ABSTRACT

Title of Dissertation: A GRAPH-THEORETIC SOLUTION TO THE CONTEXT-SENSITIVITY PROBLEM IN PROTEIN STRUCTURE PREDICTION

V V Samudrala, Doctor of Philosophy, 1997

Dissertation directed by: Professor John Moult
Molecular and Cell Biology Program

Predicting the three-dimensional conformation adopted by a protein sequence is an unsolved problem in computational molecular biology. Currently, the best technique for the prediction of structure from sequence is comparative modelling: in all known cases, two evolutionarily-related proteins with similar sequences ($> 30\%$ sequence identity) have similar three-dimensional conformations. A sequence alignment can therefore be used to construct a model for a target sequence, using the coordinates of a related parent sequence for which a structure has already been determined by experimental methods.

The predictive power of comparative modelling was objectively assessed by making *bona fide* predictions for three targets at the first meeting for the Critical Assessment of protein Structure Prediction methods (CASP1) in December,

1994. The results from this meeting show that even though comparative modelling is the best method available to predict structures, the context-sensitivity of interactions in protein structures appears to be a major hurdle preventing the construction of accurate models.

We use an algorithm based on graph theory to handle this problem. Each possible residue conformation is represented as a node in a graph. Each node is weighted based on the strength of the interaction between the side chain and the local main chain. Edges are then drawn between nodes in a self-consistent manner, and are weighted based on the strength of the interaction between the two nodes. Once a graph representing all possible conformations and their interactions is constructed, the maximal sets of completely connected nodes (cliques) the size of the protein sequence are found using a clique finding algorithm. The clique with the best weight represents the optimal combination of the main chain and side chain possibilities that are input to the algorithm, and is assumed to represent a correct native-like conformation.

We have tested this novel method objectively by making predictions at the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2) in December, 1996, and find that significant improvements have been made in the building of side chains and main chain regions in comparative modelling.

A GRAPH-THEORETIC SOLUTION TO THE CONTEXT-SENSITIVITY PROBLEM IN PROTEIN STRUCTURE PREDICTION

by

V V Samudrala

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland at College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1997

Advisory Committee:

Professor John Moult, Chairman/Advisor
Professor Osnat Herzberg
Professor Doug Julin
Professor Samir Khuller
Professor Sarah Woodson

DEDICATION

To the twenty amino acids

ACKNOWLEDGEMENTS

First and foremost, this work would not have been possible without the love of my first teacher, my mother, and her belief in my abilities. Every teacher I've had since then has played a role in bringing this work to fruition: from my high school mentors to my undergraduate and graduate professors. In particular, I thank Drs. Alan Zaring for being the guiding inspiration of my undergraduate life, Jeffrey Nunemacher for constantly pushing me to the limit of my abilities, and Gerry Goldstein, for guiding me to the right path in molecular biology.

I thank my candidacy and dissertation committee members, Drs. Dave Mount, Doug Julin, Osnat Herzberg, Samir Khuller, and Sarah Woodson, for their time and advice during the course of my candidacy and final defense.

I am grateful to Shriram Krishnamurthi for picking out the flaws in my thinking; Jan Pedersen for giving me new ideas, construc-

tive criticism, financial aid, and help in writing the thesis; Brett Milash, Michael Braxenthaler, Rui Luo, and Zhanglin Lin for engaging intellectual conversations; Hillary Gilson, Jim Given, and Ursula Pieper for critical reading of the thesis; and everyone else at CARB, USENET, and elsewhere who patiently listened to my ideas and offered me advice.

I thank John Moult, for being a mentor, a friend, and giving me a direction to proceed. The wonderful arguments, the ice cream and the movies, his energetic and jolly attitude, and his constant encouragement have had a tremendous impact on this work.

Final thanks go out to Maureen Perotto who has put up with a lot, especially in the time of writing the thesis, and has been a source of pleasant distraction while providing constant support.

This work was supported in part by a Life Technologies Fellowship to Ram Samudrala and National Institutes of Health grant GM41034 to John Moult. Some computations were performed using National Instutite of Standards and Technology computing resources.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMPS, Alignment of Multiple Protein Sequences

BLAST, Basic Local Alignment Search Tool

CASP, Critical Assessment of protein Structure Prediction methods

CDF, contact discriminatory function

CDR, complimentary determining region

CF, clique finding

crabpi, cellular retinoic acid binding protein I

csc, cucumber stellacyanin

DSSP, Dictionary of Secondary Structure Preferences

edn, eosinophil derived neurotoxin

egi, endoglucanase I

e5.2, immunoglobulin domain protein

GA, genetic algorithm

HIV, Human Immunodeficiency Virus

HMM, Hidden Markov Model

hpr, histidine-containing phosphocarrier protein

IFU, independent folding unit

IRAPDF, linearly interpolated residue-specific all-atom conditional probability discriminatory function

MP, Minimum Perturbation

ncd, neurocalcin delta

NMR, nuclear magnetic resonance

nm23, nucleoside disphosphate kinase protein

NVPDF, non-residue-specific virtual-atom conditional probability discriminatory function

PDB, Protein Data Bank

PDF, probability discriminatory function

pns1, polyribonucleotide nucleotidyl s-transferase

p450, heme protein

RAPDF, residue-specific all-atom conditional probability discriminatory function

RVPDF, residue-specific virtual-atom conditional probability discriminatory function

RMSD, root mean square deviation

SCD, self-consistent domain

SCOP, Structural Classification of Proteins

ubc9, ubiquitin conjungating enzyme

3D, three-dimensional

Chapter 1

Background and overview

1.1 From protein sequence to protein structure to protein function

Once a protein sequence has been determined, deducing its unique three-dimensional (3D) native structure is a daunting task. Experimental methods to determine protein structure in detail, such as x-ray diffraction studies and nuclear magnetic resonance (NMR) analyses, are highly labour intensive [1, 2, 3, 4, 5]. Since it was discovered that proteins are capable of folding into their unique functional 3D structures without any additional genetic mechanisms [6], over 25 years of effort has been expended into the determination of 3D structure from sequence alone, without further experimental data. Despite the amount of effort, the protein folding or protein structure prediction problem, as it has come to be known, remains largely unsolved [7, 8, 9, 10, and references therein].

Knowing the structure of a protein sequence enables us to probe the function of the protein [11, 12, 13, 14, for example], understand substrate and ligand binding [15, 16, 17, 18, for example], devise intelligent mutagenesis and bio-chemical protein engineering experiments that improve specificity and stability [19, 20, 21, 22, for example], perform rational drug design [23, 24, for example], and design novel proteins [25, 26, 27, for example]. Understanding structure has potential applications in the various genome projects being undertaken, such as mapping the functions of proteins in metabolic pathways for whole genomes [28, 29] and deducing evolutionary relationships [30].

1.2 Predicting structure through comparative modelling

The continually increasing amount of DNA and protein sequence data from genome projects makes it infeasible for NMR and x-ray crystallography techniques to rapidly provide information about the 3D structures of the sequences determined [31]. Thus there is an urgent need for predicting structure from amino acid sequence. Over the last 30 years, a simple but powerful way to make predictions concerning the 3D structure of proteins utilising evolutionary relationship among families of proteins has been developed [8, 32, 33].

Comparative modelling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each position based on an optimal structural superposition, have similar structures [34]. For example, 75 pairs of hemoglobin structures with percentage sequence identities ranging from 30% and above can be superimposed within a C_α root mean square deviation (RMSD) of 2.0 Å for all pairs of superimposable residues [35, 36]. The similarity of structures is very high in the so-called "core regions", which typically are comprised of a framework of secondary structure elements such as α-helices and β-sheets [34]. Loop regions connect these secondary structure and generally vary even in pairs of homologous structures with a high degree of sequence similarity. [37, 38].

One of the earliest comparative models was that of bovine trypsin which was built from the homologous serine protease α-chymotrypsin, where the differences in the specificity pockets between trypsin and chymotrypsin were successfully explained through the use of the model [39]. Since then, comparative modelling methods have been used in applications as diverse as constructing an approximate model of the Human Immunodeficiency Virus (HIV) protease which was built based on the identification of a catalytic triad observed in the acid proteases [40]; predicting a variety of features of lysosomal proteins such as catalytic, binding, and proteolytic cleavage sites [41]; and identifying new classes of "lead" compounds for drug development against enzymes in the malarial [42] and schistosome parasitic life cycles [43].

Comparative modelling generally has also been used to aid in the determination of phases for solving crystal structures using molecular replacement [44, 45] and has helped focus site-directed mutagenesis studies of the relationship between structure and function in macromolecules [46], as well as studies of specificity and stability in molecular recognition [47, 48].

1.2.1 The model building process

Given a target sequence and a related parent structure, the process of building a model is conceptually straightforward [8, 49]. First, an alignment is performed between the sequence for which the structure has been determined by experimental methods (the parent) with the sequence to be modelled (the target).

This sequence alignment is used to construct an initial model (sometimes referred to as a framework or template) by copying over some side chain and main chain coordinates from the parent structure based on the equivalent residue in the sequence alignment.

Figure 1.1 illustrates situations where side chains and main chains need to be built. Side chains in the alignment that are not identical, and identical side chains thought to vary between the parent and target structures, are built using a variety of side chain building methods available [50, 51, 52, 53].

Generally, a given alignment may have regions in the target sequence that do not correspond to any other region in the parent sequence. These regions represent insertions in the target relative to the parent (Figure 1.1a). Since no corresponding coordinates exist in the parent structure, the residues in these regions must be built using a variety of main chain building methods available [54, 55, 56, 57, 58, 59]. In the case of regions in the parent sequence that do not correspond to any other residue in the target sequence, which represent deletions in the target relative to the parent (Figure 1.1b), the model must be constructed in such a way that the peptide bond between the residues neighbouring the deletion is covalently plausible. Finally, there may be regions in the alignment that do not correspond to insertions or deletions but may have low level of local sequence identity. These regions can potentially vary in the target structure and thus must also be built by some means other than simply copying the atomic coordinates from the parent structure(s).

Building side chains in all cases where the main chain is built must be done simultaneously. The accuracy of the side chain prediction inevitably depends on the accuracy of the main chain prediction [60].

The accuracy of comparative models generally depends on the percent of sequence identity between the target and parent structures, with C_α RMSDs ranging from 1.0-2.0 Å for alignments with $> 40\%$ sequence identity to RMSDs higher than 6.0 Å (which is close to random for small proteins) for alignments with $< 30\%$ sequence identity [8].

1.2.2 Problems with current comparative modelling approaches

Accurate prediction of the amino acid sequence alignment, side chain, and main chain conformations is absolutely essential if the model built is to be of use in further studies regarding function. Errors in any or all of these components in the model building process can result in significant deviations between the model and target structures in important functional regions.

For example, in the case of model building of eosinophil derived neurotoxin (edn) by Sali and co-workers [61], residues 1-16 are misaligned. In this case, the misalignment affects the active site residues, glutamine 14-histidine 15, in the edn experimental structure. These residues should be aligned to glutamine 11-histidene 12 in the parent structure (ribonuclease A), but are aligned to threonine 17-serine 18. The active site histidine in the model as a result of this misalignment is more than 20.0 Å away from its correct position in the experimental structure.

In the case of the model building of the histidine-containing phosphocarrier protein (hpr) from *M. capricolum* by Delarue and Koehl [8], the side chain conformation of aspartic acid 10 which forms part of the active site was predicted incorrectly, as was the main chain region containing the other active site residues

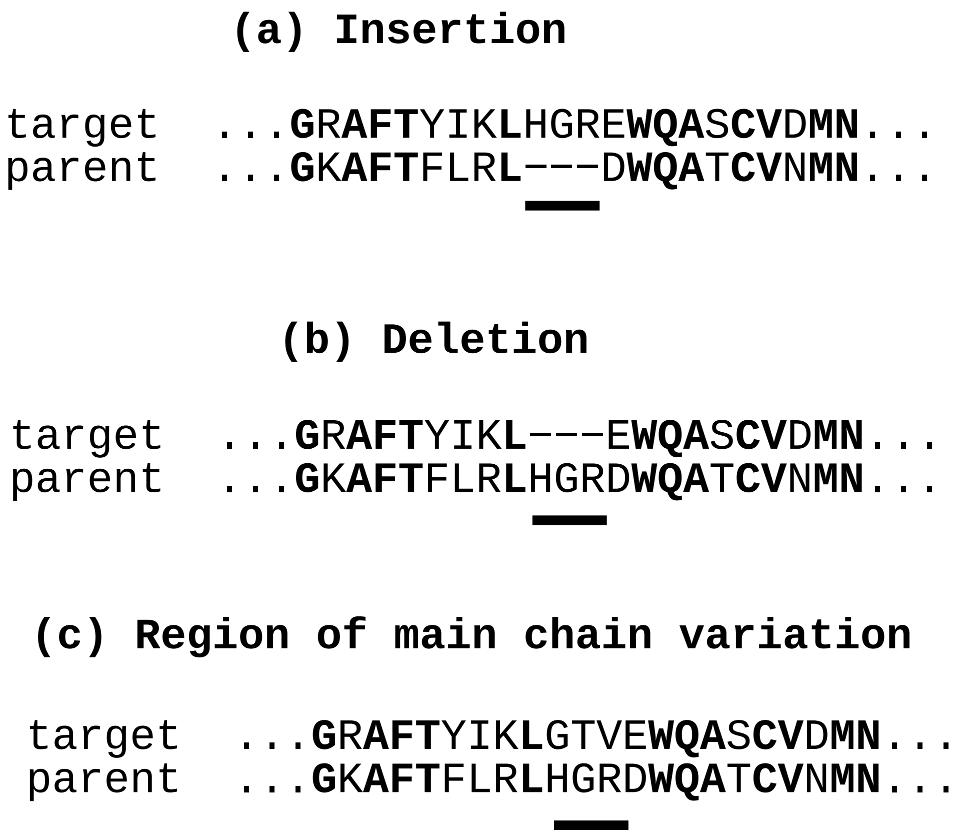

(a) Insertion
target ...GRAFTYIKLHGREWQASCVDMN...
parent ...GKAFTFLRL---DWQATCVNMN...
(b) Deletion
target ...GRAFTYIKL---EWQASCVDMN...
parent ...GKAFTFLRLHGRDWQATCVNMN...
(c) Region of main chain variation
target ...GRAFTYIKLGTVEWQASCVDMN...
parent ...GKAFTFLRLHGRDWQATCVNMN...

Figure 1.1: Main chain and side chain building scenarios in comparative modelling. An alignment is first performed between the sequence for which structure has been determined by experimental methods (the parent) and the sequence to be modelled (the target). The sequences are listed in one-letter code for the amino acid, and bold letters indicate identities in the sequence alignment. Side chains must be built for any residues in the target that does not correspond to an identity in the alignment, and for any residues where the side chain conformation is thought to vary in the target relative to the parent structure. Main chains must be built in the case of insertions (a), regions surrounding a deletion (b), and in other regions of suspected main chain variation (c), indicated above by a thick line. All other main chain and side chain conformations are simply copied from the parent structure.

histidine 15 and arginine 17. Figure 1.2 compares the predicted model to the experimental structure in these regions.

Each of the above issues in comparative modelling, accuracy of alignment,

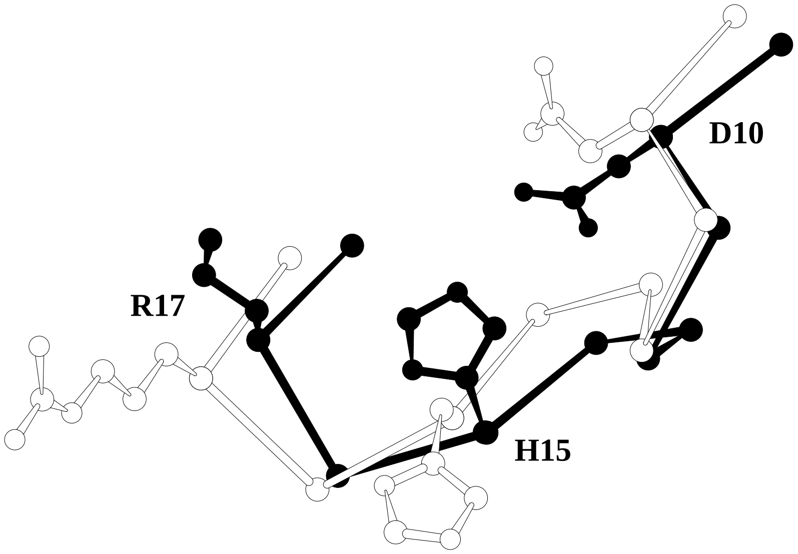


Figure 1.2: Comparison of the predicted (white) and experimental (black) active site region in the histidine-containing phosphocarrier protein (hpr) from *M. capricolum*. The predictions were made by Delarue and Koehl as part of the first meeting on the Critical Assessment of protein Structure Prediction (CASP1) [8]. The three active site residue side chains (aspartate 10, histidine 15, and arginine 17) are shown with all atoms, and the main chain region from residues 9-18 is shown as a C_α trace. The conformations of all three side chains are predicted incorrectly. The missing atoms in the arginine 18 side chain in the experimental crystallographic conformation is due to a lack of electron density associated with the side chain beyond the C_δ atom.

side chain positioning and main chain construction, has implications with regard to the usefulness of the model built. If the alignment, side chains, and main chains in the above examples had been predicted accurately, the resulting models may have been useful to a molecular biologist in designing experiments that probe function. However, if important functional regions, which can include side chains in the active site as well as "loop regions" that interact with the binding site, are not predicted accurately, the model is useless.

In each scenario, current comparative modelling methods fail because they do not take into account the environment in which model building occurs. That is, they ignore the interconnectedness, or context-sensitivity, seen in protein structures [62]. This is explored in greater detail in Chapter 2.

In the case of alignment, taking the environment into account involves constructing an initial model and checking to see if it makes sense structurally. For building side chains and main chain regions, residues present in the environment of the initial model that are in contact with the region being built must also be constructed simultaneously. Building side chains and main chain regions while keeping the environment fixed may result in inaccurate predictions for two reasons: (*i*) correct conformations, even if selected by a discriminatory function, may be sterically excluded as the environment is only approximate, and (*ii*) the environment in some cases can be grossly incorrect [8] and it is likely that the prediction of a side chain or a main chain in the context of an incorrect environment will fail due to poor energetics.

For example, in the case of modelling the active site of the *M. capricolum* hpr (Figure 1.2), the corresponding amino acid to aspartic acid 10 in the sequences of three related parent hprs, one of which was used by Delarue and Koehl as a template to build the model, is an alanine residue. The main chain region around residues 13-17 also shifts slightly in the *M. capricolum* hpr relative to the other hprs [63]. Thus a model building approach must take into account the structural context-sensitivity of the active site and build the main chain and side chain conformations simultaneously in that region for accurate prediction.

1.2.3 A solution to the context-sensitivity problem

For building side chains and main chains in an interconnected context-sensitive manner, we propose a solution based on a graph-theoretic clique finding approach. This approach assumes that the following exist:

- A sequence alignment method for aligning the target sequence to the parent sequence accurately in such a way that the sequence alignment is identical to a structure-based alignment obtained from optimal superimposition of the structures.
- A method for sampling side chain conformations.
- A method for sampling main chain conformations for building insertions, residues around a deletion, and other regions of main variation.
- A discriminatory function for distinguishing correct side chain and main chain conformations from incorrect ones.

While existing sequence alignment and main chain sampling methods are not perfect, we do not develop any new methods for alignment of sequences and for sampling main chain regions. Instead, we use pre-existing algorithms and techniques. We focus our efforts on developing a discriminatory function for distinguishing correct conformations from incorrect ones, a method for generating the most probable side chain conformations given a fixed main chain, and use this in conjunction with a main chain sampling method previously published to predict side chain and main chain conformations in comparative modelling situations.

1.2.4 Evaluation of comparative modelling methods

Even though comparative modelling has been demonstrated to have broad utility (see discussion at the beginning of Section 1.2), until recently there have been only very few cases where the modelling process has been assessed objectively. While many comparative modelling methods published in the literature have produced good results in test cases where the experimental structure is already known, the results have not been generally positive in cases where the models were built before the solution of the experimental structure [64].

In 1994, for the first time, an experiment on the Critical Assessment of protein Structure Prediction methods (CASP1) was held to objectively and rigourously assess protein structure prediction methods, including comparative modelling methods, in a large-scale manner [7, 8]. This experiment was designed to test predictive methods by seeking predictions for sequences for which the correct experimental structures were not known ahead of time. The models built for these sequences were collected before the corresponding experimental structures were published and then compared to the correct answers. The predictive powers of different methods were assessed and compared using identical criteria. In 1996, a second experiment, CASP2, was held in a similar spirit.

While we provide standard benchmarks and test our methods in cases where the answer is already known, we use the results of the first and second CASP experiments as the definitive tests of our predictive methods, and to evaluate the progress of our comparative modelling methods.

1.3 Organisation of the thesis

The next chapter introduces the context-sensitivity problem which we were forced to confront at CASP1 while building comparative models (using a combination of conventional methods) for three proteins, and provides the motivation for this work.

Chapters 3 and 4 describe a discriminatory function to distinguish correct from incorrect structures of a protein sequence and a side chain sampling method for generating the most probable side chains giving a fixed main chain. Chapter 5 illustrates how these methods are combined using a graph-theoretic clique finding approach, with a main chain sampling method previously published, to handle the context-sensitivity problems encountered in comparative modelling.

Chapter 6 assesses how the graph-theoretic clique finding method performs by making blind predictions at CASP2.

Each of the topics addressed in the above chapters in and of itself helps further our understanding of protein structure-function relationships and is therefore treated as an independent unit. Each chapter also contains a summary that outlines the main points made in that chapter and places it in the context of the rest of the thesis.

We conclude by comparing the progress of our comparative modelling approach from CASP1 to CASP2, with some ideas about future prospects.

Appendix A provides a visual comparison between the experimental structures and the models constructed by us for CASP1 and CASP2 targets.

Chapter 2

Confronting the problem of interconnected structural changes in the comparative modelling of proteins

2.1 Introduction

Our objective in this work was to test the usefulness of as many of the available computational techniques for comparative modelling as possible, and to try to see where improvements can be made. To this end, models of three of the target proteins, the histidine-containing phosphocarrier protein from *M. capricolum* (hpr; 89 residues [63]), the cellular retinoic acid-binding protein I from *M. musculus* (crabpi; 137 residues[1] [65]), and the eosinophil derived neurotoxin

[1] We constructed two models of crabpi; we only consider the model with the lower RMSD to the experimental structure in this work. The numbering of the residues in the PDB file for crabpi differs from the numbering we have used. The model structure begins at M1 whereas the experimental structure begins one residue later, at P1. The first methionine is probably not present in the protein expressed in *E. coli*.

(edn; 134 residues [66]) from *H. sapiens*, were built. We divide the modelling into three main stages: (*i*) an alignment mapping the sequence of the target protein on to a template or parent structure, (*ii*) procedures for assigning side chain positions (rotamers) in the context of the surrounding model, and (*iii*) procedures for building regions of main chain. For each stage we indicate what methods were used, what went right, what went wrong, and why (if we think we know). In the last section we discuss what we learned and what type of next generation algorithms may lead to improved model accuracy.

2.2 Methods

2.2.1 Search for parent sequences with known structure

Target protein sequences were obtained from the National Center of Biotechnology Information (NCBI) protein and nucleotide sequence database Entrez [67]. A search using the program FASTA [68] was performed on the Owl database [67] to obtain sequences that were related to the target protein. The Structural Classification of Proteins (SCOP) [30, 69] database was used to find the PDB identifiers for the known structures that belonged to the same family as the target sequence.

2.2.2 Sequence and structure alignment

A multiple sequence alignment was generated with the Alignment of Multiple Protein Sequences (AMPS) package [70, 71]. The AMPS-derived alignment was used to identify regions of variability within the target sequence family. AMPS pairwise alignments were also used to determine the degree of identity between the target sequences and the other sequences of known structure. The default PAM250 mutation matrix, which contains information about frequencies of amino acid substitutions in evolutionarily-related proteins, was used to score alignments between the target and parent sequences and select the one with the best score. A length-independent gap penalty of 8.0 was used to limit the tolerance for introducing and lengthing insertions and deletions made in the sequence alignment. Structural alignments between the template structures were generated using the G program [72] based on the alignment procedure of [73]. These alignments were used to examine the structural variation at a given position and to assess the correctness of the multiple sequence alignment.

2.2.3 Building side chains

Following the sequence alignment, an initial model was generated by mutating the residues of the template structure with the highest identity to the target sequence. This was done using a minimum perturbation (MP) technique implemented by the program MUTATE [74]. The MP method changes a given amino acid to the target amino acid preserving the equivalent χ angles, as determined by an equivalence table, between the two side chains. The χ angles not present in the model are constructed using a standard library based on the residue type. A careful environment analysis was performed by visual inspection of the initial model using interactive computer graphics. If residue A in a template structure was changed to residue B in the model, then the environments (the contacting residues, their locations, and conservation) of residue A and residue B were compared. The rules used to consider plausibility were packing (whether there was too much or too little space left after any change), favourable and unfavourable electrostatic interactions (hydrogen bonding, salt bridges) of side chains and main chain, and burial or exposure of a residue. The confidence of the model at a given position was rated qualitatively using these criteria. Alternate side chain rotamer choices were considered for regions of low confidence.

Two other methods using different χ libraries were employed in order to generate possible alternative rotamers. These were from the INSIGHT [75] and QUANTA [76] packages. In addition, a preliminary version of a self-consistent domain (SCD) method [77] was used. This method iteratively adjusts side chain conformations within a neighbourhood to find the electrostatically most favourable clash free set, and checks for consistency with adjacent and overlapping neighbourhoods.

An electrostatic energy analysis using point charge electrostatics with an intergroup cutoff distance of 5.0Å was performed on the model using the ENEANA program [78]. Residues with unfavourable electrostatic interactions were corrected by examining alternative residue conformations and selecting an energetically favourable one. Residues with unlikely burial were identified by checking the probability of observing that particular burial in an experimental protein structure and similarly corrected.

2.2.4 Building insertions and deletions

Insertions and regions flanking the deletion in the target sequences relative to the templates were rebuilt using one of four different methods (Table 2.4).

In hpr, a lengthening of the C-terminal region compared with the primary template from *B. subtilis* appeared to enable the formation of an additional short anti-parallel β-strand to pair with the N terminal strand. These residues (87-89) were rebuilt manually.

In crabpi, the main chain for residues 34-37 was manually adjusted to extend the C terminus of the second α helix. For residues 90-92 in crabpi, loops that had the same structural pattern as the region of uncertainty (two strands with a three residue loop between them with glutamatic acid as the centre residue of the loop) were obtained from a database of structures. A manual inspection of these loops was used to select the most appropriate one, which are residues 320-322 in α_1 anti-chymotrypsin (PDB code 2ach-A [79]). Residues 101-106 in crabpi were built using the SCD loop building program [55]. This method systematically generates a large set of possible main chain and side chain conformations. In this instance, too many main chain possibilities were generated, and therefore a subset had to be chosen by manual inspection.

Residues 1-5 in edn were built using *ab initio* methods described in [80] which predicted this set of residues to be partly helical. Residues 18-22 represent a deletion in edn with respect to the 7rsa template. This region was constructed manually using 1onc as a template (which results in a deletion of only 2 residues, as opposed to a deletion of 6 residues when 7rsa is used) and further refined using CONGEN [54, 56]. In this procedure, each side chain in the loop and its surroundings is spun in turn to find the lowest energy conformation. The process is iterated until the total energy has converged. For the other three loops in edn (residues 62-70, 89-96, and 112-126), distance constraints from the parent structure were used to search a database of loops [57] for matching regions. The matching loops were positioned in the model structure using the method of Martin, *et. al.* [57]. Side chains were then rebuilt as described above using CONGEN [54, 56].

