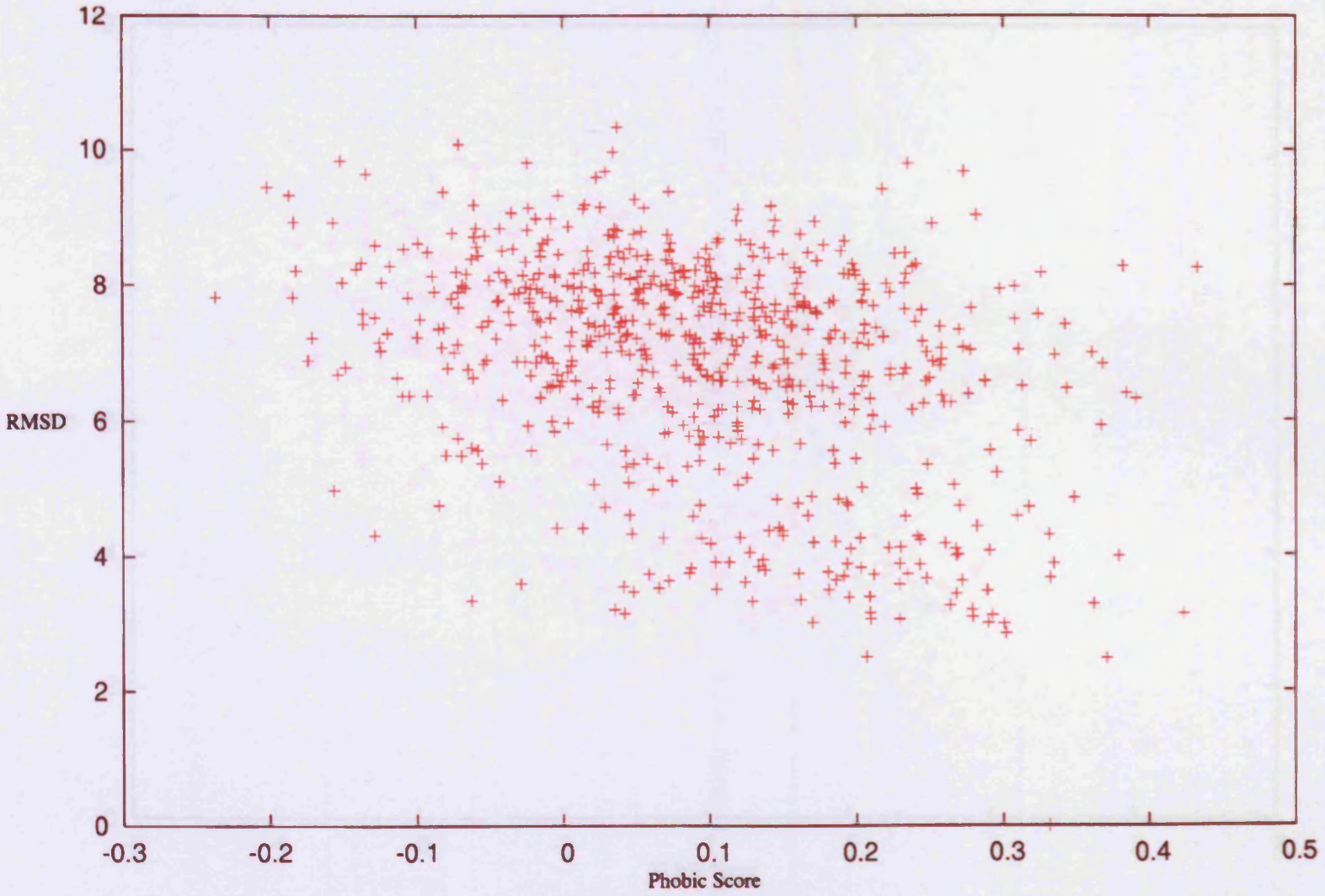
a)



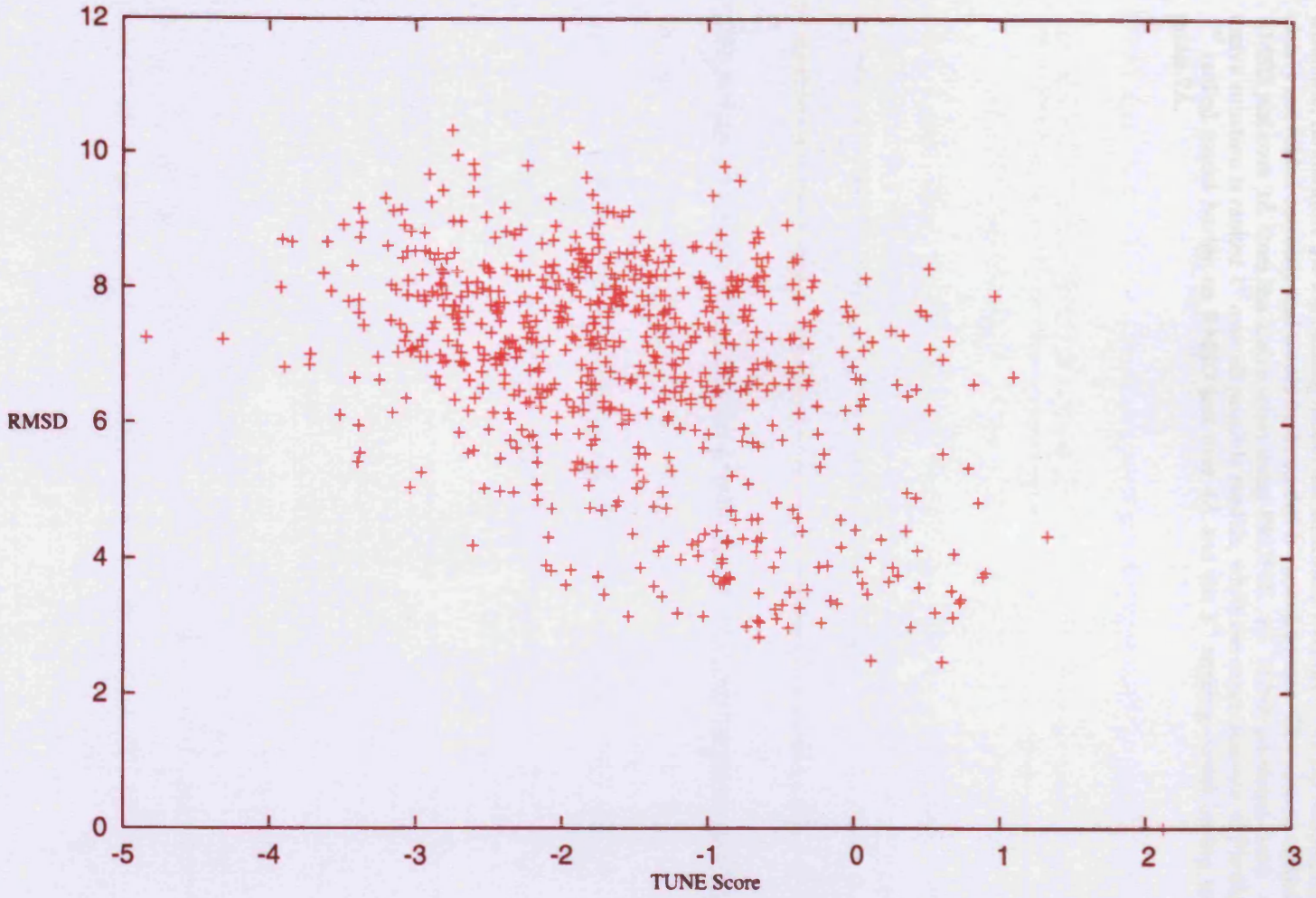
b)

Figure 4.4 Target 4PTI evaluation using Phobic and TUNE: 4PTI, classified as a small protein by SCOP, is 58 residues long. **a)** Phobic failed to identify the native structure, however the top ranking model is 3.2Å from the native using PROFIT. **b)** TUNE ranks the native structure 1st overall but the best ranking model has an RMSD just over 4Å from the native using PROFIT.

The small protein inhibitor of trypsin and trypsinogen (PDB 4PTI) is classified as a 'small' protein by SCOP. It is 58 residues in length having two α-helices composed of 12 residues and 3 β-strands composed of 15 residues. Phobic fails to identify the native structure in the top 10 models while TUNE correctly identified the structure from the ensemble of non-native structures. The best model identified by Phobic was 3.2Å from the native structure while the best structure identified by TUNE had an RMSD of 4Å. Both methods did not separate the ensemble of structures into distinct groups based on their overall similarity to the native structure, however this is not of great concern as any function which can continually identify a low RMSD structure from an ensemble is useful to the prediction pipeline (see figure 4.4).



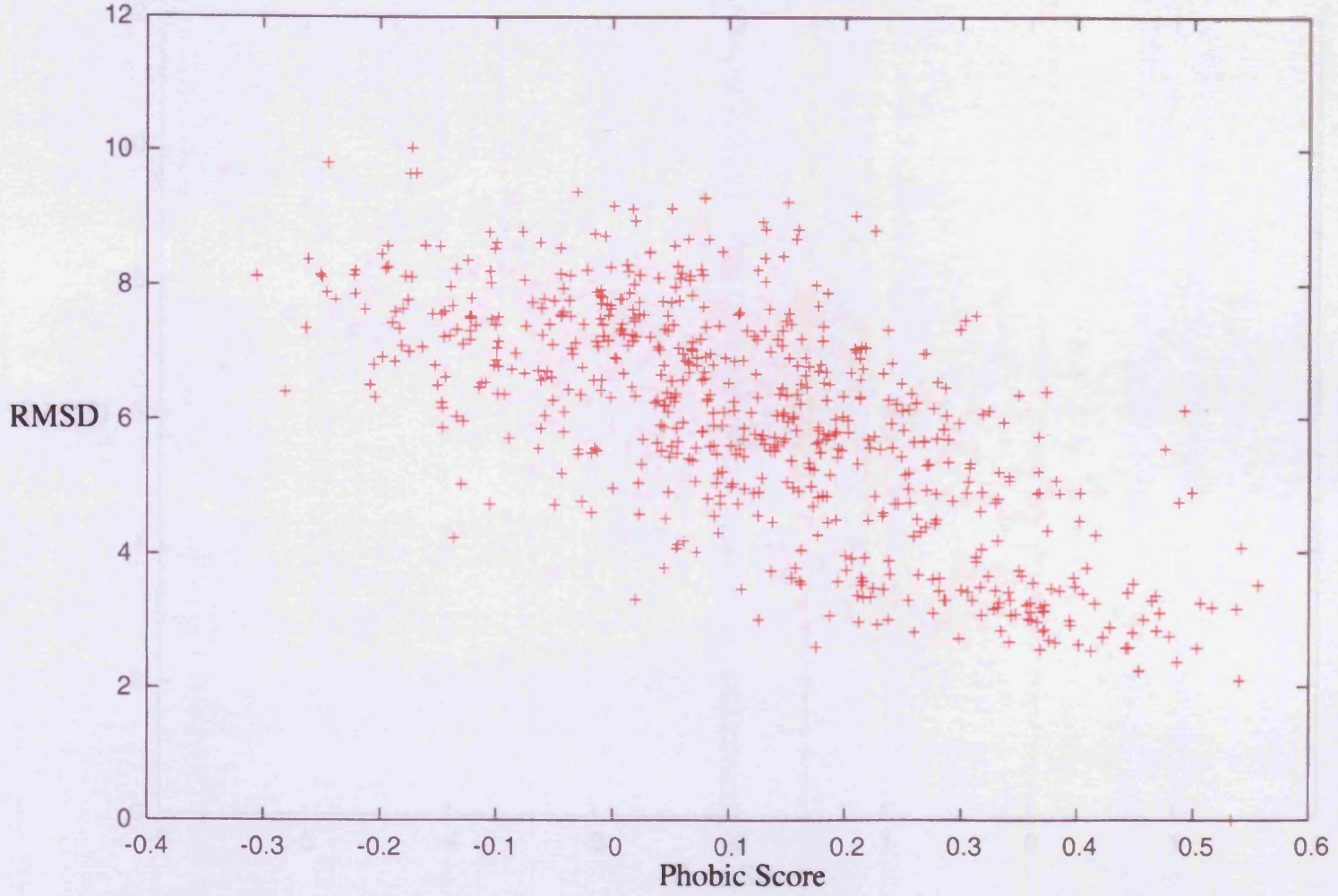
a)



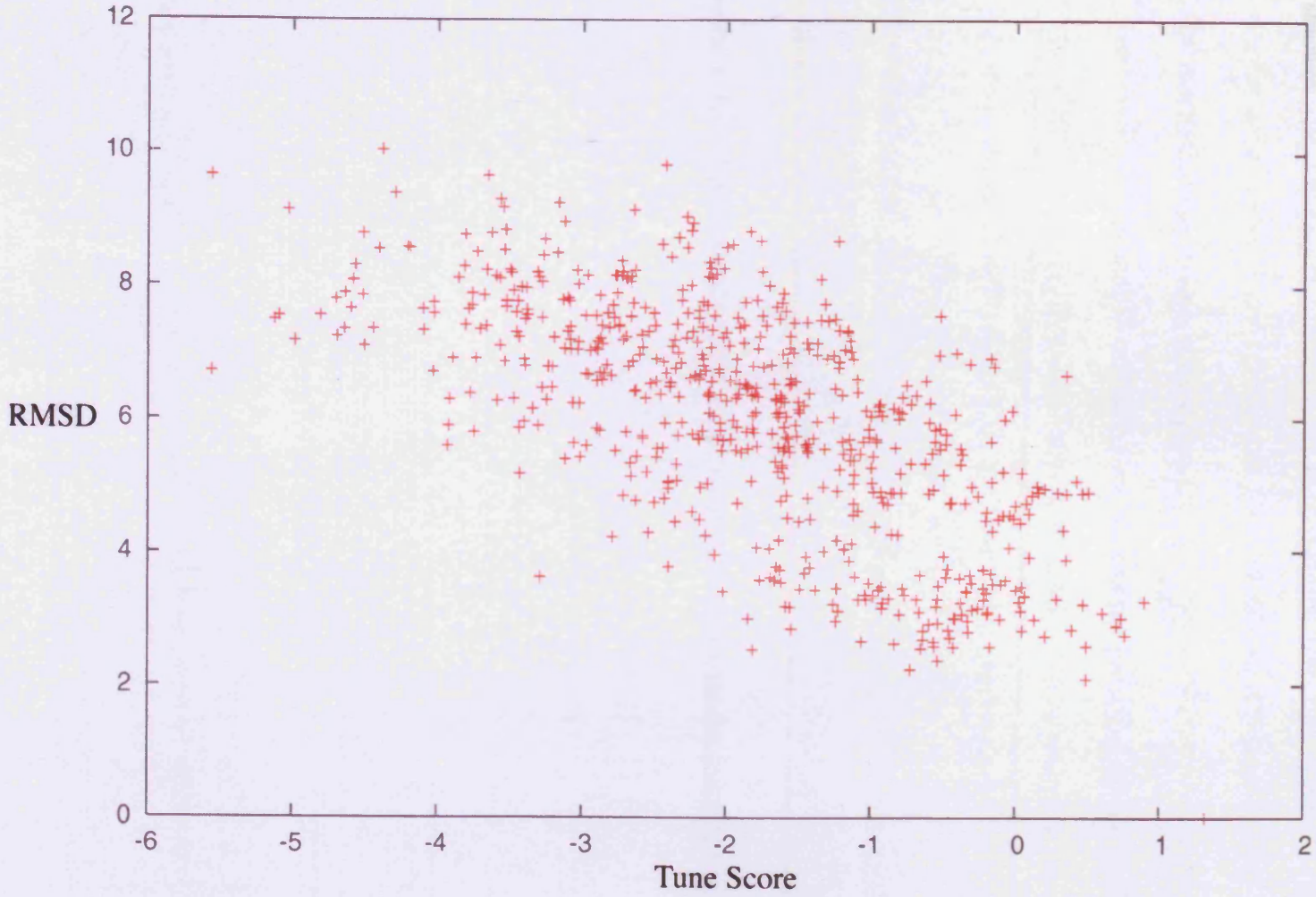
b)

Figure 4.5 Target 1SN3 evaluation using Phobic and TUNE: 1SN3 is the third protein in the set to be classified as a small protein by SCOP. It is larger than 4RXN and 4PTI at 65 residues but proves to be more of a problem for both Phobic and TUNE. a) Phobic fails to identify the native structure from the ensemble, furthermore the top ranking model has an RMSD over 8Å. The top ranking model, however, is clearly not folded correctly and would be discarded if viewed by eye. The second ranked structure has an RMSD just over 3Å from the native when using PROFIT. **b)** TUNE performs better on this target, the native structure is ranked 1st over all possible models, while the exact opposite of Phobic occurs with the 2nd ranked model having an RMSD just over 4Å and the 3rd ranking model having an RMSD slightly under 7Å.

Scorpion toxin variant (1SN3) is the third protein to be classified as "small" by SCOP, meaning that the structure has little or no ordered secondary structure and often no hydrophobic core. . It adopts a knottins fold and is 65 residues in length, composed of 1 α-helix of 8 residues and seven β-strands composed of 16 residues. While larger than 4PTI and 4RXN, 1SN3 posses the most problems for Phobic as the native structure is not identified in the top 10 structures and the highest scoring model has an RMSD over 8Å. An encouraging aspect of this evaluation is that the top scoring model does not look native when viewed by eye, something that is always done with the Taylor pipelines (chapters 3 and 5). If this model is excluded, the best model has an RMSD of just over 3Å from the native structure. The TUNE function is successful at identifying the native structure from the ensemble but, like Phobic, does not perform well at identifying a low RMSD structure from the ensemble, with the top scoring model have an RMSD greater than 4Å. Both functions also discard a number of structures both close and distant from the native where the score is less than 0 (see figure 4.5).



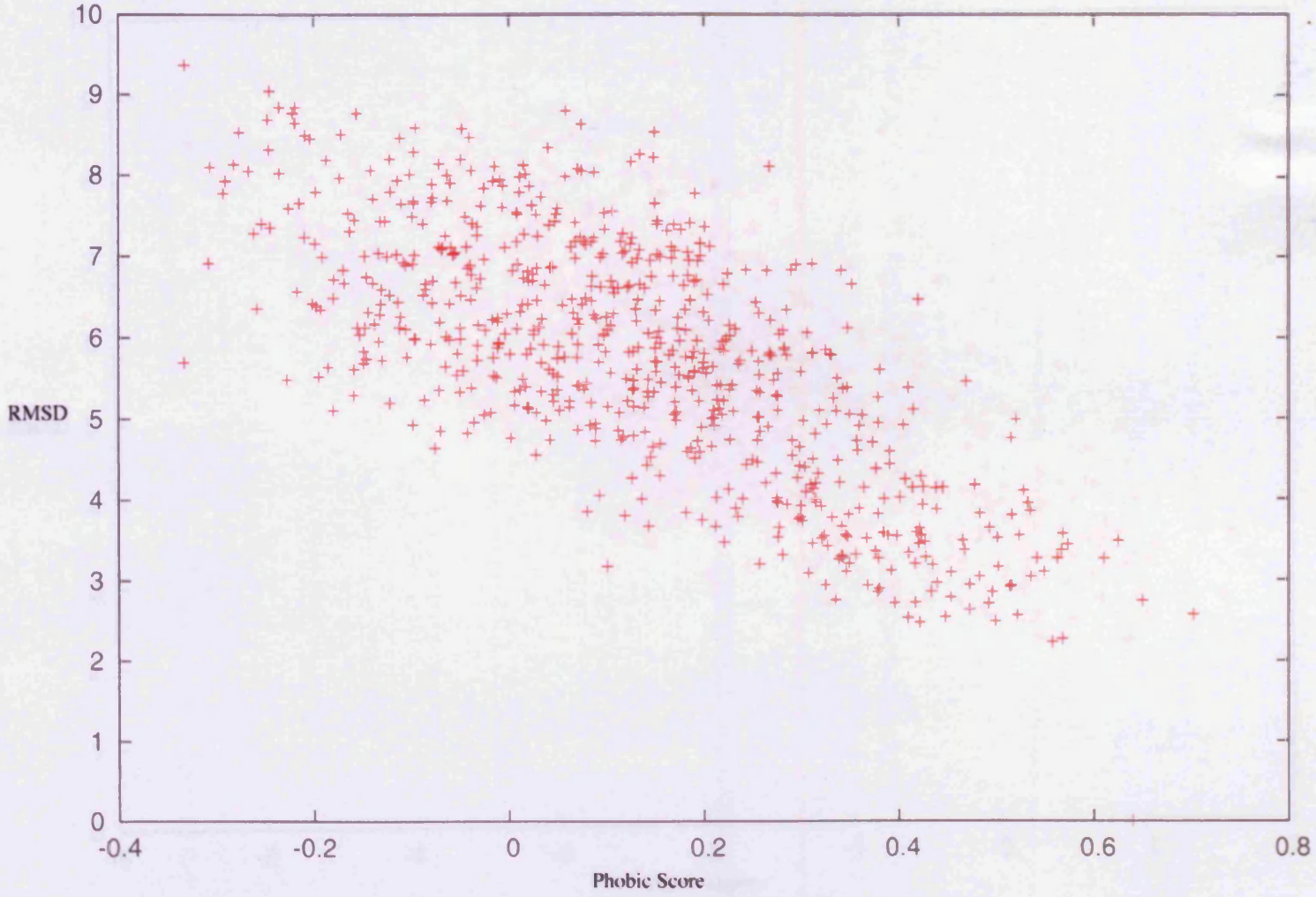
a)



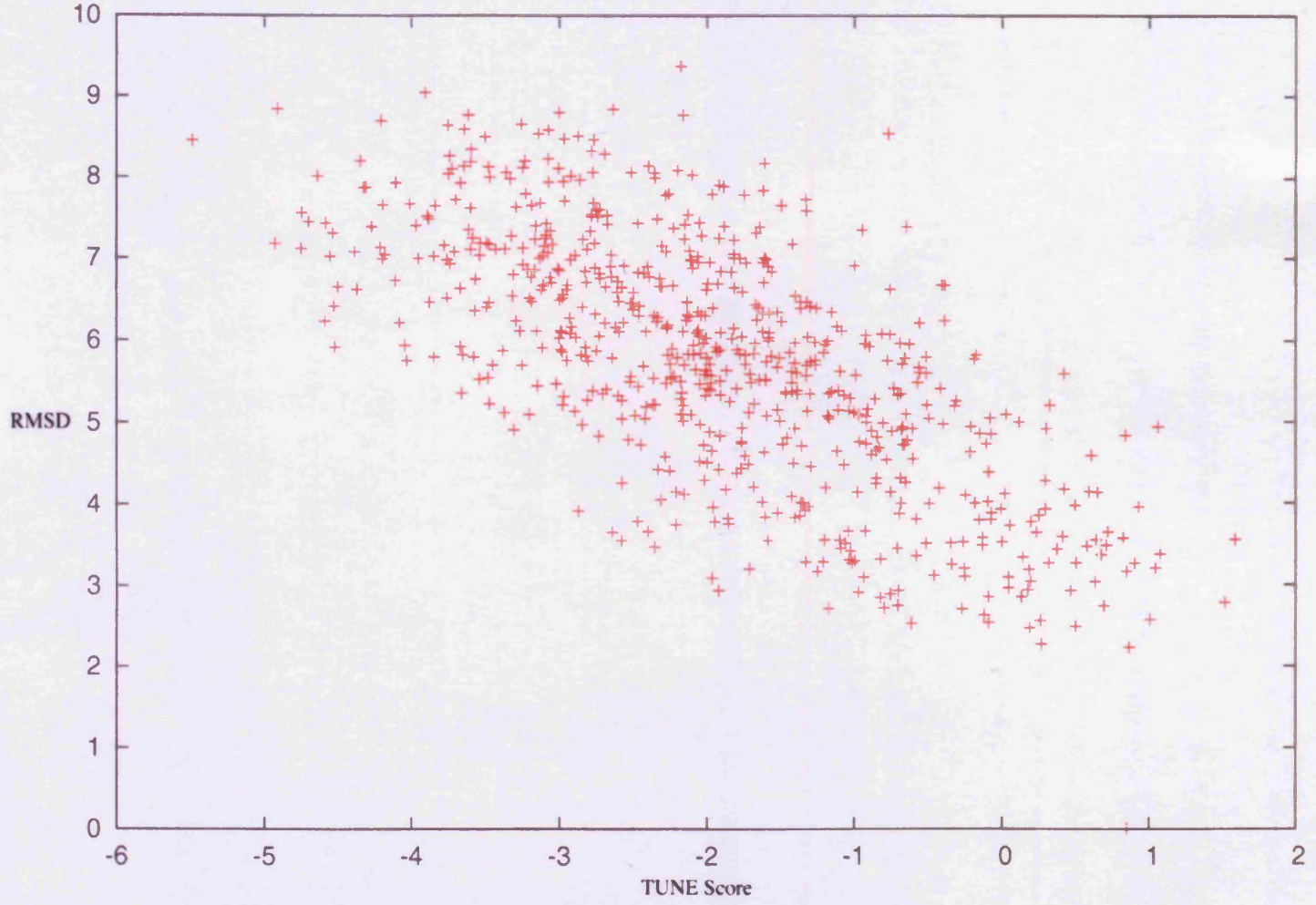
b)

Figure 4.6 Target 1CTF evaluation using Phobic and TUNE: The performance of Phobic (a) and TUNE (b) is comparable for 1CTF, a 74 residue $\alpha+\beta$ protein. The native structure is ranked 1st by TUNE and 4th by Phobic. Both functions have a top scoring model which is approximately 3Å from the native structure.

The C-terminal domain of ribosomal protein L7/L12 (PDB 1CTF) is the largest protein in the 4State set at 74 amino acids. It is classified, by SCOP, as an $\alpha+\beta$ protein and consists of a ClpS-like fold which has 4 α-helices and 4 β-strands consuming 75% of the residues in the protein. The native structure is the 4^{th} highest scoring model behind three structures between 4 and 2Å. The performance of TUNE is comparable to Phobic, with TUNE identifying the native structure ahead of all models and the top 5 scoring structures having RMSDs around 3Å from the native (see figure 4.6).



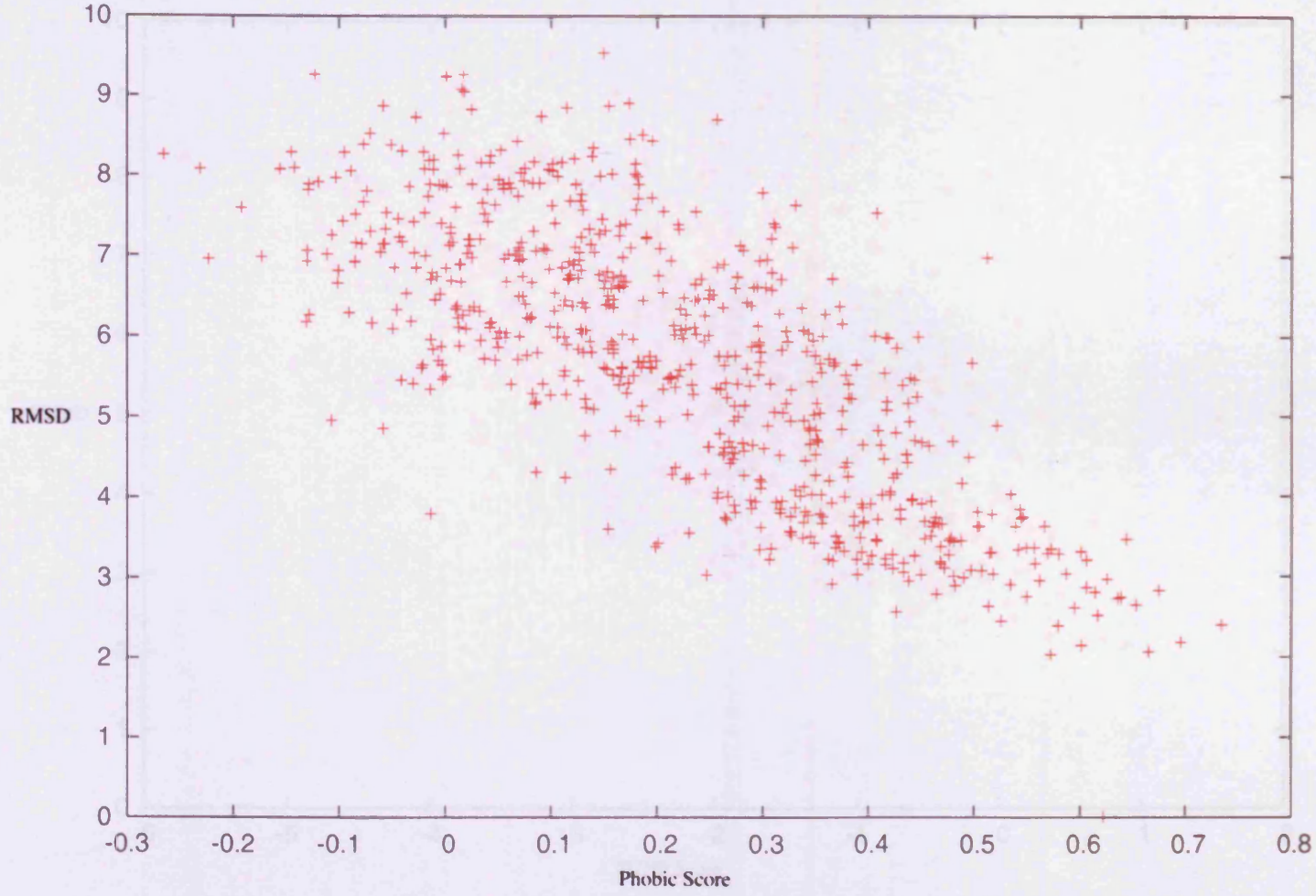
a)



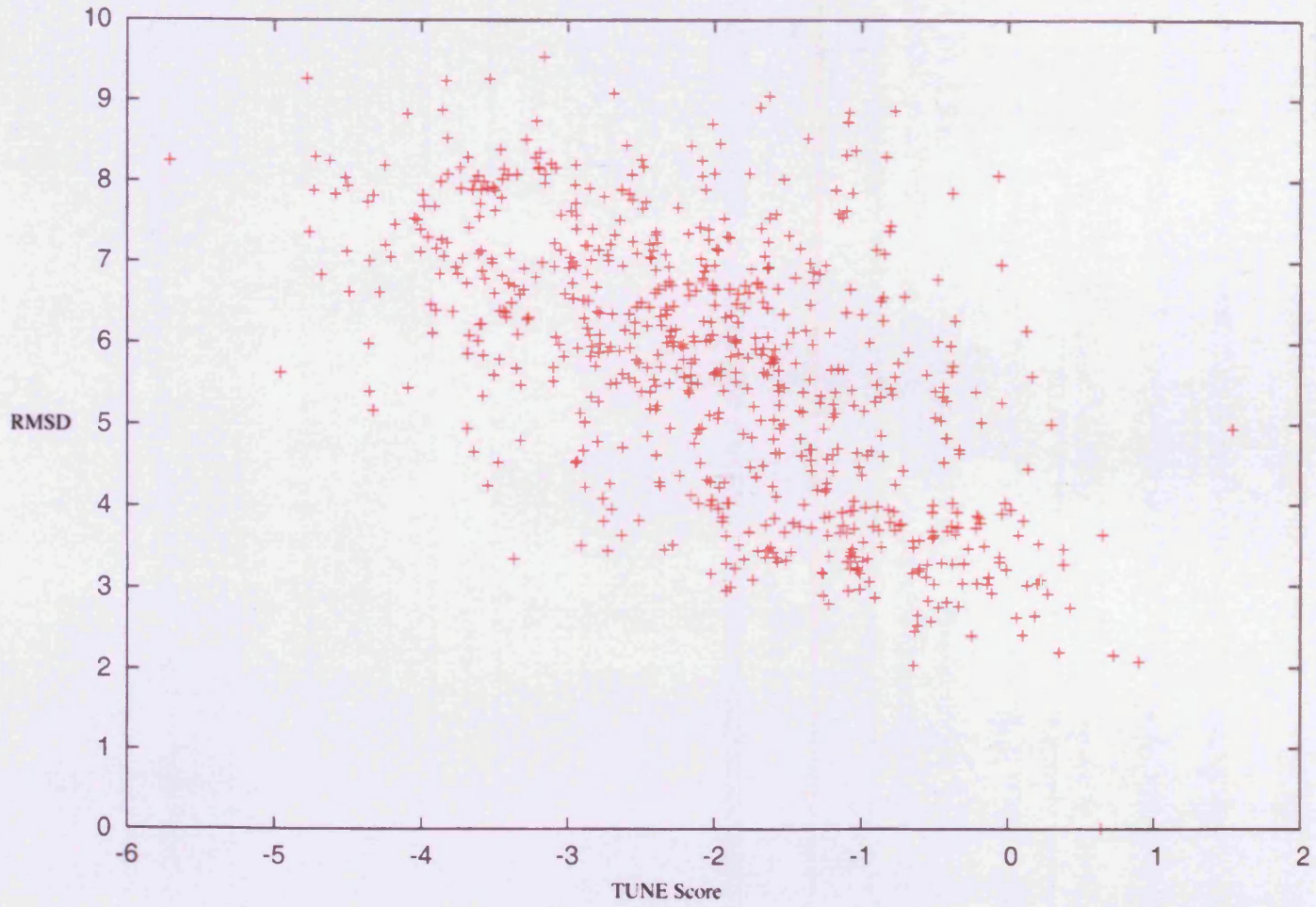
b)

Figure 4.7 Target 1R69 evaluation using Phobic and TUNE: 1R69 is 69 residues in length and is classified as all-alpha by SCOP. Both functions perform well, with Phobic (a) ranking the native structure 3rd and TUNE (b) ranking it just outside the top 10 at 13th. The top ranking Phobic structures have a lower RMSDs than those of TUNE.

The amino-terminal of phage 434 repressor, PDB 1R69, is an all alpha protein that is 69 residues in length. It adopts a lambda repressor-like DNA binding domain fold consisting of 5 α-helices which cover 57% of sequence space. The native structure is ranked third overall behind two structures under 3Å. Three of the top five models. not including the native, are under 3Å from the native using a 1:1 superposition using ProFit. TUNE did not perform as well as Phobic on 1R69. The native structure ranks outside the top 10 structures (13^{th} place) and the top 5 structures fall between 2.7Å and 4.95Å (see figure 4.7).


RMSD
Phobic Score
10
9
8
7
6
5
4
3
2
1
0
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

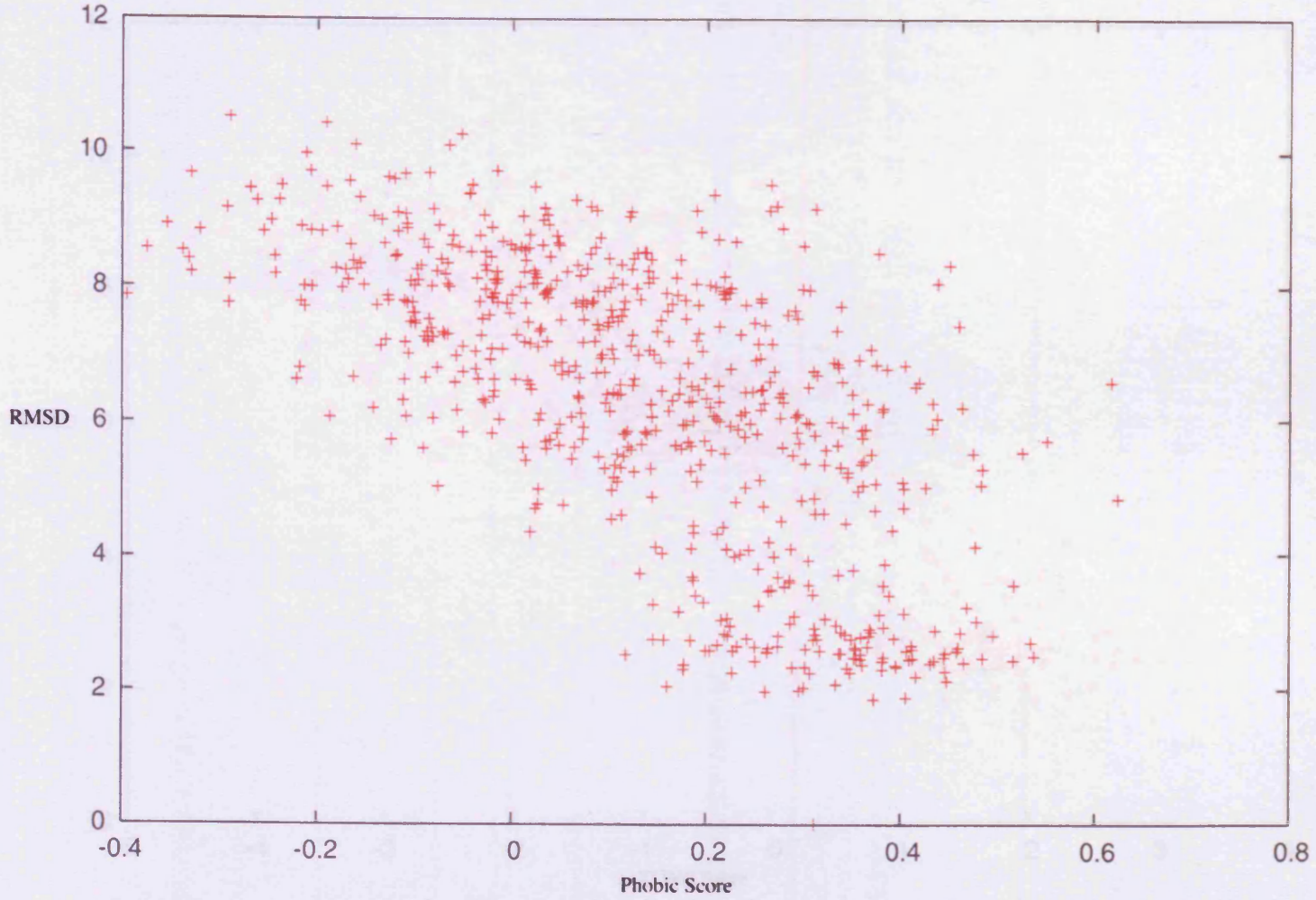
a)



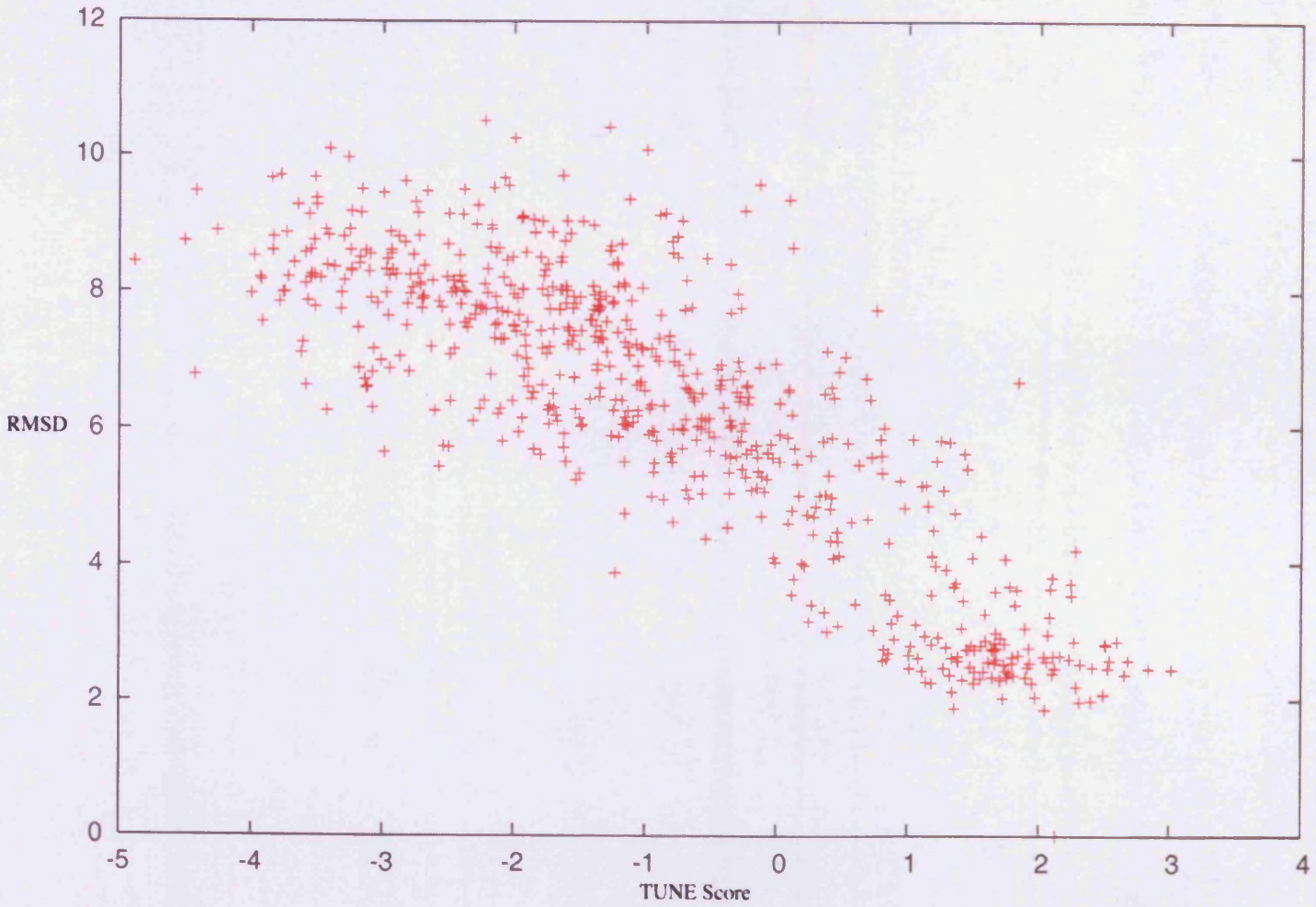
b)

Figure 4.8 Target 2CRO evaluation using Phobic and TUNE: The performances on 2CRO are fairly reasonable for both functions. While neither ranks the native structure above some models they are ranked 5th and 9th for TUNE (b) and Phobic (a) respectively. The top ranking models are closer to the native structure for Phobic than TUNE with the 1st ranked structures having RMSDs of 2.45Å and 4.96Å respectively.

Phage 434 CRO protein, PDB 2CRO, is an all-α protein 71 residues in length. The fold classification of 2CRO is lambda repressor-like DNA-binding domain which consists of 5 α-helices covering 40% of the sequence with the rest of the residues involved in connecting loops. Neither TUNE or Phobic correctly rank the native structure ahead of the bundle, being ranked 5th and 9th respectively. The top five scoring Phobic structures all fall under the 3Å threshold with the best model having an RMSD of 2.4Å. TUNE does not perform as well as Phobic with the top scoring model having an RMSD of 4.9Å from the native structure, additionally this structure looks like a reasonable structure when viewed by eye. The remaining top four structures score 2.1Å, 2.2Å, 3.7Å and 2.65Å respectively (see figure 4.8).



a)



b)

Figure 4.9 Target 3ICB evaluation using Phobic and TUNE: Phobic (a) does not perform well on 3ICB, an all-alpha protein 75 residues in length. The native structure fall outside the top 10 and the top 5 structures have RMSDs of 4.9, 6.3, 5.7, 2.5 and 2.3Å. TUNE (b) in contrast performs well at ranking models, but, like Phobic, fails to distinguish the native structure from the ensemble of models.

Vitamin D dependent calcium binding protein, PDB 3ICB, is an all alpha protein. It adopts an EF-hand-like fold which consists of four α-helices covering 57% of the sequence. The performance of Phobic on 3ICB was not outstanding as the native structure was not identified in the top 10 models, however the top 5 scoring structures had RMSDs of 4.9, 6.3, 5.7, 2.5 and 2.3Å. TUNE ranked the native structure 22nd among a group of models which were within 3.0Å of the native structure (see figure 4.9 or table 4.2).

Table 4.2:Performance Summary of TUNE and Phobic on the 4 State Decoy Set.

PDB ID	Method	Model 1 rmsd	Model 2 rmsd	Model 3 rmsd	Native in top 10
4rxn	Phobic	3.4	3	2.9	Yes
	Tune	3.4	3.5	3.2	No
4pti	Phobic	3.2	3.1	7	No
	Tune	4.1	4.7	6.2	Yes
1sn3	Phobic	8.3	3.3	6.2	No
	Tune	4.3	6.8	8	Yes
1ctf	Phobic	3.5	2.1	4.1	Yes
	Tune	3.4	3	3.2	Yes
1r69	Phobic	2.6	2.7	3.5	Yes
	Tune	3.6	2.7	3.5	No
2cro	Phobic	2.4	2.2	2.8	Yes
	Tune	4.9	2.1	2.2	Yes
3icb	Phobic	4.9	6.3	5.7	No
	Tune	2.4	2.4	2.5	No

Over this set, the performance of Phobic is consistently good with the exception of one model – the all-α 3ICB. It has been observed that scoring functions do not perform well on α-only structures because of the number of ways helices can pack together

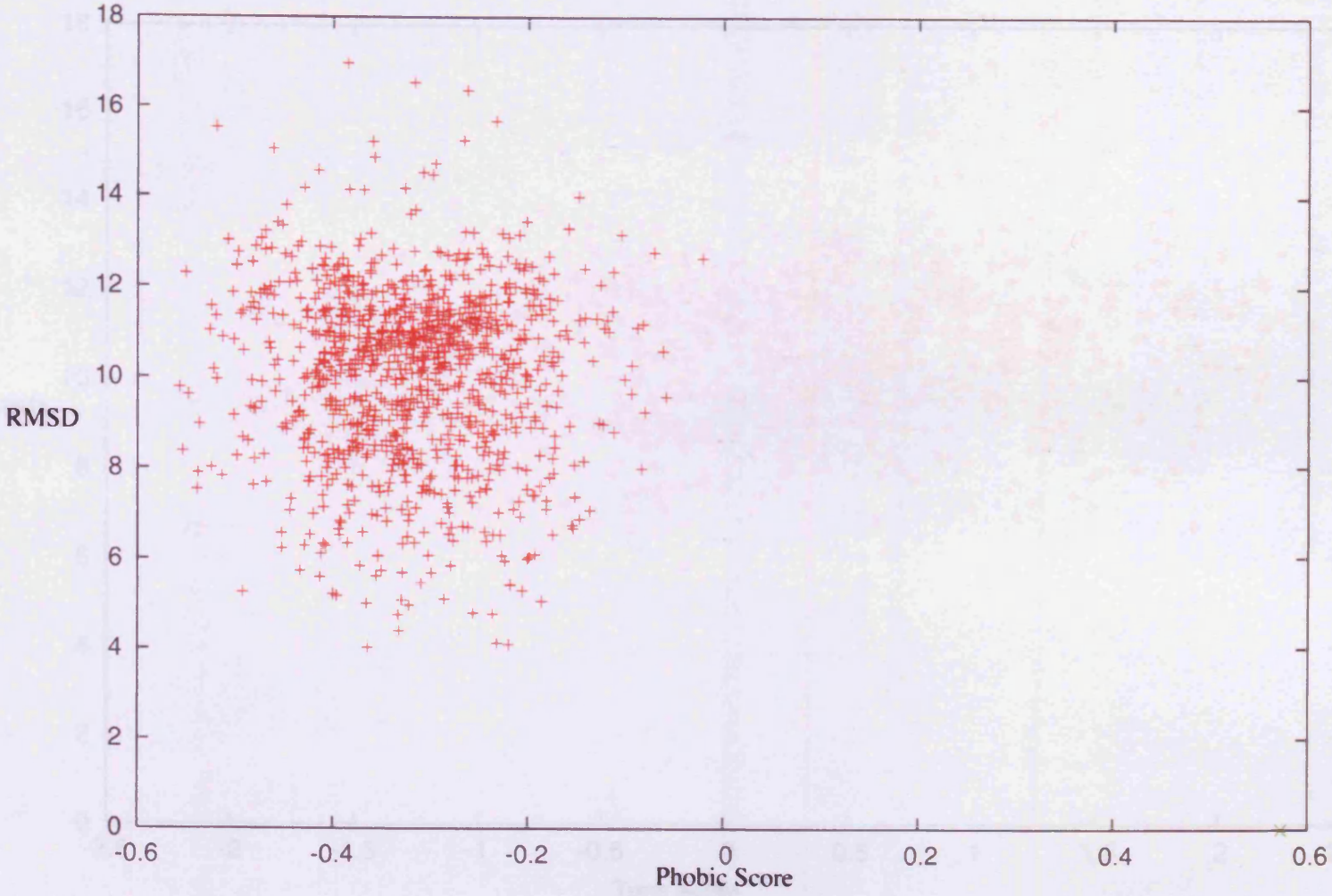
(Berglund et al., 2004), however on two other all-α structures (2CRO & 1R69) Phobic performed well. The poor performance could be due to the flexibility of the relatively large calcium binding loops between the helical structures, which in the native structure are exposed but hydrophobic, this feature was not observed in the models.

SCOP classifies three of the proteins (1SN3, 4RXN, 4PTI) in the 4state decoy set as 'small'. When training machine learning functions, such as SABLE, AccPro and Phobic, one of the first classes of proteins to be discarded are the small proteins. This is because they do not always form compact globular structures and have poorly formed secondary structure elements which are not characteristic of the larger, globular proteins. They remain present in decoy test sets because smaller structures are more amenable to physical scoring functions and dynamic modelling approaches.

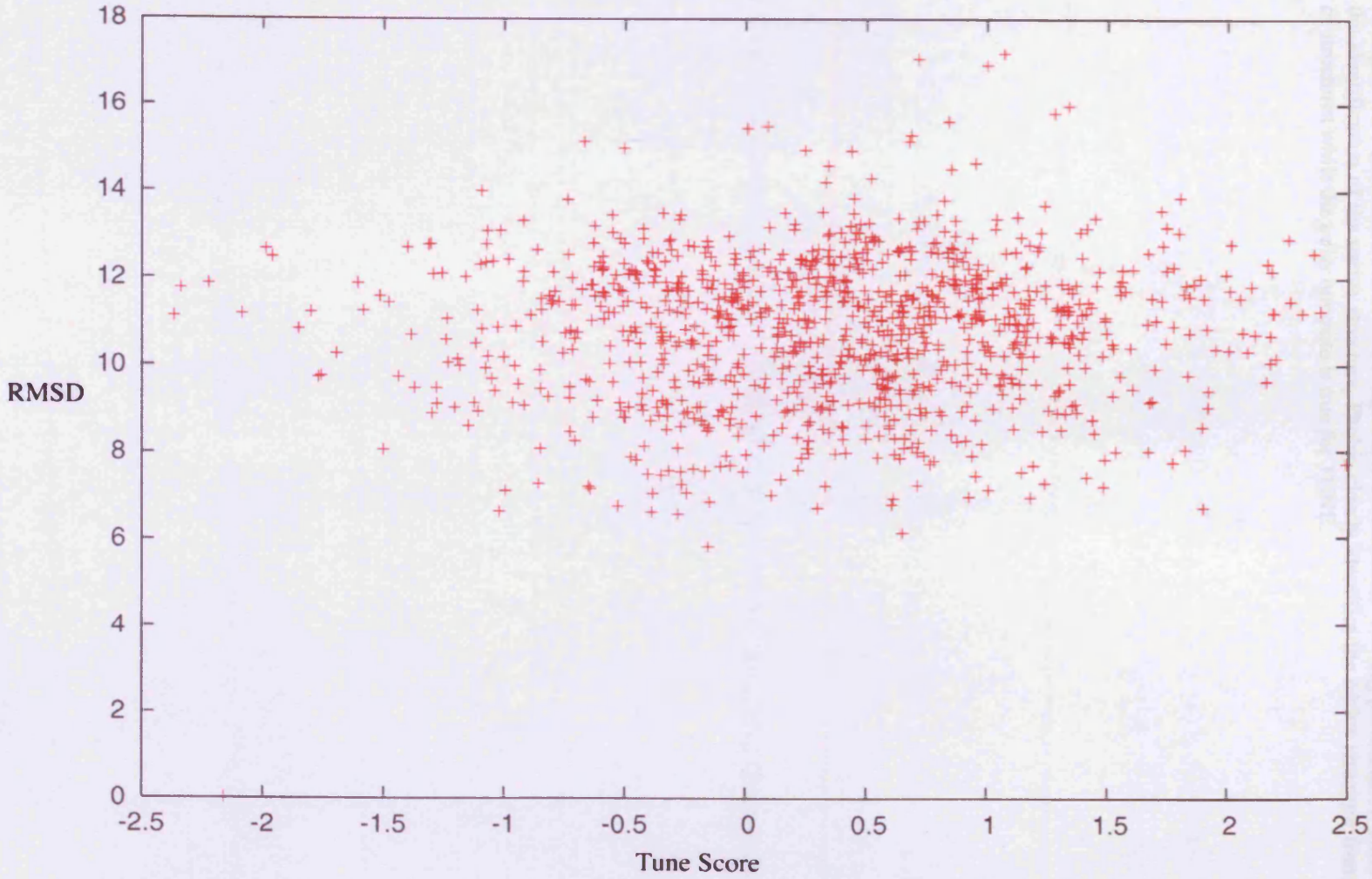
The Rosetta Decoy Set

Four structures were selected at random from the 78 proteins in the Rosetta set: 1CC5; 1C5A;1CSP; 1KTE, these structures are all larger than 70 amino acids and provide an extra evaluation step which although designed for physical scoring functions, show some interesting results.