2.2.5 Building other regions of main chain variation

For some regions, we suspected that the main chain would not faithfully follow the primary template after a preliminary environmental analysis. We built main chains for those regions by explicitly copying them from other parent structures. In the case of hpr, we copied the main chain for residues 51-55 from 1poh. In the case of crabpi, residues 1-10, 46-52 and 116-118 were copied from a secondary template, 1opa-A.

2.2.6 Model refinement

Once the final side chain rotamers and loop conformations were selected from the variety of choices available, the models were energy minimised for 100 steps using the steepest descent method and either the CHARMM or DISCOVER potentials without electrostatics [76, 75]. This procedure was intended to remove steric clashes and to produce acceptable bond lengths and angles rather than change the conformation significantly.

2.2.7 Calculation of RMSDs between the model and experimental structures

Throughout this work, the RMSD between two structures with n equivalent positions is defined as

$$\sqrt{\frac{\sum_{i=1}^{n} dx_i^2 + dy_i^2 + dz_i^2}{n}}, \tag{2.1}$$

where dx_i, dy_i and dz_i are distances in Cartesian space between two structures at position i. RMSDs were computed using the program G [72] and represent global RMSDs (i.e., RMSDs listed for specific regions are calculated after optimally

Structure (PDB code)	Source	Function	Sequence identity (%)	Resolution (Å)	Reference
Histidine-containing phosphocarrier protein - hpr					
2hpr	*B. subtilis*	phosphotransferase	40.9	2.0	[81]
1ptf	*S. faecalis*	phosphotransferase	40.2	1.6	[82]
1poh	*E. coli*	phosphotransferase	34.1	2.0	[83]
Cellular retinoic acid binding protein I - crabpi					
2hmb	*H. sapiens*	heart fatty acid binding	42.7	2.1	[84]
1opa	*R. rattus*	retinol transport	36.6	1.9	[85]
1lie	*M. musculus*	adipocyte lipid binding	34.6	1.6	[86]
2ifb	*R. rattus*	intestinal fatty acid binding	29.0	2.0	[87]
1mdc	*M. sexta*	fatty acid binding	23.8	1.6	[88]
Eosinophil derived neurotoxin - edn					
7rsa	*B. taurus*	pancreatic ribonuclease	33.9	1.3	[89]
1bsr-A	*B. taurus*	seminal ribonuclease	31.4	1.9	[90]
1onc	*R. pipiens*	pancreatic ribonuclease	29.4	1.6	[91]

Table 2.1: Percentage sequence identity between the target sequence and other homologous sequences with known structures as determined by AMPS pairwise alignments for CASP1 targets. For each target, details regarding the known homologs are given.

superimposing the complete molecules [73]).

2.3 Results

2.3.1 Template structures for modelling

Once the related sequences for each target were found, high resolution parent structures obtained using x-ray crystallography were used as template structures for the modelling. Table 2.1 shows the parent structures that were selected for each family and the percentage identity to the target protein sequence.

2.3.2 Sequence alignment

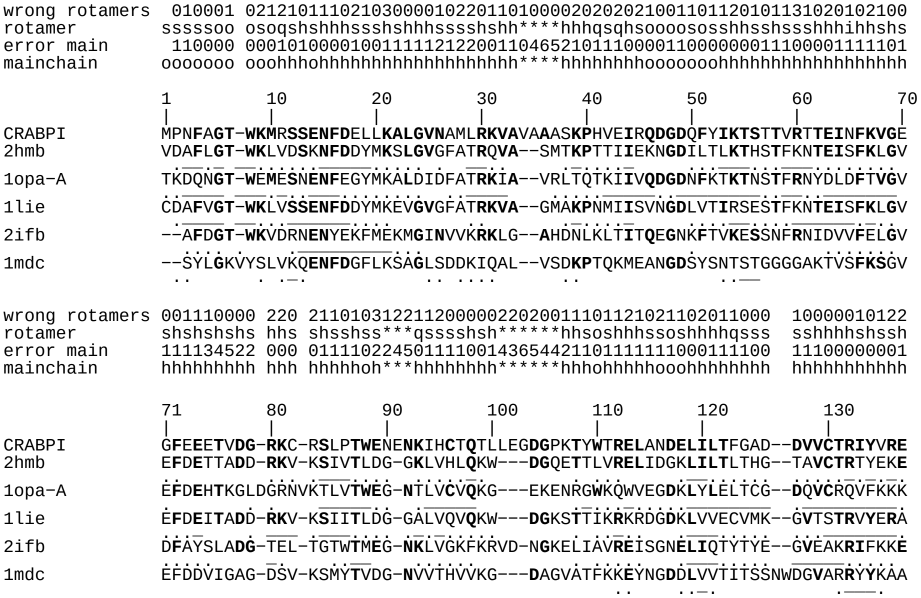
Visual inspection of the initial AMPS alignments revealed two regions where the alignment was dubious (see Figure 2.1). One of the regions is in crabpi (insertion at residue 90 which is not seen in the AMPS alignment), and the other is in edn

```
           88                105
           |                |
2hmb-final LDG-GKLVHLQKW---DG
                |    |     ||
CRABPI     WENENKIHCTQTLLEGDG
                  |        ||
2hmb-AMPS  LDGGKLVHLQKW----DG

              1              16
              |              |
7rsa-final ------KETAAAKFERQHM
                     |  || ||
EDN        ---KPPQFTWAQWFETQHI
                  |
7rsa-AMPS  KETAAAKFERQHMDSSTAA
```

Figure 2.1: Differences between the final correct sequence alignments and those generated with AMPS. Correct alignments were produced by visual inspection of the sequences and preliminary models.

(the FEQTH sequence (residues 11-15) is aligned incorrectly). The incorrect alignment in crabpi results in K93 being buried, which seemed electrostatically intolerable. In the case of edn, inspection of the alignment suggested a better alternative. Both alignments were adjusted manually. The final alignments for all proteins agree with those produced by structural superposition of the target experimental structures with the respective primary templates. Figure 2.2 shows the results of the sequence alignment for crabpi after correction using structural information.



```
wrong rotamers  010001 021210111021030000102201101000020202021001101120101131020102100
rotamer        sssssoo osoqshshhhssshshhhssssshshh****hhhqsqhsoooososshhsshssshhhihhshs
error main      110000 000101000010011111212200110465210111000011000000011100001111101
mainchain      ooooooo ooohhhohhhhhhhhhhhhhhhhhhhh****hhhhhhhoooooohhhhhhhhhhhhhhhhhh

               1        10        20        30        40        50        60        70
               |        |         |         |         |         |         |         |
CRABPI         MPNFAGT-WKMRSSENFDELLKALGVNAMLRKVAVAAASKPHVEIRQDGDQFYIKTSTTVRTTEINFKVGE
2hmb           VDAFLGT-WKLVDSKNFDDYMKSLGVGFATRQVA--SMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGV
1opa-A         TKDQNGT-WEMESNENFEGYMKALDIDFATRKIA--VRLTQTKIIVQDGDNFKTKTNSTFRNYDLDFTVGV
1lie           CDAFVGT-WKLVSSENFDDYMKEVGVGFATRKVA--GMAKPNMIISVNGDLVTIRSESTFKNTEISFKLGV
2ifb           --AFDGT-WKVDRNENYEKFMEKMGINVVKRKLG--AHDNLKLTITQEGNKFTVKESSNFRNIDVVFELGV
1mdc           --SYLGKVYSLVKQENFDGFLKSAGLSDDKIQAL--VSDKPTQKMEANGDSYSNTSTGGGGAKTVSFKSGV

wrong rotamers 001110000 220 21101031221120000022020011101121021102011000  10000010122
rotamer        shshshshs hhs shsshss***qsssshsh******hhsoshhhssoshhhhqsss  sshhhhshssh
error main     111134522 000 01111022450111100143654421101111111000111100  11100000001
mainchain      hhhhhhhhh hhh hhhhhoh***hhhhhhhh******hhhohhhhhooohhhhhhhh  hhhhhhhhhhh

               71       80         90        100       110       120         130
               |        |          |         |         |         |           |
CRABPI         GFEEETVDG-RKC-RSLPTWENENKIHCTQTLLEGDGPKTYWTRELANDELILTFGAD--DVVCTRIYVRE
2hmb           EFDETTADD-RKV-KSIVTLDG-GKLVHLQKW---DGQETTLVRELIDGKLILTLTHG--TAVCTRTYEKE
1opa-A         EFDEHTKGLDGRNVKTLVTWEG-NTLVCVQKG---EKENRGWKQWVEGDKLYLELTCG--DQVCRQVFKKK
1lie           EFDEITADD-RKV-KSIITLDG-GALVQVQKW---DGKSTTIKRKRDGDKLVVECVMK--GVTSTRVYERA
2ifb           DFAYSLADG-TEL-TGTWTMEG-NKLVGKFKRVD-NGKELIAVREISGNELIQTYTYE--GVEAKRIFKKE
1mdc           EFDDVIGAG-DSV-KSMYTVDG-NVVTHVVKG---DAGVATFKKEYNGDDLVVTITSSNWDGVARRYYKAA
```

Figure 2.2: Final alignments of the cellular retinoic acid binding protein I (crabpi) target sequence to other sequences in the family that have known structures. The first line indicates the accuracy of the predicted rotamer by listing the number of χ angles that deviated more than 30° from the experimental structure for each residue. The second line is the list of rotamer choices that were used to generate the final model—for each residue, the rotamer was selected using one of the following methods: s - standard library; i - INSIGHT; q - QUANTA; or by selecting from a template structure: h - 2hmb; o - 1opa-A. The third line lists the C_α deviation between the target experimental structure and the model (0: 0-1Å; 1: 1-2Å; ...). The fourth line indicates the parent structure from which the main chain was taken: o - 1opa-A; h - 2hmb. An '*' indicates that the main chain and/or side chain was generated using loop building techniques. In the multiple sequence alignment, conserved residues are indicated by bold letters. For each amino acid in all the sequences aligned to the target, the C_α distance between the target experimental structure and each related structure after structural alignment is given: a solid line under the one letter code indicates that the C_α distance was within 1.0 Å, a dotted line indicates that the C_α distance was within 2.0 Å, and a blank indicates that the C_α distance was greater than 2.0 Å.

	All side chains			“No excuse” side chains		
Rotamer Origin	hpr	crabpi	edn	hpr	crabpi	edn
Library	50.0% (60)	48.0% (100)	50.5% (95)	25.0% (8)	50.0% (6)	37.5% (24)
Identity	34.7% (46)	38.0% (84)	25.0% (48)	26.6% (15)	37.5%(8)	24.2% (33)
Loops	80.0% (5)	66.6% (15)	66.6% (75)	50.0% (2)	00.0% (0)	73.6% (19)
Manual	65.5% (29)	41.1% (34)	00.0% (5)	50.0% (2)	33.3% (3)	00.0% (5)
All	48.5% (142)	45.4% (233)	49.3% (223)	29.6% (27)	41.1% (17)	38.2% (81)

Table 2.2: Percentage of model χ angles that deviate more than 30° from the experimental structure, considering rotamers that were constructed using a standard library (row 1), identities (row 2), loop builders (row 3), and by other methods (row 4; see Figure 2.2). The overall percentages are given in the last line. The “no excuse” set on the right hand side omits residues that have crystallographic contacts closer than 4.0 Å to a neighbouring protein molecule and χ angles where one or more atoms have a temperature factor greater than 25.0 Å^2. The numbers in parenthesis show the total number of χ angles that were included.

2.3.3 Side chains

The percentage of model χ angles that deviated more than 30° from those in the experimental structures is given in the left hand side of Table 2.2. A number of χ values may be affected because of high temperature factors or contacts with neighbouring molecules in the crystal structure. For the purpose of evaluating the methods used, it is desirable to eliminate these effects and produce a “no excuse” set of χ angles. We thus calculated additional statistics, excluding residues that have atomic contacts of less than 4.0 Å to a neighbouring molecule and χ angles where one or more atoms had a temperature factor greater than 25.0 Å^2. The right hand side of Table 2.2 shows these results. Errors are significantly lower in this set, but still surprisingly large, even for cases where the residues in the models and template structures are identical (row 2).

Changes in the position of conserved side chains between related structures must be because of changes in other parts of the structure. To obtain more insight into these correlation effects and others, we examined the seven cases (three Library, three Identity, and one Manual) in the "no excuse" set of the crabpi model where the χ values are incorrectly predicted. This was done by introducing the model rotamers into the experimental target structure and inspecting the resulting environment. Table 2.3 shows the results of this analysis. For three of the seven rotamers, the model rotamers were not acceptable in the experimental structure because of clashes that are not present in the model. For two of these, the clashes are directly attributable to main chain differences between the experimental structure and model, so better side chain positioning algorithms would not help. Figure 2.3 illustrates one of these main chain effects for I53. Here, a difference in the main chain in the target structure relative to the template of the neighbouring I64 results in the model rotamer being unacceptable. The side chain conformation of the conserved I64 in the model is similar to that seen in the experimental structure. There is also a side chain clash between the model conformation of I53 and the experimental conformation of R112. Similarly, the side chain conformation of F123 is determined by the conformation of a loop region that was incorrectly modelled. The side chain conformation of I133 is dependent on the conformation of the side chain of R11, which forms a salt bridge with E118 in the experimental structure, but not in the model. The experimental conformation of R112 appears to interact better with solvent than the model conformation. For the other three cases, our criteria could not distinguish between the experimental and model rotamers.

χ angle	Residue	$\Delta\chi$ (°)	Effect of using the rotamer in the experimental structure
χ_1	I53	66	clash with I64 and R112 (see Figure 2.3)
χ_3	R112	74	incompatible with experimental solvent structure
χ_1	I120	34	no clashes
χ_2	I120	125	no clashes
χ_2	F123	45	clash with L121 and V77; V77 is in an incorrectly modelled loop
χ_1	I133	149	clash with R11 and S12; these residues have high temperature factors
χ_1	V135	66	no clashes

Table 2.3: Correlation of individual χ angle errors with other errors in the cellular retinoic acid binding protein I (crabpi) model. Data are for the incorrect angles in the right hand side of Table 2.2. For each χ listed, the conformation of the corresponding residue in the experimental structure was changed to adopt the model χ value and the resulting environment inspected for inconsistencies.

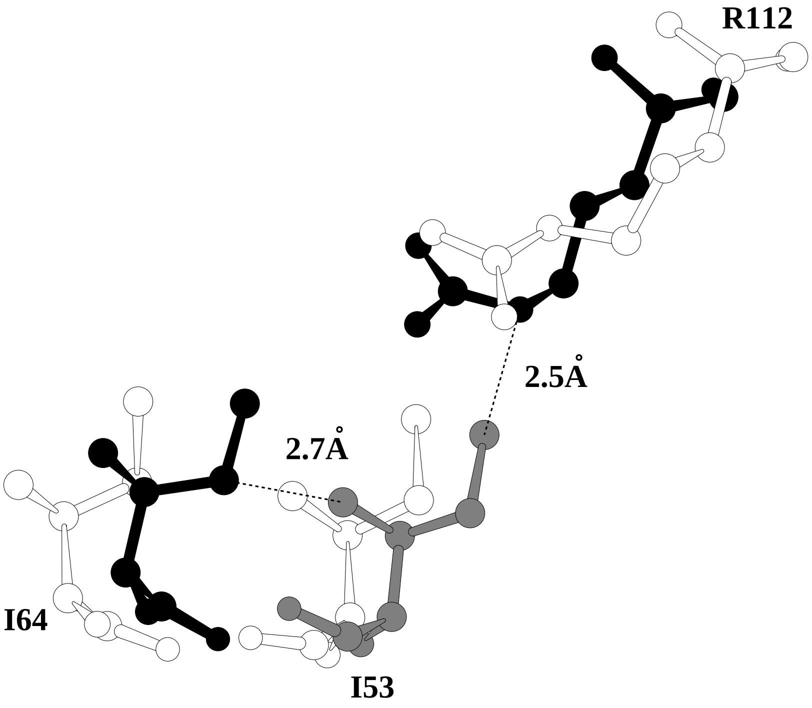


Figure 2.3: An example of a incorrect model rotamer in the cellular retinoic acid binding protein I (crabpi) that is unacceptable given the context of the experimental structure. The model structure is white, the experimental structure is black, and the model side chain of I53 placed in the experimental structure is grey. In the model, I64 is further away because of a main chain shift, so the principal clash excluding the I53 model side chain conformation is not present.

Region	RMSD (Å)	Max root error (Å)	Method	Intermolecular contacts	<B> (Å^2)
hpr					
87-89	5.5	0.6	manual	88-89	29.6
crabpi					
34-37	5.0	2.8	manual	37	32.7
90-92	4.2	2.0	pattern matching	-	39.0
101-106	5.3	2.2	systematic search [55]	-	80.0
edn					
1-5	9.7	3.3	*ab initio* [80]	3-5	13.3
18-22	5.3	3.4	manual & CONGEN [54, 56]	19,21	8.7
62-70	3.1	1.2	database [57]	66-67,69	22.9
89-96	5.2	6.1	database [57]	90-91,95	37.1
112-126	9.9	7.1	database [57]	113-114,116-117 122,124-125	15.9

Table 2.4: C_α RMSDs between the experimental structure and the model for insertions and residues flanking the single deletion (edn: 18-22). The larger of the two root C_α atom errors is given in column 4. For each region, the list of residues with at least one atom in the side chain having intermolecular contacts less than 4.0 Å is given in column 6. Column 7 lists the average temperature factor (B) for the C_α atoms.

2.3.4 Insertions and deletions

All of the regions representing insertions and deletions have final conformations with C_α RMSDs greater than 4.0 Å (Table 2.4).

In hpr, residues 87-89 were rebuilt manually. However, in the experimental structure, this region turns away from the protein surface with the last two residues involved in an intermolecular contact. Thus this conformation could be the result of a crystal packing effect.

In crabpi, the main chain for residues 34-37 does indeed adopt the conformation of an helix as we had guessed, but since the adjustment was manual, the shape of the helix is far from ideal. Experimental errors make it hard to assess what the cause of a mis-prediction of a loop is, especially in the other two cases of loop building (residues 90-92 and 101-106), since all the loops in the crabpi structure have atoms with large temperature factors (Table 2.4).

Residues 1-5 in edn were built as a helix, whereas the correct conformation in this region resembles a turn. Table 2.4 shows that all the loops in edn have contacts with neighbouring protein molecules. This factor cannot be taken in account in the modelling.

The errors in the positions of the root residues (i.e., residues flanking regions insertions or deletions) shown in Table 2.4 are large (up to 7.0Å) for many of the loops, and indicates one reason as to why the loop conformations are so poor. In such cases, the region rebuilt was not large enough and therefore no low RMSD loops could possibly be generated.

2.3.5 Other regions of main chain variation

Comparison of the experimental target structures with the primary templates used in the modelling shows other regions where the main chain conformations are significantly different. We list those regions that have a C_α RMSD greater than 3.0 Å in crabpi and edn and 1.0 Å in hpr, or regions where we explicitly changed the main chain from the primary template (Table 2.5).

Three such regions in crabpi, residues 1-10, 46-52 and 116-118, were predicted with acceptable accuracy by using 1opa-A as a template rather than 2hmb. The changes in the conformation between crabpi and 2hmb of the N terminus and the hairpin around residue 49 are correlated (Figure 2.4), and appear to be the consequence of a set of side chain differences: Two residues (F51L, W88L) are more bulky in crabpi and there is a loss of a salt bridge between residues 2 and 46. For the third region, there is a glycine in 2hmb at position 117 with ϕ/ψ values not allowed for other residue types. In crabpi and 1opa-A, there is

Region	RMSD to primary template (Å)	RMSD to model (Å)	Intermolecular contacts	<B> (Å^2)
hpr				
39	1.9	1.8	39	23.6
14-17	1.5	1.5	14,17	19.1
51-55	0.7	1.3	51-52,54-55	17.9
70-83	0.5	1.0	71-72,75-76,78-79	18.2
crabpi				
1-10	1.4	0.8	9-10	40.7
46-52	4.1	1.1	46,49	39.3
75-80	3.2	3.1	-	37.8
116-118	2.3	1.4	-	53.1
edn				
30-34	5.1	5.1	33,34	10.8
58	4.1	3.9	58	11.37

Table 2.5: C_α RMSDs for other regions of main chain variation. The RMSDs to the primary template shows how much that main chain differs from the experimental structure and the RMSDs to the model shows how accurately the variation was predicted. Three regions in crabpi were predicted well. The list of residues with at least one atom having intermolecular contacts less than 4.0 Å is given in column 4. Column 5 lists the average temperature factor (B) for the C_α atoms.

an aspartate here. These and other side chains in the core of crabpi are more similar to 1opa-A than those in 2hmb, even though the overall sequence identity is significantly lower.

All other regions of main chain variability that do not involve insertions and deletions were not identified and thus incorrectly modelled. With the wisdom of hindsight, some of these can be understood: In hpr, the region around 14-17 appears to shift from the template because of the presence of a salt bridge between H15 and D10 (see Figure 1.2). E39 has a sterically strained ϕ/ψ pair (93°,3°), and such conformational strain outside of functional regions is rare [92]. In this case it is almost certainly due to contacts with a neighbouring molecule in the crystal [63]. In crabpi, the loop around 75-80 appears to move relative to

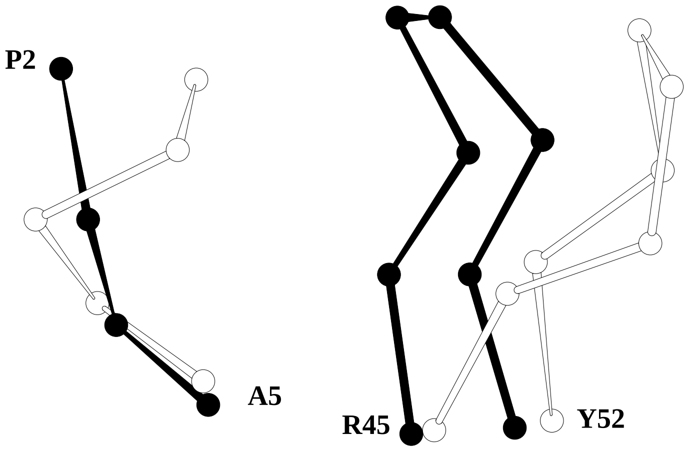


Figure 2.4: Correlated changes between the N terminus and the loop around residues 46-52 in the cellular retinoic acid binding protein I (crabpi; shown in black) compared with 2hmb (white), the closest homolog. In 1opa-A, the conformation is similar to that of crabpi, providing a better template than the primary one of 2hmb. Correlated changes of this type are common, and such regions of main chain often cannot be modelled independently.

2hmb because of the V77→A and L77→Y changes, a main chain shift of about 1.0 Å at positions 20-25, and the loss of a salt bridge between residues R59K and D78G. In edn, the change around residues 30-34 relative to 7rsa may be due to Y33T causing a clash with the conserved Y98.

Two changes in main chain were wrongly introduced. In one of these, residues

Target	C_α RMSD	All-atom RMSD
	(Å)	(Å)
hpr	1.18 (88)	1.76 (644)
crabpi	2.01 (136)	2.62 (1087)
edn	4.55 (134)	5.50 (1079)

Table 2.6: Accuracy of the models that were built compared to the experimental structures for CASP2 targets. The C_α and all-atom RMSDs between the model and the experimental structures are given. The numbers listed in parenthesis are the number of atoms considered for all residues.

51-55 in hpr, we incorrectly supposed that side chain volume changes would cause a main chain shift seen in one of the other templates. In the other case, the last helix in hpr (residues 70-83) shifts as a consequence of energy minimization done to accommodate the incorrectly built C terminus.

2.3.6 Model refinement

After energy minimization, the C_α root mean square deviation (RMSD) between the model and experimental structure increased slightly. For hpr, the increase of the C_α RMSD between the minimised and the unminimised model with respect to the experimental structure is 0.070 Å, for crabpi it is 0.014 Å, and for edn it is 0.021 Å.

2.3.7 Overall accuracies of the model compared to the experimental structure

Table 2.6 lists the RMSDs for all residues for the three models.

2.4 Discussion

The accuracy of the models is very unsatisfactory but the modelling experiment has been educational. Three common themes have emerged: The first is the usefulness of visual inspection rather than a reliance on numerical algorithms. The second is the extraordinary interconnectedness of changes between different homologous proteins. The third is the possibility, in some cases, of devising automatic procedures that may significantly improve accuracy.

2.4.1 Alignment

It has been known for some time that alignment of sequences with less than 40% identity tends to produce frequent errors in the mapping of a sequence onto a template structure [64, 93]. This is because the signal used in a mutation matrix is not strong enough to distinguish the correct structure-based alignment from other incorrect alignments. We encountered two cases of that (Figure 2.1). In one, inspection of the alignment at the amino acid sequence level suggested a better solution. In the other, inspection of the structural implications of the alignment allowed a correction. With these adjustments, the sequences of all three models were correctly aligned with the available template structures. It should be possible to develop algorithms that examine the structural implications of alternative alignments.

2.4.2 Selecting side chain rotamers

Inspection of the structural implications of default rotamer choices did lead to a small improvement in accuracy, but the error level is still very high, even for those residues not likely to be affected by crystal packing or high crystallographic temperature factors. Better methods based on consideration of interacting sets of side chains are clearly needed. Such algorithms have been published, with reported high accuracy in core regions [94, 95, 96]. However, from the analysis of the crabpi errors (Table 2.3), it is clear that these algorithms will be seriously affected by the main chain inaccuracies present in real models.

2.4.3 Insertions and deletions

Several algorithms [54, 55, 56, 57] have been shown to produce usefully accurate structures of short stretches of chain in the context of the surrounding protein. There are four obvious explanations as to why they did not work here, all related to the difference between real modelling versus algorithm development tests. The first, and in the long run most difficult to address, is the interconnectedness of the differences between related protein structures. An example of this is the interaction between the N terminal region of edn relative to ribonuclease A and the long insertion at residues 112-126. These two regions pack against each other in the experimental structure, so that predicting one in isolation from the other is likely to be very problematic (Figure 2.5). Spotting these correlated changes can some times provide the key to modelling, as in the case involving the N terminus of crabpi and the conformation of the loop around residues 46-52 (Figure 2.4). More often than not, they simply render any automatic loop builder useless. Table 2.7 shows the striking extent of the interconnectedness between the variable regions in the experimental structures.