Cytochroms C5 (PDB 1CC5), is an all-alpha protein, 83 residues in length consisting of 5 helices which adopt a cytochrome C fold. The native structure was the top scoring model for Phobic, however the method was unable to identify a low RMSD model from the ensemble with the top 5 models all having RMSDs greater than 8Å. The performance for TUNE was worse than that of Phobic, not only did TUNE fail to identify the native structure but the top five models had RMSDs greater than 10Å (see figure 4.10).

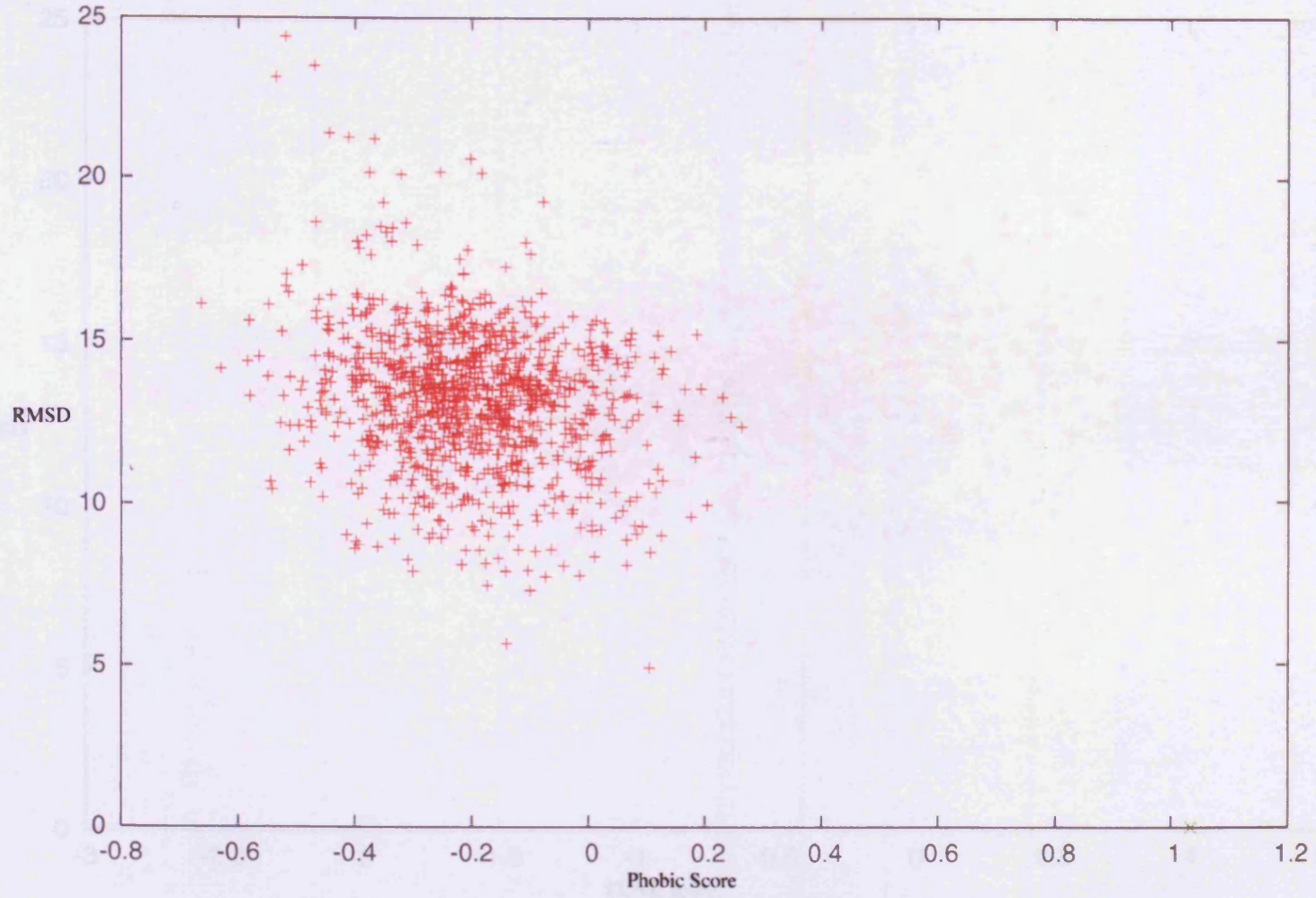


a)

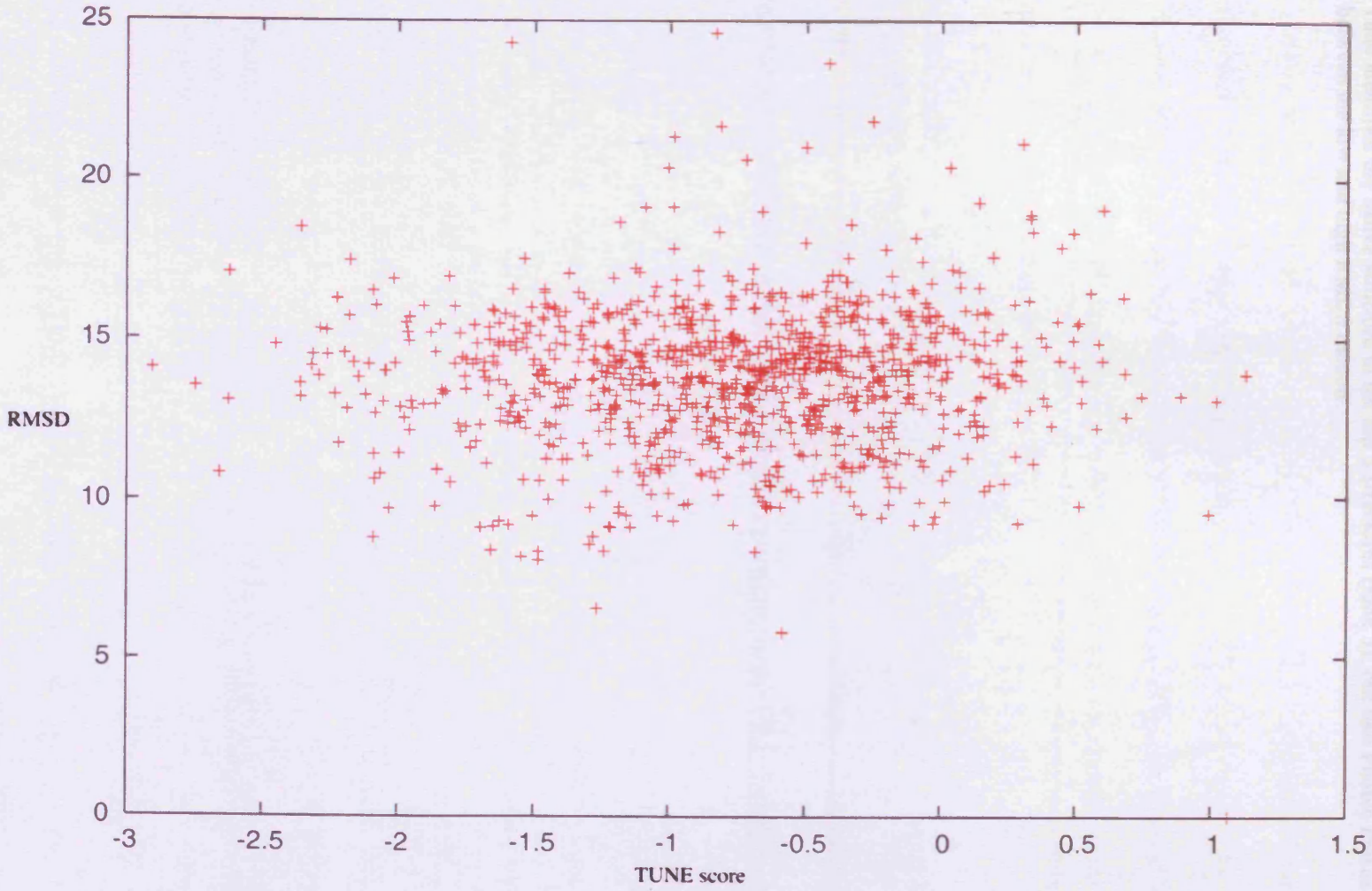


b)

Figure 4.10 Rosetta set 1CC5 Phobic and TUNE: Neither method performed well on this target, both failing to distinguish between low and high RMSD structures. The big performance difference comes in the identification of the native structure, Phobic clearly identifies the native structure from the ensemble of structures while the polar opposite is true for TUNE.



a)



b)

Figure 4.11 Rosetta set 1KTE Phobic and TUNE: Phobic clearly identifies the native structure from the ensemble, but as with previous examples it fails to discriminate between the low and high RMSDs. Tune identifies the native structure in the top 5 structures (rank 3), but like Phobic makes no distinction between the low and high RMDS models.

Thioltransferase, PDB 1KTE, proved to be another tough target. The protein is 105 residues in length and is composed on 6 helices and 6 strands which adopt a Thioredoxin fold. Both Phobic and TUNE performed well at identifying the native structure 1st and 2nd respectively. The margin on the Phobic score is considerable with the vast majority of models being ranked as non-native (Phobic scores less than 0). The same is true for TUNE where the majority of models receive a score less than 0, which marks them as non-native (see figure 4.11).

The remaining models tested 1CSP and 1C5A both produced similar results (not shown) where the native structure is correctly identified but there is no distinction made between low and high RMSD models. The top scoring models had RMSDs of 6.1Å and 11Å for 1CSA and 1CSP respectively.

The failure to discriminate between the low and high RMSD models is not unexpected for either Phobic or TUNE. Both methods assess proteins on a reduced representation, in the case of Phobic pure Cα and TUNE, a mixed Cα - pseudo Cβ model. The Rosetta decoy set is not designed for reduced representation functions, although they can obviously be used, they prove to be a very tough test. This is because reduced representation functions are unable to examine side chain packing and other 'fine grain' elements of protein structure.

Model Evaluation using TRACK and the DDT protocol

To establish how effective the Phobic function would be in the prediction pipeline several targets were run through the DDT-TRACK protocol described in chapter 3. The

performance of Phobic is shown in table 4.3 which identifies the protein and the RMSD of the top 5 scoring models. The table shows that Phobic is able to identify close approximations of the native structure from ensembles where they are present. The top three proteins,1F3R, 1MP9 and 1HKQ, appear to be poor performances however the high RMSDs are a result of poor model construction with the lowest RMSD models being 5Å, 8Å and 6.07Å respectively.

Table 4.3 Phobic performance on DDT generated models

Target	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1F3R B	7.79	6.51	8.98	8.90	7.25
1MP9 A1	9.74	9.97	10.53	10.75	10.38
1HKQ A	7.06	6.44	6.49	6.23	6.46
1OTS C1	3.78	2.79	3.69	5.33	5.59
1GXU A	5.37	5.46	11.62	5.83	5.21
1AE7	3.60	3.70	3.87	4.27	3.83
1DT9 A3	1.34	3.13	1.68	1.18	3.74
1IZ6 A2	4.19	4.31	4.03	3.96	4.42
1BXU A	4.65	4.40	4.54	4.47	4.53
1A8Y 2	2.04	3.71	1.94	2.25	1.91
1ONI A	4.69	4.72	5.21	5.26	4.97
1I9E A	4.23	5.15	4.28	4.29	4.09
1QHF A	1.67	1.42	1.41	1.45	1.38
1IVH A1	7.25	6.61	6.73	7.58	7.22

The first four characters are the PDB IDs, the remaining alphanumeric combination identifies the chain where present. The rank 1..5 are the top five rated models according to the Phobic score. For all of these models, the top five structures were all from the best representative cluster. This means for templates like 1MP9 A, no good models were produced by the DDT pipeline. The use of remote homology means that structures within 6Å of the native are considered reasonable.

As well as the above proteins, four of the 4State set proteins were run through the pipeline and then evaluated using SPREK, TUNE and Phobic as shown in figures 4.13-15. For proteins 4PTI, 3ICB and 2CRO, Phobic ranked 2nd, 3rd and 1st respectively. Unlike 1R69, the results of the remaining models were good for each of the evaluations with the top scoring models all under 3Å from the native structure.

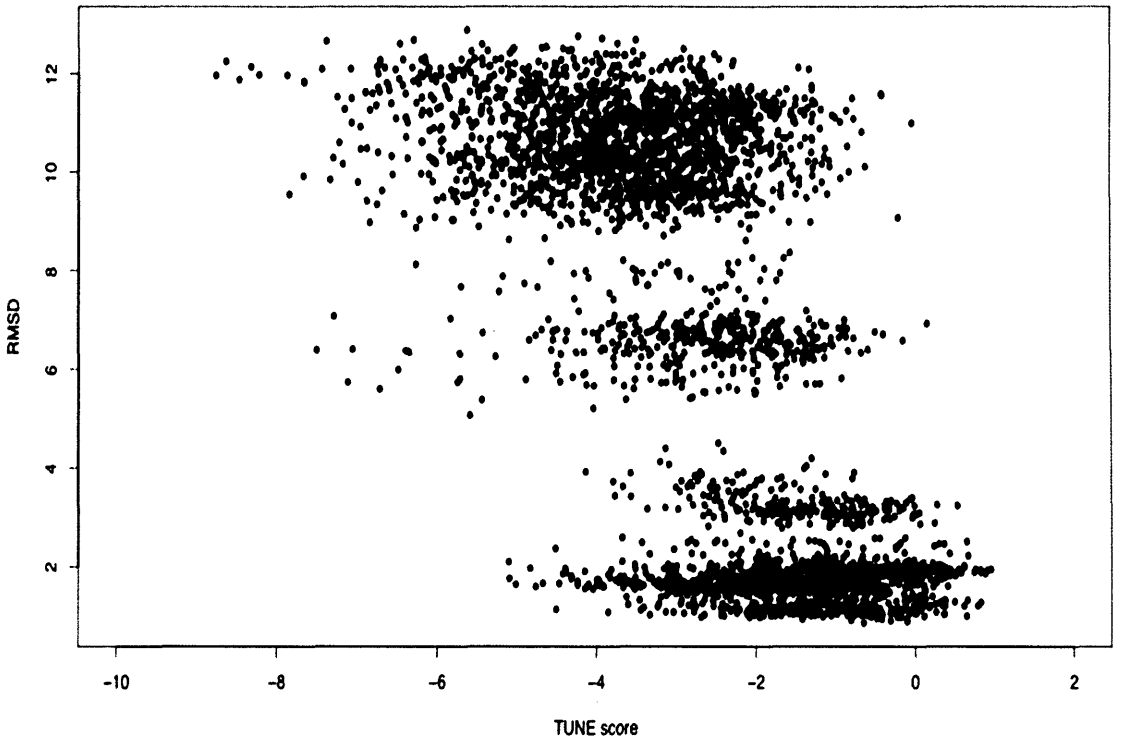


Figure 4.13 1R69 DDT-TRACK evaluation using TUNE: Ranked 2nd, the top 5 scoring models are approximately 2Å from the native structure and are based on the 1B0N_A template. With the exception of the structures based on 1R69, TUNE does not discriminate as the templates become more remote.

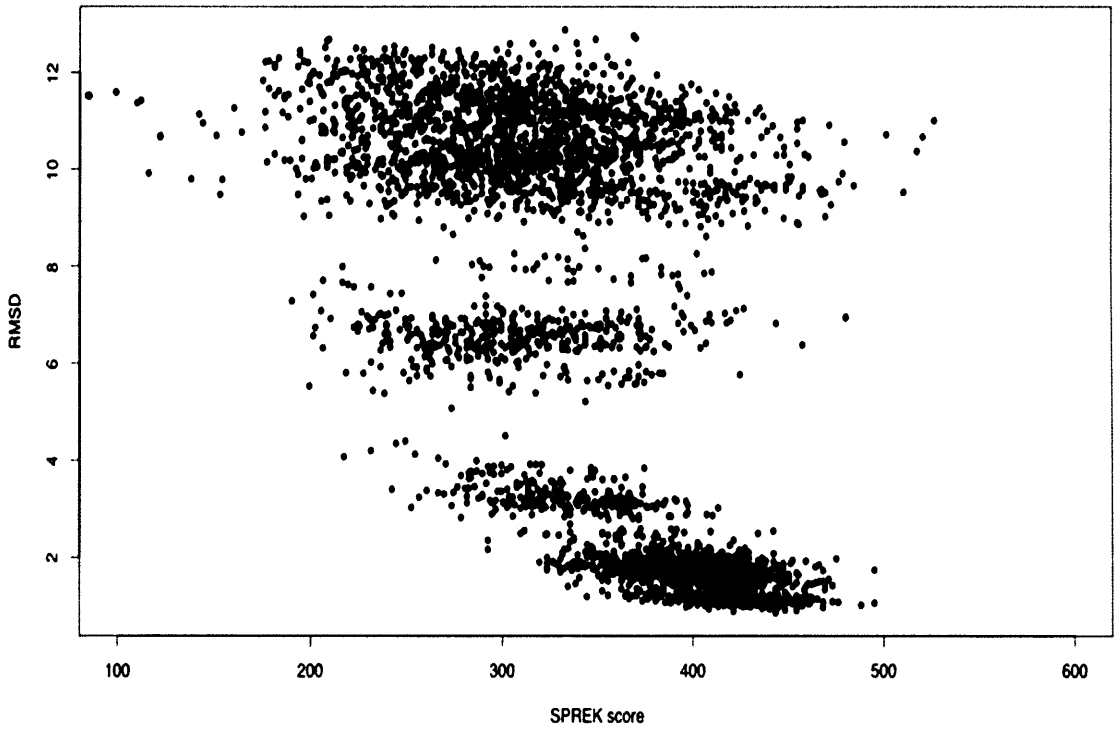


Figure 4.14 1R69 DDT-TRACK evaluation using SPREK: Ranking 3rd overall 1R69 doesn't present a good result for SPREK. The top 5 scoring models having RMSDs greater than 9Å from the native structure. However at ranks 6-8 models based on the 1R69 template are identified.

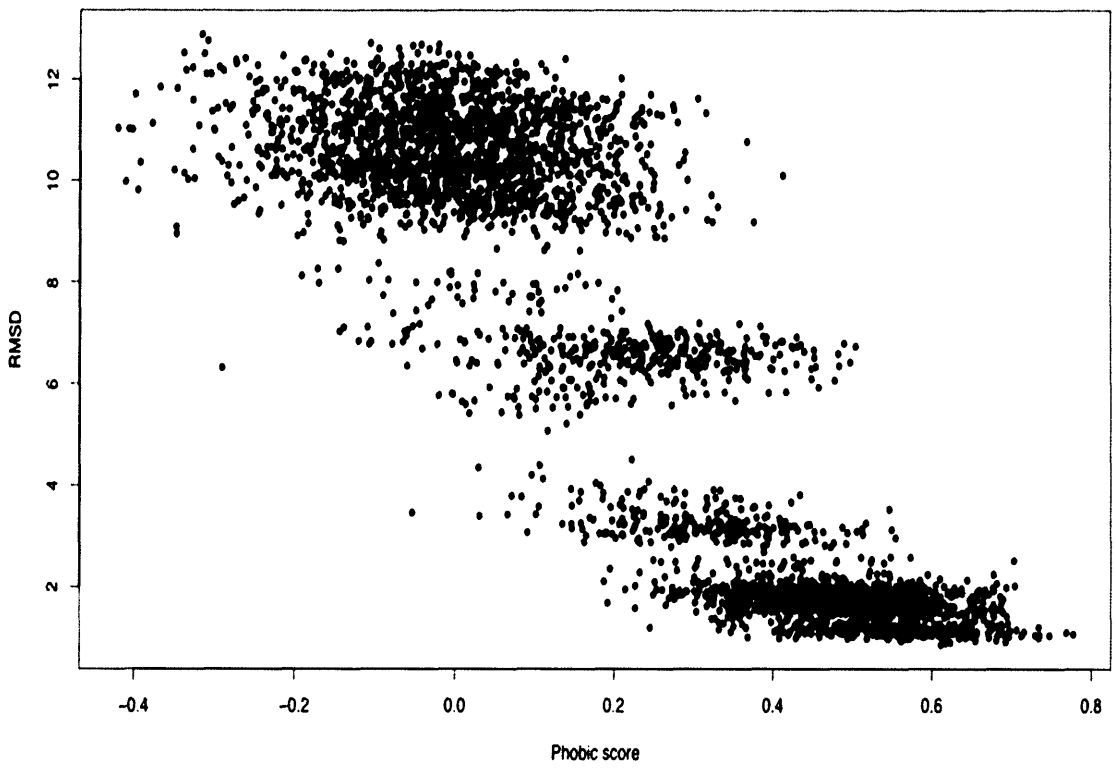


Figure 4.15 1R69 DDT-TRACK evaluation using Phobic: Ranked 1st overall, Phobic not only identifies models around 1Å from the native structure, it also discriminates among models as the templates become more remote, this is shown by the four tiers.

A common method to assess the functionality of an energy function is to calculate the correlation coefficient (cc) between the functions output and RMSD. This measure is useful but should not be considered absolute for the following reason. As demonstrated here and in other papers (Berglund et al., 2004, Lin et al., 2002) the cc does not necessarily reveal if a method consistently identifies one (or several) good models *but* fails to discriminate against the remainder. If cc is used as a sole measure of function it is clear that effective energy functions could be discarded, it is for this reason that graphical representations are presented instead of correlation coefficients or any other measure. Further more, because we are only interested in if a good models has been identified and we have no intention of comparing scores across multiple proteins, the approach taken to assess Phobic is not justifiable. While measures of enrichment are interesting, again, this work is solely concerned with assessing another physical feature of protein structure which can be combined with existing measures (TUNE, SPREK etc) to improve the current assessment routine, such as that described in chapter 3 (dynamic domain threading).

Conclusion

Compared to TUNE and SPREK, Phobic is simple. It relies on the observation of residue exposure from each predicted structure and a prediction of solvent accessibility from sequence using an artificial neural network. Performance on the 4State decoy set and models generated using DDT (chapter 3) shows that Phobic is as effective as TUNE and SPREK.

The structures in the 4state decoy set do not lend themselves to the evaluation of Phobic due to their small size, which means that they do not form compact hydrophobic cores, the very feature that Phobic 'probes' for. Despite this disadvantage, the performance of Phobic was encouraging, consistently ranking low RMSD and native structures. The exception to this performance being 3ICB which, while not as good as other targets, was acceptable as the top ranking structures are always examined by an 'expert' eye. When compared to TUNE the performance of Phobic also was good, if not better, on all but one structure (3ICB). In addition to this performance it should be noted that TUNE is not affected by the size problem which affects Phobic.

The Rosetta set was constructed for the evaluation of physical scoring functions. This means that the structures are designed to be well folded so that they have realistic intramolecular interactions such as hydrogen bonding and van der Waals forces. To present a challenge for physical scoring functions the violations found in the 4State set are too coarse, instead focusing on the intramolecular interactions. This has the knock-on effect off eliminating the types of error that TUNE and Phobic 'look' for. It is somewhat interesting then that the Phobic function was able to identify the native structure in the four instances used here.

Despite the good performance of Phobic in this instance, there are several problems with knowledge-based statistical scoring functions in general. The dependency on multiple sequence alignments is at the root of these problems. As stated in the introduction, scoring functions are subject to the axiom 'rubbish in, rubbish out' meaning that the alignment, be it sequence or structure based, can 'make or break' a function. Theoretically a 'poor' alignment used at any point in the function can

manifest in a poor prediction, be it secondary structure, solvent accessibility or disorder. The poor prediction would almost certainly produce spurious results at the end of the prediction pipeline.

When dealing with alignments, homology is also important, especially with Phobic. Where clear or even remote homology is discovered, the performance of PSI-BLAST based prediction tools is reasonable, typically in the range of 75 ± 10%. In scenarios where there is little or no information obtained from an alignment then prediction will be less accurate. In fact the purpose of completing a sequence alignment is to generate a profile which is beneficial for prediction. The use of multiple sequence alignments for prediction of solvent accessibility is contentious, some claim that prediction accuracy is increased (Adamczak et al., 2004) while others claim there is little benefit, as solvent accessibility is not well conserved across familial alignments. It therefore becomes necessary to answer the following question: "does the use of poorly formed multiple sequence alignments result in worse prediction of a feature, than the use of a single sequence?". This problem is not trivial, given that it is widely known that multiple sequence alignments are critical for secondary structure prediction, so the question then becomes 'can we identify good alignments from bad', the question posed in chapter 2 that remains unanswered.

The third point, consistency, is intrinsically linked to the previous points. It is accepted that there is currently no single function which can identify the best model produced by a prediction pipeline. In attempting to solve this problem, it is useful to combine several scoring functions, which address different protein attributes into a single function. This combination is achieved in an ad hoc fashion (Taylor et al., 2006) or by

using techniques such as artificial neural networks or partial least squares regression-like methods (Berglund et al., 2004), however, even in combination, these scores are not infallible. This is one of many problems that continues to plague protein structure prediction as well as one that has no obvious solution.

In summary, Phobic has been shown to be effective at evaluating protein Cα models in the 4State decoy set and, arguably more importantly, those produced by the DDT and TRACKS pipelines. When compared to tools already applied in our prediction pipelines it performs equally well at discriminating amongst native, native-like and non-native structures. It is currently used in the protein fragment tessellation tool (Jonassen et al., 2006) and in a novel method for *de novo* prediction of alpha/beta proteins (chapter 5).

Chapter 5

De novo Prediction of alpha/beta proteins using Ideal Forms and CASP 7

Introduction

For over thirty years it has been widely accepted that the amino acid sequence of a protein is sufficient to explain the fold of a poly-peptide chain (Anfinsen, 1973). This statement suggests that the problem of mapping the sequence to structural space is trivial, however over a period of nearly forty years there has been limited success in explaining how sequence dictates fold. The most robust way to demonstrate our understanding of protein folding is to predict the 3D structure armed only with knowledge of the sequence. Currently the only way to do this is to code a computer program that, given a sequence as input, returns the 3D coordinates as output.

There are two computational approaches to this problem. The first approach is to allow a flexible chain of twenty virtual amino acids to fold under specific physical and chemical restraints (referred to as *Ab initio* prediction). The second approach is to take numerous static 'snap-shots', where each image represents a potential conformation, and then try to pick the right one (referred to as *combinatorial* modelling) (Cohen et al., 1979, Cohen et al., 1980).

Until recently, the most successful *ab initio* method was able to predict the approximate structure of a short polypeptide (36 residues in length) with an RMSD 4.5Å from the native (Duan and Kollman, 1998). As mentioned in Chapter 1, recent advances in the field have been made by incorporating prior knowledge in the form of protein fragments (this is referred to as *de novo* prediction since it is no longer from first principles). As a result of these advances the maximum size of a protein which can be predicted has jumped to almost 100 residues (Bradley et al., 2005).

Both of the methods operate close to the limit of conventional compute power and to extend either method could increase operation time in excess of linear extrapolation with protein length. For smaller proteins consisting predominantly of alpha helices, the fold of the protein can be approximated from local packing – something of an ideal for folding simulations. However, where proteins exceed 100 residues in length and contain secondary structure elements that form non-local interactions (β-sheets) the increased time to search feature space is prohibitive.