A second and related problem is the one of the size of variable region that must be included. Both systematic and database searches are severely limited

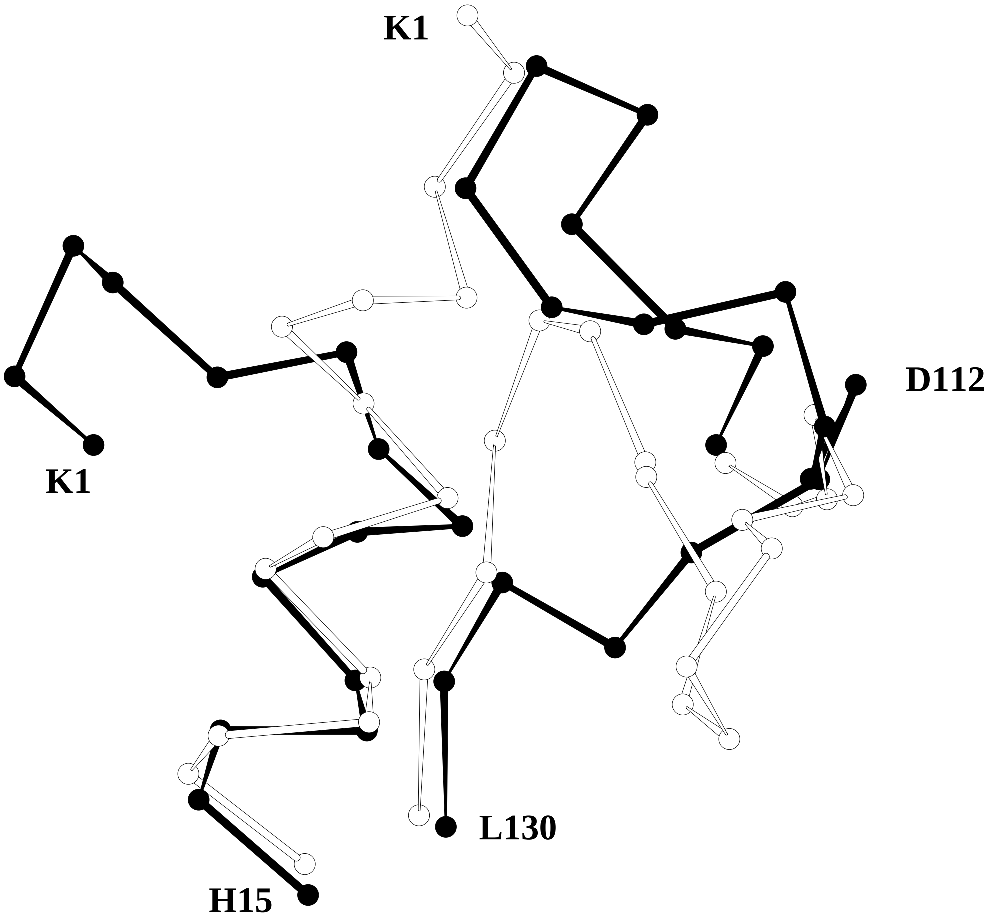


Figure 2.5: An example of an error in the building of one main chain region excluding the selection of the correct conformation of another region for the eosinophil derived neurotoxin (edn). Experimental structure of edn is black, model is white. The incorrect structure of the model N terminus occupies space needed for the loop 113-129 (shown in black).

hpr	14-17	39	51-55	70-83	87-89
14-17			3		
39					
51-55	3				
70-83					2
87-89				2	

crabpi	1-10	34-37	46-52	75-80	90-92	101-106	116-118
1-10		2	4		3		2
34-37	2						
46-52	4						
75-80						4	
90-92	3						
101-106				4			
116-118	2						

edn	1-5	18-22	30-34	58	62-70	89-96	112-126
1-5			3				3
18-22							
30-34	3					2	
58							
62-70							
89-96			2				
112-126	3						

Table 2.7: The interconnectedness of the insertions and deletions and other regions of main chain variation. The number of residue pairs that have more than one atomic contact less than 4.0Å is given.

in the size of region they can consider [59]. Effective maximum loop sizes are probably currently about seven residues, ignoring any changes in the surroundings. The short rebuilt regions that we used resulted in large errors of the root residues which led inevitably to high loop RMSDs (Table 2.4). It is apparent that insertions and deletions often cause significant main chain adjustment in the adjacent residues even where sequence conservation is high. The third problem is the need for reliable and affordable energy functions to screen possible conformations. In no case were we able to do this because of time and computing limitations. The fourth problem is knowing when to believe the reported experimental structure is relevant to the modelling. In the case of hpr, we saw one possible case of the crystal environment affecting the conformation around the C terminus. For crabpi, all three loop regions have very large temperature factors. In edn, the temperature factors are more reasonable, but all the loops are involved in intermolecular interactions in the crystal (Table 2.4). While it is very unlikely that in all cases the high RMSD between the experimental structure and the model are the result of the crystal effects, it does make it difficult to assess the individual predictions.

2.4.4 Identification of regions of main chain variability

When there are insertions and deletions in the sequence alignment, it is obvious that the local main chain conformation is unknown. But there are additional regions of main chain variability that are less easy to identify (Table 2.5).

Examination of the structural variation within the family may be useful for identifying such regions. For example, positions where the RMSD was greater than the mean RMSD within the crabpi family (2hmb, 1lie, 1opa-A, 2ifb, and 1mdc) were found to be residues 1-6, 37-40, 47-50, 58-64, 74-83, 89-91, 99-107, 116-118, 127, and 137. This would identify all regions of structural variation listed in Table 2.5 but also would identify 3 additional regions that are conserved. This analysis, together with consideration of two other factors, changes involving glycine and proline residues and the level of local sequence similarity, may help in identifying main chain changes.

As in the case of insertions and deletions, all the regions that vary extensively in main chain conformation have high temperature factors or form intermolecular contacts (see Table 2.5). For example, the conformation around residue E39 in hpr appears to be determined by crystal packing.

2.4.5 Choice of alternate templates

In the case of crabpi, we were able to significantly improve the model by recognising 1opa-A as a better choice for the main chain in three regions. Inspection of Figure 2.2 reveals other regions where the prediction could have been improved by choosing main chains from other related structures. For example, the main chain around 108-115 and 128-137 in crabpi is better modelled by using the main chain from 1opa-A. These choices depend on structural details and are difficult to automate. The usefulness of a "mix and match" approach to template selection is well known [49].

2.4.6 Long term hopes

We can see the way ahead for improvements in sequence alignment, rotamer choice and identification of main chain changes. Loop building is the most glaring and seemingly intractable problem in these results. Its successful treatment requires the development of methods for handling the interconnectedness of features in protein structures. One partial solution may be to consider pieces of chain that have their conformation determined essentially independently from the rest of the protein structure [97]. An example of the relevance of that approach is the interaction between the N terminus of edn and the region 133-129. Analysis of the surface accessibility of atoms in this region suggests that the N terminus has its conformation determined by local sequence effects [80], so it should be built first and then the long loop added.

A complete solution to the comparative modelling problem, *i.e.*, methods rivalling experiment in accuracy, requires the development of radically new approaches that handle the interconnectedness of the structural changes between related protein structures.

2.5 Summary

In this chapter, we test conventional comparative modelling methods by making blind predictions of three proteins using a variety of computational methods, heavily supplemented by visual inspection, for the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1). We consider the accuracy obtained to be worse than expected. A careful analysis of the models shows that a major reason for the poor results is the interconnectedness of the structural differences between the target proteins and the template structures they were modelled from. Side chain conformations are often determined by details of the structure remote in the sequence, and can be influenced by relatively small main chain changes. Almost all of the regions of substantial main chain conformational change interact with at least one other such region, so that they often cannot be modelled independently. Visual inspection is sometimes effective in correcting errors in sequence alignment and in spotting when an alternative template structure is more appropriate.

In the next chapters, we discuss the development of an all-atom distant-dependent conditional probability discriminatory function, a graph-theoretic method for sampling side chain conformations, and a method for building side chains and main chains in a context-sensitive manner, and see how well the improvements in our methodology work at CASP2.

Chapter 3

An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction

3.1 Introduction

Any algorithm that attempts to predict protein structure requires a discriminatory function that can distinguish between correct and incorrect conformations. These discriminatory functions can be extremely simple, like counting atomic contacts in a given conformation, or could involve elaborate calculations to determine the energy of an amino acid sequence conformation [98, 99, 100].

A class of discriminatory functions are knowledge-based. These functions compile parameters from tendencies observed in a database of experimentally determined protein structures [101, 102, 103, 104]. Knowledge-based discriminatory functions have been used to validate experimentally determined protein structures [105, 106], to recognise the fold an amino acid sequence belongs to without any sequence homology [107, 108, 109], and for *ab initio* protein structure prediction [110, 111].

Generally, knowledge-based discriminatory functions have used a simple one- or two-point-per-residue representation. That is, they usually represent each residue in a protein sequence with one or two positions in three-dimensional space. Discrimination is based on each residue's preference to be buried or exposed, its preference for a particular secondary structure conformation, and its preference to be in contact with other residues [102, 101, 103, 104]. However, to capture the finer details of atom-atom interactions in proteins, a more detailed representation is necessary [112]. For example, in a comparative modelling scenario where two possible models can be extremely close (within 1-3 Å in terms of root mean square deviation (RMSD) of the C_α atoms) to the experimentally determined structures [8], we need all the information we can possibily obtain from the two models to determine which one is more accurate. A one-point-per-residue discriminatory function may not be able to discriminate as well as an all-atom discriminatory function, which takes into account all the atoms on the side chain of the residue. In the case of comparative modelling, building side chains will not be possible using a simple representation. There exists a large degree of interconnectedness between side chains and main chains, making such a represention a necessity for the most accurate prediction (Chapter 2).

Our goal is to develop a discriminatory function that will work well at identifying the best conformation from a set of incorrect or approximate conformations. To accomplish this, we derive a pairwise distance-dependent all-atom conditional probability discriminatory function that represents atom-atom preferences in a residue specific manner. We evaluate the performance of a discriminatory function by seeing how well it correctly identifies correct conformations of an amino acid sequence from incorrect or approximate (decoy) conformations. We perform this evaluation for a wide variety of decoy types. We compare this discriminatory function to three more approximate representations to observe the effect of decreasing detail in the representation. Two of the approximate representations treat combinations of atoms as single "virtual atoms". The third approximate representation, a simple contact-based discriminatory function, is used to illustrate how much of the discriminatory information is obtained from compactness alone. We discuss the implications of these results for protein structure prediction and model refinement.

3.2 Methods

We will describe two formalisms here. The first computes the conditional probabilities, and the second computes the Boltzmann free energies, of pairwise atom-atom preferences in proteins using statistical observations of native structures. We make the observation that these two formalisms are equivalent for all practical purposes. It is however more straight-forward to think of pairwise preferences of atoms in proteins in terms of probabilities rather than in terms of free energies: the Boltzmann formalism assumes an equilibrium distribution of atom-atom preferences, the physical nature of the reference state in this formalism is not clear, and the probability of observing a system in a given state in this formalism must change with respect to the temperature.

3.2.1 The conditional probability formalism

Given a set of known structures from the Brookhaven Protein Data Bank (PDB) [113], we can make observations of atom-atom contacts in particular distance bins. The bins are discrete distance ranges whose mid-point is represented by the bin number. We compute the probability of observing atom type a and atom type b in a particular distance bin d in a native conformation F, $P(d_{ab}|F)$, like so:

$$P(d_{ab}|F) = f(d_{ab}) = \frac{N(d_{ab})}{\sum_d N(d_{ab})} \tag{3.1}$$

Here $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d. The denominator is the number of a-b contacts observed for all distance bins. We assume that the frequency distributions obtained from the database, $f(d_{ab})$, here and elsewhere, represent the probabilities.

For example, if the number of lysine N_ζ and glutamate $\mathrm{O}_{\delta 1}$ (KN_ζ-$\mathrm{EO}_{\delta 1}$) contacts within a distance range of 4.0-5.0 Å was found to be equal to 10 in the data set, and the total number of KN_ζ-$\mathrm{EO}_{\delta 1}$ contacts observed in all distance bins was 100, the frequency of KN_ζ-$\mathrm{EO}_{\delta 1}$ contacts at distance bin 4.5 is $10/100 = 0.1$.

Since we are dealing with observations of distances between pairs of atom types in compact structures, which is a subset of the sample space of all distances, we need to define an appropriate reference state to compute the conditional probabilities. We use a Bayesian approach to define the reference state as a prior distribution of contacts between pairs of atom types in *any* compact conformation, native or otherwise [114]. Our prior distribution is compiled from a set of compact conformations to obtain specific information about preferences between atom-types in the system. Using extended or random-coil conformations for compilation of the prior distribution would result in preferences for a pair of atom types to be close irrespective of the types, thus obscuring the signal in the specific preferences [115]. We assume that averaging over different atom types in experimental conformations is an adequate representation of the random arrangements of these atom types in any compact conformation. We approximate $P(d_{ab})$, the probability of finding atom types a and b in a distance bin d in *any* compact conformation, native or otherwise, to be equal to $P(d)$, the probability of seeing *any* two atom types in a distance bin d. $P(d_{ab})$ is thus computed by averaging over all atom types a and b in the native conformations:

$$P(d_{ab}) = P(d) = f(d) = \frac{\sum_{ab} N(d_{ab})}{\sum_d \sum_{ab} N(d_{ab})} \tag{3.2}$$

Here, $\sum_{ab} N(d_{ab})$ refers to the total number of contacts between all pairs of atom types in a particular distance bin d, and the denominator is the total number of contacts between all pairs of atom types summed over all the distance bins d.

We require an expression for $P(F|\{d_{ij}\})$, the probability of observing a correct conformation, F, given a specific set of n contacts between i and j atoms in the conformation $\{d_{ij}\}$. Assuming that the individual probabilities are independent of each other, we first write equations (3.1) and (3.2) as:

$$P(\{d_{ij}\}|F) = \prod_n P(d_{ij}|F) \tag{3.3}$$

and

$$P(\{d_{ij}\}) = \prod_n P(d_{ij}) \tag{3.4}$$

By applying the conditional probability chain rule [114], we notice that:

$$P(F|\{d_{ij}\}) \cdot P(\{d_{ij}\}) = P(F) \cdot P(\{d_{ij}\}|F) \tag{3.5}$$

We rewrite equation (3.5) as:

$$P(F|\{d_{ij}\}) = P(F) \cdot \frac{P(\{d_{ij}\}|F)}{P(\{d_{ij}\})} \tag{3.6}$$

Expanding (3.6) gives us the expression for computing the conditional probability of seeing a native-like conformation F given a set of n observations, $\{d_{ij}\}$:

$$P(F|\{d_{ij}\}) = P(F) \cdot \prod_{n} \frac{P(d_{ij}|F)}{P(d_{ij})} \tag{3.7}$$

More usefully, to perform the computation, we take the logarithm of both sides to obtain the summation:

$$\ln P(F|\{d_{ij}\}) = c + \sum_{n} \ln \frac{P(d_{ij}|F)}{P(d_{ij})} \tag{3.8}$$

where c is $\ln P(F)$.

We initially set all values in the numerator in equation (3.1) to one, and compile a table of the probability ratios on the right hand side (ignoring c, which is a constant) by computing the frequencies for all pairs of atom types a and b for all distance bins d using a set of experimental conformations. Given an amino acid sequence conformation, we calculate all the distances between all pairs of atom types and compute the conditional probability on the left hand side by summing up the probability ratios assigned to each distance between a pair of atom types.

3.2.2 The potential of mean force

As in the above conditional probabilty formalism, the frequencies of distances between atom types are calculated from a database of known structures available from the PDB [113]. The observed frequencies are transformed using the inverse of Boltzmann's law to yield the free energy of the interaction, as a function of some parameters (such as the distance between the atom types) [101, 103, 104]. Boltzmann's law states that a particular state s of a physical system in equilibrium is occupied with a probability $P(s)$ which is related to the free energy of that state $\Delta G(s)$:

$$P(s) = \frac{-e^{\Delta G(s)/kT}}{\sum_s -e^{\Delta G(s)/kT}} \tag{3.9}$$

where k is Boltzmann's constant, and T is the temperature. The logic used to compute statistical potentials of mean force is that given the probabilities (which can be computed from statistical observations), the free energies can be calculated using the inverse of Boltzmann's law [101].

Specifically, the free energy of interaction between atom types a and b in a distance bin d is given, using the same notation in the previous section, by:

$$\Delta G_d^{ab} = -kT \cdot \ln \frac{P(d_{ab}|F)}{P(d_{ab})} \tag{3.10}$$

The free energy for a given conformation, $\Delta G(C)$, is simply computed by summing up the individual free energies of all n pairs of atom types:

$$\Delta G(C) = \sum_n \Delta G_d^{ij} = -kT \cdot \sum_n \ln \frac{P(d_{ij}|F)}{P(d_{ij})} \tag{3.11}$$

Assuming the same reference state, the expansions for $P(d_{ij}|F)$ and $P(d_{ij})$ are the same as in equations (3.1) and (3.2) respectively. Thus, the above equation can be related to equation (3.8) in the following manner:

$$\ln P(F|\{d_{ij}\}) = c + \frac{\Delta G(C)}{-kT} \tag{3.12}$$

Ignoring the constants, c and $-kT$, we see from equation (3.12) that the conditional probability and the potential of mean force formalisms are functionally equivalent.

3.2.3 The residue-specific all-atom probability discriminatory function (RAPDF)

The conditional probabilities for the residue-specific all-atom probability discriminatory function (RAPDF) are compiled from frequencies of contacts between pairs of atom types in a database of protein structures. All non-hydrogen atoms are considered, and the description of the atoms is residue specific, i.e., the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. Contacts between atoms within a single residue are excluded from the counts. We divide the distances observed into 1.0 Å bins ranging from 3.0 Å to 20.0 Å. Contacts between pairs of atom types in the 0.0-3.0 Å range are placed in a separate bin, resulting in total of 18 distance bins. Table 3.1 lists the atom types used for this discriminatory function.

A table containing the negative log probabilities for all pairs of atom types for all distances is compiled from a database of known structures using equation (3.8). Given an amino acid sequence conformation, the probabilities of all contacts between pairs of atom types with distances that fall within the distance

C	C_α	C_β	C_δ	$C_{\delta1}$	$C_{\delta2}$	C_ϵ	$C_{\epsilon1}$	$C_{\epsilon2}$	$C_{\epsilon3}$	C_γ	$C_{\gamma1}$
$C_{\gamma2}$	CH_2	C_ζ	$C_{\zeta2}$	$C_{\zeta3}$	N	$N_{\delta1}$	$N_{\delta2}$	N_ϵ	$N_{\epsilon1}$	$N_{\epsilon2}$	NH_1
NH_2	N_ζ	O	$O_{\delta1}$	$O_{\delta2}$	$O_{\epsilon1}$	$O_{\epsilon2}$	O_γ	$O_{\gamma1}$	OH	S_δ	S_γ

Table 3.1: List of atom types used in the residue-specific all-atom probability discriminatory function (RAPDF). Each of these atom types is prefixed by the type of the residue (in one-letter code), resulting in 167 different atom types.

cutoff above are summed up to yield the total negative log conditional probability of the conformation being correct. This procedure is used for all probability discriminatory functions described in this chapter.

3.2.4 The residue-specific virtual-atom probability discriminatory function (RVPDF)

The residue-specific virtual-atom probability discriminatory function (RVPDF) uses a virtual atom approximation similar to the one developed by Head-Gordon and Brooks [116]. This representation combines a group of atoms into a single virtual atom type by averaging over the corresponding x, y, and z cartesian coordinates of the individual atoms. Aside from labelling conventions, this representation differs from the original representation in the determination of the virtual centres for virtual atoms vNH and vOH (which are taken to be represented by the positions of the N and O atoms respectively, rather than the geometric centres). The distance bins are the same as in the RAPDF. Each of the virtual atom types is prefixed by the type of the residue, resulting in 105 different virtual atom types. Table 3.2 lists the virtual atoms used for this discriminatory function and the combinations of atom types they represent.

Virtual atom	Components	Present in residue
vCO	$C + O$	all
vNH1	N	all
vNH2	$N_{\delta 2}$	N
vNH2	$N_{\epsilon 2}$	Q
vNH2	NH_1	R
vNH2	NH_2	R
vNH3	N_{ζ}	K
vNHE	$N_{\delta 1}$	H
vNHE	$N_{\epsilon 2}$	H
vNHE	N_{ϵ}	R
vNHE	$N_{\epsilon 1}$	W
vCOS	$C_{\gamma} + O_{\delta 1}$	N
vCOS	$C_{\delta} + O_{\epsilon 1}$	Q
vCSC	$C_{\gamma} + S_{\delta} + C_{\epsilon}$	M
vCCC	$C_{\beta} + C_{\gamma 1} + C_{\gamma 2}$	V
vCCC	$C_{\gamma} + C_{\delta 1} + C_{\delta 2}$	L
vC3R	$C_{\gamma} + C_{\delta 1} + C_{\epsilon 1}$	F
vC3R	$C_{\delta 2} + C_{\epsilon 2} + C_{\zeta}$	F
vC3R	$C_{\gamma} + C_{\delta 1} + C_{\epsilon 1}$	Y
vC3R	$C_{\delta 2} + C_{\epsilon 2} + C_{\zeta}$	Y
vC3R	$C_{\delta 2} + C_{\epsilon 3} + C_{\zeta 3}$	W
vC3R	$C_{\epsilon 2} + C_{\zeta 2} + CH_2$	W
vCOO	$C_{\gamma} + O_{\delta 1} + O_{\delta 2}$	D
vCOO	$C_{\delta} + O_{\epsilon 1} + O_{\epsilon 2}$	E
vOH	O_{γ}	S
vOH	$O_{\gamma 1}$	T
vOH	OH	Y
vCC	$C_{\beta} + C_{\gamma 2}$	I
vCC	$C_{\gamma 1} + C_{\delta 1}$	I
vCC	$C_{\beta} + C_{\gamma}$	K
vCC	$C_{\delta} + C_{\epsilon}$	K
vCC	$C_{\gamma} + C_{\delta}$	R
vCC	$C_{\beta} + C_{\gamma 2}$	T
vCCP	$C_{\beta} + C_{\gamma}$	P
vCCR	$C_{\gamma} + C_{\delta 1}$	W
vCCH	$C_{\gamma} + C_{\delta 2}$	H
vSH	S_{γ}	C
vCE1	$C_{\epsilon 1}$	H
vC	C_{ζ}	R
vCH3	C_{β}	A
vCH2	C_{α}	G
vCH2	C_{β}	C,D,E,F,H,L,M,N,Q,R,S,W,Y
vCH2	C_{γ}	E,Q
vCH2	C_{δ}	P
vCH	C_{α}	A,C,D,E,F,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y

Table 3.2: List of virtual atom types used in the residue-specific and non-residue-specific virtual-atom probability discriminatory functions (RVPDF and NVPDF, respectively). The table lists the virtual atom type, the atom types of the components, and the residues it is present in (in one-letter code). Aside from labelling conventions, this representation differs from the the one described in [116] in the determination of the virtual centres for virtual atoms vNH and vOH (which are taken to be represented by the positions of the N and O atoms respectively). For the RVPDF, each of the virtual atoms is prefixed by the type of the residue (in one-letter code), resulting in 105 different virtual atom types.

3.2.5 The non-residue-specific virtual-atom probability discriminatory function (NVPDF)

The non-residue-specific virtual-atom probability discriminatory function (NVPDF) differs from RVPDF only in that the virtual atom types are not residue-specific. For example, all vC_α atom types are considered the same, and all vC_β atom types are considered the same, and so on. The total number of virtual atom types considered under this approximation is 21.

3.2.6 The contact discriminatory function (CDF)

In the above three probability discriminatory functions (PDFs), the reference state is compiled from a database of native conformations, i.e., it is a compact reference state. It could be argued that the signal in these discriminatory functions arises from the fact that what the functions are really measuring is non-specific compactness and nothing more. That is, the discriminatory functions penalise conformations that are not as compact as an experimental conformation and are thus able to discriminate correct conformations from incorrect conformations. Even if this were not the case, it would be interesting to examine how well non-specific compactness alone can discriminate between correct and incorrect conformations relative to the three PDFs (RAPDF, RVPDF, and NVPDF) described above [115]. To assess this, we use a simple contact discriminatory function which assigns a negative log probability of -1.00 for every atom-atom contact within 6.0 Å in an amino acid sequence conformation, excluding contacts within a single residue.

3.2.7 The linearly interpolated residue-specific all-atom probability discriminatory function (IRAPDF)

The three PDFs described above (RAPDF, RVPDF, NVPDF) use discrete bins, to compile the conditional probabilities. This leads to a situation where, for a given distance which falls anywhere within the range of a distance bin, the negative log conditional probability of observing that distance between a pair of atom types is the same. For example, if we encountered a distance of 5.1, 5.5, or 5.9 Å for a particular pair of atom types, the magnitude of the probabilities will be identical as all those distances fall into one discrete bin with a range of 5.0-6.0 Å. In reality, preferences between atom types must vary in a continuous manner as the distances between the contacts vary. We thus evaluate the total negative log conditional probability of an amino acid sequence conformation by linearly interpolating between the negative log conditional probabilities for the discrete bins to precisely determine the specific negative log conditional probability for a given distance. To test the effect of using discrete distance bins, we evaluate all the decoy sets using the linearly interpolated PDF (IRAPDF) to see how well it performs compared to the RAPDF.

We determine the linearly interpolated conditional probability by assuming that the mid-point of the distance bins represents the actual negative log conditional probability for that distance, and that there is a linear relationship between probability values observed for neighbouring bins.

Thus, if d_a represents the actual distance encountered, and d_l represents mid-point of the closest distance bin value on the left hand side and d_r represents the mid-point of the closest distance bin value on the right hand side, then I_l represents the negative log conditional probability for d_l, and I_r represents the negative log conditional probability for d_r. To find I_a, the negative log conditional probability for d_a, we use the formula:

$$I_a = I_l + ((I_r - I_l) \cdot \frac{d_a - d_l}{d_r - d_l}) \tag{3.13}$$

3.2.8 Low counts analysis

We have investigated the effect of finite counts used to define the frequences required for a PDF (equation (3.8)), by introducing a simple model of count variation. For each set of counts obtained from observations of particular pairs of atom types in contact in the database in equations (3.1) and (3.2), we assume a Gaussian distribution for variation in counts in repeated experiments and modify the counts accordingly. The modified counts are obtained by the equation:

$$N' = (\sum_m R - 6.0) \cdot \sigma + N \tag{3.14}$$

Where N is the observed count value, N' is the modified count value, $\sum_m R$ represents the sum of m random numbers in the interval [0,1] ($m = 12$), and σ represents the standard deviation, assumed to be $N^{1/2}$.

For each pair of atom types examined, we generate 100 different sets of four counts for the 18 distance bins using the above procedure, and compare the corresponding conditional probabilities to the observed conditional probabilities computed from the counts obtained from the database.

We compare contacts between two pairs of atom types, one where the counts in equations (3.1) and (3.2) represent an average situation, and another where the counts are among the lowest encountered in the database, to assess the effect of sparse data on the conditional probabilities.

3.2.9 Construction of the structure library for obtaining conditional probabilities

Table 3.3 lists the PDB codes of the 265 structures that were used for compiling the conditional probabilities for the discriminatory functions described above. The PDB codes were initially obtained from the CATH database and are a set of non-homologous (less than 30% sequence identity between any proteins in the set) high-resolution (less than 3.0 Å) x-ray structures [117]. Structures with multiple side chain conformations have been modified such that only the side chains conformation with atoms having the highest occupancy and lowest temperature factors is used.

3.2.10 Decoy set generation

The decoy sets used were obtained from the Protein Potential Site (PROSTAR) [118] and can be divided into two classes. Decoy sets in class I discriminate between one correct and one or more incorrect or approximate conformations. Decoys sets in class II are a set of approximate conformations that vary in RMSD to the experimental conformation, excluding the experimental conformation itself.

Table 3.4 lists the decoy sets in class I. The MISFOLD decoy set, generated by Holm and Sander [119], consists of 25 examples of pairs of proteins with the same number of residues in the chain, but different conformations. Sequences were swapped between two different conformations, and side chain packing annealed using a Monte Carlo process [119]. These provide inappropriate environments for most of the side chains in the structures.

135l
1aaf
1aak
1aba
1aco
1acp
1add
1adn
1ads
1ak3A
1ala
1alkA
1aozA
1apa
1apmE
1aps
1arb
1arqA
1atnA
1atr
1atx
1ayh
1bal
1bbpA
1bbt1
1bbt2
1bbt3
1bbt4
1bds
1bgc1
1bgc2
1bgh
1bha
1bia
1bllE
1bmv1
1bmv2
1brnL
1btc
1bw3
1c5a
1cauA
1cdb
1cde
1cdg
1cewI
1cmbA
1cobA
1colA
1coy
1cpcA
1cpt
1csc
1cseE
1cseI
1ctf
1d66A
1dhr
1dmb
1eca
1ede
1egf
1etrL
1ezm
1fbaA
1fc2D
1fiaA
1fkb
1fnr
1fus
1fxd
1gal
1gatA
1gd1O
1gdhA
1gky
1glaG
1glt
1gluA
1gof
1gox
1gpb
1gpr
1hcc
1hgeA
1hgeB
1hleA
1hmy
1hoe
1hsbA
1hstA
1huw
1hyp
1ifc
1ipd
1isuA
1kst
1lab
1lct
1lfi
1lis
1lla
1lmb3
1ltsA
1ltsC
1lyaA
1mat
1mfaH
1mfaL
1minA
1minB
1mypA
1mypC
1nar
1nipA
1noa
1nrcA
1nrd
1nscA
1ofv
1oma
1omf
1ovb
1pba
1pda
1pdc
1pfkA
1pgd
1pgx
1pha
1phh
1pii
1pkp
1plc
1poa
1poxA
1ppn
1prcC
1prcH
1prcL
1ptf
1pyaA
1pyaB
1pyp
1raiA
1raiB
1rcb
1rec
1rfbA
1rhd
1ribA
1rip
1rro
1rveA
1sbp
1shaA
1shg
1sim
1sryA
1stp
1tabI
1ten
1tfi
1tgl
1thg
1tml
1tnfA
1tplA
1tpm
1ttaA
1ttf
1ula
1utg
1vil
1vsgA
1wsyA
1wsyB
1xis
1ycc
1ysaC
1zaaC
256bA
2aaiB
2bbkH
2bbkL
2bopA
2bpa1
2bpa2
2cas
2cba
2cdv
2cmd
2cpl
2ctc
2ctvA
2cyp
2dnjA
2er7E
2gstA
2hhmA
2hpdA
2ltnA
2ltnB
2mev4
2mnr
2msbA
2nckL
2ohxA
2ovo
2pia
2plv4
2pmgA
2polA
2reb
2rhe
2rn2
2sicI
2sn3
2sns
2stv
2tgi
2tmdA
2tmvP
2ts1
2tscA
2yhx
3b5c
3bcl
3blm
3cla
3cts
3dfr
3ebx
3ecaA
3gapA
3grs
3il8A
3mdsA
3monA
3monB
3pgk
3pgm
3rubS
3sc2A
3sc2B
4enl
4fgf
4gcr
4htcI
4mt2
4sbvA
4sgbI
5fd1
5p21
5pti
5rubA
5timA
6insE
7aatA
7catA
7rsa
8abp
8fabB
8rxnA
9wgaA

Table 3.3: List of PDB codes of the 265 protein chains used for compilation of conditional probabilities. In cases where a single chain of the protein is used for the compilation, the chain identifier is shown.

The first conference on the critical assessment of protein structure prediction methods (CASP1) produced a set of 42 comparative models of six different proteins [8]. These form the CASP1 decoy set. The models vary in C_α RMSD to the corresponding experimental conformation, ranging from 0.53 Å to 7.40 Å, depending on the difficulty of the model building process.

The IFU decoy set is a set of 44 peptides which are proposed to be independent folding units as determined by local hydrophobic burial and experimental evidence [97]. The set consists of the structure of the peptides as observed in the complete experimental protein structure, and a conformation of the fragment generated with a Genetic Algorithm and a physics-based potential of mean force [120, 121].

The PDBERR decoy set consists of structures determined using x-ray crystallography which where later found to contain errors, and the corresponding corrected experimental conformations [118].

The SGPA decoy set consists of two conformations generated by molecular dynamics simulations starting with the *S. griseus* Protease A experimental structure (PDB code 2sga) [122], and the 2sga experimental structure.

Among all the decoy sets referenced in this chapter, only the LOOP decoy set belongs to class II. This decoy set consists of sets of conformations for short loops (four or five residues) that were systematically generated using the methods in [55, 59]. Table 3.5 gives details about each of the loops in the LOOP decoy set.

3.2.11 Decoy set evaluation

For class I decoys sets, the ratio of the negative log conditional probabilities of the incorrect conformation and the correct conformation is determined. A

Decoy set name	Number of decoys	C_α RMSD range (Å)	Reference
MISFOLD	25	8.66 - 22.43	[51, 118]
CASP1	42	0.53 - 7.40	[8, 118]
IFU	44	0.21 - 10.02	[97, 118, 121, 120]
PDBERR	3	0.81 - 13.21	[118]
SGPA	2	1.91 - 2.06	[118, 122]

Table 3.4: Class I decoys. The name used to identify the specific decoy set, the number of decoys in the set, the C_α RMSD range of the decoys to the experimental structure, and the appropriate references are given.

#	Protein name	Residue range	Sequence	Number of conformations	All-atom RMSD range (Å)
1	3dfr	20-23	PWHL	394	0.75 - 4.58
2	3dfr	27-30	LHYF	1390	0.81 - 3.47
3	3dfr	64-68	HQED	71439	0.89 - 4.19
4	3dfr	120-124	GSFEG	474	0.57 - 2.91
5	3dfr	136-139	FTKV	10782	1.39 - 2.15
6	2sga	35-39	TNISA	15453	1.20 - 3.17
7	2sga	97-101	GSTTG	2079	0.60 - 3.34
8	2sga	116-119	YGSS	26572	0.47 - 4.91
9	2sga	132-136	AQPGD	206	0.97 - 2.58
10	2fbj	265-269	HPDSG	393	0.96 - 3.90
11	2hfl	264-268	LPGSG	339	1.11 - 2.81

Table 3.5: Class II decoys. The LOOP decoy set is a set of loop conformations that were systematically generated using the methods in [59, 55]. The name of the protein from which the loop was taken, the range of the loop residues, the sequence of the loop, the number of conformations, and the all-atom RMSD range of the conformations is given. Further details of this decoy set are given in [118].

discrimination ratio less than 1.0 (or log discrimination ratio less than 0.0) indicates that the discriminatory function is able to correctly discriminate between the correct conformation and the incorrect one. The lower the log discrimination ratio, the more reliable the discrimination.

For class II decoy sets, the all-atom RMSD of the conformation with the lowest negative log conditional probability among all the conformations is determined and is compared to the other RMSDs. The probability of choosing a conformation with an equal or lower RMSD than the one selected by the discriminatory function by chance is equal to the number of conformations that have a lower RMSD divided by the the total number of conformations. Accurate discrimination is defined to be the selection of a conformation with an all-atom RMSD within 1.0 Å of the lowest RMSD conformation present in the decoy set.

When appropriate, the percentage of decoys correctly discriminated is also used to evaluate the performance of a discriminatory function on a decoy set. Further details on the evaluation protocols and the decoy set generation are given in [118].

For each decoy set evaluation, structures in the decoy set with the same PDB codes were removed from the structure library and the probabilities were recalculated, i.e., the procedure was jack-knifed or properly cross-validated to ensure that information about a protein was not pre-included in the conditional probability tables.

3.3 Results

3.3.1 The all-atom discriminatory function performs the best across a wide variety of decoys

An ideal discriminatory function is one that correctly discriminates 100% of class I decoys and selects conformations with low all-atom RMSDs (within 1.0 Å of the conformation with the lowest RMSDs) in the LOOP decoy set. In addition, the average discrimination ratios, which indicate the difference on average between the negative log conditional probability for the correct and incorrect conformations in the decoy set, must be statistically significant.

The RAPDF comes close to achieving this goal, particularly in comparison to the RVPDF and NVPDF. Figure 3.1a and Figure 3.1b show that the RAPDF has the best average discrimination ratio and the largest percentage of decoys correctly discriminated across a range of decoy sets. In the case of the MISFOLD, PDBERR, and SGA decoy sets, it correctly discriminates 100% of the decoys in the set. Further, the average discrimination ratios show that the negative log conditional probabilities for the correct conformations in the MISFOLD, PDBERR, and SGPA decoy set are on average lower (better) by 60%, 50% and 25% respectively, compared to the negative log conditional probabilties for the incorrect conformations.

In the case of the CASP1 decoy set, the percentage of decoys correctly discriminated by the RAPDF is 93%. The RAPDF performs slightly worse in terms of the average discrimination ratio than two other discriminatory functions in two specific instances (Figure 3.2, under nm23 and hpr), but this is due to the fact that the approximate conformations in these cases are very close to the experimental conformation, and the all-atom discriminatory function overwhelmingly identifies these approximate conformations as being better than the experimental one, thus skewing the average values. The difference, on average, between the negative log conditional probabilities for the correct conformations and the incorrect conformations for the RAPDF is 15%.

For the IFU decoy set, the percentage of conformations correctly discriminated by the RAPDF is 73%. The difference, on average, between the negative log conditional probabilities for the correct conformations and the incorrect conformations for the RAPDF is 10%.

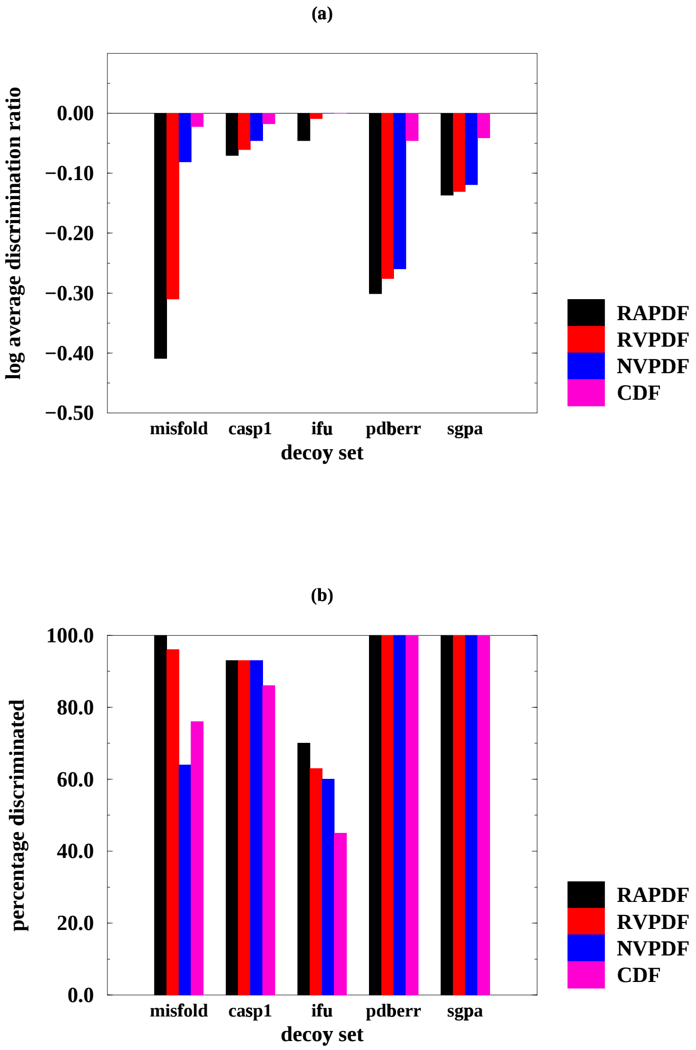


Figure 3.1: Comparison of the performances of the residue-specific all-atom probability discriminatory function (RAPDF), the residue-specific virtual-atom probability discriminatory function (RVPDF), the non-residue-specific virtual-atom probability discriminatory function (NVPDF), and the contact discriminatory function (CDF) for class I decoy sets. The log of the average discrimination ratios between incorrect and correct conformations (a) and percentage of decoys correctly discriminated (b) for the five decoy sets in class I is shown. The lower the log average discrimination ratio in (a), the better the discrimination. (b) shows the percentage of decoys that were accurately discriminated within a decoy set.

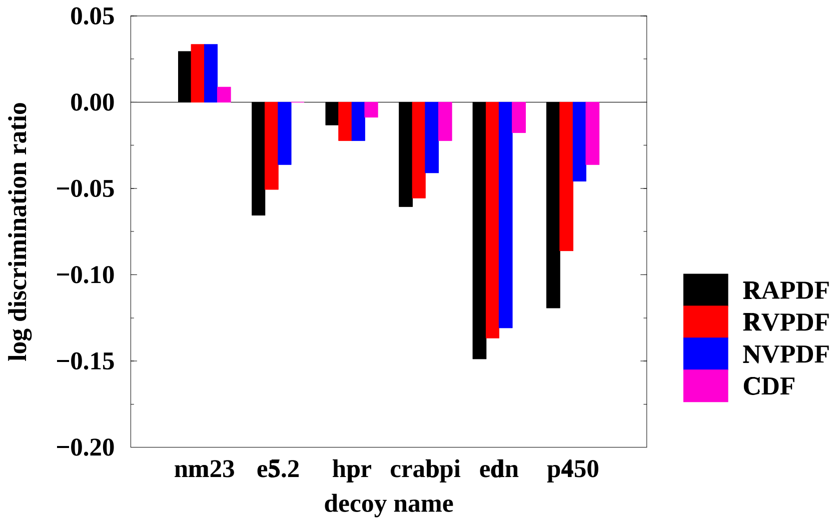


Figure 3.2: Comparison of the performances of the five discriminatory functions (IRAPDF, RAPDF,RVPDF, NVPDF, and CDF) for selected decoys in the CASP1 set. The log discrimination ratios between the experimental conformation and the model is shown. The model with the lowest C_α RMSD to the corresponding experimental conformation is chosen from a given set of models for this evaluation. The identifiers used to label the decoys are the same as in [8].

It is less obvious which discriminatory functions perform best from the log probability and all-atom RMSD data for the LOOP decoy set (Figure 3.3). However, if we examine the actual RMSD values, we note that for 10/11 loops (93%), the RAPDF picks a conformation that is within 1.0 Å of the lowest RMSD conformation in the sample space. RVPDF, NVPDF, and CDF have ratios of 5/11, 9/11, and 8/11 respectively.

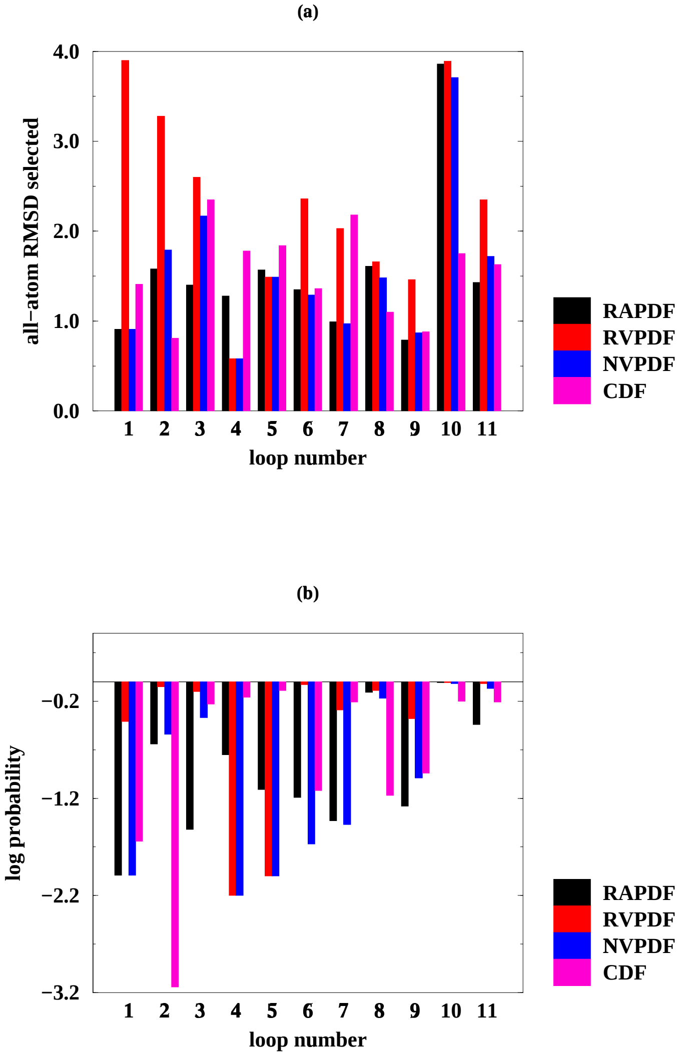


Figure 3.3: Comparison of the performances of the residue-specific all-atom probability discriminatory function (RAPDF), the residue-specific virtual-atom probability discriminatory function (RVPDF), the non-residue-specific virtual-atom probability discriminatory function (NVPDF), and the contact discriminatory function (CDF) for the LOOP decoy set. The all-atom RMSD (a) and log probabilities of observing an equal or lower RMSD by chance (b) is shown. The loop numbers in the horizontal axis corresponds to the numbers in Table 3.5, column 1.

3.3.2 Discriminatory power decreases upon successive approximations

There is generally a successive degradation of the signal going from the all-atom discriminatory function to the CDF, as the description gets more and more approximate (Figure 3.1). One major exception is the LOOP decoy set (Figure 3.3) where the NVPDF performs significantly better than the RVPDF. The other exception can be noticed by examining Figure 3.1b, where the CDF does better in terms of the percent correct discriminations than the NVPDF in the case of the MISFOLD decoy set.

3.3.3 The compactness term alone is useful for discriminating between correct and incorrect conformations

The contribution of the compactness term, which is measured by the CDF, is better than some of the other PDFs for certain decoy sets (Figure 3.1b under MISFOLD, and Figure 3.3). Further, in a majority of the decoy sets, it is adequate to distinguish between correct and incorrect conformations most of the time (Figure 3.1b under PDBERR and SGA, and Figure 3.3).

3.3.4 Using a large distance cutoff helps in discrimination

As shown by the plot of the percentage of decoys correctly discriminated for the decoy sets at different distance cutoffs (5.0 Å, 10.0 Å, 15.0 Å, and 20.0 Å) (Figure 3.4), there is a significant advantage overall to using a larger distance cutoff. A distance cutoff of at least 15.0 Å is necessary to accurately discriminate all the 25 decoys in the MISFOLD decoy set. In the cases of the PDBERR and

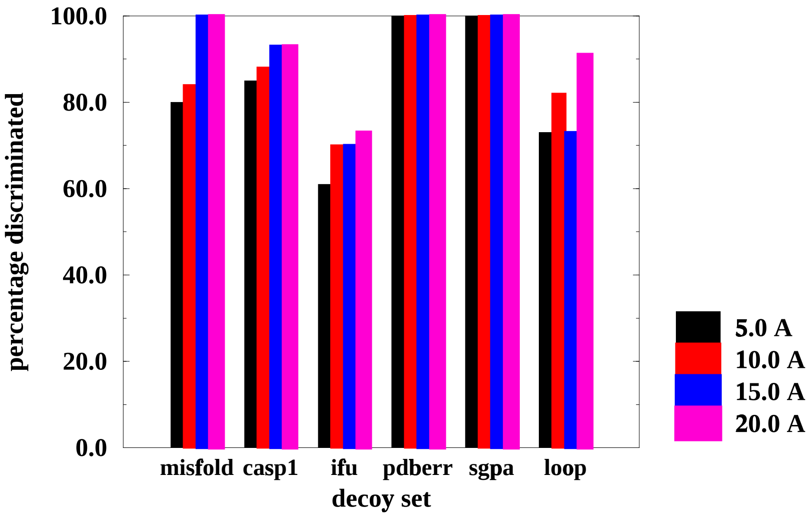


Figure 3.4: Comparison of the performance of the residue-specific all-atom probability discriminatory function (RAPDF) at different cutoffs. The percentages of structures correctly discriminated for six decoy sets at four different cutoffs is shown. In the case of the LOOP decoy set, "correct" discrimination is defined to be the selection of conformation that is within 1.0 Å of the lowest all-atom RMSD conformation for each loop.

SGPA decoy sets, it does not appear to make a difference which cutoff is chosen in terms of percentage of decoys correctly discriminated.

3.3.5 Comparison of the contribution of electrostatics and non-electrostatics terms

To determine the nature of the signal in the RAPDF, we partition the discriminatory function according to contributions from electrostatic and non-electrostatic contacts. Any four possible combinations of N and O atoms are defined to be electrostatic in nature. All other contacts are considered non-electrostatic. Figure 3.5 compares the accuracy (by measuring the percentage of conformations

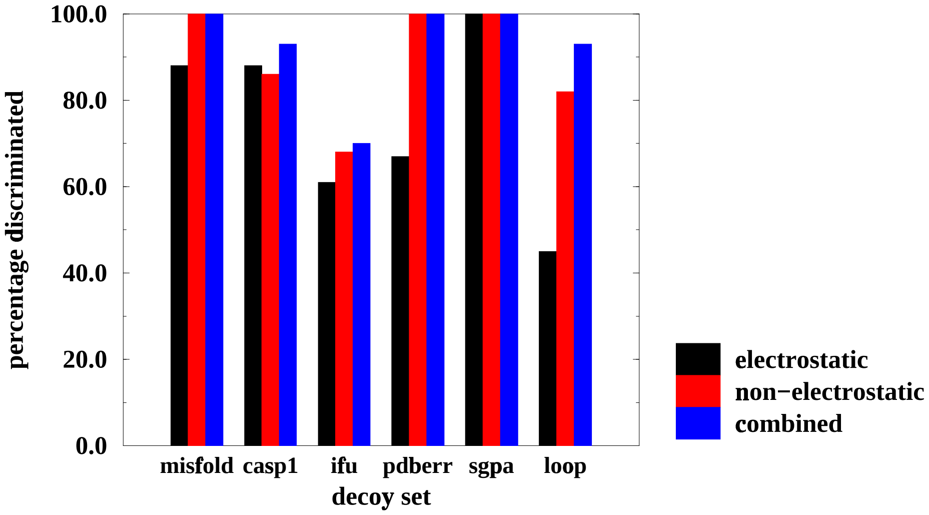


Figure 3.5: Comparison of electrostatic, non-electrostatic, and combined terms in the residue-specific all-atom probability discriminatory function (RAPDF). The percentages of structures correctly discriminated for various decoy sets is shown. In the case of the LOOP decoy set, "correct" discrimination is defined to be the selection of conformation that is within 1.0 Å of the lowest all-atom RMSD conformation for each loop.

correctly discriminated) of using only the electrostatics and non-electrostatics terms, relative to the combined PDF. Even though the non-electrostatic terms alone are adequate for correct discrimination in most cases, the electrostatic terms play a significant role in enhancing the signal. This is particularly noticeable in the CASP1, IFU, and LOOP decoy sets where the combined signal leads to discrimination of more decoys than the individual signals by themselves.

3.3.6 Linear interpolation improves discrimination

As shown in Figure 3.6, comparison between the IRAPDF and the RAPDF shows that linear interpolation helps discrimination between correct and incor-

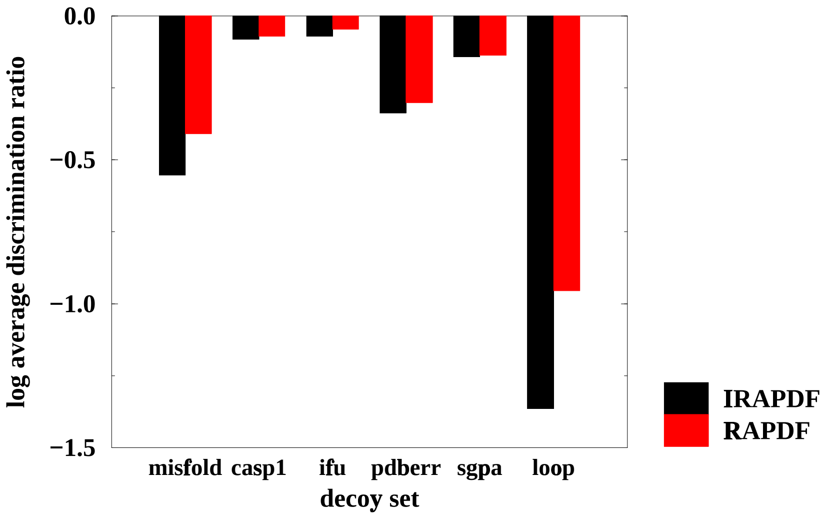


Figure 3.6: Comparison of the residue-specific all-atom probability discriminatory function (RAPDF) to the linearly-interpolated version of the RAPDF (IRAPDF). For five of the decoy sets, the bars represent the log average of the ratio of the probabilities between the correct and incorrect structures. For the LOOP decoy set, the bars represent the log average of the probabilities of finding at least one structure with a lower all-atom RMSD than the one with the best discrimination by chance (i.e., the sum of log probabilities divided by the number of loops).

rect conformations. This is most obvious in the LOOP decoy sets where the improvement is quite dramatic, but for each decoy set there is some positive improvement upon using the IRAPDF to evaluate the conformations.

3.3.7 The problem of sparse data for compilation of probabilities is negligible

There are generally two sorts of problems in knowledge-based discriminatory functions that arise from inadequate data that lead to errors in the conditional probabilities. These can be illustrated best by examining the expression for the

Term	Minimum counts	Average counts
$N(d_{ij})$	1	648
$\sum_d N(d_{ij})$	1525	11,708
$\sum_{ij} N(d_{ij}$	1,011,903	9,143,295
$\sum_d \sum_{ij} N(d_{ij})$	164,980,971	164,980,971

Table 3.6: Details of the raw counts obtained when compiling the conditional probabilities (see the METHODS section for more detail on the compilation process). For each term in equation (3.15), the minimum counts and the average counts are given. $\sum_d \sum_{ij} N(d_{ij})$, the denominator in the expression for $P(d_{ij})$ is always a constant for any combination of i and j.

probability of seeing two atom types, i and j, in contact in distance bin d in a correct conformation, $P(d_{ij}|F)$:

$$P(d_{ij}|F) = \frac{P(d_{ij}|F)}{P(d_{ij})} = \frac{N(d_{ij})/\sum_d N(d_{ij})}{\sum_{ij} N(d_{ij})/\sum_d \sum_{ij} N(d_{ij})} \tag{3.15}$$

Detailed explanations for these terms is given in the METHODS section. To begin our analysis, let us examine Table 3.6 for the nature of the raw counts that we encounter in our observations for each of the four terms in the above expression. Typically if the numerator in equation (3.15) has low counts for both its terms, one can have significant differences in the probabilities due to minor statistical fluctuations. For example, $P(d_{ij}|F)$, in such a situation, could represent 1/2 or 2/2, and the difference in the probabilities is a factor of 2. This situation never arises in our formalism because three of the four terms in the equation (3.15) have large counts relative to the $N(d_{ij})$ term, as shown in Table 3.6, column 2. This is because we do not partition our counts based on the sequence separation and the directionality of the polypeptide chain [101, 112].

However, there could be problems due to errors in the counts of atom types i and j in a particular distance bin d. We analyse two atom-atom preferences, cysteine N-tryptophan O (CN-WO) which represents a minimum counts situation, and isoleucine C_{α}-leucine $C_{\delta 2}$ (IC_{α}-$LC_{\delta 2}$) which represent an average counts situation. These pairs of atoms were selected for analysis based on the counts in Table 3.6.