The combinatorial approach is not subject to the same compute problems as the *ab initio* method where large proteins are involved. The ability to tackle these proteins is derived from the analysis of structure at the higher level of secondary structure elements. However, the approach does have a caveat – it requires an accurate secondary structure prediction as well as a suitable framework on which to over lay the prediction. Provided these requirements can be met, it is reasonable to assume that this method would provide a solution to the prediction of large proteins (in excess of 100 residues).

Improved computer resources and the classification of proteins into a Periodic Table-like system ((Taylor, 2002) referred to as PT from this point) provided an opportunity to review and update combinatorial structure prediction. To this end a new system was devised that did not make direct use of structural information that was specific to the target protein. Instead the predicted protein structures are placed onto all frameworks in the PT. The resulting models are refined and evaluated before applying a conventional threading method that is dependent on matching secondary structure predictions and framework elements as demonstrated in (Taylor et al., 2006). Throughout the work

sequence homologues were explicitly excluded ensuring that the method was truly *de novo*.

Materials and Methods

An outline of the *de novo* prediction pipeline is given in figure 5.1.

Generation of Multiple Sequence Alignments

Multiple sequence alignments were generated automatically using the Multal-Mulsel method described in chapter 2. A sequence database was prepared using the non-redundant database as a template. All low complexity, coiled coil and trans-membrane sections were masked using PFILT version 1.3 (Jones et al., 1994). Sequences alignments were generated using three PSI-BLAST iterations (-j 3) with an e-value threshold (-h) of 0.001 for inclusion in a multipass model.

Secondary Structure Prediction

Secondary structure (SS) was predicted using PSI-PRED (McGuffin et al., 2000) and YASPIN (Lin et al., 2005), both tools derive sequence alignments from a standard PSI-BLAST (Altschul et al., 1997) search against the nrdb. Despite the introduction of multiple sequence alignments to SS prediction and an average accuracy of 75% over three states, a non-standard approach was taken. To circumvent error associated with secondary structure prediction for each sequence in the alignment predictions of secondary structure were made. The predictions were then pooled to create variation

for the Ideal Forms. The variation over the 10-20 sequences in the alignments was sufficient such that at least one was a close approximation of the true secondary structure.

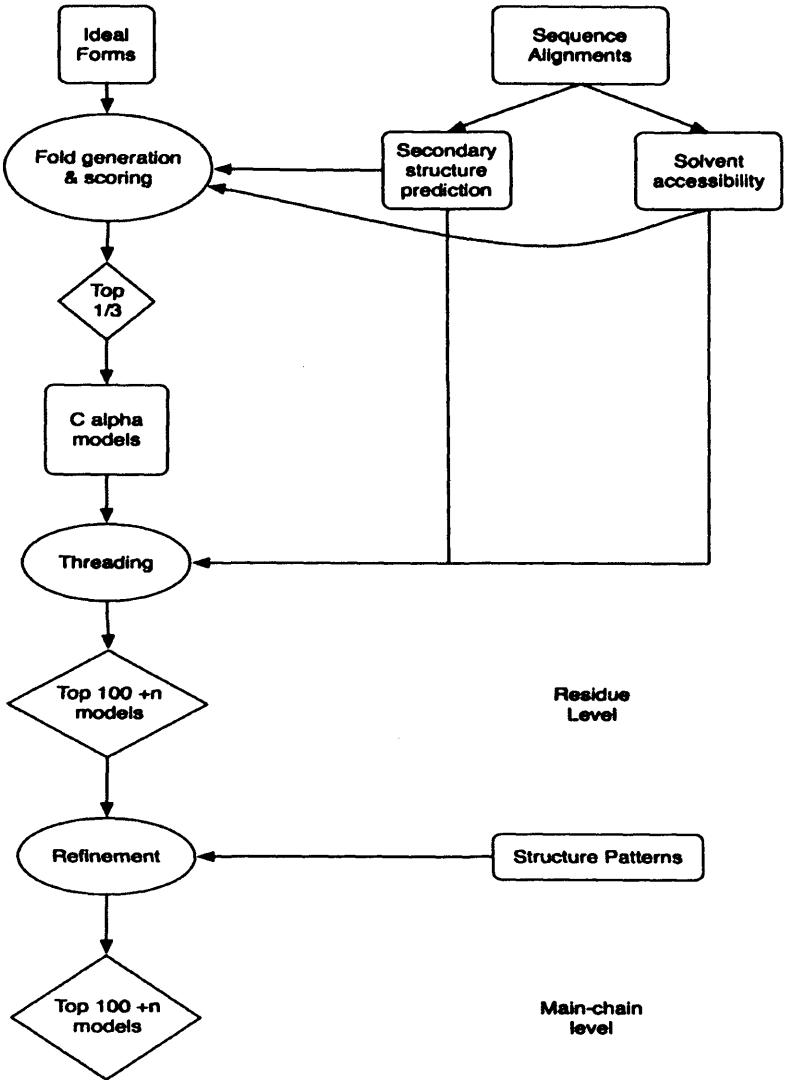


Figure 5.1 *de novo* prediction pipeline: The prediction pipeline described in this chapter is a development of the DDT method introduced in chapter 3 (Taylor et al., 2006). The ideal forms provide the starting point for the process, forming the framework onto which sequences are threaded. Following this step several rounds of model construction and evaluation are completed using the techniques described in previous chapters.

Ideal Forms

The ideal forms were derived from Taylor's "stick" models (Taylor, 2001). In these domain models, a layered structure is imposed by hydrogen-bonded links across β-sheets. The layers can consist of either α-structures or β-structures. Models are limited to four layers in a single domain with each layer consisting of one type of secondary structure.

Forms were represented following the approach of Chothia and Finkelstein (Chothia and Finkelstein, 1990) where each form is represented as a packed box with the α-structures taking the form of a square section 4*4 and the β-structures half the thickness (2*4). The assumption that the secondary structures have an equal depth allows volume and surface area to be estimated. The area (V), the perimeter (A) were calculated and a 'compactness' score (*cpn*) was calculated as $10V/A$. In all architectures a bias was imposed so that an even number of α-structures occurred above and below β-sheets. Asymmetry was penalised by a factor which was incremented by 10 for each unbalanced α-helix.

Another measure of solvent exposure was made for each element in the ideal forms. For the helices this ranges from 0 ... 16 (4*4) and for sheets 0 ... 8 (2*4 as each sheet is packed to an adjacent strand giving 8 rather than 12). The sum of the exposed edges was then normalised into the range ±5. To supplement this, a conserved measure of hydrophobicity was used (Taylor et al., 2006). This value was then summed over each secondary structure element and normalised by the square root of the number of

elements in the section. This value was also scaled to the range ±5 to match the previous measure.

The combination of these two values gives a score for how well the predicted segments matches the degree of burial in the ideal form. A double weight was applied to the β-sheets to reflect their importance in specifying the overall fold. The score for the α-segments is denoted α and the score for the β-segments is denoted β.

To maintain native-like connections between sections left-handed (between β-sheets) and crossing loop connections as well as knotted topologies were eliminated. Parallel connections were also penalised where edge strands or helices were not involved. The penalty (e) was initialised at 0.5 and incremented by 0.5 with each violation of the restraint. The final restraint was placed on the length of connections between segments. Longer connections were penalised using a Gaussian function which decays slowly, the result is that the penalty increases slowly and therefore is small. The function takes the form $dist = \sum 1 - \exp(-\delta^2 / 10^2)$, the value δ is defined by the amount the connection exceeds the 10Å limit.

All of these elements were combined, arbitrarily, into a final function *f(s)* such that $s = 10w/(5 + \alpha + \beta + cpn + dist + e)$. Because the values in this function are small five was added.

Generating Folds

For each sequence identified in the refined alignment, secondary structures were generated as described previously. A limited number of these sequences, typically 50 (restricted by compute time and availability), were then mapped onto ideal forms. Because the arrangement of secondary structure elements is limited, so to are the number of ideal forms tested. The limited accuracy of secondary structure prediction (approximately 75% ±10%) meant that further variation in secondary structure was introduced by the ablation of weak α and β predictions. This means that if five α and five β sections are predicted rather than having a possible three ideal forms (0-5-5), (1-5-4) and (2-5-3) the choice is limited to (2-5-3) and (2-4-3). In the case of ambiguous predictions, such as (---HHHHHEEEEE---), rather than exclude the section two variations were allowed, one pure α the other pure β.

Ideal Forms specify the overall architecture but do not define how secondary structure elements are connected. There are numerous ways to connect a number of secondary structure elements, to limit the compute time, restraints were used to make connections between strands in the same sheet, right handed; that two surface loops seldom cross and that protein knots are rare. Despite this the number of possible combinations remains prohibitively large for a protein of ~200 residues. To further reduce compute time the methods described above were implemented for a second time in combination with the hydrophobicity of each element. Using this score Cα models were constructed using the method described in (Taylor, 1993).

In order to generate realistic models steps are taken to make Cα models less 'ideal'. This is achieved using each Cα model as a template for threading. Two scores were used in the threading process, one to optimise the fit of the secondary structure and the model, the second to assess the hydrophobic packing of the model (chapter 5). To reduce compute time this score was first applied to the template because it is unlikely that a poor template can give rise to a good model(s). The remaining templates were then used to generate a number of models which were scored using the same method as the initial templates so that the best models were identified. Unfortunately the number of models that could be used in the following step is limited by compute resources. To bypass this problem the top 100 proteins plus the length (L) of the target are selected and assessed using the observed and predicted secondary structure, Phobic (predicted/observed exposure) and SPREK (residue packing).

The 100 + L proteins were further refined using the program Furball (Jonassen et al., 2006). Furball encodes each model as a series of fragment patterns, each pattern describes the environment of a residue and it's environment. These patterns are then scanned against a database of known proteins. In order to maintain the guise of minimal sequence identity and fit the *de novo* profile all targets are scanned against the Furball sequence database. Sequences (and patterns) that match the targets were removed from the Furball database.

Models were then further refined by the inclusion of main-chain atoms so that the extent of hydrogen bonding could be estimated. These models were then subjected to another round of scoring – a combination of SPREK, Phobic and the number of hydrogen bonds

where each bond in a β-structure counted twice. As in the previous step, the top 100 + *L* models were then assessed using the Periodic Table (Taylor, 2002).

Pipeline Evaluation

Despite attempts to reduce compute time, including serialisation across 50 compute nodes, a target approximately 100 residues in length takes over 12 hours to complete. The prolonged run-time meant that optimisation was problematic and, as such, five proteins, referred to as *the Fives*, were selected to cover several different folds and lengths. In addition to these five, the pipeline was applied in the 7th round of CASP and on 1auo, a 218 residue Rossmann fold protein.

a)

b)

Figure 5.2 Native Structures of 3CHY & 1F4P: The structures of chemotaxis Y protein PDB 3CHY (*a*) and flavodoxin PDB 1F4P (*b*). Both proteins are approximately 150 residues in size and belong to the SCOP α/β class. The overall structure for each protein is called the flavodoxin fold.

Figure 5.3 Native Structure 1COZ A: 1COZ, chain A is 126 residues in length, it is covered by the 2-5-3 Ideal Form but has different connections to the previous structures. The overall topology is that of the Rossmann fold. Images were generated using PYMOL.

Figure 5.4 Native Structure of 1DI0 A and 2TRX A: Lumazine synthase (PDB 1DI0, *a)* is 147 residues in length and has a Rossmann fold topology, it assumes the Ideal Form 2-4-2, *b)* Thioredoxin (PDB 2TRX) is 108 residues in length and is the smallest protein in the test set. It adopts the Ideal Form 2-5-2 and has a helical connection between two anti-parallel β-strands which is not, currently, supported by the Ideal Forms. Images were generated using PYMOL.

In figure 5.2 the protein on the left (labelled *a*) is bacterial chemotaxis Y protein (3chy). It is 128 residues in length and matches the Ideal From 2-5-3. The protein on the right (*b*) is a flavodoxin mutant (tyrosine 98 → tryptophan). Flavodoxin (1f4p) is 158 residues in length and adopts the 2-5-3 Ideal Form. It is larger than 3chy and has longer loops between secondary structure elements. Both proteins share the common flavodoxin fold.

Glycerol-3P cytidyltransferase (1coz) is 126 residues in length and assumes the Ideal Form 2-5-3. The difference between this protein and the others in the set is that the strand order in the sheets and loops is different. In addition to this ,1coz has a small C-terminal helix that does not pack on the sheet (see figure 5.3).

The last two proteins are 1di0 (figure 5.4 *a*) and 2trx (figure 5.4 *b*). 1di0 or lumazine synthase is 147 residues in length and assumes the Ideal Form 2-4-2, it only has four strands which are packed against long α-helices. Thierodoxin (2trx) at 108 residues in length is the smallest protein in the test set. It assumes the Ideal Form 2-5-2 and has a helical connection between two anti-parallel β-strands. This feature is not well represented in the lattice models and is the subject of further work.

Results

Four sets of proteins were used for construction and evaluation of the *de novo* pipeline. The first set are the Fives mentioned above; second. a number of proteins less than 150 residues in length; third, a group of proteins in excess of 150 residues long; fourth, the CASP7 proteins identified as α/β. The results are presented as four pooled runs per target.

The 'Fives'

Chemotaxis Y protein (3chy) produced the most consistent results. In all but one of the four runs only one incorrect fold ranked greater than 25th. Models that deviated from the native structure by under 5Å are said to be correct. For this target, the highest scoring model had an RMSD of 4.4Å when calculated as a 1:1 structure alignment. Using SAPit (Taylor et al., 2000), which allows the structure alignment to 'slip' into the 'best' position, the RMSD decreases to 3.8Å. The use of SAPit indicated that three of the β-strands had slipped by one position as shown in figure 5.5.

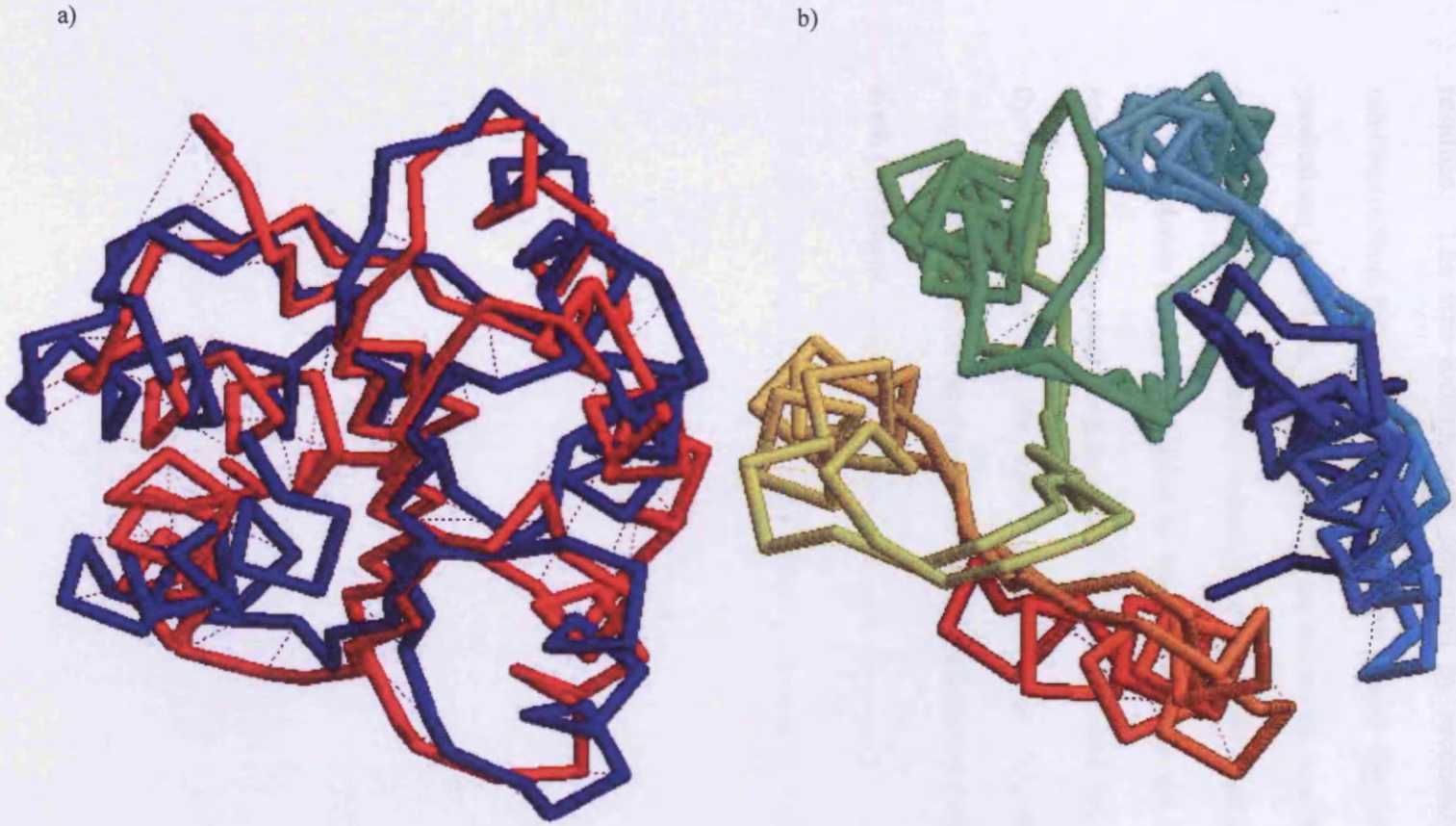


Figure 5.5 Structure Superposition of Chemotaxis Y protein (3CHY) and Top Scoring Model: a) The blue structure is the native conformation and red is the top ranking model. The RMSD as measured by SAP is 3.7Å over 121 matched residues. Although the 2D representation is not ideal it is clear that the model is a close approximation of the native structure. b) Structural superposition coloured from N-terminal (blue) to C-terminal (red), the structures are identical but shown from a different angle.

Flavodoxin (1f4pA, figure 5.6) shares the same fold as 3chy but is longer by 20 residues. The size difference is absorbed by secondary structure elements and the interconnecting loops. By nature of being longer the diversity of secondary structure predictions increased, this resulted in an increased number of models being generated and evaluated. The increase in number of possible folds also meant that the number of incorrect folds increased relative to 3chy. Even with the increase in incorrect folds, the correct fold was identified and ranked 2nd twice and third once. The RMSD between the best structure and the native structure was ~5Å, this was a result of the loop connecting β-strands three and four being sequestered to the edge of the sheet forming a sixth β-element.

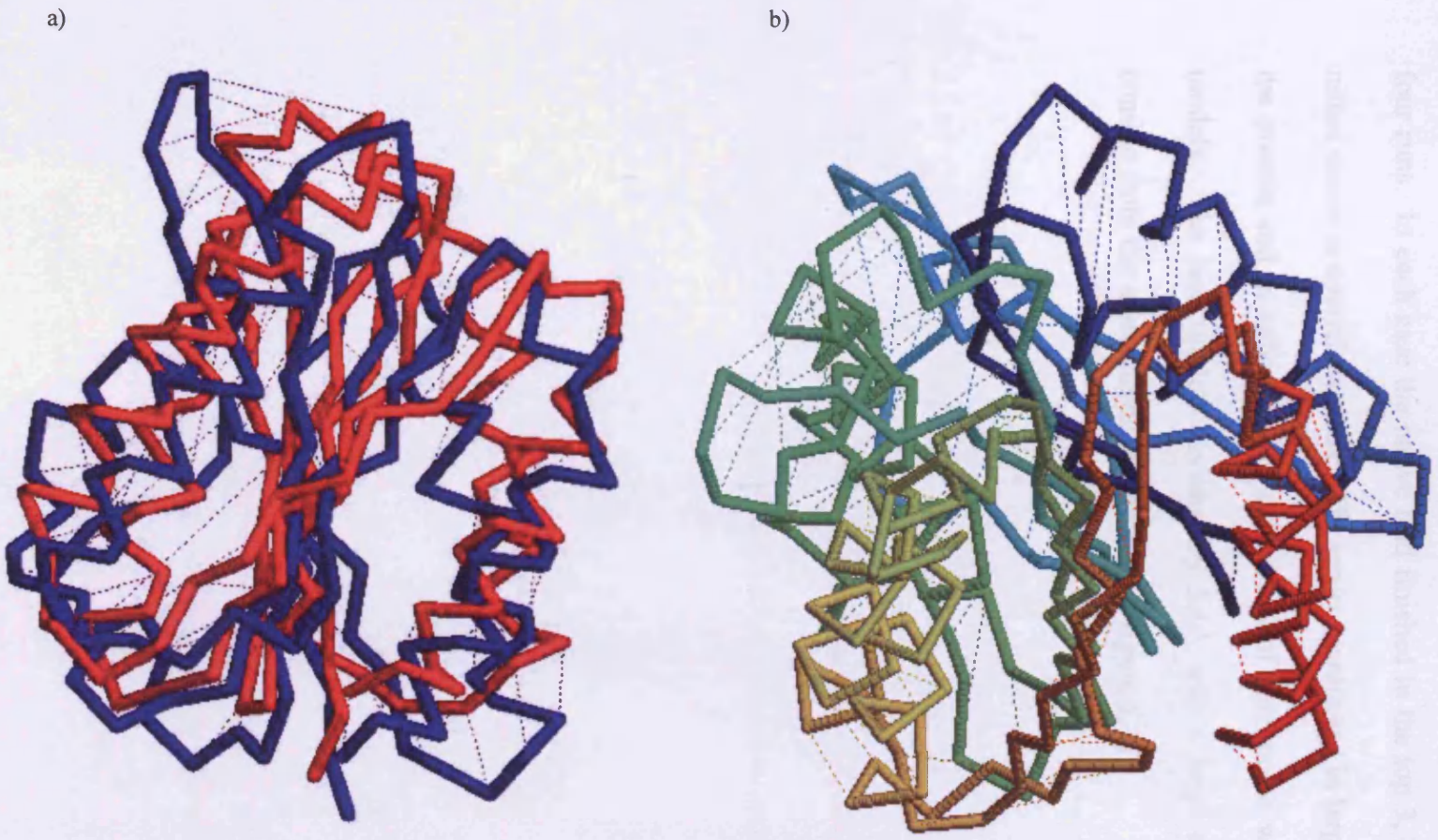


Figure 5.6 Flavodoxin (1F4P_A): a) The blue structure is the native conformation and red is the top ranking model. The RMSD, as calculated by SAP is 5.169Å. The large variations tend to be found in the loop regions at the north and south poles of the image. b) The structural superposition coloured from N-terminal (blue) to C-terminal (red), the structures are the same as in *a* but shown from a different angle.

Predictions for glycerol-3P-cytidyltransferase (1coz_A) were consistently good over the four runs. In each case the native fold finished in the top 3, however there were some minor errors in overall structure. The errors manifested in larger loops on the surface of the protein and in helices which drifted out of position or were absent from the final models. The best models deviated by 5.1Å with a large amount of this difference coming from the aforementioned errors (see figure 5.7).

a) b)

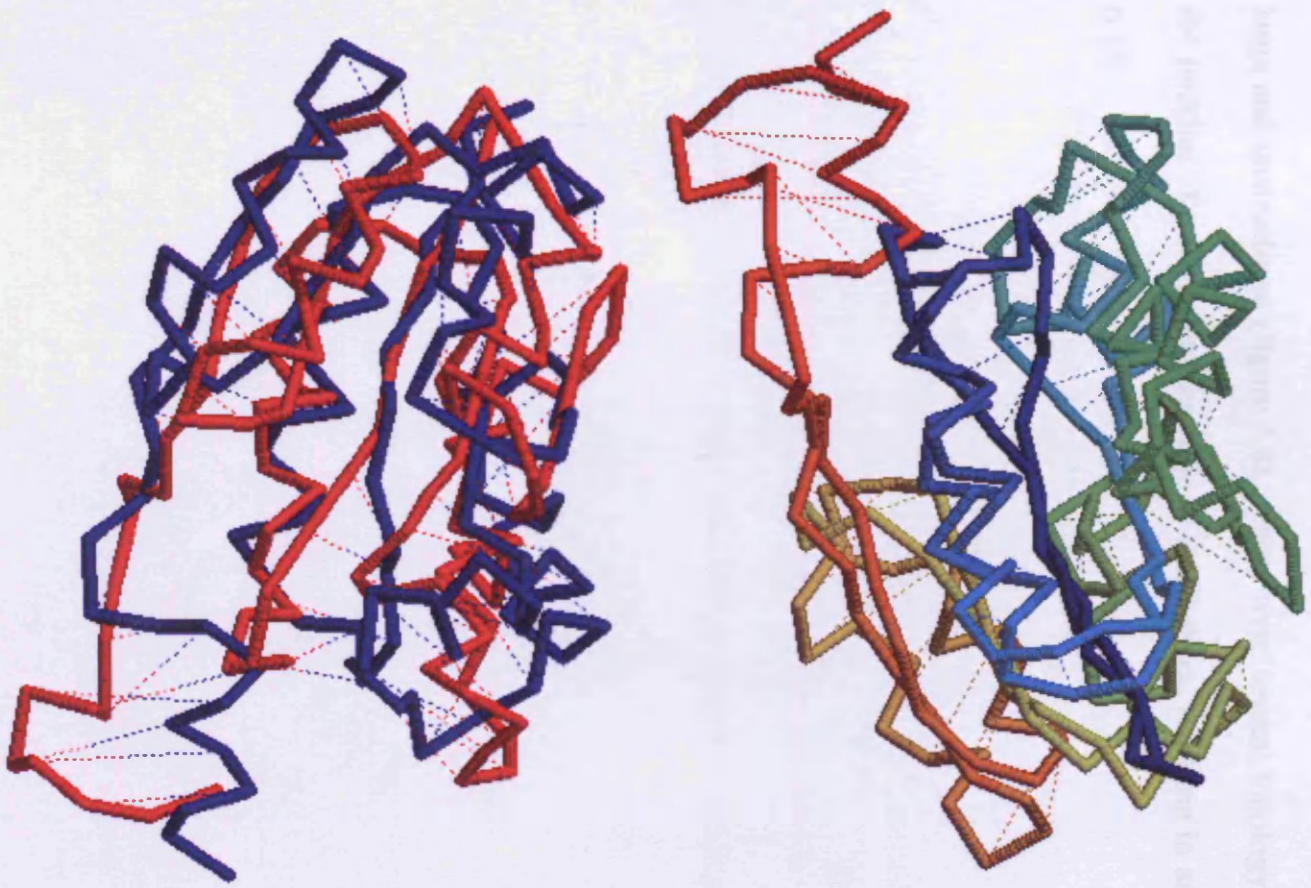


Figure 5.7 Glycerol-3P-cytidyltransferase (1COZ A): a) The blue structure is the native conformation and red is the top ranking model. The overall RMSD calculated by SAP is 5.255Å over 123 residues. The errors are manifested in the large loops on the surface of the protein, there is also a large error at the C-terminal end of the protein. In the above image divergence of the structures seems large but is actually a curse of the 2D representation. b) The structural superposition coloured from N-terminal (blue) to C-terminal (red), the structures are the same as in *a* but shown from a different angle.