Figure 3.7 compares the effect of uncertain count values (generated using a weighted random number generator, described in the METHODS section) on the conditional probabilities for the preferences between two pairs of atom types. We can see from the two plots that there is significantly more error in the conditional probabilities for the worst case (CN-WO) than for the average case (IC_{α}-$LC_{\delta 2}$). In the average case, the absolute error in the negative log conditional probability when varying the individual counts randomly is typically less than 0.1 and has a maximum value of about 1.0 (in the 0-3 Å distance bin). In the worst case, the absolute error in the negative log conditional probability is typically less than 0.5 and has a maximum value of about 1.0.

To compute the negative log conditional probability of a conformation, we sum over a large number of probabilities at a large number (18) distance cutoffs (up to 20.0 Å). The error due to sparse data in the worst case situations is mitigated by this summation of terms.

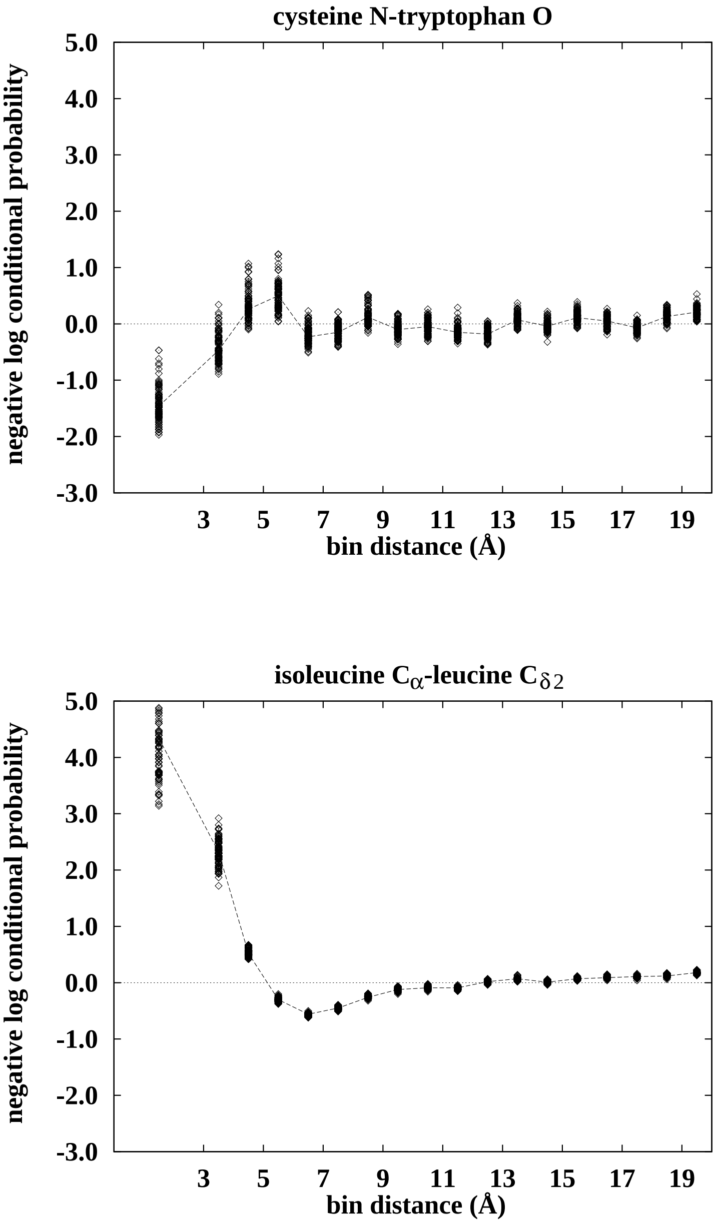


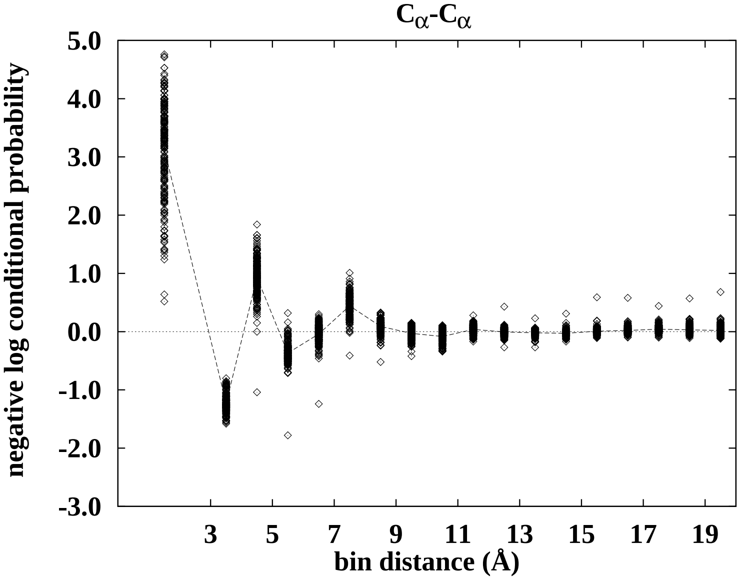
Figure 3.7: Comparison of the effect of counting uncertainties on the conditional probabilities for two pairs of atom types, cysteine N-tryptophan W (CN-WO), where the counts are among the lowest in the database, and isoleucine C_α-leucine $C_{\delta 2}$ (IC_α-$LC_{\delta 2}$), where the counts are similar to the average counts in Table 3.6. The dashed line connects observed conditional probabilities and the points around the dashed line represent the variation in the conditional probabilities due to the uncertainty in the counts (see the METHODS section for the generation of variation in the counts).

3.3.8 Relationship between the conditional probabilities and the nature of physical interactions in proteins

To examine the relationship between the conditional probabilities and the energetics of atom-atom preferences in proteins, we select a set of atom pairs and plot the RAPDF conditional probabilities for the 18 distance bins. The atom pair probabilities are shown for all C_α-C_α and C_β-C_β pairs (Figure 3.8), all main chain nitrogen-main chain oxygen (N-O) pairs and all main chain nitrogens to the $O_{\delta 1}$ atom in aspartic acid (N-$DO_{\delta 1}$) pairs (Figure 3.9), alanine C_α-alanine C_α (AC_α-AC_α), and valine C_α-valine C_α (VC_α-VC_α) pairs (Figure 3.10), and aspartate N-lysine O (DN-KO) and proline N-tryptophan O (PN-WO) pairs (Figure 3.11).

Comparing the C_α-C_α to the C_β-C_β curve (Figure 3.8) shows that the two pronounced minima for C_α-C_α atom-atom contacts are in the 3.0-4.0 Å and 5.0-6.0 Å distance bins, whereas the minimum for C_β-C_β curve is in the 5.0-6.0 Å distance bin.

The C_α-C_α minimum in the 3.0-4.0 Å bin is due to the presence of C_α-C_α contacts between $i, i+1$ (neighbouring) residues. The C_α-C_α minimum in the 5.0-6.0 Å bin is due to the presence of C_α-C_α contacts between $i, i+2$ and $i, i+3$ residues in alpha-helices. The C_β-C_β minimum in the 5.0-6.0 Å occurs as a result of the presence of C_β-C_β contacts between $i, i+1$ residues in both α-helices and β-sheets. These observations are supported by counting the number of C_α-C_α and C_β-C_β contacts in helices and sheets in a set of 100 proteins selected from Table 3.3, taking into account the sequence separation, which shows that largest counts fall into the bins with the minima. Likewise, the slight rise in the negative log conditional probabilities in the 7.0-8.0 Å bin in the C_α-C_α plot and the slight



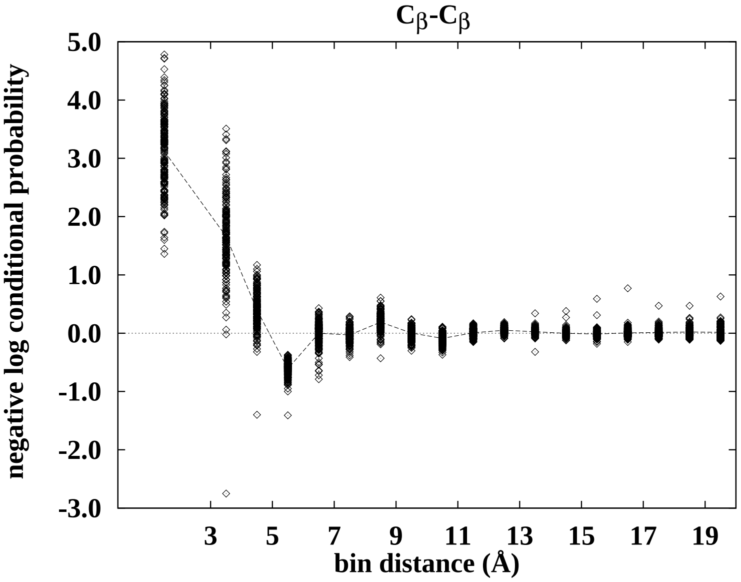
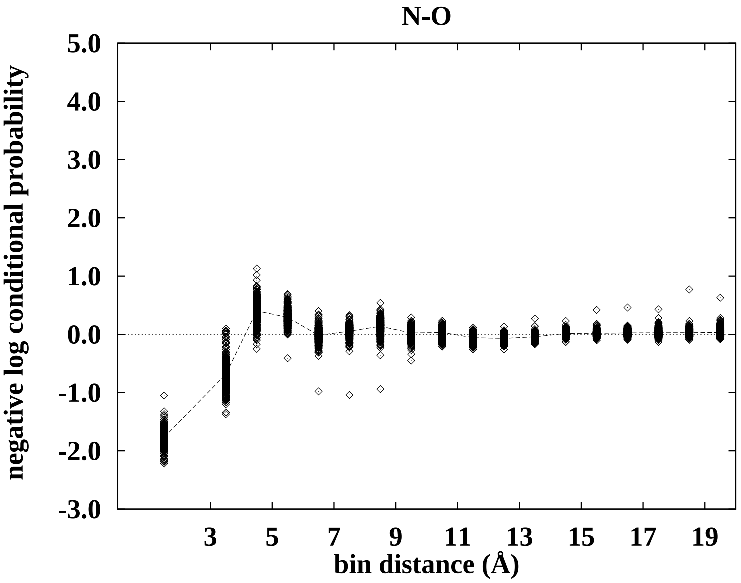


Figure 3.8: Plot illustrating the conditional probabilities encountered in the 18 distance bins for all C_α-C_α contacts and all C_β-C_β contacts. For each pair of atom types, the negative log conditional probabilities are plotted against the 18 distance bins. The spread at a given distance bin illustrates the differences in probabilities for the various atom types within that bin. The average of the negative log conditional probabilities for each bin is connected by the dashed line.



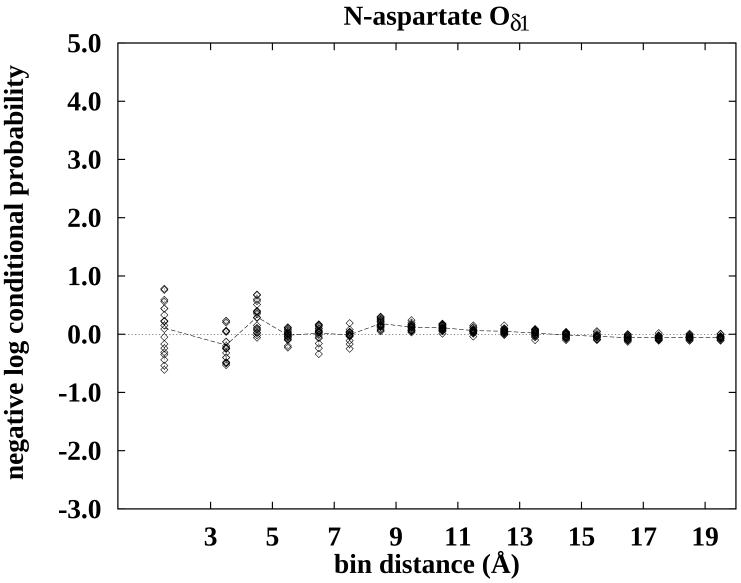
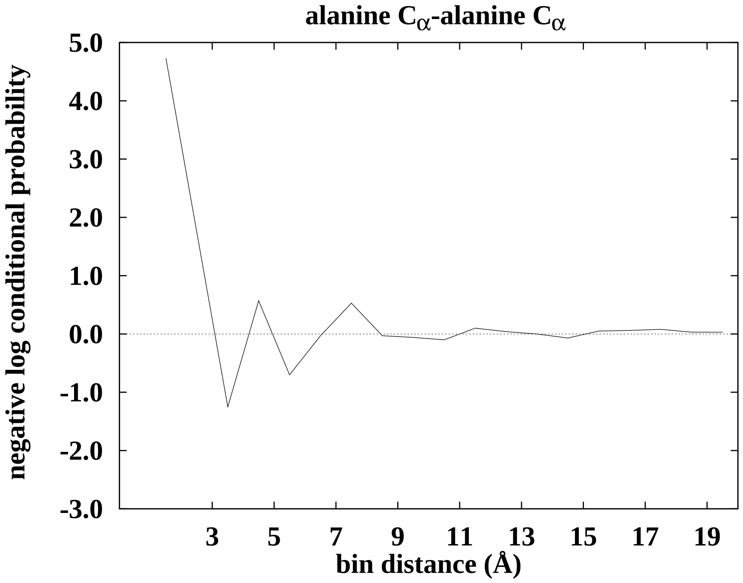


Figure 3.9: Plot illustrating the conditional probabilities encountered in the 18 distance bins for all N-O contacts and contacts between all main chain nitrogens and aspartic acid $O_{\delta 1}$. For each pair of atom types, the negative log conditional probabilities are plotted against the 18 distance bins. The spread at a given distance bin illustrates the differences in probabilities for the various atom types within that bin. The average of the negative log conditional probabilities for each bin is connected by the dashed line.



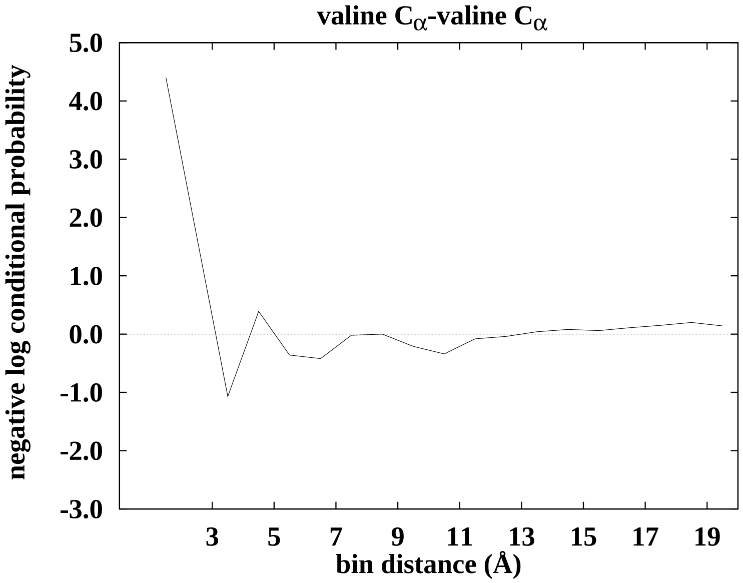
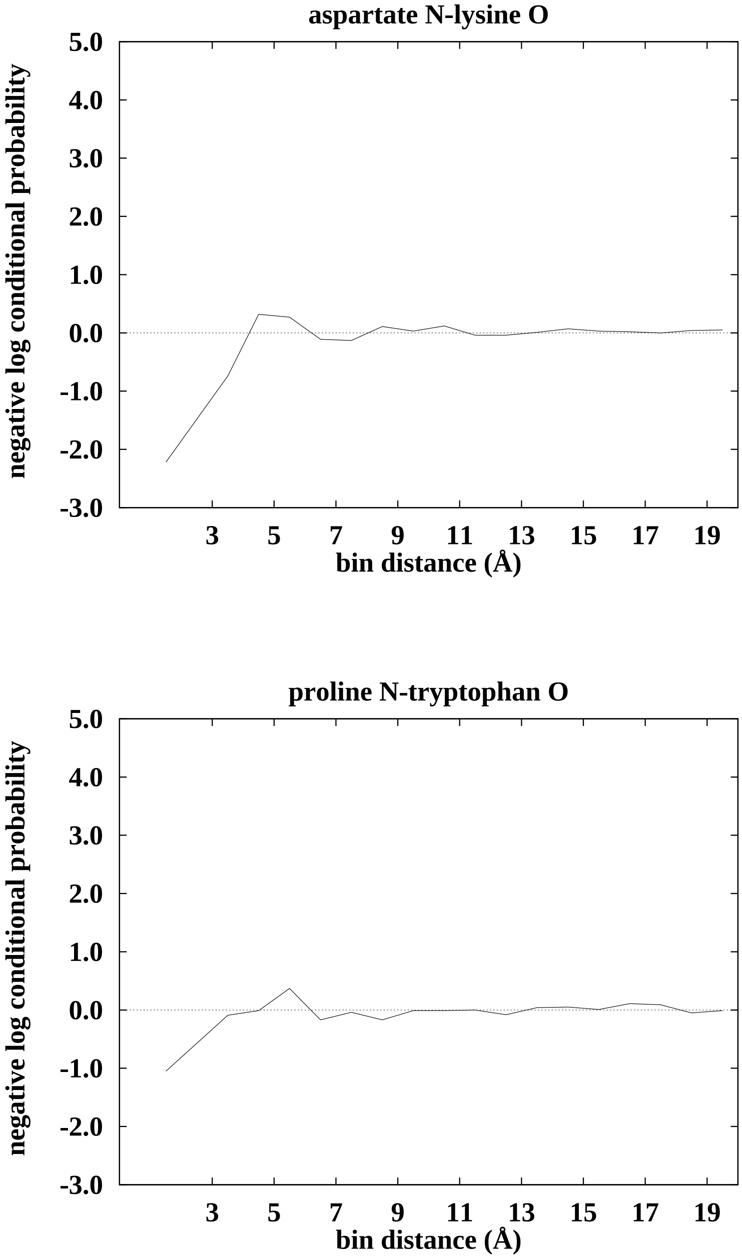


Figure 3.10: Plots illustrating the conditional probabilities encountered in the 18 distance bins for alanine C_α-alanine C_α (AC_α-AC_α) and valine C_α-valine C_α (VC_α-VC_α) contacts. The negative log conditional probabilities are plotted against the 18 distance bins.



aspartate N-lysine O

proline N-tryptophan O

Figure 3.11: Plots illustrating the conditional probabilities encountered in the 18 distance bins for apartate N-lysine O and proline N-tryptophan O contacts. The negative log conditional probabilities are plotted against the 18 distance bins.

rise in the negative log conditional probabilities in the 8.0-9.0 Å bin in the C_β-C_β plot is because very few contacts occur in these distance ranges, particularly in helices.

The large spread in the C_α-C_α and C_β-C_β plots (Figure 3.8) in the 0.0-3.0 Å distance bin is observed not because of variation in the number of contacts between the atom types, but because of the variation in the counts when the total over all distances for that pair of atom types is considered. That is, the value $\sum_d N(d_{ij})$ in equation (3.15) is generally the only term that varies the most during the computation of probabilities for this distance range. The denominator in equation (3.15), corresponding to $P(d_{ij})$, is fixed for a given distance bin and a pair of atom types. The term $N(d_{ij})$ is initialised to 1 for all values of i and j before compilation of probabilities, and this is generally not incremented *except in the case of pairs of residues forming cis-pepides* where the C_α distance between $i, i+1$ residues is below 3.0 Å. This is clearly reflected in the C_α-C_α plot: one of the largest negative log conditional probabilities (of 4.7) is observed for AC_α-AC_α contacts in the 0.0-3.0 Å bin, reflecting the tendency of alanine residues to occur very frequently in proteins (the value for $N(d_{i,j})$, where $d < 3.0$ Å and i and j are alanine C_α atoms, is 1 in the computation of this probability). The smallest negative log conditional probability (of 0.57) is observed for PC_α-PC_α contacts in the 0.0-3.0 AA bin, reflecting the tendency of proline residues to be found in the *cis* conformation (the value for $N(d_{i,j})$, where $d < 3.0$ Å and i and j are proline C_α atoms, is 17 in the computation of this probability).

The extreme outliers in the N-O, C_α-C_α, C_β-C_β plots in Figures 3.8 and 3.9 represent contacts between pairs of atom types in cysteine residues. These probably reflect the tendency of cysteine residues to form disulphide bonds, thus constraining the choices of the other atom types. For example, in the C_β-C_β plot (Figure 3.9) in the 3.0-4.0 Å bin, the large negative log conditional probability for C_β-C_β contacts between two cysteine resides reflects the distance between C_β atom types of cysteines involved in disulphide bonds.

Comparing the N-O curve to the N-$DO_{\delta 1}$ curve (Figure 3.9), we see that it is main chain hydrogen bonding that is commonly observed, and there is a smaller preference for main chain to side chain hydrogen bonding. However, the spread of conditional probabilities is quite distinctive in the 0.0-3.0 Å bin in the N-$DO_{\delta 1}$ plot, clearly indicating differences in preferences for the main chain nitrogen of various residues to form hydrogen bonds with aspartate $O_{\delta 1}$. The highest negative log conditional probability observed is for proline nitrogen and aspartate $O_{\delta 1}$ contacts, and the lowest negative log conditional probability observed is for asparagine nitrogen and aspartate $O_{\delta 1}$ contacts. This illustrates the importance of separating the atoms based on the residue types.

The plots of AC_α-AC_α, and VC_α-VC_α contacts (Figure 3.10) illustrate the differences in preferences for alanine and valine residues to be in α-helix and β-sheet secondary structures. The preference for a pair of alanine C_α atoms to be within a distance bin of 5.0-6.0 Å is significantly greater than the preference for a pair of valine C_α atoms to be within the same distance bin. This reflects the preferences of alanine residues to occur frequently in α-helices (the distances between C_α atoms for $i, i+2$ and $i, i+3$ residues in α-helices generally fall in this range), whereas the minimum in the VC_α-VC_α plot in the 6.0-7.0 Å bin reflects valine preferences for β-sheets (the distances between C_α atoms in $i, i+2$ residues in β-sheets fall in this range). The slight minimum in the VC_α-VC_α plot in the 10.0-11.0 Å bin reflects C_α contacts between $i, i+3$ residues in β-sheets.

The DN-KO and PN-WO plots (Figure 3.11) represent the lowest and highest negative log conditional probabilities in the 0.0-3.0 Å distance bin in the plot of all N-O contacts (see Figure 3.9). Proline nitrogens generally have the highest negative log conditional probabilities for contacts with other oxygen atoms, reflecting the fact that proline nitrogens lack a hydrogen atom and are thus unable to form hydrogen bonds. DN-KO contacts have the lowest negative conditional probabilities, possibly reflecting $i, i+4$ salt bridges in α-helices and between opposite residues in β-sheets.

3.4 Discussion

3.4.1 Performance of the all-atom residue-specific probability discriminatory function

The all-atom residue-specific probability discriminatory function (RAPDF) selects the correct conformation 87% of time in cases where the decoy set pair consists of one correct conformation and one incorrect conformation (class I decoys). This suggests that the RAPDF can prove to be useful tool in model refinement, identifying the "best" conformation among a set of possibilities in a modelling situation (say, in a comparative modelling scenario where many models have been built and picking the correct model is necessary).

The RAPDF also selects a conformation that is within 1.0 Å of the lowest RMSD conformation for 93% of the loops in the LOOP decoy set (where the experimental structure is not included). This suggests that the signal in the discriminatory function is valid for a range of conformations of an amino acid sequence. Plotting the RMSD vs. negative log probability for a set of conformations from the LOOP decoy set (Figure 3.12) shows that the RAPDF produces a worse signal as the all-atom RMSD gets worse. This suggests that this discriminatory function can be useful in simulations that attempt to get closer to the native conformation starting from a distant conformation.

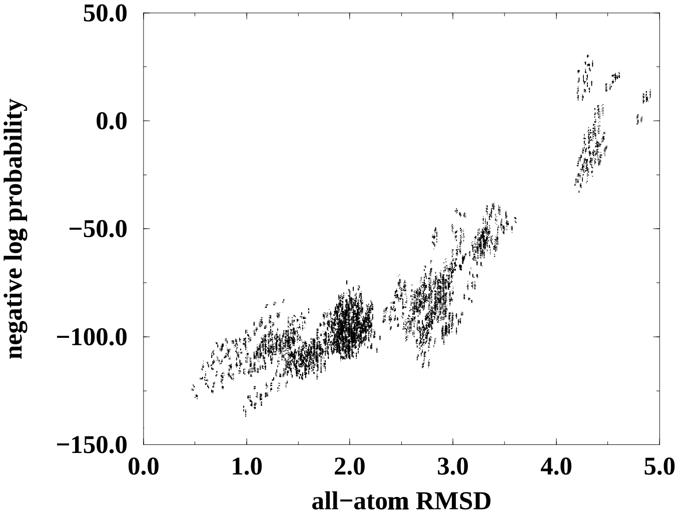


Figure 3.12: Performance of the residue-specific all-atom probability discriminatory function (RADPF) for a selected loop in the LOOP decoy set. The all-atom RMSD vs. the negative log conditional probability of the 26,572 conformations for the 2sga 116-119 LOOP set is shown. The plot shows how the negative log conditional probability increases as the RMSD progressively gets worse.

3.4.2 Effect of approximating the detail in the discriminatory function representation

The approximate discriminatory functions (RVPDF, NVPDF, CDF) are all able to discriminate between the correct and incorrect structures to some degree, but for the most accurate discrimination across a range of different decoys, an all-atom representation is necessary (Figures 3.1a and 3.1b).

3.4.3 Effect of the compactness term on predictive power

Particular types of signals are effective for distinguishing correct from incorrect conformations. The compactness term alone (measured by the CDF) is able to correctly distinguish correct from incorrect/approximate conformations quite often, which is striking in the case of the MISFOLD and the LOOP decoy sets (Figures 3.1b and 3.3). This suggests that some discriminatory functions might appear to be working reasonably well on certain types of decoys because they are measuring non-specific compactness. However, the compactness term performs poorly, when the average discrimination ratio of the negative log conditional probabilities of incorrect and correct conformations is considered for the various decoy sets, compared to the other PDFs (Figure 3.1a) that are parameterised on a compact reference state (i.e., the observations are made on a set of structures in the PDB). This illustrates the importance of taking specific atom-atom preferences into account, and the necessity of testing a discriminatory function on several different decoy sets to measure its effectiveness.

3.4.4 Effect of using a large distance cutoff

Using a 20.0 Å distance cutoff results in the most accurate discrimination for the RAPDF in comparison to smaller distance cutoffs (Figure 3.4). The signal from each atom-atom interaction is extremely weak at such large distances (see Figures 3.8, 3.9, 3.10, and 3.11), but each atom has a very large number of contacts, so that the combined signal still has an impact. It is unlikely that the energy of interaction is significant, except perhaps for between charged groups. However, the overall tendency of proteins to be organized “hydrophobic inside, hydrophilic outside” may result in a significant signal [123].

3.4.5 Contributions of electrostatics and non-electrostatics terms

Comparing the contributions of the electrostatic and non-electrostatic terms (Figure 3.5), we see that electrostatic terms alone are inadequate for the largest percentage discrimination of correct from incorrect conformations in the decoy sets, but contribute significantly to the RAPDF's ability to discriminate between correct and incorrect or approximate conformations in the the CASP1, IFU, and LOOP decoy sets.

3.4.6 Effect of linear interpolation and the problem of sparse data

We constructed the discriminatory functions described here using the simplest possible models. A more sophisticated model would take into account effects of low counts in the computation of the frequencies and would also perform some sort of "smoothing", or interpolation, between the discrete conditional probabilities obtained using the simple model.

Due to the fact that we do not partition the observed counts based on the sequence separation and the directionality of the polypeptide chain and the fact that we sum over a large number of probabilities and a large number (18) of distance cutoffs, the problem of low counts is negligible, as demonstrated in Figure 3.7.

We note that linear interpolation of the conditional probabilities in the RAPDF results in better discrimination for these decoy sets (Figure 3.6). This leads to the possibility that the discrete points for each distance bin can be represented by a continuous function and used in a protein folding simulation technique that requires a continuous differentiable function.

3.4.7 Effect of artifacts in the decoy sets

In the case of certain decoys sets, a discriminatory function may be able to select the correct conformation due to subtle differences of the incorrect conformations in a decoy set which distinguish it from an experimentally determined structure. For example, refinement of structures determined using x-ray crystallography is usually done with programs (like X-PLOR) which may restrain particular distances for atom-atom interactions, such as hydrogen bonds, seen in proteins. Since the PDFs are parameterised on high resolution x-ray crystallography structures, it is important to demonstrate that discrimination of correct conformations from incorrect ones is not achieved due to subtle details (i.e., that result highly precise interatomic distances) in the experimental structures.

The results from the LOOP decoy set (Figures 3.3 and 3.12) show that this is not the case for that particular decoy set, as the criteria for correctness depends on selecting a low RMSD conformation to the experimental structure. In the case of PDBERR decoy set, which consists of structures determined using x-ray crystallography which where later found to contain errors and the corresponding corrected experimental conformations, this is unlikely as both the correct and incorrect conformations were refined using similar refinement procedures.

To test whether such an artifact is responsible for accurate discrimination in the CASP1 decoy set consisting of homology models and their corresponding experimental structures, we took two models and the corresponding experimental structure and energy minimised them with 1000 steps of steepest descent using

INSIGHT/DISCOVER [75]. The discrimination ratio of negative log conditional probabilities between the unminimised model and the unminimised experimental structure is 0.82 and 0.87 for the two cases. The discrimination ratio of the minimised model to the minimised experimental structure is 0.78 and 0.83 for the two cases. The difference in the discrimination ratios in both cases is less than 0.1 even though there is a slight decrease in the negative log conditional probabilities in the minimised forms for both the approximate and experimental conformations. While this is not an exhaustive test, it suggests that the signal that separates correct and incorrect conformations is not due to fine details in the experimental structures in both this decoy set and the SGPA decoy set, where approximate conformations were produced by molecular dynamics simulations of the *S. griseus* Protease A experimental structure.

In the case of the MISFOLD decoy set, the main chains of both the correct and incorrect conformations are from structures determined using x-ray crystallography. But the percentages of structures correctly discriminated using non-specific compactness, measured by the CDF (Figure 3.1a) indicates that the side chain packing in the incorrect conformations does not generate as many atom-atom contacts within 6.0 Å as would normally be observed in an experimental conformation for that sequence. Thus, the signal in this case may be partially due to fine detailed differences between correct and incorrect conformations in a decoy set. This underscores the importance of testing a discriminatory function on a variety of decoy sets.

3.4.8 Limits on the resolution of the discriminatory functions

There appear to be definite limits as to what the discrimination functions can achieve. For example, in the CASP1 decoy set (Figure 3.2), the RAPDF in one case (nm23) fails to discriminate between the correct and approximate conformations, and in another case (hpr) performs worse than than RVPDF and NVPDF. In both these cases, the approximate conformations are close (within 0.53 Å and 1.05 Å C_α RMSD respectively) to the experimental structure. This suggests that the discriminatory function is unable to discriminate accurately when the conformations are close (around 1.0 Å C_α RMSD) to the experimental conformation. However, for purposes such as building comparative models that rival experimental nuclear magnetic resonance (NMR) methods, this level of accuracy is adequate.

3.4.9 Effect of experimental accuracy

In the case of the CASP1 decoy set consisting of comparative models of a nucleoside diphosphate kinase (nm23) and the corresponding experimental structure, where the RAPDF is unable to discriminate the experimental conformation from the approximate conformation, the experimental structure has been solved to 2.8 Å resolution with an R-factor of 0.24. The parent structure used for the comparative modelling of nm23, PDB code 1ndl, with a percentage sequence identity of 77%, has been solved to a 2.4 Å resolution with an R-factor of 0.16. The fact that the RADPF produces a better (lower) negative log conditional probability for the model might simply reflect the moderate resolution and incomplete refinement of the experimental structure relative to the structure used to construct the nm23 model.

3.4.10 Relationship between the conditional probabilities and the nature of physical interactions in proteins

The discriminatory function compiled using statistical observations averages over different environments. As a result, it displays features not observed in a direct way in a physics-based energy function, as shown in Figures 3.8, 3.9, 3.10, and 3.11.

Some of these features are obvious: the largest minima in the plot of C_α-C_α contacts (Figure 3.8) reflects the geometrical constraints imposed by the covalent structure for the amino acid sequence, and the fact that contacts between proline nitrogens and other main chain oxygen atoms have the lowest negative log conditional probabilities reflects the absence of the hydrogen atom in proline nitrogens (Figure 3.9).

Some features are less obvious: the C_α-C_α minimum in the 5.0-6.0 Å bin is due to the presence of C_α-C_α contacts between $i, i+2$ and $i, i+3$ residues in alpha-helices, and the C_β-C_β minimum in the 5.0-6.0 Å bin is due to the presence of C_β-C_β contacts between $i, i+1$ residues in both α-helices and β-sheets.

These observations, described in detail in the RESULTS section, suggest that the conditional probability formalism described in this chapter can be used to elucidate properties about atom-atom preferences without a potential of mean force analysis [124].

However, caution should be used when interpreting the conditional probability or potential of mean force data in physical terms due to the averaging of environments that occurs during the compilation of the probabilities. To illustrate with an extreme example, the minimum in the N-O plot (Figure 3.9) in the 0.0-3.0 Å bin averages over hydrogen bonds between N and O atoms in $i, i+4$ residues in α-helices and between N-O distances in $i, i+1$ (neighbouring) residues. It is thus difficult to ascertain exactly where the signal is coming from given the two different environments.

3.4.11 Availability of conditional probability tables on the World Wide Web

The conditional probability tables are available on the World Wide Web at [118].

3.5 Summary

We present a discriminatory function formalism to compute the conditional probability of an amino acid sequence conformation being native-like given a set of pairwise atom-atom distances. The formalism is used to derive three discriminatory functions with different types of representations for the atom-atom contacts observed from a database of protein structures. These functions include two virtual atom representations and one all-atom representation. When applied to six different decoy sets containing several correct and incorrect conformations of amino acid sequences, the all-atom distance-dependent discriminatory function is able to identify correct from incorrect more often than other discriminatory functions which approximate the detail in the representation. We illustrate the importance of using a detailed atomic representation for the most accurate discrimination, and the necessity for testing discriminatory functions against a wide variety of decoys. The discriminatory function is also shown to be capable of capturing the fine details of atom-atom preferences. These results suggest an all-atom residue-specific distance-dependent representation with a large distance cutoff is necessary for the most accurate discrimination for use in protein structure prediction and model refinement.

In the next chapters, we show how the all-atom discriminatory function can be used to select the most probable side chain rotamers, to assign weights to nodes and edges in our graph-theoretic representation of protein structure, and to select the most native-like conformation of an amino acid sequence from a set of conformations in *bona fide* comparative modelling scenarios.

Chapter 4

An analysis of side chain preferences in protein structures

4.1 Introduction

Given a protein main chain conformation, constructing side chains by exploring all possible rotamer conformations simultaneously is a computationally intractable problem. Several approaches have been developed to reduce the number of possibilities. These include conformational searching using Monte Carlo and simulated annealing methods [50, 51], using main chain dependent rotamer libraries to construct side chains [52], and matching local main chain coordinates to a database of side chain/main chain combinations [53, 125, 126].

The need to build side chains from a fixed main chain often arises in the case of comparative modelling, where an initial main chain of the sequence to be modelled (the target) is obtained from copying the main chain coordinates of a related sequence for which the structure has already been determined using experimental methods (the parent) [8, 49]. Alignment of the target and parent sequences is used to determine the equivalent residue positions for which the main chain in the parent structure can be copied over to the main chain of the target structure. While the copied main chain is generally not identical to the main chain of the target structure, it is quite similar in regions where the sequence is conserved [34]. Thus side chain building methods have generally been evaluated by re-building side chain conformations on an experimental structure main chain.

We introduce a method that will reduce the number of conformational choices for a given side chain based on a given environment, such as the local main chain. We use the conditional probability based discriminatory function previously described in Chapter 3 to find the negative log conditional probability of a side chain conformation being correct. These probabilities are used to rank the different side chain conformations sampled using a discrete rotamer library. We perform an analysis of the accuracy of side chain construction using only the local main chain (up to ± four residues, total of nine), using the entire main chain of the protein, and building side chains in a pairwise manner. We compare the change in accuracy as the environment used for the construction of side chains is changed. We evaluate the effect of the rotamer library approximation, and compare our results to other side chain building methods. We illustrate how side chain construction using only the local main chain can be combined with other search techniques to explore the conformational space of multiple protein side chains in the context of comparative modelling.

4.2 Methods

4.2.1 Description of discriminatory functions

Our objective here is to evaluate the strength of the interaction of a side chain conformation with different environments. To do this, we introduce an all-atom distance dependent conditional probability-based discriminatory function which is used to calculate the conditional probability of contacts between pairs of atom types in a given protein conformation. The conditional probabilities for the residue-specific all-atom probability discriminatory function (RAPDF) are compiled by counting frequencies of distances between pairs of atom types in a database of protein structures. All non-hydrogen atoms are considered, and the description of the atoms is residue specific, i.e., the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. We divide the distances observed into 1.0 Å bins ranging from 3.0 Å to 20.0 Å. Contacts between atom types in the 0.0-3.0 Å range are placed in a separate bin, resulting in a total of 18 distance bins.

We compile a table of negative log conditional probabilities for all possible pairs of the 167 atom types for the 18 distance ranges using the expression for the probability of seeing two atom types, a and b, in contact in distance bin d in a native conformation, $P(d_{ab}|F)$:

$$P(d_{ab}|F) = \frac{P(d_{ab}|F)}{P(d_{ab})} = \frac{N(d_{ab})/\sum_d N(d_{ab})}{\sum_{ab} N(d_{ab})/\sum_d \sum_{ab} N(d_{ab})} \tag{4.1}$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d, $\sum_d N(d_{ab})$ is the number of a-b contacts observed for all distance bins, $\sum_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atoms types a and b in a particular distance bin d, and $\sum_d \sum_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types a and b summed over all the distance bins d.

The table of conditional probabilities is compiled from a set of 265 non-homologous ($< 30\%$ sequence identity between any proteins in the set) high-resolution (< 3.0 Å) x-ray structures [117, 118]. A detailed description of this formalism, along with the proteins used in the compilation process, is given in Chapter 3.

For observations of contacts between pairs of atom types within a single residue, a separate table of negative log conditional probabilities, which are different from the ones observed for inter-residue contacts, is compiled using the same formalism but with a different distance cutoff. We divide the distances observed for atoms within a residue into 10 1.0 Å bins ranging from 0.0 Å up to 10.0 Å.

Given a set of n distances in an amino acid sequence that fall within the 20.0 Å distance cutoff, we can calculate the negative log conditional probability of the conformation being native-like given a set of distances, $P(F|\{d_{ij}\})$, using the expression:

$$\ln P(F|\{d_{ij}\}) = \sum_n \ln P(d_{ij}|F) + c \tag{4.2}$$

where c is a constant which is ignored in practice.

For evaluating the probability of a single side chain conformation, the set of distances between atoms in the side chain to atoms in the environment and within the residue are calculated. The negative log conditional probabilities based on these distances and the atom types (obtained by looking up the appropriate

Residue	Number of χs	χ_1	χ_2	χ_3	χ_4
C	1	N-C_α-C_β-S_γ			
D	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$O_{\delta1}$		
E	3	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-$O_{\epsilon1}$	
F	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta1}$		
H	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$N_{\delta1}$		
I	2	N-C_α-C_β-$C_{\gamma1}$	C_α-C_β-$C_{\gamma1}$-$C_{\gamma2}$		
K	4	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-C_ϵ	C_γ-C_δ-C_ϵ-N_ζ
L	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta1}$		
M	3	N-C_α-C_β-C_γ	C_α-C_β-C_γ-S_δ	C_β-C_γ-S_δ-C_ϵ	
N	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$O_{\delta1}$		
Q	3	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-$O_{\epsilon1}$	
R	4	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-N_ϵ	C_γ-C_δ-N_ϵ-C_ζ
S	1	N-C_α-C_β-O_γ			
T	1	N-C_α-C_β-$O_{\gamma1}$			
V	1	N-C_α-C_β-$C_{\gamma1}$			
W	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta1}$		
Y	2	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta1}$		

Table 4.1: Definitions for the χ angles for all amino acids excluding alanine, glycine, and proline. For each residue indicated in one letter code, the number of χ angles and the names of atom types that define the rotamer used to calculate the χ angle is given.

table) are summed up to give the strength of interaction of the side chain with its environment.

4.2.2 Definition of χ angles

Table 4.1 gives the definitions of the χ angle(s) for each residue having one or more χ angles (alanine, glycine, and proline residues are not included in the library). χ angles are defined by the positions of four atoms which comprise a "rotamer", with the middle two atoms forming a vector around which the other side chain atoms are rotated.

4.2.3 Description of rotamer library

Table 4.2 describes the main chain independent rotamer library used to sample the side chain conformations. For each torsion angle, between two to three χ angle values ("rotamers") are defined. The library values are compiled by observing preferences of side chains to be in discrete rotamer value bins in a database of protein structures.

4.2.4 Selection of the protein structures for testing side chain building

To devise a test set of proteins we first took a list of 487 proteins whose amino acid sequences were less than 25% identical to each other as determined by the PDB SELECT tool [127]. From this set, all structures determined using NMR methods, all structures determined using x-ray crystallography having a resolution greater than 1.50 Å or an R-factor of greater than 0.20, and all structures that were used in the compilation of the conditional probabilities for the atom type preferences were eliminated. Table 4.3 gives the details of the remaining fifteen structures that were selected using this process.

4.2.5 Exploration of side chain conformations

For each rotamer in each residue side chain (excluding alanine, glycine, and proline residues), all possible χ angle values in the rotamer library (Table 4.2) are explored systematically. For example, in the case of valine which has three possible values for its one χ angle, there are three possible side chain conformations. In the case of lysine, which has four χ angles with three possible values for each χ angle, there are $3^4 = 81$ possible side chain conformations. Each possible conformation is assigned a negative log conditional probability based on the contacts between the atom types in the side chain and the atom types in the environment. The conformations with the lowest negative log conditional

Residue	Rotamer	Angle 1 (°)	Angle 2 (°)	Angle 3 (°)
C	χ_1	60.0	180.0	300.0
D	χ_1	60.0	180.0	300.0
D	χ_2	0.0	90.0	
E	χ_1	60.0	180.0	300.0
E	χ_2	60.0	180.0	300.0
E	χ_3	0.0	90.0	
F	χ_1	60.0	180.0	300.0
F	χ_2	0.0	90.0	
H	χ_1	60.0	180.0	300.0
H	χ_2	60.0	180.0	300.0
I	χ_1	60.0	180.0	300.0
I	χ_2	60.0	180.0	300.0
K	χ_1	60.0	180.0	300.0
K	χ_2	60.0	180.0	300.0
K	χ_3	60.0	180.0	300.0
K	χ_4	60.0	180.0	300.0
L	χ_1	60.0	180.0	300.0
L	χ_2	60.0	180.0	300.0
M	χ_1	60.0	180.0	300.0
M	χ_2	60.0	180.0	300.0
M	χ_3	60.0	180.0	300.0
N	χ_1	60.0	180.0	300.0
N	χ_2	60.0	180.0	300.0
Q	χ_1	60.0	180.0	300.0
Q	χ_2	60.0	180.0	300.0
Q	χ_3	60.0	180.0	300.0
R	χ_1	60.0	180.0	300.0
R	χ_2	60.0	180.0	300.0
R	χ_3	60.0	180.0	300.0
R	χ_4	60.0	180.0	300.0
S	χ_1	60.0	180.0	300.0
T	χ_1	60.0	180.0	300.0
V	χ_1	60.0	180.0	300.0
W	χ_1	60.0	180.0	300.0
W	χ_2	0.0	90.0	270.0
Y	χ_1	60.0	180.0	300.0
Y	χ_2	0.0	90.0	

Table 4.2: Main chain independent rotamer library used to sample side chain conformations. For each residue in one letter code, the rotamer and between two to three values of angles (in degrees) are given.

Protein PDB code	Number of residues	Resolution (Å)	R-factor	Name
1bab-B	146	1.50	0.16	Hemoglobin (Human)
1cbn	46	0.83	0.11	Crambin
1ccr	111	1.50	0.19	Cytochrome C
1cus	197	1.25	0.16	Cutinase
1pmy	123	1.50	0.20	Pseudoazurin (Cupredoxin)
1ptx	64	1.30	0.15	Scorpion Toxin II
1wfb-A	37	1.50	0.18	Antifreeze Protein Isoform Hplc6
1xnb	185	1.49	0.17	Xylanase
1xso-A	150	1.49	0.10	Cu, Zn Superoxide Dismutase
2end	137	1.45	0.16	Endonuclease V
2hbg	147	1.50	0.13	Hemoglobin (Bloodworm)
2ihl	129	1.40	0.17	Lysozyme (Japanese Quail)
3sdh-A	145	1.40	0.16	Hemoglobin I
2sga	181	1.50	0.13	Proteinase A
9rnt	104	1.50	0.14	Ribonuclease T1

Table 4.3: List of proteins selected to test side chain construction. The PDB code of the protein, the size of the protein, the resolution, the R value, and the name are given. The proteins are selected based on their high resolution (<= 1.5 Å) and uniqueness (less than 25% sequence identity to each other) and are not used in the compilation of the residue specific all-atom conditional probability discriminatory function (RAPDF).

probabilities are used to assess the accuracy of this approach.

4.2.6 Generation of side chain conformations using only the local main chain

All possible side chain conformations for each residue (excluding alanine, glycine, and proline) were explored. The top five conformations with the lowest negative log conditional probability based on the contacts between atoms in the side chain to the local main chain (up to ± four residues, total of nine) were selected for evaluation.

4.2.7 Generation of side chain conformations using the entire main chain

All possible side chain conformations for each residue (excluding alanine, glycine, and proline) were explored. The top five conformations with the lowest negative log conditional probability based on the contacts between atoms in the side chain to the main chain of the entire protein were selected for evaluation.

4.2.8 Generation of side chain conformations in a pairwise manner

For a given experimental structure main chain, all pairs of possible side chain conformations that have at least one interatomic contact within a distance of 6.0 Å are explored (excluding any pairs that have an alanine, glycine, or proline residue in the pair). The top five pairs of conformations with the lowest total negative log conditional probabilities, evaluated by summing the probabilities of the contacts between atom types of each of the two side chains with their respective local main chains (up to ± four residues), and the probabilities of the contacts involving atom types between the two side chains, were selected for evaluation.

We also select a single best side chain conformation for each residue based on the pairwise construction so we can compare the accuracy of side chain construction when pairwise information is added to the other cases where side chains were built using only the local main chain and the entire main chain. The best side chain conformation for each residue is obtained by examining the pairs of conformations that residue interacts with, selecting the pair of conformations with the lowest negative log conditional probability, and selecting, from that pair, only the side chain conformation of the residue of interest.

4.2.9 Evaluation of side chain construction

Once the top five conformations in each situation described above were selected, they were compared to the experimental structure conformation. All the rotamers in a given side chain must agree with the experimentally observed rotamer conformation (i.e., *all* the rotamers for a given side chain must be within ± 60° or ± 45° of the corresponding experimental rotamers depending on the residue type) in order for a side chain conformation to be considered correct. We do not consider alanine, glycine, or proline residues in the evaluation. A side chain is considered to be correctly built if one of the members of the set of up to five conformations satisfies this test.

4.2.10 Comparison to other methods

Evaluating the method by checking to see if a built side chain conformation and the experimental conformation fall into the same rotamer library "bin" is useful for comparison of the different ways we construct side chain conformations in different environments, and circumvents the problem of using an approximate rotamer library. However, it does not indicate exactly how accurate the conformations we generate are in an absolute sense, so they can be compared to other methods. To this end, we generate the side chain conformation with the lowest negative conditional probability between the side chain atoms and the local main chain atoms (± four residues, total nine where available) for a set of ten structures that have been used by others to build side chains. We compare our methods to those of Dunbrack and Karplus [52], Holm and Sander [51], Laughton [53], and Lee and Subbiah [50], by calculating the percentage error in χ_1 angles using a 30° cutoff and the RMSD of the side chain atoms (including the C_β atom) between the built side chain conformation and the experimental conformations. These criteria were selected for comparison based on the criteria used in published papers describing the methods, with the intent of being able to compare the approach described here with the largest number of methods.

The set of ten proteins for which side chains are rebuilt using the new evaluation criteria are different from the test set used previously. Since there have been different proteins tested by different methods, we selected a set of proteins that have been used to build side chains previously by at least two of the methods described in [50, 51, 52, 53]. Details regarding this set are given in Table 4.4. In cases where different methods have used the same protein but with a different PDB structure (for example Lee and Subbiah [50] have used 1rn3 instead of 7rsa for Ribonuclease A), we test our method using the PDB structure used by a method (other than our own) that gives the best results for that protein.

The conditional probability discriminatory functions compiled originally using the set of experimental structures in Chapter 3 were recompiled by removing all the proteins and homologs for which side chains are being constructed.

The methods we choose are representative of the diverse set of methods available for side chain construction: Dunbrack and Karplus generate a main chain dependent library for side chain conformations (based on the ϕ/ψ values adopted by the main chain) and use it to construct side chain conformation initially, and then use a minimisation scheme to reorient side-chains that conflict with the main chain or other side chains after initial placement [52].

Holm and Sander use a Monte Carlo algorithm together with the rotamer library of Tuffery, et al. [128] with simulated annealing and a simple potential energy function to optimise the packing of side chains on a given main chain [51].

Laughton compares the local environments of each side chain conformation to be built to a database of local environments for the same side chain type constructed from an analysis of protein structures. The database description consists of a list of C_α coordinates and residue type for each residue in the protein that has at least one atom within 4.0 Å of a side chain atom of the residue of interest. Side chain conformations that match the local environment criteria the best are input to a Monte Carlo procedure to give a final structure [53].

Lee and Subbiah apply a simulated annealing algorithm to the optimisation of side chain packing interactions, using a simple van der Waals potential function [50].

4.2.11 Effect of rotamer library approximation

Since we sample only between two to three angles per rotamer in a given side chain (Table 4.2), it is possible that our results using the percentage error measure with a 30° degree cutoff or the side chain atom RMSD are influenced by the non-ideal χ values in experimental structures as well as the ability of the RAPDF to distinguish between correct and incorrect rotamers. To evaluate the effect of restrictions imposed by using the approximate library, we calculate the rotamer library value nearest to the experimental structure value for each rotamer in the experimental structure and generate conformations for all side chains for the

Protein PDB code	Number of residues	Resolution (Å)	R-factor	Name
1crn	46	1.5	0.11	Crambin
1ctf	68	1.7	0.17	L7/L12 ribosomal protein
1lz1	130	1.5	0.18	Lysozyme (Human)
3apr	325	1.8	0.15	Rhizopuspesin
2cro	65	2.4	0.20	λ cro repressor
3app	323	1.8	0.14	Pencillopepsin
3tln	316	1.6	0.21	Thermolysin
3fxn	138	1.9	0.21	Flavodoxin
5pti	58	1.0	0.20	Pancreatic tripsin inhibitor
7rsa	124	1.3	0.15	Ribonuclease A

Table 4.4: List of proteins selected to compare side chain construction against other methods. The PDB code of the protein, the size of the protein, the resolution, the R-factor, and the name are given. The proteins that have been used previously to test side chain building in at least two of the methods described in [50, 51, 52, 53] were selected.

ten proteins in Table 4.4 using these values. We compare the accuracy of the model generated with the percentage error and side chain atom RMSD measures, excluding alanine, glycine, and proline residues.

4.3 Results

4.3.1 Construction of side chains using only the local main chain

Figure 4.1 shows five different percentages of side chains accurately constructed using only the local main chain for the fifteen proteins in the test set. The set size is the number of conformations considered (based on the negative log conditional probability score). A set size of two indicates that the top two conformations, as ranked by the negative log conditional probability score, were checked to see if one of them was correct, and a set size of one indicates only the top ranking conformation was checked to see if it was correct. The percentages are determined by computing the number of side chains accurately constructed for a given set size over the total number of possible side chains for each of the structures.

The average percentages of side chains accurately constructed using only the local main chain for the fifteen proteins in the top five set sizes are 51.9%, 67.8%, 78.5%, 83.3%, and 85.5% respectively.

4.3.2 Accuracy of individual residue side chain construction using only the local main chain

Figure 4.2 shows the percentages of side chains constructed for the 17 different amino acids using only the local main chain and the RAPDF. Figure 4.3 shows the difference in the percentage accuracy between side chains constructed for the 17 different amino acids using only the local main chain in cases where the residue adopts a α-helix or β-sheet secondary structure as classified by the program DSSP [129], and percentage accuracy of side chains constructed regardless of secondary structure adopted. The data to calculate the percentages in Figures 4.2 and 4.3 is obtained by calculating the percentage accuracy for each of the seventeen side chain types by building the side chain conformations for the fifteen structures in Table 4.3.