In all four lumazine synthase runs (1di0_A) the correct fold was identified by the scoring functions. The highest scoring model had an RMSD of 4.7Å which, when considering the size of the protein (147 residues), is rather good as the loop regions are large and unstructured (figure 5.8). There were several topology violations, including the packing of a loop onto the edge of the β-sheet, resulting in an increased RMSD of 6.7Å.

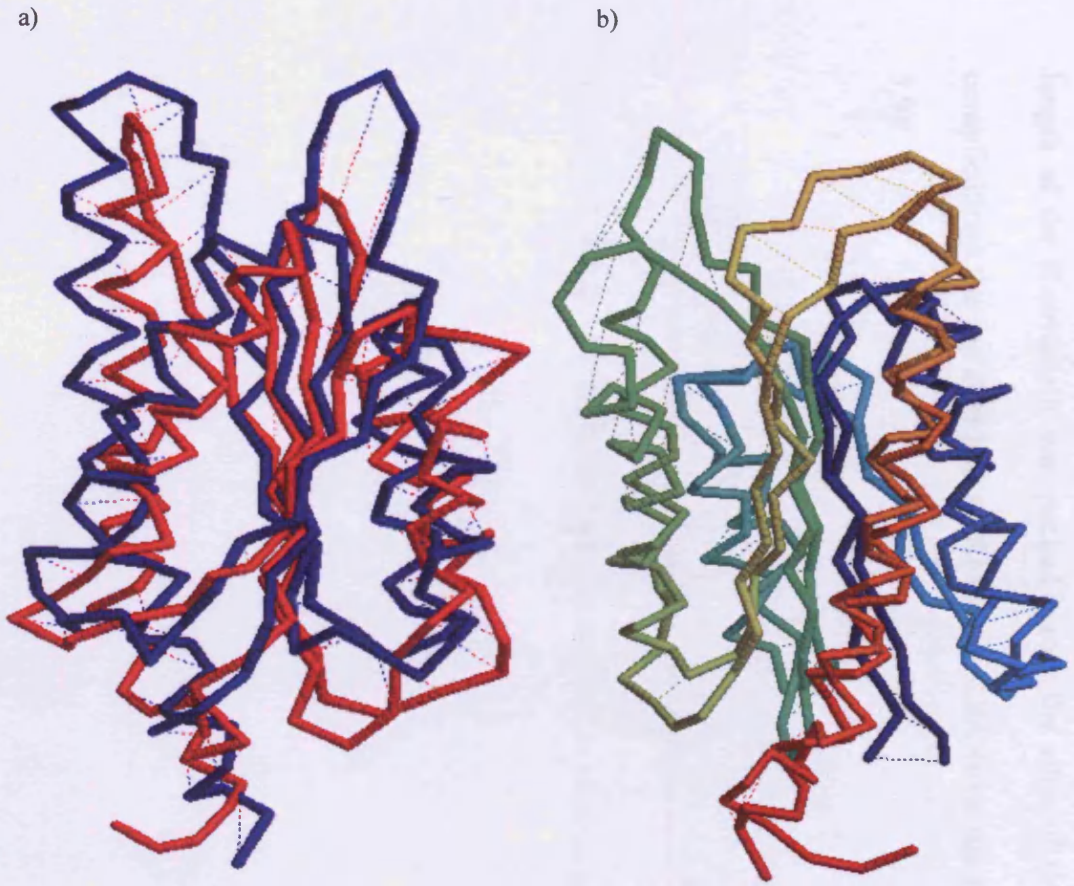


Figure 5.8 Lumazine synthase (1DI0 A): a) The blue structure is the native conformation and red is the top ranking model. The best model achieves a RMSD of 4.392Å using SAP. The overall structure is good however there are clear overextensions of loops and the long helix that packs along the back of the sheet is broken. b) The structural superposition coloured from N-terminal (blue) to C-terminal (red), the structures are the same as in *a* but shown from a different angle.

Thioredoxin (2trx_A) was also encouraging despite some minor errors. Fold competition was fierce and resulted in models with an incorrect series of connections scoring highest. It appears that this error is the result of poor modelling of the helix that bridges strands three and four. Rather than assuming the βαβ connection along the length of the β-strands it was packed across the edge of the sheet. Even with these complications the top fold had an RMSD of 4.8Å from the native structure (see figure 5.9).

a)

b)

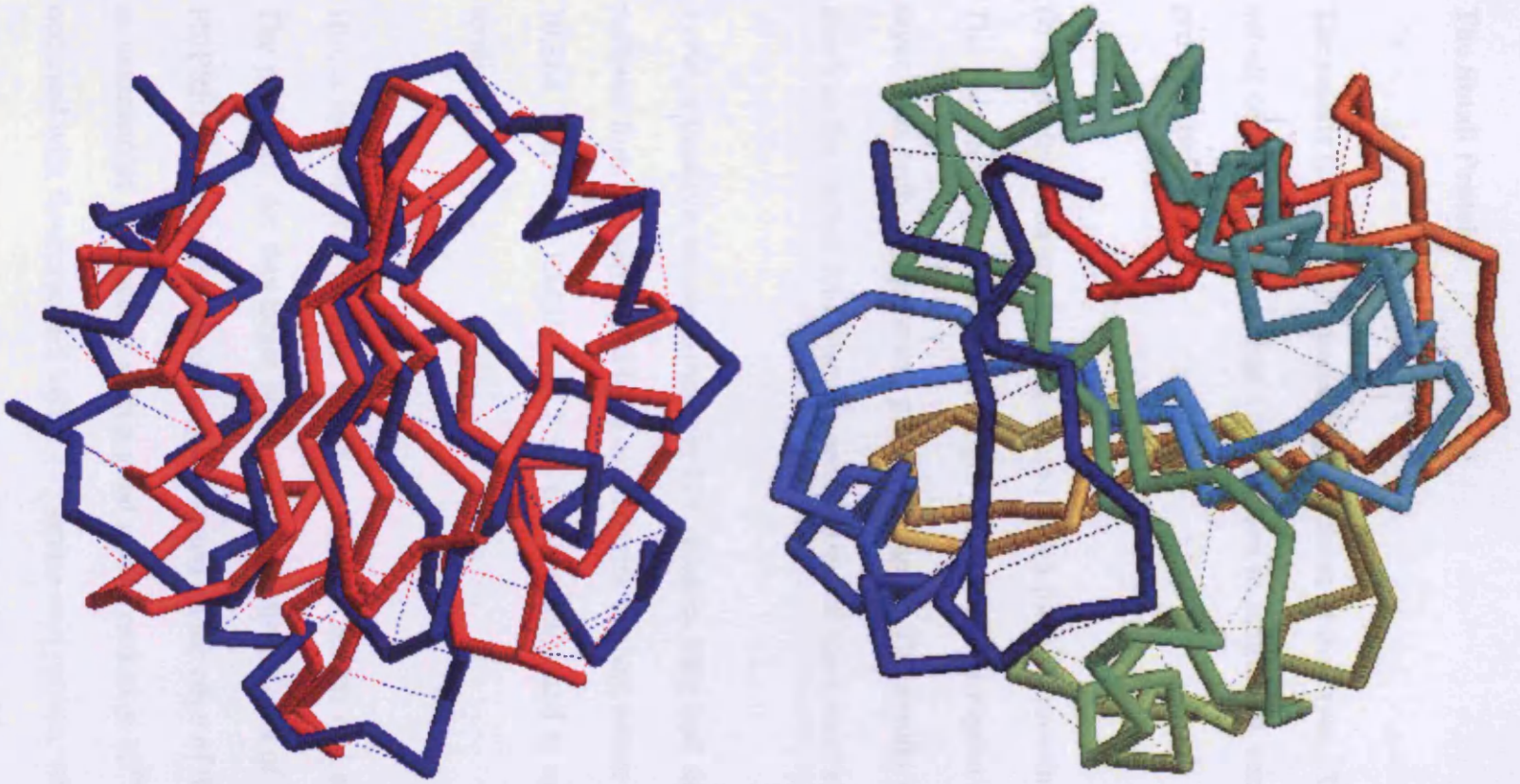


Figure 5.9 Thioredoxin (2TRX A): a) The blue structure is the native conformation and red is the top ranking model. The top scoring model has an RMSD of 4.815Å over 106 aligned residues. Despite there being some prediction errors with the helix that connects β-strands three and four the overall model is decent. b) The structural superposition coloured from N-terminal (blue) to C-terminal (red), the structures are the same as in *a* but shown from a different angle.

The Small Proteins

The results for the small proteins mirror those of the Fives. There were five proteins in the set all of which are less than 150 residues in length. A succinct review of the results is provided below.

Of all proteins the smallest considered was a putative protein from *Aquifex aeolicus* (1t6t). The protein is 108 residues in length and assumes a toprim domain fold which is a three layer αβα with a four strand parallel β-sheet. The results were similar to 1coz_A in as much as the correct fold was top ranked however there were errors in the C-terminal helix.

1v9w, a putative mouse protein, is 119 residues long and assumes a thioredoxin fold. It suffered from the same problems as 2trx with the best native ranked model finishing 9th of 10,364. The final RMSD was 7.5Å which is attributed to unstructured residues at the N-terminus.

1tjn, a hypothetical protein, is 135 residues in length and assumes a chelatase-like fold. The problems for this target arose during the first stages of the pipeline with Yaspin and PSIPRED being unable to identify a β-strand on the edge of the sheet. The results were not as catastrophic as expected with a good model ranked at 12th position. The same problem occurred with flavoprotein (1rlj), a 135 amino acid protein, where only two of 45 secondary structure predictions were accurate. Despite this problem the results were encouraging, with the top scoring model deviating by 4.6Å from the native structure.

The final structure in the small set is 1vk9, an ADP-phosphorylated protein from *Thermotoga maritima* which is 136 residues long and proved to be too much of a challenge for the pipeline. Secondary structure prediction was a complete failure with no correct combination of secondary structure elements occurring. There were no plausible predictions made for this structure.

The Large Proteins

The large protein set consisted of proteins greater than 150 amino acids in length. The results were similar to those of the small proteins where correct predictions of secondary structure were absent, to those where the top model was 4.5Å from the native structure. The typical outcome was that the top fold was not the native, but a similar fold in which a pair of β-strands had swapped places within the sheet (a buried-buried swap).

Methenyltetrahydrofolate synthetase (1sbq) proved to be the hardest target. At 189 residues it is one of the largest targets consisting of a number of short β-strands – DSSP shows 10 β-strands covering 35 residues and 7 helices covering 56 residues. In addition to this, the structure posses a large N-terminal helix that packs across the β-sheet instead of along it, a feature which is not accommodated in the Ideal Forms. The result of these features meant that none of the predictions were close to the native structure. When ranked on RMSD the 'best' model was 13Å from the native.

N5-glutamine methyltransferase (1vq1) is 178 residues in length. Unlike other targets the secondary structure was almost perfectly predicted. The exception was the C-terminal strand, which in the native structure is next to the edge of the sheet but, in models, was placed adjacent to the sheet in a more buried position. The overall hydrophobicity of the strand is greater than the one with which it had swapped. It is for this reason that it is understandable that the prediction and evaluation functions would favour this model. The RMSD of the top scoring structure was 7Å with much of the error derived from a large loop region where the chain is fragmented in the X-ray structure.

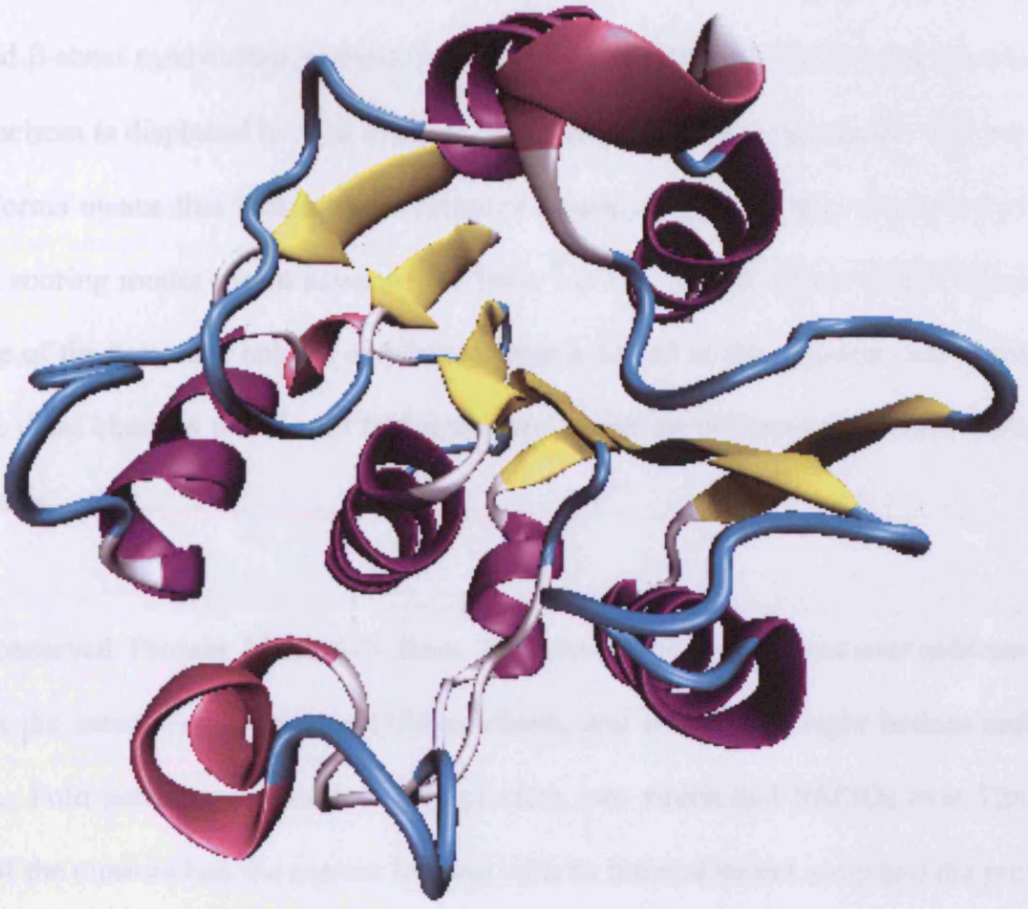


Figure 5.10 1UXO crystal structure of YDEN gene product: The structure consists of a six stranded β-sheet sandwiched between two sets of α-helices. On one side of the sheet there are two helices and on the other there are five. One helix pulls away from the structure, projecting 'out' of the page, and almost forms a new layer.

The YDEN gene product (1uxo figure 5.10) is 186 residues long and suffered from the same Ideal Form limitations that affected 1sbq. The native structure consists of a six stranded β-sheet sandwiched between two α-helices on one side and five on the other, one of the helices is displaced to such a degree that it almost forms a new layer. The use of the Ideal Forms means that helices were balanced around the sheets, this was reflected in the highest scoring model which assumed the form 3-6-3. The side-effect of this behaviour is that one of the helices is ablated and that another is forced to the opposite side of the sheet. Despite these changes the overall fold was correct with the differences manifesting in large loop shifts.

The Conserved Protein MTH1675 from *Methanobacterium thermoautotrophicum,* PDB 1t57, is the same length as 1uxo (186 residues), and consists of eight helices and seven strands. Fold assessment revealed three clusters, two which had RMSDs over 10Å. The lower of the clusters had the correct fold but with an internal strand swap and the prediction of the C-terminal strand as an α-helix. Further errors were introduced by a sequence of three helices, all which extend away from the body of the protein and are involved in subunit packing in the native multimeric state.

The final structure in the large set was that of Uracil-DNA glycosylase (TM0511) from Thermotoga maritima (1vk2). The highest scoring fold had some similarity to the native but with several strand swaps. The fifth ranked model was mostly correct with the exception of a β-hairpin inversion on the edge of the sheet at the C-terminus. The overall result of these minor changes is an RMSD of just under 10Å.

Performance at CASP7.

The approach to CASP7 differed to that described in chapter 3. There was no effort invested in comparative or fold recognition modelling, nor in targets which were less than 100 or greater than 200 residues in length. Rigorous checks were also made to remove traces of homology from sequence and structure resources, this was achieved using GenThreader (Jones, 1999a) and PSI-BLAST (Altschul et al., 1997). Sequences that returned 'significant' hits from PSI-BLAST searches were removed and only structures returning 'LOW' and 'GUESS' were considered as not violating the *de novo* threshold. From the possible 10 targets a further three were discarded because they were identified as all-alpha, one more was predicted as $\alpha+\beta$ and another remains unsolved.

In addition to running the *de novo* approach and using the Ideal Forms a second approach, similar to threading, was used. This method used native structures as templates instead of the Ideal Forms. The second approach was necessary to overcome the limitation of the Ideal Forms which currently only represents a small number of native structures.

Another difference between the above set, especially the fives, and the CASP targets is the absence/presence of multi-domain proteins. Multi-domain proteins pose problems for structure prediction because, while they fold and function semi-independently, the multi-domain structure can be dramatically different to the single domain. Before building models there were two problems to be aware of: the native templates reduce the options available to the *de novo* pipeline, while poor domain definitions disrupt prediction accuracy

when compared to the results reported earlier. To allow comparison of the original *de novo* method, the first set of results will cover target prediction using the Ideal Forms approach and the second set will cover the template based approach.

Target T0273 is a multi-domain protein. The domain definition used in this work was placed at the 150th residue and, unfortunately, contained some of the following domain. The 120 residues which were correctly identified by the domain definition assumed the correct fold, however differences were identified in the first edge strand, while the following helix and β-strand were not predicted. Using the SAPit program, an RMSD of 8Å was calculated over the 120 residues.

Target T0299, the structure of conserved bacterial protein SP0830 (2hiy), is 180 residues and consists of 10 α-helices covering 73 residues and 8 β-strands covering 45 residues. The N-terminal domain of the protein was correctly predicted, ranking 3rd overall, and included the location of the domain swapped C-terminal helix. The native structure includes an internal duplication with each domain having its own distinct sheet. Although the method does not accommodate this feature, the overall RMSD of the best model is 6.5Å over 100 residues.

Target T0357, the NMR solution structure of UPF0107 protein AF_0055 (2hi6), is 132 residues long and consists of 3 α-helices and 11 β-strands. The structure assume an αββ architecture which is correctly predicted by the pipeline and identified as the top prediction.

Target T0383 (2hng) is 124 residues long and consists of three α-helices and 5 β-strands. Each of the edge strands faced in the opposite direction to the native structure but, despite this, the core topology of the protein was correctly predicted.

Target T0353 was a failure due to an error in secondary structure prediction which missed a β-strand. Despite this a reasonable prediction, with an adjacent strand swap, was found at rank 23. The final two targets, T0319 & T0350 were dominated by alpha-helix packing. Furthermore T0319 could reasonably be excluded from the set because of the secondary structure prediction composition. T0250 had disordered termini which reduced the size of the protein to under 100 residues, thus eliminating it from the set. Predictions for these two targets was poor.

CASP7 and Native Forms

As mentioned previously, some CASP targets could not be approached using the *de novo* method described above. This is because, in their current state, the Ideal Forms have only been constructed for the α/β class of proteins, additionally the approach is designed around proteins in excess of 100 residues but less than 250 – targets that would be prohibitive for any other *de novo* method. To avoid this limitation and allow for more participation in the CASP7 event, the method was altered. Without the Ideal Forms there exists no lattice on which to build the initial structures, to solve this problem a 'standard' threading approach is used. As described previously all sequence and structure information was removed to avoid reducing the problem to comparative modelling (as described in chapter 2).

GenThreader was used to identify reasonable templates – those which did not share overwhelming structural similarity with the target. GenThreader returns a list of potential templates with a *p*-value which infers a likelihood of the fold being incorrectly assigned. In this work, templates with a score of low ($1\% < p < 10\%$) and guess ($p > 10\%$) were considered as suitable templates. These templates were then used in place of the Ideal Forms.

Using the modified approach, described above, 20 targets were attempted. Table 5.1 shows the Cα RMS deviation (cRMSD) of each target from the solved structure using a sequence independent structural alignment. The most interesting looking result is that of domain one of target T0356 which has a cRMSD of 4.93Å, however this is somewhat misleading as the model represents only a small fraction of the target, consisting of a long helix and loop which incorrectly bridges the two domains in the native structure.

Table 5.1 CASP7 Template Based Prediction Results

Target ID	Target & Domain ID	Target Length	cRMSD
T0348	T0348	68	9.35
T0353	T0353 D1	83	9.15
T0350	T0350 D1	91	11.13
T0300	T0300	102	16.77
T0347	T0347 D2	71	10.76
T0304	T0304 D1	101	13.59
T0382	T0382 D1	119	13.02
T0309	T0309	70	15.57
T0307	T0307 D1	123	13.59
T0356	T0356 D1	90	12.73
T0361	T0361 D1	158	19.22
T0314	T0314 D1	103	15.33
T0356	T0356 D3	99	14.42
T0321	T0321 D2	155	15.7
T0287	T0287	199	18.63
T0319	T0319	135	16.49
T0347	T0347	205	15.38
T0296	T0296	214	12.74
T0321	T0321	251	18.38
T0356	T0356 D1	20	4.93
T0356	T0356	168	16.96
T0296	T0296	231	16.34
T0356	T0356	119	15.79
T0356	T0356	92	14.59
T0356	T0356	67	16.87

The Target ID is the standard CASP target nomenclature. cRMSD is the carbon-α root mean squared deviation from the native structure, in the case of NMR structures the first chain in the file.

A number of the targets (13) fail to meet the size restraints either being less or greater than the desired length (100-200 residues) but are still run through the pipeline. While the cRMSDs are large 9-19Å the same observations were made for the template based approach as for those based on the Ideal Forms. A summary of some of the more interesting structures is provided below.

T0300 – the native structure has a long helix which forms the axis of the protein, at either end there are loops which connect to short helices. The predicted structure breaks the long helix into two, one half of which double backs and packs against the first half and a correctly predicted helix. Visually the structure appeared to be native-like, possessing a compact structure. Had the helix axis helix not been split the overall structure would have been a good approximation.

T0353 - the native structure consists of three α-helices and four β-sheets packed in an anti-parallel sheet under which the helices pack. The overall prediction was good with the exception of the loop modelling which is where the large deviations are identified. The odd helix packing for this structure is not currently supported by the Ideal Forms.

The native structure of T0350 consists of a beta sheet which is packed against three helices, however unlike the balanced ideal forms the helices all pack on one side of the sheet. When run through the prediction pipeline the ideal forms attempt to balance the combination resulting with a 1-3-2 form.

T0296 is a large structure, over 400 residues in length. The prediction, which only covers a small section of the structure, has elements representative of globular proteins but does not accurately account for interconnecting loops. The result was a top scoring model with an RMSD of 12.47Å from the native structure.

T0382 - an all alpha protein, 121 residues in length consists of 6 helices. As with previous predictions, the overall structure is native-like with packed helices. The error comes from the unstructured N and C terminals as well as the miss-prediction of two helices and interconnecting loops. The inaccurate prediction of secondary structure resulted in one set of helices being extended and the other reduced.

T0307 - the overall structure of the protein is not well predicted as a result of poor secondary structure prediction. The poor predictions result in stunted α-helices connected by extended loop regions. The problem with a helix bundle is the number of ways that the helices can pack together.

T0361 – is another helix bundle, DSSP identifies 11 helices which are poorly predicted. Additionally the native structure is dimeric and involves a helix swap between domains. The model produced by the pipeline, while not an accurate prediction of the target, does look native-like. The predicted secondary structure, rather than indicating 11 separate helices, combines them into three extended structures which pack together.

In summary the models, while not perfect, all exhibit native-like protein features including compact overall structure and biologically realistic partitioning of the hydrophilic/hydrophobic amino acids. The test of the pipeline using the native structures in place of the Ideal Forms, the base on which the method was constructed, was foolhardy as it replaces one of the most important aspects of the prediction pipeline. Further to this, several of the targets within the size range are classed as all-α, which unlike the α/β

proteins had not undergone optimisation. It has also been suggested that the all-α class proteins prove particularly challenging for evaluation functions because of the number of ways in which helices can pack together (Berglund et al., 2004). Despite these observations the level of accuracy across the groups of proteins is consistent with the original 'fives' set.

How could performance be improved?

Table (5.3) shows the results for the free modelling targets presented at CASP7, these are the targets that should have been approached using a fold recognition or *de novo* approach. The table itself presents the target id, the group whose model was highest ranked, the rank of our model and the number of models submitted. The data shows that the performance of the *de novo* method had mixed success, on some targets it performs better than other methods (T0296, T0356_D1), while on other targets, it performs poorly (T0309, T0361_D1).

The performance of our method compared to other groups, was distinctly average as shown in table 5.2. As described previously, the approach to each target was to manipulate the starting information such that it could be attempted as a true *de novo* target this instantly introduces a handicap for where a suitable template can be identified using threading-like techniques. Thus the methodology, while it suited our purposes, will produce misleading comparison between our method and others, simply because template fragment based modelling approaches may have been better suited.

Table 5.2: Rank of Free Modelling Targets using the *de novo* prediction pipeline

CASP Target	Top ranked group name	Taylor group rank	Number of models submitted
T0287	Pcons6	37	121
T0296	mGen-3D	2	134
T0300	karypis.srv.4	94	135
T0304_D1	Zhang-Server	69	137
T0307_D1	panther2	41	136
T0309	mGen-3D	113	138
T0314_D1	POMYSL	83	133
T0319	SAM-T02	68	132
T0321	nFOLD	61	124
T0321_D2	FEIG	86	120
T0347	SAM-T99	31	135
T0347_D2	FORTE1	41	128
T0348	Akagi	27	142
T0350_D1	SAM-T02	47	138
T0353_D1	SAM-T02	14	147
T0356	nFOLD	9	137
T0356_D1	Jones-UCL	2	118
T0356_D3	UNI-EID_expm	28	115
T0361_D1	AMU-Biology	103	132
T0382_D1	SAM-T02	82	141

Where comparisons are made it is important to remember that during CASP7 the pipeline was in its infancy, possessing only a small number of the required ideal forms, some of which were subsequently used inappropriately. As I have already stated, the performance of our method appears to be consistently average, perhaps a more interesting question to address is "how could the method be improved?".

The most obvious way to improve the current method would be to invest time in the development of a full 'periodic table' of ideal protein folds as proposed in (Taylor, 2002). Of course, before using the ideal forms, an accurate prediction of secondary structure is

required and this is one thing that is difficult to improve. At the time of writing, secondary structure prediction is performed using two methods, PSI-PRED and Yaspin, which limits the amount of variation the method requires. It may then be of worthwhile designing a new method to predict secondary structure as detailed in chapter 6 or, as is done by several server based predictors, use what is termed a metaserver to obtain predictions of secondary structure from multiple sources. Both methods would introduce needed variation, the second considerably more. One problem that is associated with increasing the amount of variation in secondary structure is the amount of time required per target. It is possible that a considerable amount of time could be saved by optimising the code base of the prediction pipeline, this includes removing many of the scripts which tie the existing modules together. A more complicated improvement would be to introduce a final all-atom phase at the end of the pipeline, this procedure would bring the method into line with other, better performing, structure prediction methods (Simons et al., 1999a). More work would be required than with the aforementioned changes as it would be necessary to introduce side chain addition and adjustment algorithms i.e. SCWRL (Canutescu et al., 2003), refinement functions and new evaluation routines such as those used in (Simons et al., 1999a, Qiu et al., 2007).