The average percentage accuracy for all residues for the conformation with the lowest negative log conditional probability (set size 1) based on secondary structure type is 52.6% for α-helix, 42.2% for β-sheet, and 42.0% for residues not in α-helix or β-sheet. The average percentage accuracy for set size 1 conforma-

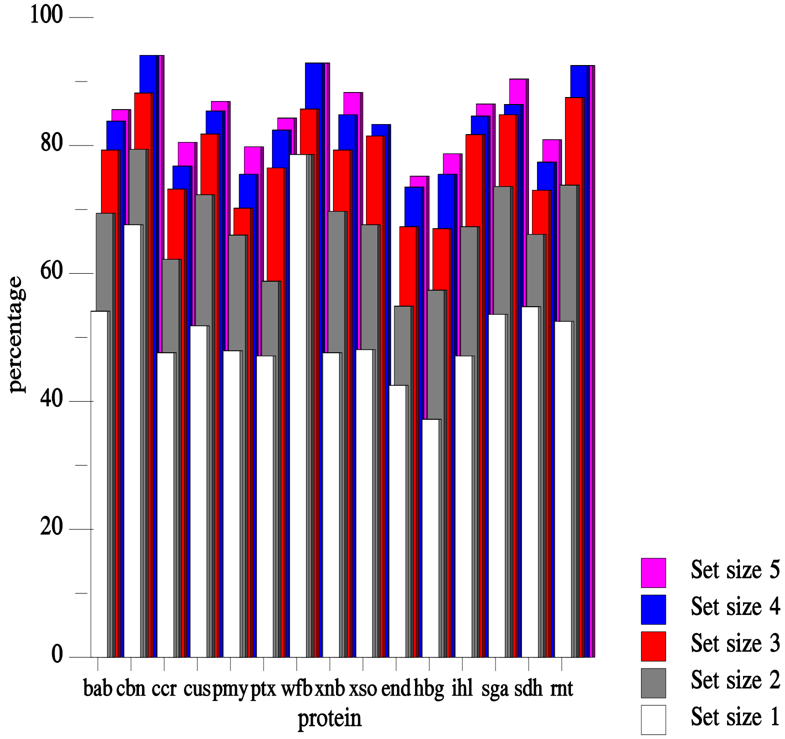


Figure 4.1: Results of building side chain conformations for fifteen proteins using the local main chain and the residue-specific all-atom conditional probability discriminatory function (RAPDF). The bars represent the percentage of side chain conformations accurately constructed for different set sizes. The lowest bars (set size 1) represent the percentage of side chain conformations accurately constructed considering only the conformation with the lowest negative log conditional probability as evaluated by the RAPDF. The highest bars (set size 5) represent the percentage of side chain conformations accurately built considering the five conformations with the lowest negative log conditional probabilities.

tions regardless of secondary structure type is 44.8%. From Figures 4.2 and 4.3, it is evident that certain residues are more easily built based on the secondary structure adopted by the main chain. For example, phenylalanine in a α-helix is constructed accurately 66.6% of the time by the RAPDF using only the local main chain, whereas in a β-sheet the percentage accuracy is 80.0%. Valine in a α-helix is constructed to 88.0% accuracy, whereas in a β-sheet, it is constructed

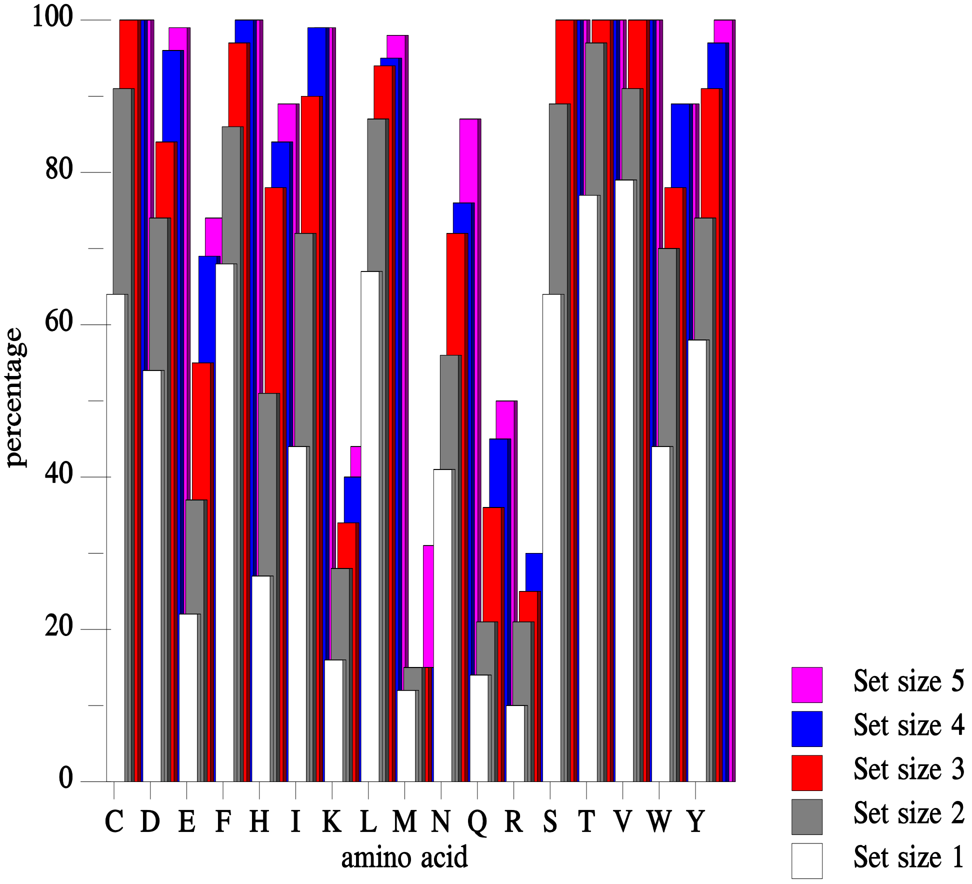


Figure 4.2: Results of building side chain conformations for seventeen amino acid types using the local main chain and the residue-specific all-atom conditional probability discriminatory function (RAPDF). The bars are as in Figure 4.1.

to 76.9% percent accuracy. Some of the more dramatic differences include tryptophan (85.7% in α-helix, 30.7% in β-sheet), and aspartic acid (82.1% in α-helix, 44.4% in β-sheet). The side chains that are the most difficult to build are the ones with the most χ angles and therefore the most degrees of freedom, such as glutamic acid, lysine, methionine, glutamine, and arginine.

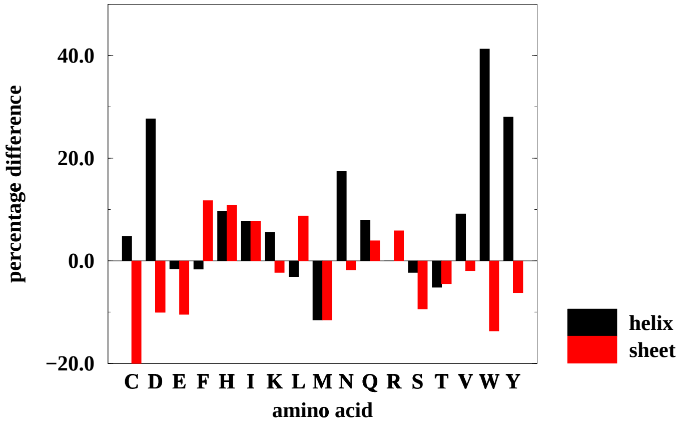


Figure 4.3: Differences in the accuracy of building seventeen amino acid types using the local main chain and the residue-specific all-atom conditional probability discriminatory function (RAPDF) as a function of main chain secondary structure. The bars represent the differences between the percentage of side chain conformations accurately constructed where the amino acid main chain is in a α-helix and β-strand secondary structure as classifed by the program DSSP [129] and the percentage of side chain conformations accurately constructed for all amino acid types regardless of the secondary structure type of the main chain. A positive percentage difference (bar above the axis) indicates that the percentage accuracy of side chain conformations built was more accurate in cases where the main chain adopted the relevant secondary structure compared to the percentage accuracy of side chain construction ignoring the secondary structure of the main chain. Only the conformations with the lowest negative log conditional probability are selected for this evaluation.

4.3.3 Construction of side chains using the entire main chain

Figure 4.4 shows the percentages for side chains accurately constructed using the entire protein main chain for the fifteen proteins in the test set. The percentages are given for each of the top five set sizes. The average percentages of side chains accurately constructed using the entire main chain for the fifteen proteins in the top five set sizes are 57.8%, 73.2%, 80.3%, 84.1%, and 86.7%, respectively. Comparing the data from Figure 4.1, which shows the results of side chain construct using only the local main chain, we see that that using the entire main chain improves the accuracy at best by only 5.9%.

4.3.4 Construction of side chains in a pairwise manner

Figure 4.5 shows the percentages for pairs of side chains accurately constructed taking into account the negative log conditional probabilities of the interatomic contacts between the two side chains as well as the probabilities of the interatomic contacts between each of the side chain conformations and the corresponding local main (± four residues). In this case, the percentage represents the number of pairs of side chains built accurately for both side chain conformations over the total number of possible pairs of side chain conformations.

While the results here are biased because of the larger number of conformations used in the percentage evaluation, the average percentages of pairs of side chains accurately constructed for the fifteen proteins in the top five set sizes are 32.3%, 45.5%, 52.0%, 56.8%, and 60.3% respectively.

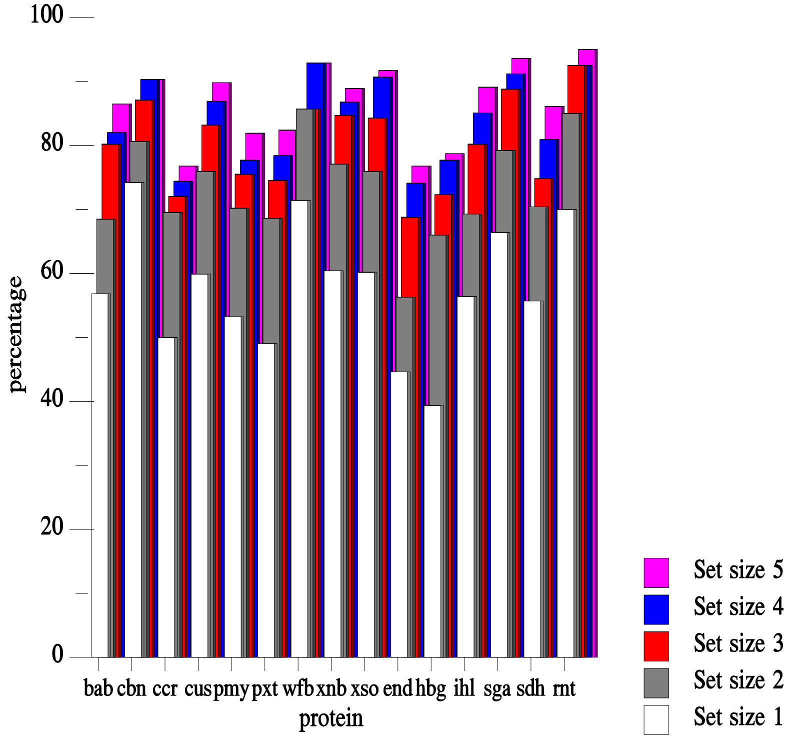


Figure 4.4: Results of building side chain conformations for fifteen proteins using the entire protein main chain and the residue-specific all-atom conditional probability discriminatory function (RAPDF). The bars are as in Figure 4.1.

4.3.5 Comparison of side chain construction at a single residue level using local and pairwise information

We wish to assess the increase in accuracy in side chain construction achieved by adding the influence of a single side chain to that of the local main chain.

Instead of measuring percentage accuracy for pairwise construction by checking to see if both side chain conformations are built accurately, we select, for a given residue, its interaction pair with the lowest negative log conditional probability and assess whether the conformation of the given residue is constructed accurately. For example, if a valine at position 13 interacts with ten other

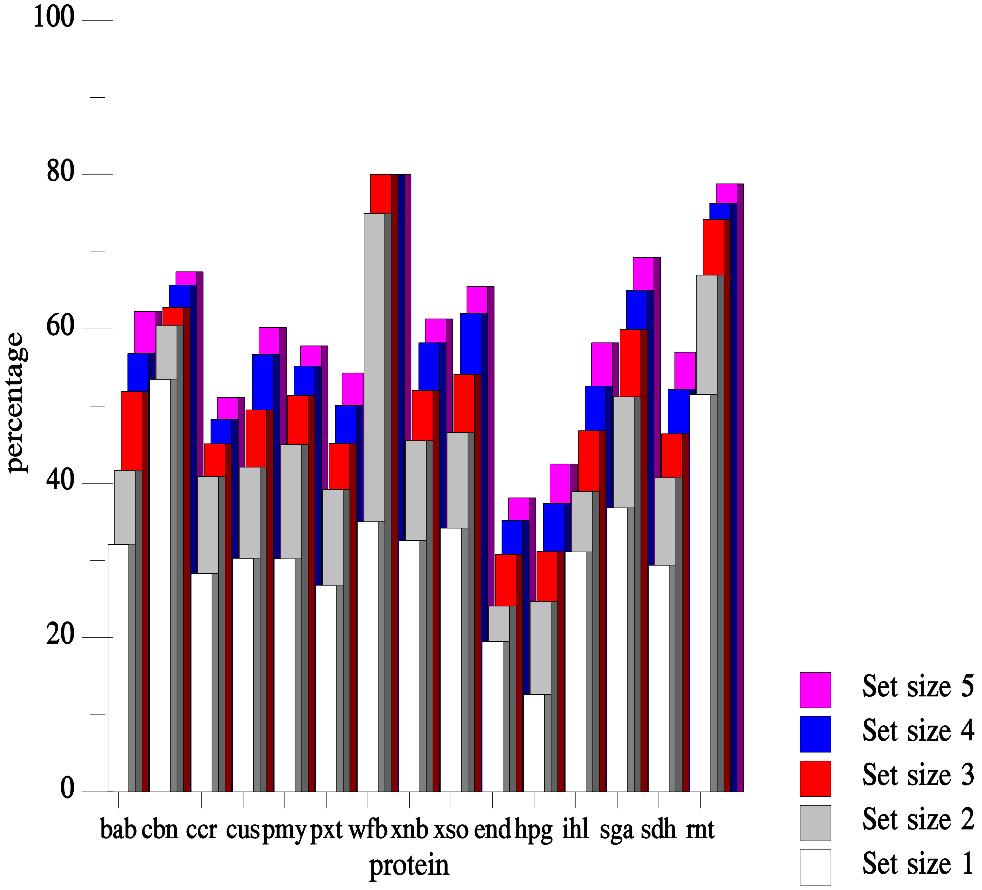


Figure 4.5: Results of building side chain conformations for fifteen proteins in a pairwise manner using the residue-specific all-atom conditional probability discriminatory function (RAPDF). The bars are as in Figure 4.1. In this case, for a conformation to be considered "correct", both the side chain conformations in a pair must be built correctly.

residues, the best pairwise conformation for each of the ten pairs is obtained by selecting the conformation with lowest negative log conditional probability for all possible conformations of that pair. Among these ten pair probabilities, we select the pair of conformations with the lowest negative log conditional probability and check to see only if the valine conformation is constructed accurately.

This process is repeated for all pairs of residues in the fifteen proteins in the test set.

Figure 4.6 shows the results of this comparison. Both on a protein (Figure 4.6a) and on an individual amino acid (Figure 4.6b) level we see that there is about a 10% improvement in the construction of the side chains on average. The average percentage accuracy over the fifteen proteins (Figure 4.6a) and for individual amino acid construction (Figure 4.6b) for side chain conformations selected using the above process is 60.4% and 53.5% respectively. This is an improvement over both using only the local information and using the entire main chain for construction of side chains. In only one protein (1wfb-A) does building the side chains in a pairwise manner worsen percentage accuracy (Figure 4.6).

The residues built with a percentage accuracy of more than 80%, when evaluated individually, are also generally the ones with high percentage accuracies when pairs of side chain conformations are evaluated simultaneously: Phenylalanine-threonine, valine-threonine, and cysteine-threonine have the largest pairwise percentage accuracies (of 75.0%, 78.1%, and 90.0% for the pair with the lowest negative log conditional probability) among all pairs of residue types. These four residues are also the ones with the highest percentage accuracies when evaluated individually and are also the ones with the greatest improvement over using only the local main chain information (Figure 4.6)

4.3.6 Comparison to other methods

Table 4.5 compares the results using our method for a set of ten proteins for which side chains have been constructed by other methods. The measures used

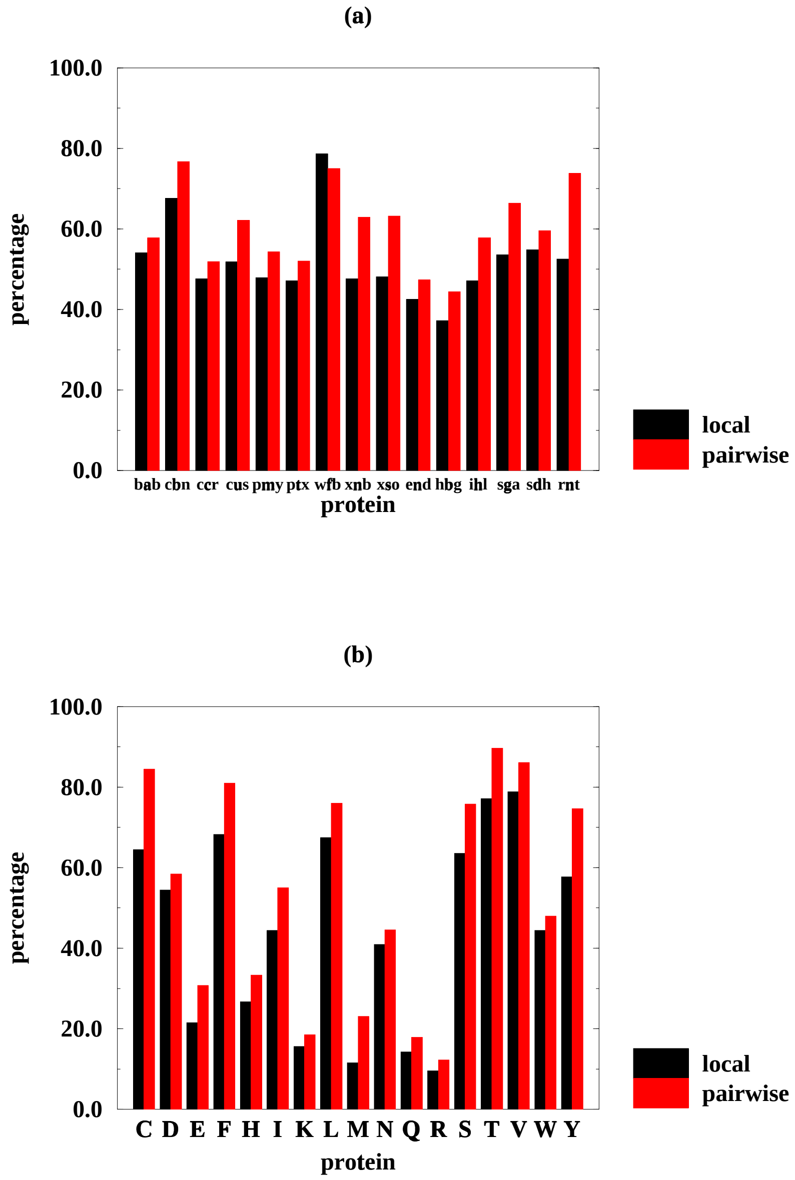


Figure 4.6: Comparison of side chain construction using only local main chain information and that plus pairwise information. The side chain conformation with the lowest negative log conditional probability using only the local main chain (± four residues) and the side chain conformation in a pair of interacting side chains with the lowest negative log conditional probability using both the local main chain and pairwise information for each residue are compared with the experimental conformation to obtain the percentage of side chains constructed correctly. The comparison is made for the fifteen proteins in the test set by protein (a) and by amino acid (b).

for comparison are the percentage error in the χ_1 angles (i.e., the fraction of built conformations where the deviation in the χ_1 angle is greater than 30° times 100) and the side chain atom RMSD (including the C_β atom) for all the residues in the protein (excluding alanine, glycine, and proline residues).

Out of the ten proteins for which side chains were built, five of the proteins have the lowest, or one of the lowest, percentage error in the χ_1 angles, and six of the proteins have the lowest side chain atom RMSD. Different methods have used slightly different criteria for calculating the percentage error in the χ_1 angle and the side chain RMSD. We compare the performance of our method to each of those methods, taking into account the individual criteria used:

Dunbrack and Karplus [52] used a different cutoff of 40° for measuring the error in the χ_1 torsion angles and they include proline residues in their calculation of the percentage of χ_1 angles correctly constructed. Taking the larger cutoff and the proline residues into account in calculating the percentage error, the results for Lysozyme and Pancreatic Trypsin Inhibitor using our method are identical to theirs. For two of the proteins (Rizopuspepsin and Thermolysin), the percentage error is lower, and in two cases (Ribonuclease A and Crambin), the percentage error is higher.

Holm and Sander [51] use a cutoff of 30° for the χ_1 torsion angle and they include proline residues in their calculation of the percentage of χ_1 angles correctly constructed. In this case, the inclusion of proline residues does not appear to change the relative performance of the methods for the nine cases where they build side chains. In four cases (Pancreatic Trypsin Inhibitor, Flavodoxin, Rhizopuspepsin, and λ cro repressor) the percentage error in the χ_1 angles is lower with our method, in two cases (Thermolysin and Pencillopepsin) the percentage error is the same, and in three cases (Lysozyme, Ribonuclease A, and L7/L12 ribosomal protein), it is worse. Holm and Sander [51] also include the side chain atom RMSD (including the C_β atom) to a single digit precision. In four cases (Pancreatic trypsin inhibitor, Thermolysin, Flavodoxin, and Pencillopepsin) our method produces lower side chain atom RMSDs. In three cases (L7/L12 ribosomal protein, Rhizopuspepsin and λ cro repressor) the RMSDs are about the same, in two cases (Lysozyme and Ribonuclease A), the RMSDs are worse.

Comparing our results to the side chain atom RMSDs provided by Laughton [53] for the eight structures whose side chain conformations are constructed, in six cases (Lysozyme, Pancreatic Trypsin Inhibitor, Crambin, Thermolysin, Flavodoxin, and Pencillopepsin) the RMSDs are better, in one case (Ribonuclease A) the chain atom RMSDs are identical, and we fail to have a lower RMSD only in one case (L17/L12 ribosomal protein).

Lee and Subbiah [50] produce lower RMSDs than the method described here for three out of seven structures for which the side chain atom RMSDs can be compared (Lysozyme, Ribonuclease A, and Pancreatic Trypsin inhibitor).

4.3.7 Effect of rotamer library approximation

Table 4.6 shows the results of using the approximate rotamer library to sample side chain conformations. The average percentage error in χ_1 angles is 5.4%, and the average side chain RMSD error is 0.92 Å. The largest percentage error for χ_1 angles (of 12.1%) and the largest side chain atom RMSD (1.17 Å) is observed for 3fxn. 1crn is the structure with the lowest percentage error (0%) and the lowest side chain RMSD (0.70 Å). These values represent the maximum accuracy the methods described here can achieve.

Name of protein	PDB code	χ_1 $> 30°$ (%)	side chain atom RMSD (Å)	Dunbrack & Karplus (%)	Holm & Sander (%/Å)	Laughton (Å)	Lee & Subbiah (Å)
Crambin	1crn	13	1.40	8	-	1.43	1.65
L7/L12 ribosomal	1ctf	28	1.69	-	19/1.7	1.59	1.86
Lysozyme	1lz1	24	1.97	23	12/1.6	2.22	1.62
λ cro repressor	2cro	34	2.29	-	43/2.3	-	2.39
Pencillopepsin	3app	19	1.20	-	19/1.4	1.22	-
Rhizopuspesin	3apr	15	1.44	18	16/1.4	-	-
Flavodoxin	3fxn	37	1.76	-	39/1.9	1.96	1.90
Thermolysin	3tln	23	1.62	26	23/1.7	1.72	-
Trypsin inhibitor	5pti	21	1.73	15	22/1.9	2.61	1.49
Ribonuclease A	7rsa	33	2.02	21	21/1.8	2.02	1.86

Table 4.5: Comparison of side chain construction using the local main chain and the residue-specific all-atom probability discriminatory function to four other previously published methods. The names and PDB codes of the ten structures chosen for comparison, the percentage error in χ_1 angles (using a 30° cutoff) excluding proline residues, the side chain atom RMSD (including the C_β atom) are given. For the method of Dunbrack & Karplus, we list the percentage error in the χ_1 as given in [52], which includes prolines and uses a 40° cutoff; for Holm & Sander, we list the percentage error in the χ_1 angles (which includes prolines and uses a 30° cutoff) and the side chain atom RMSD as given in [51]; for Laughton and Lee & Subbiah, the side chain atom RMSD, as listed in [53] and [50] respectively, is given.

PDB code	Number of χ_1 rotamers	% $\chi_1 > 30°$ (%)	side chain RMSD (Å)
1crn	32	0.0	0.70
1ctf	46	6.5	0.88
1lz1	103	6.8	1.03
2cro	52	5.7	1.06
3app	247	3.2	0.88
3apr	243	3.2	0.86
3fxn	115	12.1	1.17
3tln	244	9.0	0.99
5pti	42	4.7	0.87
7rsa	105	2.8	0.83

Table 4.6: Effect of using a discrete rotamer library approximation to sample side chain conformations. For each structure, side chain conformations with rotamer library values that are the closest to the experimental rotamer values are generated and compared with the experimental structure. The number of χ_1 angles considered, the percentage error in χ_1 angles (using a 30° cutff) and the side chain atom RMSD (including the C_β atom) is given.

4.3.8 Effect of experimental uncertainty

The percentage accuracies may be influenced by crystallographic interatomic contacts between neighbouring molecules and high temperature factors. We rebuilt side chains using the local main chain for all the fifteen proteins in the test set excluding side chains having one or more atoms with a temperature factor greater than 30.0 Å^2. Since we use high resolution structures with low R-factors, the numbers of side chains excluded by this filter does not significantly change the results (less than 5% average percentage accuracy improvement for the fifteen proteins).

4.4 Discussion

4.4.1 Effect of environment on side chain construction

We find that using local main chain information alone is enough for the residue-specific all-atom conditional probability discriminatory function (RAPDF) to select the correct side chain conformation from incorrect ones with an average percentage accuracy of about 45% when individual amino acids are considered in a set of fifteen proteins. More significantly, when the top five ranking conformations, as sorted by the negative log conditional probability score, are considered, the correct side chain conformation is selected 82% of the time (Figure 4.1). Using the entire experimental structure main chain as the environment to predict conformations increases accuracy to 53% and 86% for the top one and the top five conformations respectively (Figure 4.4). Including the effect of the single most influential side chain along with the local main chain improves the percentage accuracy on average by about 10% (Figure 4.6).

This signifies that the contribution of the non-local main chain for determining the conformation of a specific residue side chain is not as influential as the conformation of a single residue which interacts most favourably with that specific residue.

4.4.2 Choice of criteria for evaluation of side chain building methods

There are several side chain building methods published in the literature that use different evaluation methods to determine whether a side chain conformation is "correct". We used two types of evaluation criteria in this work: one method simply checks to see if all the χ angle values for a built conformation of a side chain are the closest to the experimental structure among all the possible library values for that χ angle. This translates to a maximum error of ± 60° for most χ angles; the maximum error for aspartic acid χ_2, glutamic acid χ_3, phenyalanine χ_2, and tyrosine χ_2 is ± 45° (see Table 4.2).

Using this method for evaluation, we do not deal with the problem of checking to see if an error in the side chain building is due to the approximation in the discrete rotamer library. However, for comparison with other methods published in the literature, the percentage of side chains χ_1 angles built within 30° for all residues (excluding alanine, glycine, and prolines), and the side chain atom RMSD (including the C_β atom) between the experimental structure side chains and the built side chains are used.

4.4.3 Effect of secondary structure and residue type on side chain construction

The accuracy of side chain building generally depends on the secondary structure adopted by the local main chain [130, 131]. In our case, the average percentage accuracy for individual amino acids based on secondary structure type is 52.6% and 42.2% for α-helix and β-sheet secondary structures.

The accuracy of side chain building also depends on the residue type [52, 131]. It is perhaps trivial to select the right rotamer conformation in the case of a side chain with a single χ_1 angle with three degrees of freedom (such as valine) compared with a side chain with four χ angles with a total of $3^4 = 81$ degrees of freedom (such as lysine) since random selection alone will yield percentage accuracies of 33.3% and 1.2% respectively for the two side chain types. However, even when comparing residues with similar degrees of freedom, ignoring secondary structure of the residue, there are differences in the percentage accuracy (Figure 4.2): isoleucine has a percentage accuracy of 44.1%, whereas leucine has a percentage accuracy of 67.2%. Serine has a percentage accuracy of 63.6% whereas threonine and valine have percentage accuracies of 77.2% and 78.8% respectively.