Conclusion

The method described here marks a successful return to combinatorial modelling, enabling the prediction of larger proteins than was previously possible. The structures produced by this method, both like and unlike the target's true structure, display features that are typically found in globular proteins. It would be easy to dismiss such models as complete failures, however they are important as they show that the pipeline is explores realistic fold space. The results for 'the fives' were interesting as, when pooled, they were more reliable. This suggests that increased sampling can lead to improved overall accuracy, an observation which was also made in chapter 3, however it should be noted that a 'brute force' method is not thought to be a realistic solution to protein structure prediction.

Consistent errors in models are typically a result of misprediction of secondary structure or limitations in the number of Ideal Forms. This suggests that the α/β Ideal Forms require some extension – allowing for 'off lattice' features which are necessary for protein-protein, or other, interactions. Further problems are encountered as a result of secondary structure prediction. Incorrect predictions can be partially solved by using each sequence in the multiple sequence alignment to predict secondary structure, this procedure typically resulted in at least one prediction being a close approximation of the native structure. As mentioned in the discussion, further variation may be included through the application of further prediction methods and allowing more sequences into the multiple sequence alignment.

In conclusion, the *de novo* method has been shown to accurately predict the three dimensional structure of proteins in excess of 100 residues, it also marks a successful return to combinatorial modelling for such proteins. For those targets where structurally remote models were produced, the errors were often the result of an interchange of two elements between buried environments, suggesting that the method samples realistic fold space. We believe that the inclusion of side-chain/all atom scoring functions and side-chain adding tool(s) will help improve models produced by this method in the future.

Chapter 6

Algorithmic Protein Structure Prediction: Improving pipeline performance

Introduction

The prediction of two dimensional structure started in the mid to late 1970s with the prediction of protein secondary structure and conformation (Sternberg and Thornton, 1978, Chou and Fasman, 1974, Chou and Fasman, 1978). Today it remains an active field both in isolation and as part of 3D structure prediction, and since the seventies more progress has been made in the 2D field than in 3D. Indeed it has been suggested that secondary structure has reached its theoretical limit at approximately 75% ± 10 while new fields have been identified such as disorder and contact prediction. 2D features are not independent and address characteristics that are useful for experimentalists as well as theoreticians.

As the array of predictable structural features has increased so to has the array of methods at our disposal, these include simple approaches, like the nearest neighbour methods, to more complex machine learning approaches. Among the very first techniques for prediction of secondary structure was that of Chou and Fasman (Chou and Fasman, 1978). This technique relied upon the probability parameters determined from relative frequencies of each amino acids appearances in each secondary structure type. By modern standards it is basic and this is reflected in a prediction of accuracy of 50-60%. This was quickly followed by Garnier, Osguthorpe and Robinsons' method (Garnier et al., 1978) which utilises the probability of an amino acid being in a particular structure as well as the conditional probability of its neighbours assuming the same structure. The incorporation of this extra information gained a 5% increase in accuracy which is attributed to much improved alpha helix prediction at the expense of beta sheet prediction. These methods

brought a close to the first generation of methods – those that relied on single amino acids and their propensities for particular structures (Rost and Sander, 1993, Rost and Sander, 2000). The methods that followed applied similar ideas to segments of adjacent residues but no matter what underlying algorithm was used, prediction accuracies stuttered at 60%. This problem was largely solved through the introduction of sequence variation in the form of multiple sequence alignments (Dickerson et al., 1976) and one of the first people to capitalise on this was Zvelebil (Zvelebil et al., 1987) whom incorporated multiple sequence alignments into an automatic prediction method. Rost refers to this group of methods as the, very brief, second generation of secondary structure prediction tools that struggle to break the 70% barrier (Rost and Sander, 2000). It is then, the third generation where the final breakthrough occurs through the use of sequence profiles, larger databases and new algorithms. It was the widespread adoption of new techniques and data sources that lead the current accuracies which approach 80% (based on three state prediction: alpha helix; beta sheet; coil/rest). One of the first 'new' algorithms was the artificial neural network (ANN) (Minksy and Papert, 1969, Rosenblatt, 1988, Widrow and Hoff, 1988, Minksy and Edmonds, 1954) which was brought to the attention of the wider community by Rumelhart (Rumelhart et al., 1986), the ANN was first applied in secondary structure prediction by Qian and Sejnowski (Qian and Sejnowski, 1988). The next ten years saw the growth of databases accompanying variation on the neural network theme (Rost, 1996) as well as the new application of 'old' methods including Bayesian statistics (Thompson and Goldstein, 1997). What is generally agreed as one of the most significant steps forward was made by David Jones using a combination of the ANN and PSI-BLAST in a method that he called PSIPRED (Jones, 1999b). In constructing PSIPRED, Jones was not only one of the first

people to effectively apply the PSI-BLAST position specific scoring matrix (pssm) profiles in secondary structure prediction, but also the first to apply a rigours culling of the sequence databases to avoid pollution of the PSSMs through spurious hits to unrelated proteins (Jones, 1997). While PSIPRED is a popular choice, it is worth noting that at the same time, Karplus *et al.,* introduced an alternative method that used hidden Markov models to search sequence databases for remote homology before making predictions (Orengo et al., 1999, Karplus et al., 1998, Karplus et al., 1999). Over the last eight years there have been many more methods for secondary structure (Cuff and Barton, 2000) and solvent accessibility (Adamczak et al., 2004, Ahmad et al., 2003) prediction (more often than not, the same tools are used for each). Over the last five years the fashions in structure prediction have changed, this has included the increased usage of a machine learning tool called support vector machines (Hu and Li, 2007, Ward et al., 2003, Shamim et al., 2007, Kajan and Rychlewski, 2007, von Grotthuss et al., 2003, Ginalski and Rychlewski, 2003, Ginalski et al., 2003). Another popular method is the combination of existing methodologies into what are often referred to as metaservers, of which 3D-jury (von Grotthuss et al., 2003, Ginalski and Rychlewski, 2003, Ginalski et al., 2003) is one of the better (Kajan and Rychlewski, 2007).

In this chapter I will introduce a novel method for prediction of secondary structure and solvent accessibility. The method consists of two parts, a fuzzy *k* nearest neighbour (f*k*NN) algorithm and support vector classification (SVC) machine. This approach is different to the aforementioned methods sitting at the interface of the meta-servers and the single method predictors. The aim of this work was to provide an additional method for

prediction of secondary structure for α/β structure prediction and an new alternative method for prediction of solvent accessibility. Both of these features had been identified as playing a crucial role in the overall prediction of three dimensional structure (see chapters 3, 4 and 5).

Methods and Materials

Three datasets were used for construction and evaluation of the method used in this work. To be included in each set the following criteria had to be met: The maximum pairwise sequence identity across all proteins had to be less than 25%, this is standard protocol for construction of machine learning tools; the structure had to be determined by X-ray crystallography to a resolution better than 2.5Å and contain no chain breaks or missing atoms; the protein had to be globular – to this end all proteins which were not identified in one of the four major SCOP classes were discarded - this includes small proteins and transmembrane proteins; all proteins less than 60 or greater than 500 residues were also discarded.

The first structure set was derived from the Representative PDB (Noguchi et al., 1997) and consisted of 764 proteins. The second set consisted of 1094 proteins which were derived from the DSSP select 25 list. The third and final set were identified using the PISCES server (Wang and Dunbrack, 2003), this list formed the final protein set comprising of 1024 proteins.

Sequence Alignments

Sequences were extracted from the PDB files and alignments were generated using the PSI-BLAST program (Altschul et al., 1997), a standard command was used:

-j 3 –h 0.001 –e 0.001 –F T –i <input file> –d <database> -Q <PSSM FILE>

Where –j controls the number of iterations, -e defines the expectation value, -F T switches filtering with SEG on – this was not required as the nr database was prefiltered using pfilt but was left on as a precaution. Several scoring matrices were tested but the BLOSUM62 matrix was applied as initial tests did not yield an obvious advantage in using other matrices.

Vectors

Two sets of vectors were constructed for this work. Both are based on the information from within the PSSM but the second uses information from the Taylor colour scheme (Taylor, 1997b) as a supplement. Each set is described below and were constructed for each dataset.

Set 1: The transition matrix.

Sequence vectors were constructed from the raw PSSM values. Each amino acid is represented by a 21 element vector. The first 20 elements pertain to the transition from the residue at the current position to each of the other amino acids. The 21st element indicates an unknown position, this can be an unknown or missing residue as well as a pseudo-residue. The window scheme used in this work means that sampling occurs outside sequence space (i.e. beyond the N and C termini), to make this possible pseudo-residues are used. In a pseudo-vector the first 20 positions are set to zero and the 21st is set, arbitrarily, to 0.5. In this work a window length of 15 residues (7 residues either side of the central residue) was used, resulting in feature vectors of 15 * 21 = 315 dimensions.

Set 2: Transition matrix and entropy measures.

Raw data was extracted from PSSMs as in described above. The vector were supplemented using two measures, sequence and hydrophobic entropy. The equation used is shown in 6.1. The information content of a position (x_i) is measured in bits, the lower the bit value the more conserved a position is. This is a simple measure of sequence conservation but not so arbitrary for hydrophobic entropy.

$$\sum_{i=1}^{l} p(x_i)\log_2(1/p(x_i)) \qquad (6.1)$$

To address the hydrophobic problem all amino acids identified by Taylor (Taylor, 1986) as hydrophobic (A G C T K H Y W F M I L V) were grouped into one class and the remaining amino acids into another (P S N D E Q R) (Taylor, 1986). No attempt was made to optimise the window length for entropic features. The resulting vectors were of 345 (315 + 15 + 15) dimensions.

Prediction Methods

Two methods were applied during this work, the fuzzy *k* nearest neighbour algorithm (f*k*NN) which had not been applied to solvent accessibility or secondary structure and support vector classification with novel vector encoding. As well as trying each method independently the methods were combined into a combination f*k*NN-SVC for prediction of secondary structure and solvent accessibility as shown in figure 6.1.

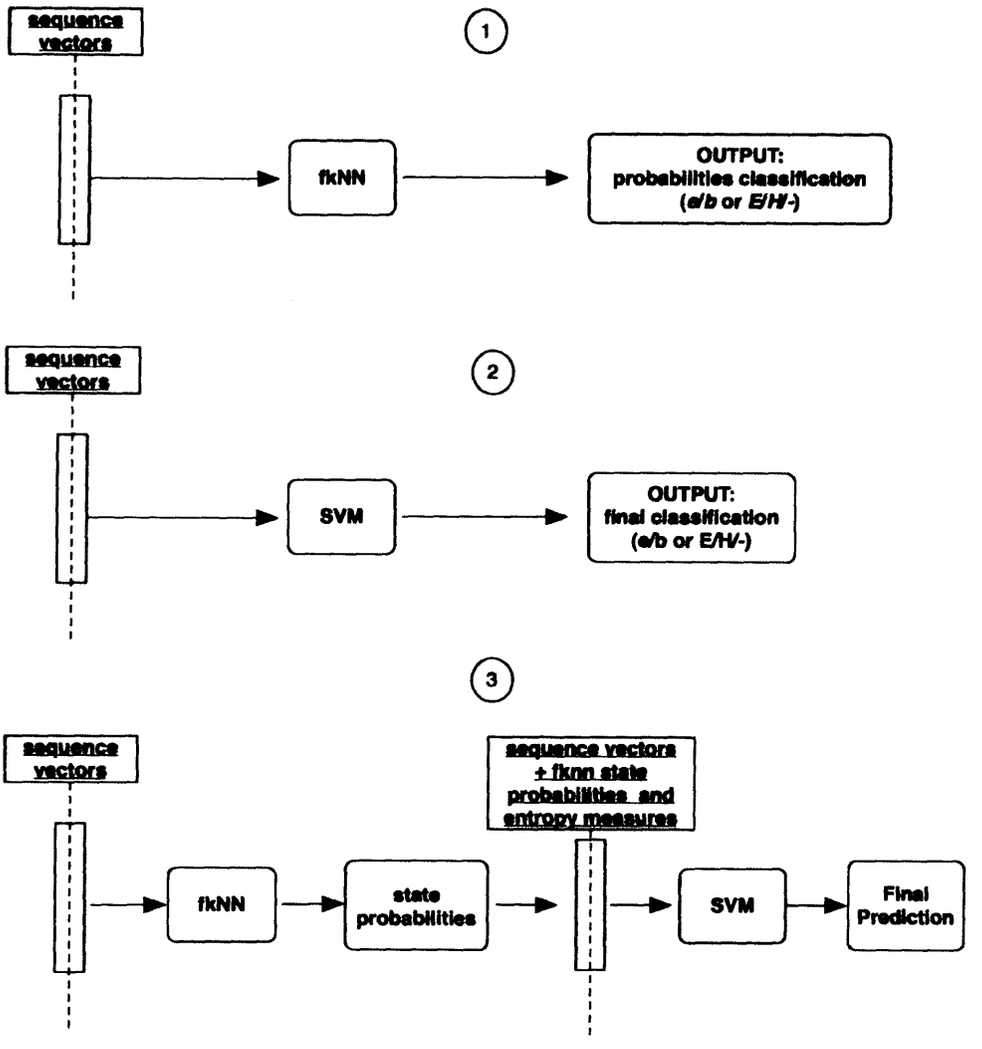


Figure 6.1 An outline of prediction methodology: 1 & 2) The first approach was to use pssm information as direct input to both the *fk*NN and an SVM. The *fk*NN produces a final prediction in the form of a probability while the SVM produces a distinct answer (buried, exposed, helix, sheet or coil). 3) the *fk*NN and SVM are combined so that the predicted probabilities produced by the *fk*NN are combined with the original input vectors and entropy measures before being passed to the SVM. The SVM then produces a distinct output of buried, exposed, helix, sheet, coil.

***k* and fuzzy-*k* nearest neighbour: Predicting solvent accessibility and secondary structure.**

The fuzzy *k* nearest neighbour algorithm is a simple technique for assigning a class or a value to an unknown quantity. It is derived from the *k* nearest neighbour algorithm (*k*NN) (equation 6.2), the difference being the weight parameter (equation 6.3).

$$c_n(x_i) = \frac{\sum_{j=1}^{k} c_n(d_{ij})}{\sum_{j=1}^{k} d_{ij}} \qquad (6.2)$$

In a standard *k*NN each neighbour is given an equal weight – this is as simple as counting *k* neighbours and assigning class based on the most numerous known class. The f*k*NN adds a weight to each of the *k* so that the closer the k^{th} element is to the unknown sample (*i*), the greater the contribution to the classification of *i*.

$$c_n(x_i) = \frac{\sum_{j=1}^{k} c_n(d_{ij}^{-2/(m-1)})}{\sum_{j=1}^{k} d_{ij}^{-2/(m-1)}} \qquad (6.3)$$

The weights are achieved by adding the $-2/(m-1)$. The parameter m is often called the 'fuzzifier' and must be optimised along with k. The d_{ij} parameter is a measure of distance, there are many potential ways to do this, however in this work, two methods were used – Euclidean Distance (equation 6.4) and Manhattan City Block (CB) distance (equation 6.5). While the Euclidean distance is more accurate than the Manhattan distance, mainly because of the square boundaries imposed by the later, however it has consequences for compute time. Requiring both a power and a square root function imposes a time penalty on operation speed, and at least theoretically becomes prohibitive when a large number of calculations have to be made.

$$d_{AB} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2} \qquad (6.4)$$

The Euclidean distance between 2 points (A & B), where A = (a_1, a_2, … a_n) and B = (b_1, b_2, … b_n).

$$d_{AB} = \sum_{i=1}^{n}|a_i - b_i| = |(a_1 - b_1) + (a_2 - b_2) + \ldots + (a_n - b_n)| \qquad (6.5)$$

The City Block distance between 2 points (A & B), where A = (a_1, a_2,…, a_n) and B = (b_1,b_2,…, b_n).

Support Vector Classification.

SVMs are used to construct optimal class separating hyperplanes in a high dimensional feature space. Most architectures are able to deal with sample sizes greater than 100,000 instances and lend themselves well to biological application. In the introduction the concept on support vector machines was introduced in the form of a linearly separable

problem. In real-world problems linear separation is rarely a reality and two advanced features of SVMs have to be exploited: the handling of miss-classified instances, this is achieved through the introduction of 'slack variables' (ξ); the projection of data into a higher dimension, this is achieved using the 'kernel trick' (see chapter 1).

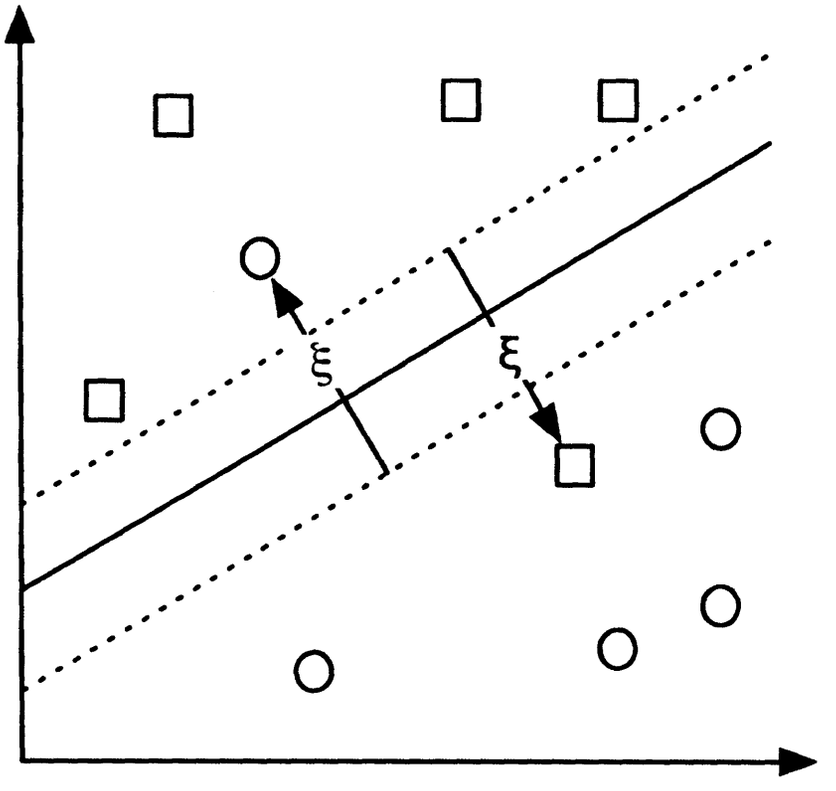


Figure 6.2 Classification: The Application of Slack Variables: In the classification the slack variables measure the violation of the support hyperplanes as shown. In regression the slack variables measure the deviation outside the support hyperplanes. As with all previous examples this diagram illustrates the point using a linearly separable problem.

The introduction of slack is achieved as follows and is illustrated above in figure 6.2. The primal form (linear programming problem) for the maximal margin is stated as:

$$\text{minimise}_{\text{w,b}} \ \langle w \cdot w \rangle,$$
$$\text{subject to } y_i\left(\langle w \cdot x_i \rangle + b\right) \geq 1, i = 1,\ldots,\ell. \qquad (6.6)$$

where w is the weight vector, b is the bias, y_i is the label of the current instance and ℓ is the length of the feature vector. To optimise the margin slack vector, the slack variable have to be incorporated to allow margin constraint violation:

$$\text{minimise}_{\text{w,b}} \ \langle w \cdot w \rangle,$$
$$\text{subject to } \begin{matrix} y_i\left(\langle w \cdot x_i \rangle + b\right) \geq 1 - \xi, i = 1,\ldots,\ell, \\ \xi \geq 0, i = 1,\ldots,\ell \end{matrix} \qquad (6.7)$$

Including the C parameter, the optimisation problem is re-written:

$$\text{minimise}_{\xi,\text{w,b}} \ \langle w \cdot w \rangle + C\sum_{i=1}^{\ell} \xi_i^2$$

$$\text{subject to } \begin{matrix} y_i\left(\langle w \cdot x_i \rangle + b\right) \geq 1 - \xi_i, i = 1,\ldots,\ell, \\ \xi_i \geq 0, i = 1,\ldots,\ell \end{matrix} \qquad (6.8)$$

The C parameter, or coefficient, affects the trade off between complexity and proportion of non-separable samples – the margin and the size of the slack variables. Shawe-Talor and

Cristianini note that it has no intuitive meaning and that it must be optimised by the user (Cristianini and Shawe-Taylor, 2000).

To address the problem of non-linearly separable problems Boser, Guyon & Vapnik (Bernhard et al., 1992) suggested a way to create non-linear classifiers by applying the *kernel trick* to maximum margin hyperplanes. The resulting algorithm is similar to the original method with the exception that each dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum margin hyperplane in a transformed feature space. There a several functions to complete transformations, in this work the Radial Basis Function (RBF) (equation 6.9) was implemented as it is equally able to deal with linear and non-linearly separable problems.

$$\exp\left(-\gamma\left\|x - x'\right\|^2\right) \text{ for } \gamma > 0. \qquad (6.9)$$

In this work C and γ were optimised using five fold cross-validation. For secondary structure prediction the optimal value of C was 4, for solvent accessibility C was found to be 5. For both secondary structure and solvent accessibility the optimal value of the γ parameter was found to be 0.001. A combination of the libsvm (Chih-Chung and Chih-Jen, 2001) and *b*SVM tools were used for secondary structure prediction while libsvm and svm^{light} (Joachims, 1999) was used for solvent accessibility.

Solvent Accessibility

Traditionally solvent accessibility has been treated as a classification problem. The work of Thompson and Goldstein (Thompson and Goldstein, 1996) is a prime example of this. Such an approach requires the definition of a threshold; this is the 'value' at which a residue can be defined as exposed or buried. To do this the absolute solvent-accessible area (asa) is determined using DSSP and NACS (correlation coefficient 0.98); then using equation 6.10 the relative solvent accessibility (rsa) is calculated – in the case of DSSP, Gromiha's (Ahmad and Gromiha, 2002) maximum solvent accessible areas (Max_x) were used.

$$RSA(x) = \left(\frac{ASA_x}{Max_x}\right) \cdot 100 \qquad (6.10)$$

The typical procedure is to use two or three values for the threshold, one of which separates the data such that 50% of the residues are classified as buried or exposed, paradigmatically this is 20-30% range. The balanced sets make training easier, when classes become unbalanced extra penalties should be imposed such that a misclassification of a minority class in more heavily penalised than that of the majority. Like the C and γ parameters this is something that has to be optimised.

Secondary Structure

Secondary structure was determined using DSSP. The standard eight-state to three state conversion was used where H, G, I are α-helical structures, E is β-structures and the remaining classes are grouped to represent coils. Beta-bridges were included as the 'random' coil, because, as described in the introduction, secondary structure elements are defined by repetitive phi-psi angles. Isolated predictions of α or β structure were left unaltered.

Results

The following sections give results for prediction of secondary structure and solvent accessibility including the accuracy of the f*k*NN and SVM as well as the combination of the two. The results presented below for the f*k*NN are based on the transition vectors only while the SVM results are based on the transition-entropy vectors combined with the f*k*NN output unless stated otherwise.

Secondary Structure.

The overall effect of varying k is shown in figure 6.3. The optimal value of k in the leave one out validation was determined as 60, however a range of values from 50 to 60 achieves almost identical results as shown in the table 6.1. The overall accuracy is determined to be 75% which is standard performance for secondary structure prediction and is 1% less than that of the stand-alone SVM on the transition vectors.

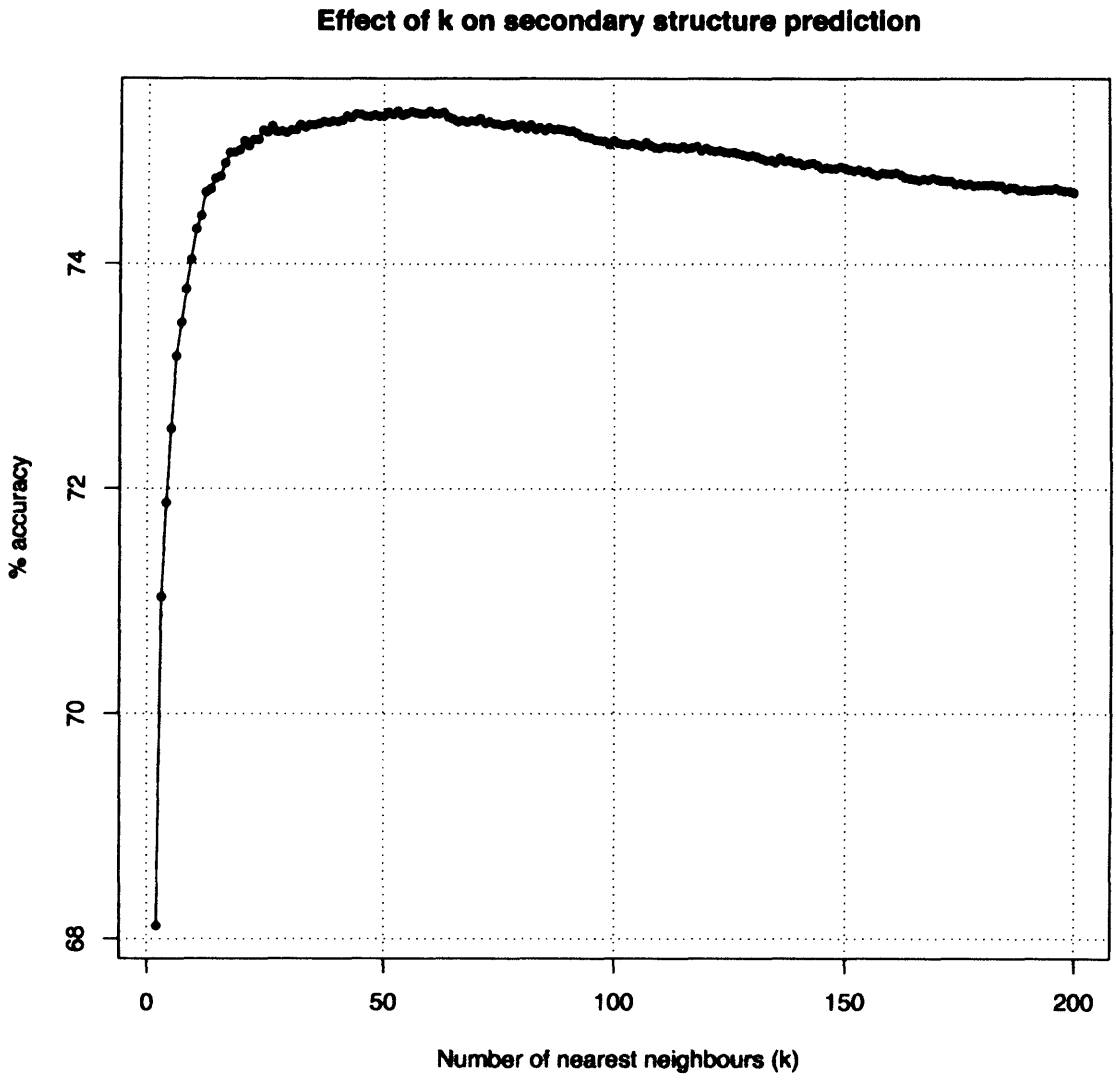


Figure 6.3 The Effect of *k* on Secondary Structure Prediction: The two parameters to be optimised with the fuzzy *k* nearest neighbour (f*k*NN) – *k* and *m*. With *m* fixed at 2, *k* can be varied such that an optimal prediction is obtained on a leave-one-out cross validation. The plot shows how the overall accuracy of the f*k*NN method changes with the number of *k*s used to assign a class to the neighbour. The optimal value of *k* is 60, however the accuracies around $k = 60$ are fairly similar as shown in table 6.1.

Table 6.1 Accuracy of fkNN on secondary structure prediction

Number of k	60	53	56	51
Accuracy	75.363	75.361	75.356	75.352

The table shows the accuracy of the f*k*NN on a leave one out cross validation of 1024 proteins. Despite the optimal value of *k* being 60, almost identical results are obtained using values of 53, 56 and 51, this is also shown in figure 6.3.

Combination of f*k*NN and SVM

The combination of the f*k*NN and the SVM yielded slight improvements over each of the individual techniques resulting in an average accuracy of 78.8%. The Q3 scores, the accuracy of prediction of each state, are as follows: Q_3- 80.5%; Q_3H 82%; Q_3E 71.7, these results are comparable to that of YASPIN.

Solvent Accessibility.

The original aim was to predict RSA by using the f*k*NN to approximate the RSA using a weighted mean. Due to problems establishing a suitable weighting scheme changes were made to the method to predict RSA using a threshold approach similar to that described in chapter 4. The first step in the method was to establish the threshold at which 50% of residues are classified as buried or exposed. The results are shown in figure 6.4 and clearly indicated that this point occurs approximately at the 20% threshold.

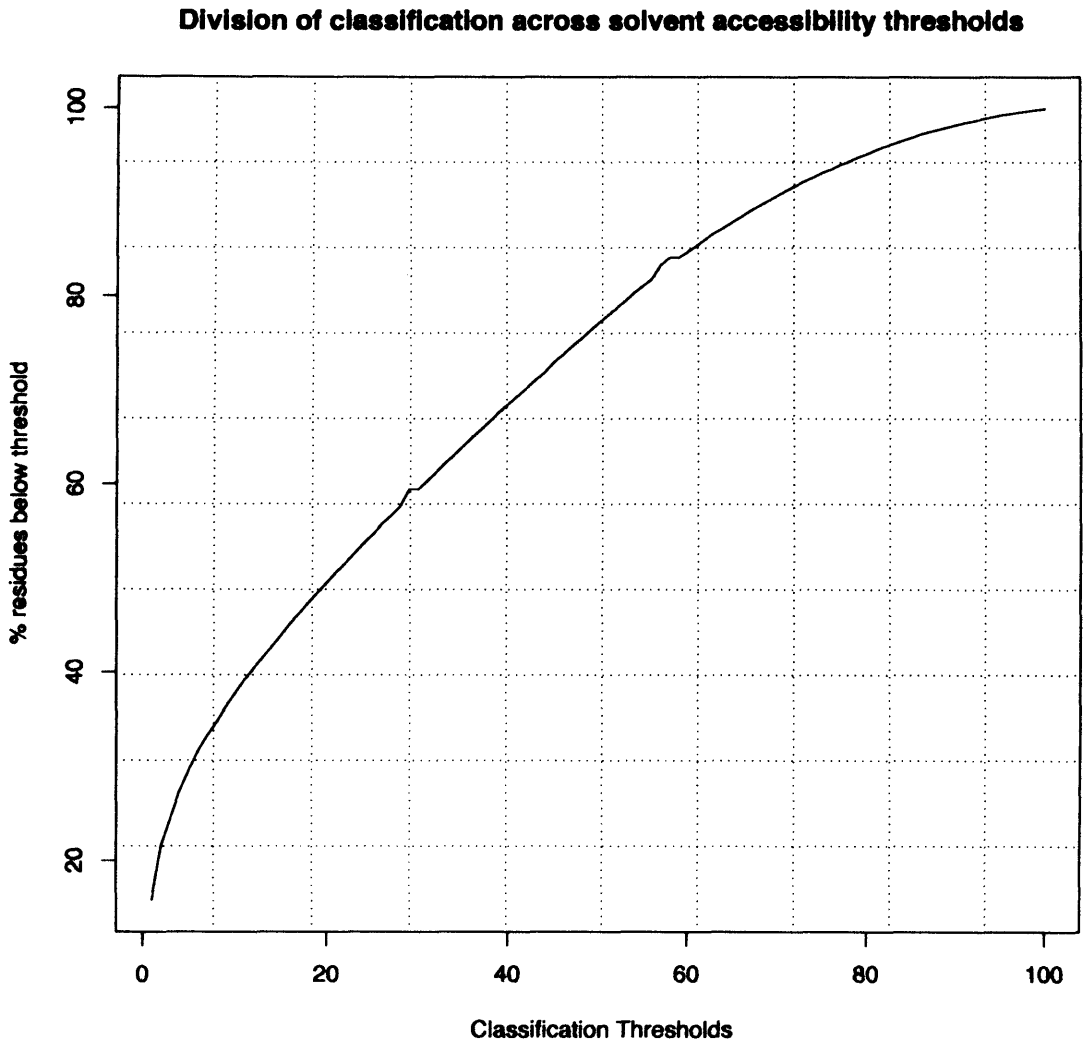


Figure 6.4 Division of Classification Across Solvent Accessibility Thresholds: For each dataset used to predict solvent accessibility the optimal threshold, the point where 50% of residues are buried and 50% are exposed, should be identified. For the 1024 proteins presented here the optimal threshold is approximately 20%. The optimal threshold presents the point where prediction is 'hardest' whilst straying either side makes the problem 'easier' because one class is over represented. In this work solvent accessibility was predicted in 5% increments from 5% to 95% (from totally buried to totally exposed).

For the solvent accessibility work a comparison was drawn between the f*k*NN, the SVM and the combination of each method. Additionally for the SVM step a further comparison was made between the basic vectors and the combination vectors (including the entropy measure). The SVM alone, using the transition vectors, achieves an accuracy of 76.88% at the 20% threshold which is comparable to the 77.8% achieved by the f*k*NN under the same circumstances (shown in figure 6.5 and table 6.4). By incorporating the entropy measure, slight performance increases are gained with the method achieving 78.16%. When combined the prediction accuracy using the transition vectors, the entropy measure and the f*k*NN predictions accuracy increases to 78.72% with precision and recall being 73.96% and 82.60% respectively.

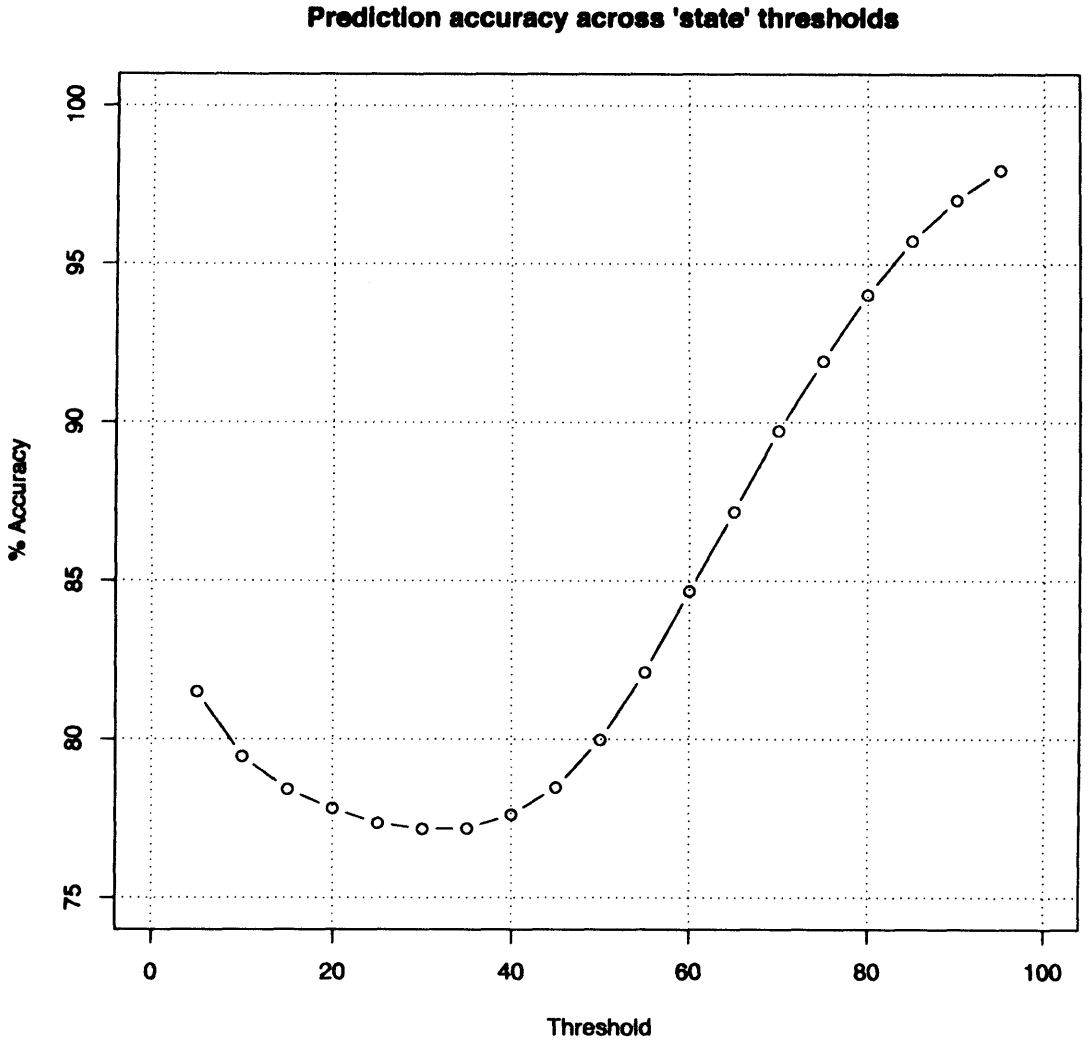


Figure 6.5 Prediction Accuracy for Solvent Accessibility Thresholds: The x-axis shows the solvent accessibility threshold at which residues were classified as buried (b) or exposed (e). The y-axis shows the overall accuracy, based on a leave one out cross validation on 1024 proteins, of the f*k*NN approach. The optimal threshold for prediction is approximately 20% as shown in figure 6.3.

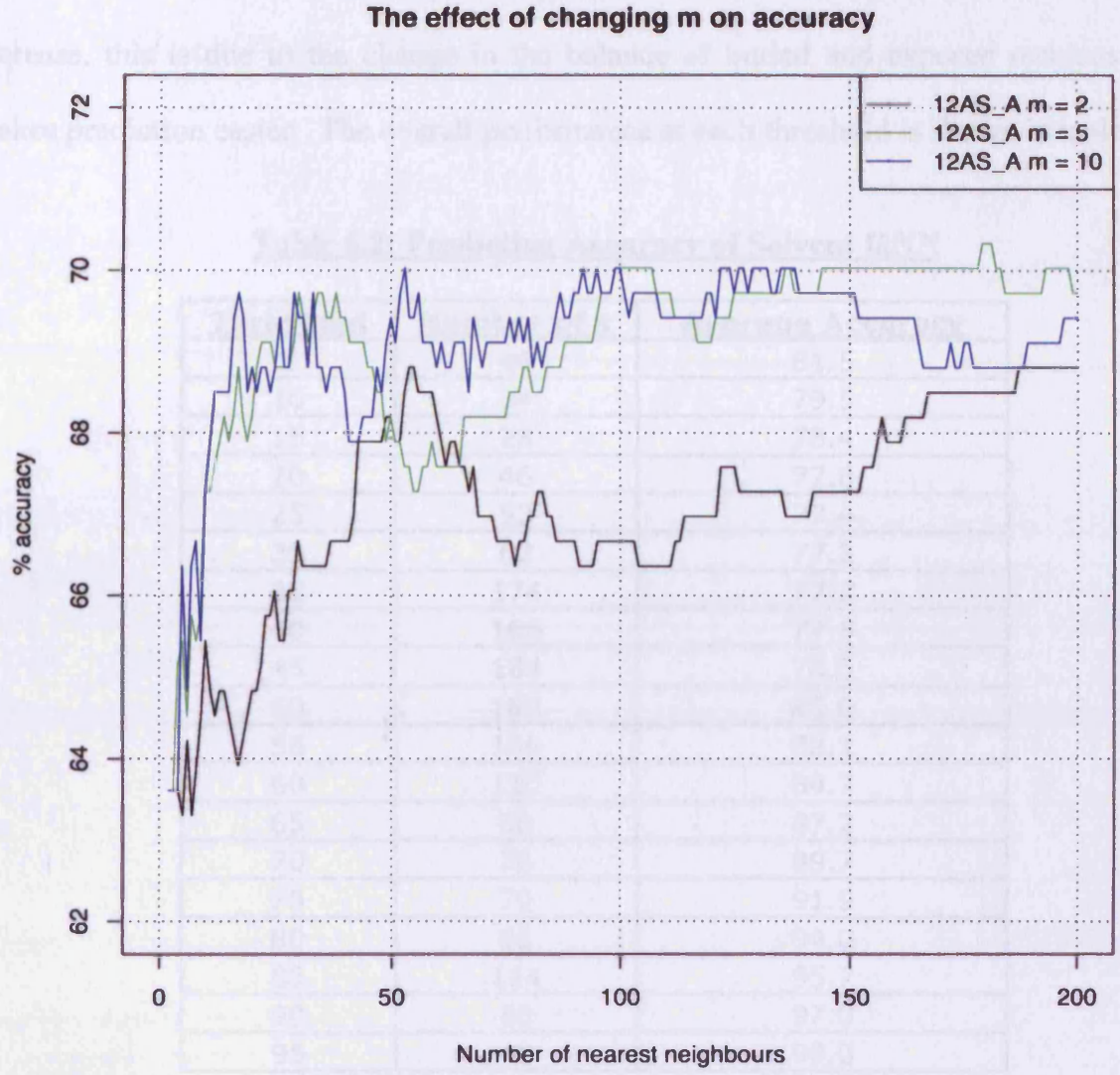


Figure 6.6 The Effect of the Fuzzy Parameter (*m*): The *m* parameter controls the effect distance has on the overall contribution of each *k* to the class membership of the k^{th} match. The greater the distance the k^{th} match is, the less the contribution to the overall class membership. The figure shows the effect of three values of m (2,5,10) on prediction accuracy for the asparagine synthetase (PDB code 12AS, chain A). A leave one out cross-validation showed minimal overall difference on the 1024 protein set with *m* being set to 2 in final runs.

For the remaining thresholds the performance accuracy at the f*k*NN stage remain similar or increase, this is due to the change in the balance of buried and exposed residues which makes prediction easier. The overall performance at each threshold is shown in table 6.2.

Table 6.2: Prediction Accuracy of Solvent f*k*NN

Threshold	Number of *k*	Average Accuracy
5	46	81.5
10	34	79.5
15	28	78.4
20	46	77.8
25	53	77.4
30	62	77.2
35	174	77.2
40	166	77.6
45	184	78.5
50	182	80.0
55	106	82.1
60	110	84.7
65	68	87.2
70	78	89.7
75	70	91.9
80	63	94.0
85	114	95.7
90	80	97.0
95	48	98.0

When combined with the entropy measures and the initial f*k*NN predictions, the accuracy increases but only marginally. The prediction accuracies are shown in table 6.3 below.

Table 6.3: Prediction Accuracy of Solvent Combination f*k*NN-SVM

Threshold	Accuracy	Precision	Recall
5	81.53	75.43	61.75
10	80.38	78.37	69.09
15	79.67	81.05	72.06
20	79.43	83.17	74.54
25	79.18	84.41	76.76
30	79.16	85.14	79.23
35	79.58	85.24	82.85
40	79.73	84.59	86.4
45	80.79	84.33	90.68
50	82.2	84.49	94.35
55	83.7	84.28	98.28
60	84.84	84.86	99.91
65	87.69	87.69	100
70	90.31	90.31	100
75	92.6	92.6	100
80	94.53	94.53	100
85	96.18	96.18	100
90	97.42	97.42	100
95	98.24	98.24	100

Table 6.3 shows that the precision and recall values increase as residues become more buried, the threshold increases, but decrease as residues become exposed. This is most likely a result of non-optimal SVM parameters which could be solved using a fine-grain optimisation. The problem with a fine-grain approach is the amount time required to complete all cross validations because of the large data set and limited compute resources.

Discussion and Conclusion

The application of a fuzzy *k* nearest neighbour approach combines a simple classification function with the output of a PSI-BLAST search. By today's standards the f*kNN* is one of the simplest techniques for assigning class to an unknown element. Despite its simplicity it performs remarkably well at predicting solvent accessibility (SA) and secondary structure (SS) and requires only three things: a measure of distance; the definition of *k;* and the identification of an optimal weight (*m*) parameter. In this work the weight parameter has little noticeable effect on the overall accuracy of the predictor and as such will not be discussed further.

The choice of distance measure is important for the f*k*NN as thousands of calculations have to be made per site. Earlier speculation was that the distance measure could have a negative impact on execution time and overall performance. In practise these concerns appear unfounded, by changing the original city-block measure to the 'slower' Euclidean function did not appear to result in a prohibitive increase in execution time or in overall performance. Changes in execution time may have gone unnoticed as the programs were run on a cluster rather than a standard desktop computer.

Initially the aim of the work was to predict RSA using a weighted mean applied to the *k* nearest neighbours, however this was not achievable as a good solution to the weight problems could not be found, resulting in consistent prediction of residues being totally buried (rsa = 0%) or totally exposed (rsa = 100%). A reasonable, although not ideal

solution to this, was to apply the same approach as described in chapter 4, where a binary classification was completed at a number of predefined thresholds. The benefit of this approach is that the technique was also suited to secondary structure prediction as it is a simple prediction of state.

For both features the f*k*NN achieves reasonable results, 76% for secondary structure and 77% for solvent accessibility. The overall results are good, but not better than published methods (although this comparison is almost entirely fruitless as described below). While this work was being completed another group, that of Julian Sim, described a similar method for prediction of solvent accessibility (Sim et al., 2005). Their method has two variations: the application of a weighted contribution from each position in the current window and the use of a larger dataset of 3644 proteins. The method described by Sim et al., achieves an accuracy of 78.5% at the 25% threshold, however the group does not report the point at which the residues are optimally separated and as such, the accuracy at 25% should be treated with some suspicion. A test was conducted on the contribution of the window weighting scheme using the method described above. The inclusion of a weighted contribution from each residue, dependent upon its distance from the central residue, did not yield an increase in the overall accuracy.

The second part of this method was the application of the support vector machine using the original f*k*NN input combined with its output as the SVM input space. For the two state prediction of solvent accessibility the svm^{light} (Joachims, 1999) and libsvm toolboxes were used. The secondary structure problem could have been addressed, as described above

using *b*SVM, or by using a one class against all approach. The *b*SVM method was used because it was designed for multi-class problems and does not require three models (helix, sheet and coil) to be trained and additional evaluation code to be written.

The results of using the SVM step were good, increasing the accuracy of the overall results. Although this method is not ground breaking, in terms of accuracy, it does provide another source of secondary structure predictions, something that was highlighted as necessary in chapter 5. The introduction of a third method not only increases the variation but allows for the creation of a consensus prediction, which may at a later date prove useful.

This method also lends itself to the *de novo* prediction pipeline because of its transparency and ease of retraining. The methods used in the pipelines (chapters 3 & 5) are black boxes and cannot be easily retrained, as such performing guaranteed *de novo* predictions is difficult in as much as it is unclear which proteins were used in training the systems and how similar they are to the target proteins. The process of identifying similar sequences in this method comprises of a scan against two sequence databases, the removal of sequences from the *fk*NN library and the possible retraining of the SVM.

The solvent accessibility method, despite not fulfilling the original aim, could have been used to replace AccPro in the Phobic function (chapter 4) were it not for one minor problem. This problem manifests in the inconsistent prediction of burial and exposure across thresholds, a problem that does not affect AccPro. The assignment of RSA based on the maximum state at which the residue is exposed is troublesome when a residue is

exposed at 30%, buried at 35 and 40%, but exposed again at 45%. While it may be possible to devise a solution to this problem, the overall accuracy of this method compared to AccPro did not warrant the time and effort that would be required.

As with all methods there are limitations, this is especially true for those which rely on multiple sequence alignments (MSAs). Problems arise with the generation and use of poor alignments to the absence of homologous sequences and structures. While it is widely accepted that MSAs do improve overall SS prediction accuracy the same is not true for solvent accessibility. Previous work has shown that MSAs do not aid the prediction of solvent accessibility (Przybylski and Rost, 2002) because it is not a feature well conserved across familial alignments, yet other groups (Adamczak et al., 2004) have shown that MSAs can improve prediction accuracy by up to 5%. Chapter 2 showed, using contact number (CN), that sequence alignments did not appear to play a large part in prediction accuracy despite the identification of CN being well conserved (Hamelryck, 2005). Despite this problem, the method made use of sequence profiles generated using PSI-BLAST. The fact that this method will not work without an MSA means that it cannot contribute to this debate.

One concern that should not be neglected, but almost invariably is, is that of the database which is used to train a method. The method presented here and that of Sim (Sim et al., 2005) provides a good basis on which to comment. Both methods can be made identical with the exception of the dataset – this work uses a set of 1024 proteins while that of Sim *et al.,* uses 3644. When comparing the accuracy the difference is just over 1% (77.2% &

78.5% respectively), while if repeated using a set of 764 proteins the accuracy decreases to 75%. With the variation in datasets the comparison of methods which use alternative training sets seems meaningless, with the improvements, often within the 1-2%, most probably coming from larger datasets than from the methods.

In summary this chapter has presented a novel combination of the f*k*NN algorithm and support vector classification to predict solvent accessibility and secondary structure to an accuracy similar to that of state of the art methods. While it does not break the 80% accuracy threshold, it provides a much needed supplementary method for use in the DDT and *de novo* prediction pipelines described in chapters 3 and 5.

Chapter 7

Discussions, Conclusions and the Future

Summary

In this short chapter I will summarise all of the work described in the previous chapters. I would also like to draw attention to several issues I think need to be addressed to improve the development process of prediction tools such as those described in chapter 6 and I will devote some time to describing improvements to the prediction pipelines (chapters 3 & 5) which could improve overall success.

For the prediction of tertiary structure there are experiments, such as CASP, EVA (Eyrich et al., 2001) & LiveBench (Bujnicki et al., 2001) which assess the state of the art – referred to as benchmarking. The aims of each are to provide continual assessment that highlights weak, strong and stagnant areas – in CASP6 secondary structure prediction was closed. While successful at evaluation, none of the benchmarking utilities provide a resource which acts in a regulatory manner, providing a robust dataset which is large enough to test and train new tools. It maybe that this is not possible, or at best very challenging, for 3D structure but it would be possible for the prediction of 2D features. Considering that much of the 'ground work' for prediction of tertiary structure is based on the prediction of 2D features this could result in improvements at the final 3D stage. As shown in chapter 6 a small change in the number of sequences in a training database can result in a drop in accuracy – a decrease in one correlates with a decrease in the other. A common method of analysing predictive tools is to use a set of proteins, such as the Rost and Sander (Rost and Sander, 1993) or Manesh (Naderi-Manesh et al., 2001) datasets to determine if method *A* is better than method *B* – where *A* and *B* could be any single or combination of methods. This

approach is logical if *A* and *B* are trained on exactly the same set of data, be it sequence or structure based, and allows for a direct comparison. Additionally the commonality of the data means that a simple statistical test, such as the t-test, can be used to determine if the improvement of *A* over *B* is statistically significant. The lack of statistical validation may seem odd given the mathematical heritage of the field, however it can be explained by the invalidity of comparing methods in which not only the datasets vary but the techniques themselves. In addition to this, another 'test' would be to take both tools (*A&B*) and see if the application of either improved overall structure prediction, again this is something which is rarely performed but tells us if the new tools really provides anything that existing tools do not. The simplest way to overcome this issue would be to provide a resource to which groups could upload their training and test sets as well as download other groups data, allowing a comparison of all methods and for the effects of features such as window length to be conclusively evaluated. Each group would then be able to choose the best methods for the prediction of the required structure from solvent accessibility to contact order, this approach would be particularly useful to the α/β method introduced in chapter 5 where secondary structure prediction proved to be problematic. An additional benefit would be the trivial nature of establishing whether a protein used to evaluate a prediction pipeline was used in the training of, for example, a secondary structure prediction tool and hence avoid any potential bias in the final outcome. Such a project would require large computer resources to store and distribute data as well as the cooperation of the prediction community and would be a bioinformatics project in itself.

In addition to the aforementioned improvements it may also be beneficial, as far as the methods described in chapters 3 & 5 are concerned, to alter the model construction process such that the final constructs include side chains. Both prediction pipelines produce Cα models which are less complex to construct in part because the potential for constraint violation is less than an all atom construct and this allows for the generation of a large ensemble of structures which has been identified as beneficial for these approaches. Generation of the same number of models using an all atom approach would be prohibitive because of the extra evaluation steps required at each phase. The modular approach described in chapter 5 could easily accommodate full atom models either post-threading or post top 100 + *n* range (see figure 5.1). At these points the ensembles of models would still be diverse but many of the less 'fit' structures would have been removed by the low level functions. The inclusion of the side chains would lead to steric clashes which would either result in models being excluded or backbone remodelling. It would then be possible to use a finer grain version of Phobic as well as other new or existing scoring functions. These models could then be refined using more sophisticated, minimalisation-like methods.

The previous chapters have described the design, development and application of tools for the construction and evaluation of protein models. Each chapter is intrinsically related to the others describing two evaluation functions, two 3D structure prediction pipelines and a 2D structure prediction architecture. The methods provide state of the art performance for each structural feature and as well as offering solutions to problems which have plagued the field, such as domain definition in threading and *de novo* prediction of large proteins.

References

ADAMCZAK, R., POROLLO, A. & MELLER, J. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins,* 56, 753-67.

AHMAD, S., GROMIHA, M., FAWAREH, H. & SARAI, A. (2004a) ASAView: database and tool for solvent accessibility representation in proteins. *BMC Bioinformatics,* 5, 51.

AHMAD, S. & GROMIHA, M. M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics,* 18, 819-24.

AHMAD, S., GROMIHA, M. M. & SARAI, A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins,* 50, 629-35.

AHMAD, S., GROMIHA, M. M. & SARAI, A. (2004b) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics,* 20, 477-86.

AIZERMAN, A., BRAVERMAN, E. M. & ROZONER, L. I. (1964) Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control,* 25, 821-837.

ALDEN, C. J. & KIM, S. H. (1979) Solvent-accessible surfaces of nucleic acids. *J Mol Biol,* 132, 411-34.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215, 403-10.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25, 3389-402.

ANDREEVA, A., HOWORTH, D., BRENNER, S. E., HUBBARD, T. J., CHOTHIA, C. & MURZIN, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research,* 32, D226-9.

ANFINSEN, C. B. (1972) The formation and stabilization of protein structure. *Biochem J,* 128, 737-49.

ANFINSEN, C. B. (1973) Principles that govern the folding of protein chains. *Science,* 181, 223-30.

APWEILER, R. (2008) The universal protein resource (UniProt). *Nucleic Acids Res,* 36, D190-5.

APWEILER, R., BAIROCH, A., WU, C. H., BARKER, W. C., BOECKMANN, B., FERRO, S., GASTEIGER, E., HUANG, H., LOPEZ, R., MAGRANE, M., MARTIN, M. J., NATALE, D. A., O'DONOVAN, C., REDASCHI, N. & YEH, L. S. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res,* 32, D115-9.

ARONSZAJN, N. (1950) Theory of reproducing kernels. *Transactions of the American Mathematical Society,* 3, 337-404.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. &

SHERLOCK, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25, 25-9.

BARKER, W. C., GARAVELLI, J. S., MCGARVEY, P. B., MARZEC, C. R., ORCUTT, B. C., SRINIVASARAO, G. Y., YEH, L. S., LEDLEY, R. S., MEWES, H. W., PFEIFFER, F., TSUGITA, A. & WU, C. (1999) The PIR-International Protein Sequence Database. *Nucl. Acids Res.,* 27, 39-43.

BATEMAN, A., BIRNEY, E., DURBIN, R., EDDY, S. R., HOWE, K. L. & SONNHAMMER, E. L. (2000) The Pfam protein families database. *Nucleic Acids Res,* 28, 263-6.

BENNET, K. P. & MANGASARIN, O. L. (1992) Robust linear programming discrimination of two linearly inseparable sets. *COLT 1992.* New York, ACM Press.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & WHEELER, D. L. (2005) GenBank. *Nucleic Acids Res,* 33, D34-8.

BERGLUND, A., HEAD, R. D., WELSH, E. A. & MARSHALL, G. R. (2004) ProVal: a protein-scoring function for the selection of native and near-native folds. *Proteins,* 54, 289-302.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28, 235-42.

BERNHARD, E. B., ISABELLE, M. G. & VLADIMIR, N. V. (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory.* Pittsburgh, Pennsylvania, United States, ACM.

BOECKMANN, B., BAIROCH, A., APWEILER, R., BLATTER, M.-C., ESTREICHER, A., GASTEIGER, E., MARTIN, M. J., MICHOUD, K., O'DONOVAN, C., PHAN, I., PILBOUT, S. & SCHNEIDER, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.,* 31, 365-370.

BONDUGULA, R. & XU, D. (2007) MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins,* 66, 664-70.

BOSER, B. E., GUYON, I. & VAPNIK, V. (1992) A Training Algorithm for Optimal Margin Classifiers. *Computational Learning Theory.*

BOUTET, E., LIEBERHERR, D., TOGNOLLI, M., SCHNEIDER, M. & BAIROCH, A. (2007) UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. *Methods Mol Biol,* 406, 89-112.

BOWIE, J. U., LUTHY, R. & EISENBERG, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* 253, 164-70.

BRADLEY, P., MISURA, K. M. & BAKER, D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science,* 309, 1868-71.

BRENNER, S. E., KOEHL, P. & LEVITT, M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res,* 28, 254-6.

BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S. & KARPLUS , M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem,* 4, 187-217.

BUJNICKI, J. M., ELOFSSON, A., FISCHER, D. & RYCHLEWSKI, L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci,* 10, 352-61.

CANUTESCU, A. A., SHELENKOV, A. A. & DUNBRACK, R. L., JR. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci,* 12, 2001-14.

CAVALLO, L., KLEINJUNG, J. & FRATERNALI, F. (2003) POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res,* 31, 3364-6.

CHIH-CHUNG, C. & CHIH-JEN, L. (2001) LIBSVM : a library for support vector machines. IN CHIH-CHUNG, C. & CHIH-JEN, L. (Eds.) *LIBSVM : a library for support vector machines.*

CHOTHIA, C. & FINKELSTEIN, A. V. (1990) The classification and origins of protein folding patterns. *Annu Rev Biochem,* 59, 1007-39.

CHOU, P. Y. & FASMAN, G. D. (1974) Prediction of protein conformation. *Biochemistry,* 13, 222-45.

CHOU, P. Y. & FASMAN, G. D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology & Related Areas of Molecular Biology,* 47, 45-148.

COHEN, F. E., RICHMOND, T. J. & RICHARDS, F. M. (1979) Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J Mol Biol,* 132, 275-88.

COHEN, F. E., STERNBERG, M. J. & TAYLOR, W. R. (1980) Analysis and prediction of protein beta-sheet structures by a combinatorial approach. *Nature,* 285, 378-82.

CRISTIANINI, N. & SHAWE-TAYLOR, J. (2000) *An introduction to Support Vector Machines and other kernel based learning methods,* Cambridge, Cambridge University Press.

CUFF, J. A. & BARTON, G. J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins,* 40, 502-11.

DICKERSON, R. E., TIMKOVICH, R. & ALMASSY, R. J. (1976) The cytochrome fold and the evolution of bacterial energy metabolism. *J Mol Biol,* 100, 473-91.

DILL, K. A. (1990) Dominant forces in protein folding. *Biochemistry,* 29, 7133-55.

DOOLITTLE, R. F. (1986) *of URFs and ORFs: A primer on how to analyze derived amino acid sequences*, University Science Books, Mill Valley, CA, USA.

DUAN, Y. & KOLLMAN, P. A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science,* 282, 740-4.

EDGAR, R. C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5, 113.

EYRICH, V. A., MARTI-RENOM, M. A., PRZYBYLSKI, D., MADHUSUDHAN, M. S., FISER, A., PAZOS, F., VALENCIA, A., SALI, A. & ROST, B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics,* 17, 1242-3.

FEIG, M. & BROOKS, C. L., 3RD (2002) Evaluating CASP4 predictions with physical energy functions. *Proteins,* 49, 232-45.

FRATERNALI, F. & CAVALLO, L. (2002) Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome. *Nucleic Acids Res,* 30, 2950-60.

GALPERIN, M. Y. (2007) The Molecular Biology Database Collection: 2007 update. *Nucl. Acids Res.,* 35, D3-4.

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. *Journal of Molecular Biology,* 120, 97-120.

GEORGE, D. G., BARKER, W. C. & HUNT, L. T. (1986) The protein identification resource (PIR). *Nucleic Acids Res,* 14, 11-5.

GIBBS, A. J. & MCINTYRE, G. A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem,* 16, 1-11.

GINALSKI, K., ELOFSSON, A., FISCHER, D. & RYCHLEWSKI, L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics,* 19, 1015-8.

GINALSKI, K. & RYCHLEWSKI, L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res,* 31, 3291-2.

GRIBSKOV, M., MCLACHLAN, A. D. & EISENBERG, D. (1987) Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A,* 84, 4355-8.

GUEX, N. & PEITSCH, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modelling. *Electrophoresis,* 18, 2714-23.

HAMELRYCK, T. (2005) An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins,* 59, 38-48.

HASEL, W., HENDRIKSON, T AND STILL, W (1988) A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput. Methodol.,* 1, 103-116.

HENRICK, K. (2006) CAPRI - Critical Assessment of PRediction of Interactions. EMBL-EBI.

HIGGINS, D. G. & SHARP, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene,* 73, 237-44.

HIGGINS, D. G. & TAYLOR, W. R. (2000) Multiple sequence alignment. *Methods Mol Biol,* 143, 1-18.

HOBOHM, U., SCHARF, M., SCHNEIDER, R. & SANDER, C. (1992) Selection of representative protein data sets. *Protein Sci,* 1, 409-17.

HOLLAND, J. H. (1975) *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence,* Ann Arbor, Mich., University of Michigan Press.

HOLM, L., OUZOUNIS, C., SANDER, C., TUPAREV, G. & VRIEND, G. (1992) A database of protein structure families with common folding motifs. *Protein Sci,* 1, 1691-8.

HOLM, L. & SANDER, C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol,* 233, 123-38.

HOLM, L. & SANDER, C. (1994) The Fssp Database of Structurally Aligned Protein Fold Families. *Nucleic Acids Research,* 22, 3600-3609.

HOLM, L. & SANDER, C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem Sci,* 20, 478-80.

HU, X. Z. & LI, Q. Z. (2007) Prediction of the beta-Hairpins in Proteins Using Support Vector Machine. *Protein J.*

HUBBARD, S. J. A. T., J.M. (1993) 'NACCESS', computer program. Department of Biochemistry and Molecular Biology, University College London; 1993. .

HUBBARD, T. J., AILEY, B., BRENNER, S. E., MURZIN, A. G. & CHOTHIA, C. (1998) SCOP, Structural Classification of Proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallographica Section D-Biological Crystallography,* 54, 1147-54.

HUBBARD, T. J., AILEY, B., BRENNER, S. E., MURZIN, A. G. & CHOTHIA, C. (1999) SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research,* 27, 254-6.

JOACHIMS, T. (1999) *Making large-Scale SVM Learning Practical.,* Cambridge, Mass, MIT Press.

JONASSEN, I., EIDHAMMER, I., GRINDHAUG, S. H. & TAYLOR, W. R. (2000) Searching the protein structure databank with weak sequence patterns and structural constraints. *J Mol Biol,* 304, 599-619.

JONASSEN, I., KLOSE, D. & TAYLOR, W. R. (2006) Protein model refinement using structural fragment tessellation. *Comput Biol Chem,* 30, 360-6.

JONES, D. T. (1997) 'The Pfilt filter', computer program, Department of Biochemistry and Molecular Biology, University College London.

JONES, D. T. (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol,* 287, 797-815.

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

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992a) A new approach to protein fold recognition. *Nature,* 358, 86-9.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992b) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci,* 8, 275-82.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* 33, 3038-49.

KABSCH, W. & SANDER, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* 22, 2577-637.

KAJAN, L. & RYCHLEWSKI, L. (2007) Evaluation of 3D-Jury on CASP7 models. *BMC Bioinformatics,* 8, 304.

KARPLUS, K., BARRETT, C., CLINE, M., DIEKHANS, M., GRATE, L. & HUGHEY, R. (1999) Predicting protein structure using only sequence information. *Proteins,* Suppl 3, 121-5.

KARPLUS, K., BARRETT, C. & HUGHEY, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* 14, 846-56.

KARPLUS, K., KARCHIN, R., DRAPER, J., CASPER, J., MANDEL-GUTFREUND, Y., DIEKHANS, M. & HUGHEY, R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins,* 53 Suppl 6, 491-6.

KINJO, A. R., HORIMOTO, K. & NISHIKAWA, K. (2005) Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins,* 58, 158-65.

KLOSE, D. & TAYLOR, W. (2007) Protein Structure. IN BALDING, D., CANNINGS, C. & BISHOP, M. (Eds.) *Handbook of Statistical Genetics.* 3 ed., Wiley Press.

LAZARIDIS, T. & KARPLUS, M. (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol,* 288, 477-87.

LAZARIDIS, T. & KARPLUS, M. (2000) Effective energy functions for protein structure prediction. *Curr Opin Struct Biol,* 10, 139-45.

LEE, B. & RICHARDS, F. M. (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol,* 55, 379-400.

LEINONEN, R., DIEZ, F. G., BINNS, D., FLEISCHMANN, W., LOPEZ, R. & APWEILER, R. (2004) UniProt archive. *Bioinformatics,* 20, 3236-7.

LEINONEN, R., NARDONE, F., ZHU, W. & APWEILER, R. (2006) UniSave: the UniProtKB sequence/annotation version database. *Bioinformatics,* 22, 1284-5.

LEVITT, M. & CHOTHIA, C. (1976) Structural patterns in globular proteins. *Nature,* 261, 552-8.

LIN, K., KLEINJUNG, J., TAYLOR, W. R. & HERINGA, J. (2003) Testing homology with Contact Accepted mutatiOn (CAO): a contact-based Markov model of protein evolution. *Comput Biol Chem,* 27, 93-102.

LIN, K., MAY, A. C. & TAYLOR, W. R. (2002) Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility. *Bioinformatics,* 18, 1350-7.

LIN, K., SIMOSSIS, V. A., TAYLOR, W. R. & HERINGA, J. (2005) A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics,* 21, 152-9.

LO CONTE, L., AILEY, B., HUBBARD, T. J., BRENNER, S. E., MURZIN, A. G. & CHOTHIA, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Research,* 28, 257-9.

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 Research,* 30, 264-7.

LUTHY, R., MCLACHLAN, A. D. & EISENBERG, D. (1991) Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins,* 10, 229-39.

MANAVALAN, P. & PONNUSWAMY, P. K. (1978) Hydrophobic character of amino acid residues in globular proteins. *Nature,* 275, 673-4.

MARCHLER-BAUER, A., PANCHENKO, A. R., SHOEMAKER, B. A., THIESSEN, P. A., GEER, L. Y. & BRYANT, S. H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research,* 30, 281-3.

MARSDEN, B. & ABAGYAN, R. (2004) SAD--a normalized structural alignment database: improving sequence-structure alignments. *Bioinformatics,* 20, 2333-44.

MARTI-RENOM, M. A., STUART, A. C., FISER, A., SANCHEZ, R., MELO, F. & SALI, A. (2000) Comparative protein structure modelling of genes and genomes. *Annu Rev Biophys Biomol Struct,* 29, 291-325.

MARTIN, A. C. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics,* 21, 4297-301.

MAXAM, A. M. & GILBERT, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A,* 74, 560-4.

MCGUFFIN, L. J., BRYSON, K. & JONES, D. T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics,* 16, 404-5.

MINKSY, M. & EDMONDS, D. (1954) Stochastic Neural Analogue Reinforcement Calculator. Princeton.

MINKSY, M. & PAPERT, S. (1969) *Perceptrons,* Cambridge, MIT Press.

MIYAZAWA, S. & JERNIGAN, R. L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* 18, 534-552.

MIZUGUCHI, K., DEANE, C. M., BLUNDELL, T. L., JOHNSON, M. S. & OVERINGTON, J. P. (1998a) JOY: protein sequence-structure representation and analysis. *Bioinformatics,* 14, 617-23.

MIZUGUCHI, K., DEANE, C. M., BLUNDELL, T. L. & OVERINGTON, J. P. (1998b) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Science,* 7, 2469-71.

MOULT, J., PEDERSEN, J. T., JUDSON, R. & FIDELIS, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins,* 23, ii-v.

MURZIN, A. G., BRENNER, S. E., HUBBARD, T. & CHOTHIA, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* 247, 536-40.

NADERI-MANESH, H., SADEGHI, M., ARAB, S. & MOOSAVI MOVAHEDI, A. A. (2001) Prediction of protein surface accessibility with information theory. *Proteins,* 42, 452-9.

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

NISHIKAWA, K. & OOI, T. (1980) Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int J Pept Protein Res,* 16, 19-32.

NISHIKAWA, K. & OOI, T. (1986) Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J Biochem,* 100, 1043-7.

NOGUCHI, T., ONIZUKA, K., AKIYAMA, Y. & SAITO, M. (1997) PDB-REPRDB: a database of representative protein chains in PDB (Protein Data Bank). *Proc Int Conf Intell Syst Mol Biol,* 5, 214-7.

NOVOTNY, J., BRUCCOLERI, R. & KARPLUS, M. (1984) An analysis of incorrectly folded protein models. Implications for structure predictions. *J Mol Biol,* 177, 787-818.

NOZAKI, Y. & TANFORD, C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J Biol Chem,* 246, 2211-7.

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, 149-70.

ORENGO, C. A., FLORES, T. P., JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1993a) Recurring structural motifs in proteins with different functions. *Curr Biol,* 3, 131-9.

ORENGO, C. A., FLORES, T. P., TAYLOR, W. R. & THORNTON, J. M. (1993b) Identification and classification of protein fold families. *Protein Eng,* 6, 485-500.

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, 1093-108.

OVERINGTON, J. P., ZHU, Z. Y., SALI, A., JOHNSON, M. S., SOWDHAMINI, R., LOUIE, G. V. & BLUNDELL, T. L. (1993) Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins. *Biochem Soc Trans,* 21 (Pt 3), 597-604.

PARK, B. & LEVITT, M. (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol,* 258, 367-92.

PEARL, L. H. & TAYLOR, W. R. (1987) A structural model for the retroviral proteases. *Nature,* 329, 351-4.

PETERSEN, K. & TAYLOR, W. R. (2003) Modelling zinc-binding proteins with GADGET: genetic algorithm and distance geometry for exploring topology. *J Mol Biol,* 325, 1039-59.

POGGIO, T. & GIROSI, F. (1990) Networks for approximation and learning. *Proceedings of the IEEE.*

POLLASTRI, G., BALDI, P., FARISELLI, P. & CASADIO, R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins,* 47, 142-53.

PRF (2007) The PRF databases. The Protein Research Foundation, Osaka, Japan http://www.prf.or.jp/en/index.shtml.

PRZYBYLSKI, D. & ROST, B. (2002) Alignments grow, secondary structure prediction improves. *Proteins,* 46, 197-205.

QIAN, N. & SEJNOWSKI, T. J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol,* 202, 865-84.

QIU, J., SHEFFLER, W., BAKER, D. & NOBLE, W. S. (2007) Ranking predicted protein structures with support vector regression. *Proteins.*

RICE, D. W. & EISENBERG, D. (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol,* 267, 1026-38.

RICHARDSON, J. S. (1976) Handedness of crossover connections in beta sheets. *Proc Natl Acad Sci U S A,* 73, 2619-23.

ROHL, C. A. & BAKER, D. (2002) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc,* 124, 2723-9.

ROSE, G. D. & ROY, S. (1980) Hydrophobic basis of packing in globular proteins. *Proc Natl Acad Sci U S A,* 77, 4643-7.

ROSENBLATT, F. (1988) The perception: a probabilistic model for information storage and organization in the brain. *Neurocomputing: foundations of research.* MIT Press.

ROST, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol,* 266, 525-39.

ROST, B. & SANDER, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol,* 232, 584-99.

ROST, B. & SANDER, C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins,* 20, 216-26.

ROST, B. & SANDER, C. (2000) Third generation prediction of secondary structures. *Methods Mol Biol,* 143, 71-95.

RUMELHART, D. E., HINTON, G. E. & WILLIAMS, R. J. (1986) Learning internal representations by error propagation. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations.* MIT Press.

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

SANGER, F. & COULSON, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol,* 94, 441-8.

SANGER, F., NICKLEN, S. & COULSON, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74, 5463-7.

SANGER, F. & TUPPY, H. (1951a) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J,* 49, 481-90.

SANGER, F. & TUPPY, H. (1951b) The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J,* 49, 463-81.

SCHOLKOPF, B. & SMOLA, A. (2002) *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press.

SCHULTZ, J., MILPETZ, F., BORK, P. & PONTING, C. P. (1998) SMART, a simple modular architecture research tool: Identification of signalling domains. *PNAS,* 95, 5857-5864.

SCHWEDE, T., KOPP, J., GUEX, N. & PEITSCH, M. C. (2003) SWISS-MODEL: An automated protein homology-modelling server. *Nucleic Acids Res,* 31, 3381-5.

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

SHRAKE, A. & RUPLEY, J. A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol,* 79, 351-71.

SIM, J., KIM, S. Y. & LEE, J. (2005) Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics.*

SIMONS, K. T., BONNEAU, R., RUCZINSKI, I. & BAKER, D. (1999a) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins,* Suppl 3, 171-6.

SIMONS, K. T., RUCZINSKI, I., KOOPERBERG, C., FOX, B. A., BYSTROFF, C. & BAKER, D. (1999b) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins,* 34, 82-95.

SIPPL, M. J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol,* 213, 859-83.

SMITH, F. W. (1968) Pattern classifier design by linear programming. *IEEE Transactions on Computers,* 4, 367-372.

SMITH, T. F. & WATERMAN, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol,* 147, 195-7.

SMOLA, A. (1997) 'The pr_LOQO algorithm', The Statistical Machine Learning Program, National ICT, 216 Northbourne Avenue, Canberra, Australia.

SRINIVASAN, J., MILLER, J., KOLLMAN, P. A. & CASE, D. A. (1998) Continuum solvent studies of the stability of RNA hairpin loops and helices. *J Biomol Struct Dyn,* 16, 671-82.

STADEN, R. (1988) Methods to define and locate patterns of motifs in sequences. *Comput Appl Biosci,* 4, 53-60.

STEBBINGS, L. A. & MIZUGUCHI, K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Research,* 32, D203-7.

STERNBERG, M. J. & THORNTON, J. M. (1976) On the conformation of proteins: the handedness of the beta-strand-alpha-helix-beta-strand unit. *J Mol Biol,* 105, 367-82.

STERNBERG, M. J. & THORNTON, J. M. (1977) On the conformation of proteins: the handedness of the connection between parallel beta-strands. *J Mol Biol,* 110, 269-83.

STERNBERG, M. J. & THORNTON, J. M. (1978) Prediction of protein structure from amino acid sequence. *Nature,* 271, 15-20.

STILL, W., TEMPCZYK, A., HAWLEY, R AND HENDRIKSON, T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.,* 112, 6127-6129.

SUZEK, B. E., HUANG, H., MCGARVEY, P., MAZUMDER, R. & WU, C. H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* 23, 1282-8.

TANAKA, S. & SCHERAGA, H. A. (1975) Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc Natl Acad Sci U S A,* 72, 3802-6.

TAYLOR, W. R. (1986) The classification of amino acid conservation. *Journal of Theoretical Biology,* 119, 205-18.

TAYLOR, W. R. (1987) Multiple sequence alignment by a pairwise algorithm. *Comput Appl Biosci,* 3, 81-7.

TAYLOR, W. R. (1993) Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data. *Protein Eng,* 6, 593-604.

TAYLOR, W. R. (1997a) Multiple sequence threading: an analysis of alignment quality and stability. *Journal of Molecular Biology,* 269, 902-43.

TAYLOR, W. R. (1997b) Residual colours: a proposal for aminochromography. *Protein Eng,* 10, 743-6.

TAYLOR, W. R. (1999a) Protein structural domain identification. *Protein Eng.,* 12, 203-216.

TAYLOR, W. R. (1999b) Protein structure comparison using iterated double dynamic programming. *Protein Sci,* 8, 654-665.

TAYLOR, W. R. (2001) Defining linear segments in protein structure. *J Mol Biol,* 310, 1135-50.

TAYLOR, W. R. (2002) A 'periodic table' for protein structures. *Nature,* 416, 657-60.

TAYLOR, W. R. (2006) Decoy models for protein structure comparison score normalisation. *J Mol Biol,* 357, 676-99.

TAYLOR, W. R., BARTLETT, G. J., CHELLIAH, V., KLOSE, D., LIN, K., SHELDON, T. & JONASSEN, I. (2008) Prediction of protein structure from ideal forms. *Proteins,* 70, 1610-1619.

TAYLOR, W. R., FLORES, T. P. & ORENGO, C. A. (1994) Multiple protein structure alignment. *Protein Science,* 3, 1858-70.

TAYLOR, W. R., HERINGA, J., BAUD, F. & FLORES, T. P. (2002) A Fourier analysis of symmetry in protein structure. *Protein Eng,* 15, 79-89.

TAYLOR, W. R. & JONASSEN, I. (2004) A structural pattern-based method for protein fold recognition. *Proteins,* 56, 222-34.

TAYLOR, W. R., LIN, K., KLOSE, D., FRATERNALI, F. & JONASSEN, I. (2006) Dynamic domain threading. *Proteins,* 64, 601-14.

TAYLOR, W. R. & ORENGO, C. A. (1989a) A holistic approach to protein structure alignment. *Protein Eng,* 2, 505-19.

TAYLOR, W. R. & ORENGO, C. A. (1989b) Protein structure alignment. *J Mol Biol,* 208, 1-22.

TAYLOR, W. R., SAELENSMINDE, G. & EIDHAMMER, I. (2000) Multiple protein sequence alignment using double-dynamic programming. *Comput Chem,* 24, 3-12.

THOMAS, P. D. & DILL, K. A. (1996) Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol,* 257, 457-69.

THOMPSON, J. D., PLEWNIAK, F. & POCH, O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics,* 15, 87-8.

THOMPSON, M. J. & GOLDSTEIN, R. A. (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins,* 25, 38-47.

THOMPSON, M. J. & GOLDSTEIN, R. A. (1997) Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Sci,* 6, 1963-75.

TSAI, J., BONNEAU, R., MOROZOV, A. V., KUHLMAN, B., ROHL, C. A. & BAKER, D. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins,* 53, 76-87.

VAPNIK, V. (1998) *Statistical learning theory,* Wiley-Interscience.

VAPNIK, V. & LERNER, A. (1963) Pattern recognition using generalize portrait method. *Automation and Remote Control,* 774-780.

VON GROTTHUSS, M., PAS, J., WYRWICZ, L., GINALSKI, K. & RYCHLEWSKI, L. (2003) Application of 3D-Jury, GRDB, and Verify3D in fold recognition. *Proteins,* 53 Suppl 6, 418-23.

WANG, G. & DUNBRACK, R. L., JR. (2003) PISCES: a protein sequence culling server. *Bioinformatics,* 19, 1589-91.

WARD, J. J., MCGUFFIN, L. J., BUXTON, B. F. & JONES, D. T. (2003) Secondary structure prediction with support vector machines. *Bioinformatics,* 19, 1650-5.

WESSON, L. & EISENBERG, D. (1992) Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci,* 1, 227-35.

WIDROW, B. & HOFF, M. E. (1988) Adaptive switching circuits. *Neurocomputing: foundations of research.* MIT Press.

WILSON, D. L. (1972) Asymptotic Properties of nearest neighbor rules using edited data. *SMC,* 2, 408-421.

WODAK, S. A. J., J. (1980) Analytical approximation to the accessible surface area of proteins. *Proc Natl Acad Sci USA,* 77, 1736-1740.

YEE, D. P. & DILL, K. A. (1993) Families and the structural relatedness among globular proteins. *Protein Sci,* 2, 884-99.

YUAN, Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics,* 6, 248.

ZEMLA, A. (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res,* 31, 3370-4.

ZVELEBIL, M. J., BARTON, G. J., TAYLOR, W. R. & STERNBERG, M. J. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol,* 195, 957-61.