Considering the effect of the combination of residue type and secondary structure of the main chain on side chain building also leads to interesting observations (Figures 4.2 and 4.3: isoleucine has an identical percentage accuracy of 52.2% in both α-helices and β-sheets, whereas leucine has a percentage accuracy of 63.9% and 75.8% in α-helices and β-sheets respectively. Threonine has a similar percentage accuracy in both α-helices and β-sheets (72.0% and 72.8%), but serine has an accuracy of 61.3% in α-helices and an accuracy of 54.2% in β-sheets.

Some of the observations are consistent with our understanding of the geometry of side chains and the geometry of secondary structure main chain [131]. The side chains for more than half the amino acid type, are built more accurately in α-helices than in β-sheets, with the exception of histidine, isoleucine, threonine, and arginine where the percentage accuracies are similar and leucine and phenylalanine, where the accuracy is better in sheet than in helix. This is presumably because the main chain conformation in helix regions reduces the number of degrees of freedom a residue side chain conformation can explore [132, 133], thus making it easier for the discriminatory function to distinguish correct from incorrect conformatons.

The secondary structure of the local main chain, along with the type of the residue being built, must thus be considered carefully when building side chains.

4.4.4 Comparison to other methods

It is difficult to compare different methods because the conformations are built with different goals and using different criteria for accuracy. Since there is insufficient detail provided, we have tried to make our criteria as rigourous as possible and handle exceptions on a case-by-case basis (see the RESULTS section). The method described here, based on using the RAPDF to select side chain conformations with the lowest negative log conditional probability, using only the local main chain, compares favourably to the other methods [50, 51, 52, 53] published in the literature. All the four methods chosen for comparison in turn compare their methods to other methods and produce similar or slightly better results.

4.4.5 Effect of rotamer library approximation

The discriminatory function and the methods described here do not rely upon any particular rotamer library. We used a discrete rotamer library to minimise the number of conformations explored. We find that using up to three χ angle values per rotamer does not drastically affect the maximum possible accuracy of the method. The average percentage error for χ_1 angles (using a 30° cutoff) is only 5.4%, and Table 4.6 shows the limit of what the most accurate chain construction method can achieve given the rotamer library (Table 4.2) for the set of ten proteins.

4.4.6 Building side chains in a realistic modelling situation

The similarity in results using different methods, some of which are highly computer intensive [50], and some that require only a few seconds for a protein of any size [52], suggests that it is not too difficult to reproduce the correct side chain conformations with the experimental main chain to an average percentage accuracy of $\simeq$ 75% in the χ_1 angles.

However, it is highly likely that in approximate environments, the side construction methods tested in idealised environments will not perform as well. For example, in a comparative modelling scenario, the main chain is approximate ($\simeq$ 1.0 Å RMSD) and sometimes incorrect ($>$ 3.0 Å RMSD) even when there is a high ($>$ 50%) degree of local sequence identity between pairs of homologous structures (Chapters 2 and 6). Chung and Subbiah have shown that as the main chain RMSD between homologous proteins rises to above 2.0 Å, and the percentage sequence identity between them is in the twilight zone (20-30% sequence identity), the average side chain RMSD error for the side chains built on the approximate main chain is 3.1 Å, and the average percentage accuracy for χ_1 angles is 22% (using a 40° cutoff), in buried residues [60].

The increase in error in side chain construction as the main chain varies between homologous structures is because the main chain and side chain conformations are intimately interconnected (Chapter 2). A proper treatment to handle the problem of interconnectedness in protein structures would be to vary both the side chains and the main chains simultaneously (Chapters 5 and 6). The approach and the analyses described here helps by reducing the number of side chain choices for a given region of local main chain.

Figures 4.2 and and 4.3 show the power of the RAPDF to select correct conformations from incorrect ones using the local main chain in different secondary structure environments for seventeen amino acids. We use this information to reduce the number of side chain conformations sampled per residue and perform a limited combinatorial search using a graph-theoretic clique finding method in the next chapter. At present this allows us to sample side chain conformations for 15-30 residues simultaneously, which enables us to construct small cores and small regions of insertions and deletions in comparative modelling in a context-sensitive manner.

The results presented here for side chain construction using the entire main chain is viable only in an ideal scenario where the entire main chain is correct. However, building side chains in a pairwise manner is possible in a comparative modelling scenario for all pairs of side chains. Selecting single side chains by first building them in a pairwise manner appears to improve the accuracy of side chain construction (Figure 4.6). Future work will include an algorithm that will generate multiple side chain conformations for single residues using pairwise information, thus reducing the number of conformations that need to be explored per residue position.

4.5 Summary

A discriminatory function based on a conditional probability formalism to distinguish between correct and incorrect conformation of protein structures is used for selecting side chain rotamers on a fixed main chain. The conditional probability discriminatory function allows us to rank different possible side chain conformations based on contacts between side chain atoms with atoms in the environment. We compare the differences in constructing side chain conformations using only the local main chain, using the entire main chain, and by building pairs of side chains simultaneously on experimental structure main chains. Using only the local main chain allows us to construct side chains with an average percentage error of 24.7% on the χ_1 angles using a 30° cutoff, and an average side chain atom RMSD of 1.72 Å for a set of ten proteins. The results of constructing side chains for the ten proteins are compared to the results of other side chain building methods previously published. The comparison shows similar accuracies for reconstruction of side chain conformations on the experimental structure main chain. An advantage of this approach is that it can be used to reduce the number of side chain conformations considered per residue position, thus enabling limited combinatorial searches for building multiple protein side chains simultaneously.

We use this approach to reduce the number of nodes in our graph theoretic representation, which is introduced in the next chapter, in conjunction with the discriminatory function described in the previous chapter, to model side chain and regions of main chain for making *bona fide* comparative modelling predictions (Chapter 6).

Chapter 5

A graph-theoretic approach to protein structure prediction

5.1 Introduction

Any algorithm that attempts to predict the three-dimensional (3D) structure of a protein sequence must be able to handle the combinatorial explosion that occurs in the search space. Side chains and main chains of residues in a protein have many degrees of freedom; if we assume there exists a perfect discriminatory function that can distinguish a native-like structure from a non-native one in all cases, and we allow only a few degrees of freedom per residue in the ϕ/ψ angle space of the main chain and χ angle space in the side chain for all the residues in a protein sequence, then there are an astronomically large number of possible conformations that need to be explored to guarantee that the native-like structure will be found [134, 135].

To overcome this computational intractability, i.e., where large amounts of computation times are required even for relatively small problems, approaches for searching only a limited subset of conformational space of a protein sequence have been developed. These approaches include a variety of Monte Carlo-based methods [136, 110, 137, 138, 139, 120] and Genetic Algorithms simulations [111, 140, 121]. These methods usually rely on some discriminatory energy function that distinguishes correct from incorrect conformations [118] as the sampling occurs and "guides" them toward the native structure. Each of these methods have their own limitations: sampling only a subset of the conformational space still limits the number of total conformations that can be explored [120, 121] and improper or inadequate sampling makes it difficult for some methods to "jump through" conformational space because different conformational states may be separated by high energetic barriers [51]. In some cases, the calculation of the fitness of a conformation can be a prohibiting factor computationally when evaluating a large ($> 10^8$) number of conformations.

The fundamental reason an exponential number of possible conformations of an amino acid sequence needs to be evaluated (in the worst case) to select the native-like conformation is due to the context-sensitivity of the interactions in proteins. That is, each residue conformation in a sequence cannot be built in isolation of other residue conformations, as different regions of the sequence influence each other in the 3D structure. Thus, an algorithm that attempts to mix and match between possible side chain and main chain conformations to find the optimal conformation must take into account the web of interactions that occur at any position in the protein structure.

In Computing Science, the notion of a "graph" has been used to describe many systems that are made up of such interconnected networks [141]. These include laying out the shortest combination of railroad segments between a network of cities (finding minimal spanning trees), finding the shortest paths between any two cities in a network of cities, and finding the shortest path in a city network which involves passing through all the cities exactly once (the famous Travelling Salesman problem). In computational chemistry and biology, graph-theoretic approaches have been used to enumerate chemical isomers [142] and for protein structure comparison [143, 144].

Our goal is to find the best set of interactions in a protein structure given a variety of side chain and main chain choices for each position in the structure. We present an algorithm based on graph theory that will find the optimal arrangement of all these choices, as measured by some discriminatory function, while adequately considering the context-sensitivity seen in protein structures. This representation gives us the control over the choices for possible side chain and main chain conformations for each residue position, enabling us to select the sample space in an intelligent manner. We use pairwise discriminatory functions to speed up the calculation of the fitness of a given conformation by adding up the weights of the nodes and the edges, which can provide an order of magnitude improvement compared to calculating the weight of each conformation separately.

Specifically, we represent possible conformations of an amino acid sequence as weighted maximal completely connected graphs (cliques) and enumerate all the cliques the size of the protein to find the ones with the best weight (which is assumed to represent a native-like conformation). This is the equivalent of systematically exploring every combination of side chain and main positions that is input, and thus limits the number of choices for each position. We therefore apply the approach presented here to comparative modelling problems where the number of residues that need to be searched combinatorially at a given time is much lesser than the size of the protein being modelled.

Comparative modelling is a special case of model building where we exploit the fact that two proteins related by evolution have similar 3D structures. Generally an alignment between the sequence to be modelled (the target) and a related sequence with known structure (the parent or the template) is first constructed [8, 49]. Given such an alignment, an initial model is built by copying the main chain coordinates for equivalent residues and copying side chain coordinations for residue identities which are thought to be conserved. We describe how the graph-theoretic clique finding method can be used in a comparative modelling scenario to build side chains and regions of main chain representing insertions, deletions, and main chain variations between the target and parent structures, and mix and match between different parent homolog structures.

5.2 Methods

5.2.1 General description

Each possible conformation of a residue in an amino acid sequence is represented using the notion of a node in a graph. Edges are then drawn between pairs of residues/nodes that are consistent with each other. Edges and nodes are weighted according to some fixed criteria. Once the entire graph is constructed, all the maximal sets of completely connected nodes (cliques) are found using a clique finding (CF) algorithm. The cliques with the best weight are considered to be similar to the native structure. Figure 5.1 illustrates how the CF method is used to model structures.

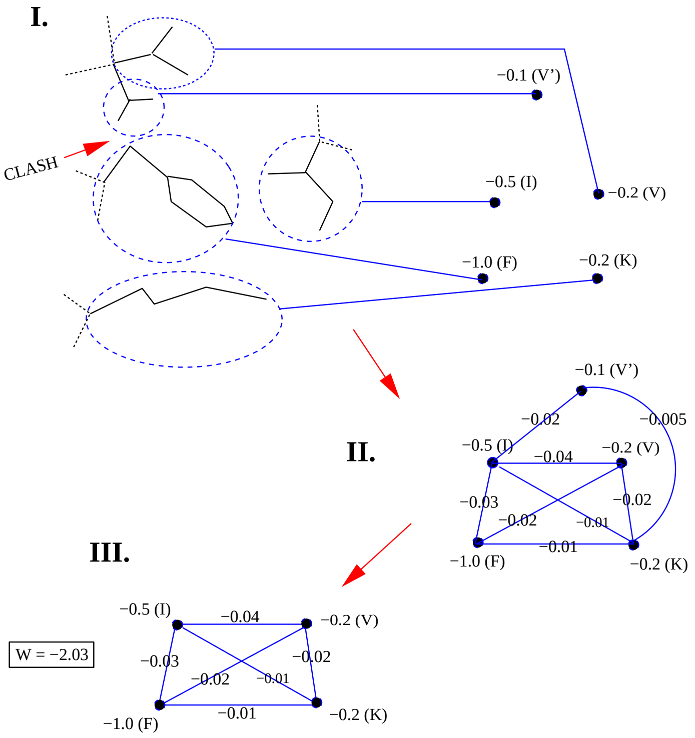


Figure 5.1: Illustration of the graph-theoretic clique finding (CF) method for protein structure prediction. In the first step, possible side chain and main chain conformations of residues are represented as nodes in a graph based on interactions between a single side chain conformation and the local main chain. In this idealised example, three residue positions (isoleucine (I), lysine (K), phenylalanine (F)) with a single possible conformation and one residue (valine) with two possible conformations (V and V') are shown, resulting in five nodes with different weights. In step II, edges are drawn between consistent nodes (see the METHODS section for details). In the example, the inconsistent pairs of nodes are the ones representing the two different valine conformations V and V' (a residue cannot have two conformations simultaneously) and a clash that occurs between V' and F; edges are not drawn between these pairs of nodes. Edges are drawn between all other pairs of nodes and each edge is assigned a weight based on the interaction between the pair of residue conformations (nodes). In the third step, all maximal completely connected subgraphs, or cliques, the size of the amino acid sequence, where every node is connected to every other node, are found and the total weights of the cliques are calculated by summing the weights of the nodes and the edges. Each clique represents a plausible conformation of the entire amino acid sequence and the clique with the best weight is assumed to represent the correct structure. In this example, there is only one clique with nodes {I,V,K,F}. A potential clique I,V',K,F is not considered because of the clash between V' and F.

5.2.2 Description of nodes

Each possible conformation of a residue (side chain and main chain) represents a node in the graph. Nodes have weights based on the strength of the interaction between the side chain atoms within a residue and between the side chain atoms and the local main chain atoms. The main chain atoms of up to four residues on either side of the residue position representing the node are considered for calculating the weights.

5.2.3 Description of edges

Edges are drawn between pairs of nodes. Edges are weighted based on the strength of interaction between the atoms of the pair of residues representing the nodes. Edges are drawn in a consistent manner. Thus any clique containing a set of edges will represent a consistent set of conformations for all residues. In this particular work, packing consistency is maintained by not drawing edges between nodes whose atoms clash (a contact less than 2.0 Å) with each other. Covalent consistency is maintained by partitioning a complete protein conformation into crossover points. If two residue positions are within a main chain region being built that is between two consecutive crossover points, then both conformations must be connected by a single covalently linked main chain conformation for an edge can be drawn between them (see Figure 5.2 for an illustration of covalent consistency). Edges are also not drawn between different possible conformations of the same residue.

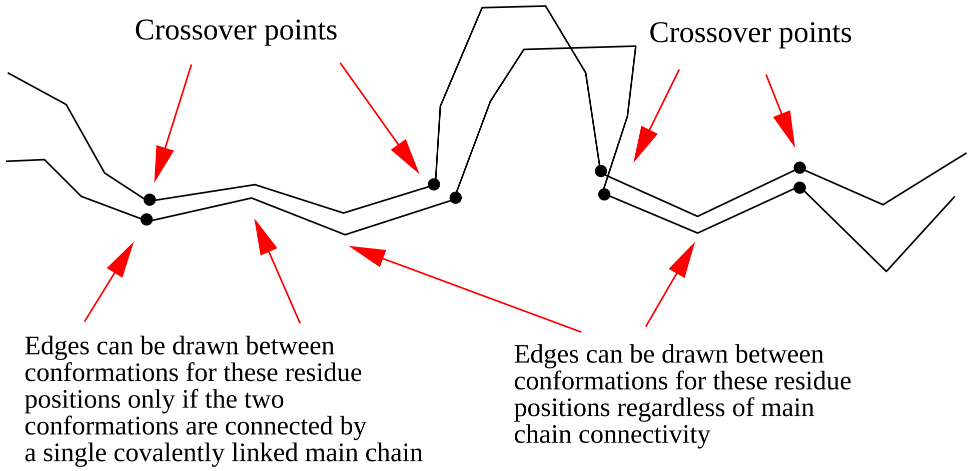

Crossover points
Crossover points
Edges can be drawn between conformations for these residue positions only if the two conformations are connected by a single covalently linked main chain
Edges can be drawn between conformations for these residue positions regardless of main chain connectivity

Figure 5.2: Definition of covalent consistency in the graph-theoretic clique finding (CF) approach. Covalent consistency is maintained by assigning crossover points to pairs of possible main chains for the entire protein. Possible conformations of different residues (nodes) in a main chain region between two consecutive crossover points must be connected by a single continuous covalently linked main chain conformation before an edge can be drawn between them.

5.2.4 Description of the discriminatory function

Our objective here is to assign weights to nodes and edges by determining the strength of the interaction of a side chain in a node to the local main chain and by determining the strength of interaction between two nodes/residues that form an edge. To do this, we use an all-atom distance dependent conditional probability-based discriminatory function to calculate the conditional probability of contacts of a given distance between pairs of atom types for a given conformation of interest. The conditional probabilities for the residue-specific all-atom probability discriminatory function (RAPDF) are compiled by counting frequencies between pairs of atom types in a database of protein structures. All non-hydrogen atoms are considered, and the description of the atoms is residue specific, i.e., the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. We divide the distances observed into 1.0 Å bins ranging from 3.0

Å to 20.0 Å. Contacts between atom types in the 0.0-3.0 Å range are placed in a separate bin, resulting in total of 18 distance bins.

We compile a table of negative log conditional probabilities for all possible pairs of the 167 atom types for the 18 distance ranges using the the expression for the probability of seeing two atom types, a and b, in contact in distance bin d in a native conformation, $P(d_{ab}|F)$:

$$P(d_{ab}|F) = \frac{P(d_{ab}|F)}{P(d_{ab})} = \frac{N(d_{ab})/\sum_d N(d_{ab})}{\sum_{ab} N(d_{ab})/\sum_d \sum_{ab} N(d_{ab})} \tag{5.1}$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d, $\sum_d N(d_{ab})$ is the number of a-b contacts observed for all distance bins, $\sum_{ab} N(d_{ab})$ refers to the total number of contacts between all pairs of atoms types a and b in a particular distance bin d, and $\sum_d \sum_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types a and b summed over all the distance bins d.

The table of conditional probabilities is compiled from a set of non-homologous (less than 30% sequence identity between any proteins in the set) high-resolution (less than 3.0 Å) x-ray structures [117]. A detailed description of this formalism, along with the proteins used in the compilation process is given in Chapter 3.

For observations of pairs of atom types that belong within a single residue, a separate table of negative log conditional probabilities is compiled using the same formalism, but with a different distance cutoff. We divide the distances observed for atoms within a residue into 18 1.0 Å bins ranging from 0.0 Å to 18.0 Å.

Given a set of n distances in an amino acid sequence that fall within the

20.0 Å distance cutoff, we can calculate the negative log conditional probability of the conformation being native-like given a set of distances, $P(F|\{d_{ij}\})$, using the expression:

$$\ln P(F|\{d_{ij}\}) = \sum_{n} \ln P(d_{ab}|F) + c \tag{5.2}$$

where c is a constant which is ignored in practice. The weight of an edge or a node is assigned by the calculating all the distances involved. In the case of node, it is the set of distances between the atoms in the side chain conformation and the local main chain, including distances between atoms within a residue. In the case of an edge between nodes, it is the set of distances between pairs of atoms in the two residues.

5.2.5 Description of side chain sampling methods

Side chain conformations for a given residue position were generated by exploring all the possible side chain conformations and selecting the most probable conformations based on the interactions of a given conformation with the local main chain. For each χ angle in a side chain conformation, up to three rotamers were considered based on the rotamer library described in Table 4.2. For each possible side chain conformation, the interactions between the atoms in the side chain and the local main chain (± four residues, if available) were evaluated using the conditional probability discriminatory function described above. The side chain conformations with the lowest negative log conditional probability were taken to represent the most probable conformations.

Depending on the number of residues sampled, we selected up to six conformations per residue based on the strength of the interaction with the local main chain. That is, the top six highest weighted side chain conformations are selected. We have shown that the correct side chain rotamer is present in one of the top five conformations (sorted according to the strength of the interaction with the local main chain) more than 80% of the time (Chapter 4).

The number of actual side chain conformations sampled for a given residue was chosen to minimize the size of the resulting graph and thus ensure the tractability of the CF algorithm.

5.2.6 Description of main chain sampling methods

For building short ($\simeq$ 15 residues) regions of main chain such as loop regions, insertions, and residues flanking deletions, an initial model or a framework, consisting of the main chain coordinates for the rest of the protein other than the region being built must exist. The initial model must also consist of additional residues in the region where the main chain will be built purely for the purposes of obtaining distance constraints and fitting the sampled regions onto the initial model. This means that at least an additional four residues, two on N-terminal side and two on the C-terminal side of the region that represents an insertion, deletion or main chain variation, must be included as part of the region being built.

We use a database method to generate main chain conformations. The database method takes a set of C_α distance constraints and finds a set of main chain conformations in a database of 520 protein structures that match those constraints [58]. We use three constraints for generating main chain conformations and their specification is the same as in [58]: if the main chain region being built is n residues spanning residue positions p to q, then the constraints used are $d(p,q)$, the C$_\alpha$ distance between residues p and q, $d(p,q-1)$, and $d(p+1,q)$.

A database conformation is considered to fit the distance constraints if $d(p,q)$ differs by less than ± 1.0 Å from the corresponding distance in the initial model, and $d(p,q-1)$ and $d(p+1,q)$ differ by less than ± 2.0 Å from the corresponding distance in the initial model. Thus this can result in deviations of up to 2.0 Å for each of the two terminal C$_\alpha$ position in the main region being built (residues p and q), and in deviations of up to 4.0 Å for C_α positions $p+1$ and $q-1$ and $p+1$ and q between the experimental structure and the conformations obtained from the database. The root residues, i.e. the residues flanking the region being built, are defined to be residues $p-1$ and $q+1$.

If a cluster of main chain conformations found by the database search all have ϕ/ψ torsion angle values that are within some cutoff (generally 30°) then only one conformation in that cluster is used. The conformation selected from the cluster is the one that represents the mean conformation of the cluster, i.e., the conformation that is closest in terms of ϕ/ψ angles to all the other conformations in the cluster. The main chain conformations found are then positioned in the initial model or the framework using the methods described in [57, 58]. At this point, the conformations of the residues in the initial model used to generate the distance constraints and for fitting purposes (residues p, $p+1$, q, $q-1$) are removed and only the database conformations are used for these regions (residues p through q). A preliminary screening is done to exclude any main chain conformation that clashes (any interatomic contact less than 2.0 Å) with the main chain of the rest of the modified initial model.

5.2.7 Description of the clique finding method

The clique finding (CF) algorithm we use was developed by Bron and Kerbosch [145]. This algorithm combines a recursive backtracking procedure with a branch and bound technique to eliminate searches that cannot lead to a clique. The recursive procedure is self-referential: finding a clique of length n is accomplished by finding a clique of length $n-1$ and finding another node that is connected to all the nodes in the clique of size $n-1$. This is made possible by defining some terminating condition, and having the procedure that implements the algorithm reference itself until the terminating condition is reached. The branch and bound technique makes use of rules that allow us to determine in advance certain cases for which possible combinations of nodes and edges will never lead to a clique.

There are three sets that are essential for this algorithm:

1. `potential-clique` — is the set of nodes where every node is connected to every other node. Each recursive call will either extend this set by one node or reduce it by one node.

2. `candidates` — is the set of candidates that are eligible for addition to the `potential-clique` set.

3. `already-found` — is the set of nodes that have already served as an extension to the present configuration of `potential-clique` and are now explicitly excluded. That is, all possible extensions of `potential-clique` containing any point in this set have already been generated.

The algorithm operates recursively on each of the sets by generating all extensions of a given configuration of `potential-clique` that it can make with the given set of candidates and that do not contain any of the nodes in already-found in this manner:

At the beginning of each recursive step, the algorithm picks a candidate-node which becomes part of potential-clique. All nodes connected to the selected node become candidates for addition to the clique. The algorithm then analyses each node in candidates, making it smaller with each recursive call by removing all nodes not connected to the selected candidate-node. A necessary condition for finding a clique, and for finding a completely-connected subgraph, is that the nodes in the set candidates must all be connected to each other. That is, at the end of the recursive calls, the set candidates will be empty. However this does not guarantee that potential-clique is maximal, i.e., it does not guarantee that potential-clique contains the largest possible set of nodes where every node is connected to every other node. In order to do so, the set of nodes that have previously served as an extension for the present configuration of potential-clique is maintained in already-found. This set is also made smaller with each recursive call by removing all nodes not connected to the selected candidate-node. If any node in the set already-found is connected to all nodes in candidates, then we know that potential-clique is not maximal and therefore will not lead to a clique since we have already observed this node in a larger clique. This is the branch and bound step of the algorithm. In the pseudocode implementation below, we see how the procedure implementing the algorithm find-cliques references itself in the middle with the sets new-candidates and new-already-found.

```
begin procedure find-cliques(potential-clique, candidates, already-found)

  if a node in already-found is connected to all nodes in candidates then
    no clique can ever be found (branch and bound step)
  else
    foreach candidate-node in candidates do
      move candidate-node to potential-clique
      create new-candidates by removing nodes in candidates not connected to candidate-node
      create new-already-found by removing nodes in already-found not connected to candidate-node
      if new-candidates and new-already-found are empty then
        potential-clique is a maximal-clique
      else
        find-cliques(potential-clique, new-candidates, new-already-found)
      endif
      move candidate-node from potential-clique to already-found
    endfor
  endif

end procedure find-cliques
```

Initially, the set `candidates` contains all the nodes in the graph and the sets `potential-clique` and `already-found` are empty. Bron and Kerbosch select their nodes in a clever manner by choosing nodes with the largest number of edges to reach the branch and bound condition as soon as possible. This leads to the larger cliques being found first and generates sequentially cliques having a large common intersection. More details of this algorithm, including a pseudocode implementation, are given in [145].

5.2.8 Application to a comparative modelling scenario

In a comparative modelling situation only those main chain and side chain conformations that are thought to vary significantly ($>$ 2.0 Å RMSD) from the parent structure are sampled using the methods described above. Main chain regions that are not thought to vary are simply copied over from the parent. Side chain conformations thought to be conserved were built using minimum perturbation (MP) method implemented by the program MUTATE [74]. The MP method changes a given amino acid to the target amino acid preserving the equivalent χ angles, as determined by an equivalence table between the two side chains. The χ angles not present in the model are constructed using a library based on the residue type (Chapter 4).

Using the CF method in comparative modelling leads to a natural definition for the crossover points. In families of homologous structures, there are usually regions of main chain that are very similar to each other (C_α atoms within 1.0 Å of each) and main chain regions that are variable, or represent insertions and deletions. We define a set of crossover points connecting regions of main chains built with the CF method and those copied from a parent structure. Figure 5.2 represents a typical scenario in comparative modelling. In cases where we mix and match between parent structures, we define crossover points based on a structural superposition. Mixing and matching of main chain regions between crossover points ensures that for every edge representing a pair of residue conformations, the corresponding residue positions have a covalently linked main chain connecting them.

Once we have a set of possible side chain and main chain choices, we can generate a graph using the representation described here, and find cliques which will represent candidates for the final model.

5.2.9 Building side chains in a comparative modelling scenario

To illustrate how the clique finding method performs in terms of building side chains, we select a comparative modelling target and a corresponding model from the First Meeting on the Critical Assessment of Protein Structure Prediction methods (CASP1). The target is the histidine-containing phosphocarrier protein (hpr) from *M. capricolum*, which is an 89 residue protein [63]. In the model we built for CASP1, 27 of 67 χ_1 angles deviated more than 30° relative to the experimental structure. We rebuild the 27 side chains using the discriminatory function, side chain sampling, and clique finding methods described above. We compare the accuracy of building the 27 side chains on the correct experimentally-determined main chain and on the approximate model main chain (which is copied over from the parent 2hpr). The side chains are built in the context of the structure: in the case of building side chains on the correct experimental structure main chain, the experimental side chain conformations for the residues not built by the CF method are used. In the case of building side chains on the model, the side chain conformations as modelled for CASP1 are used for residues not built by the CF method.

For each of the 27 residues, we sample as many conformations as necessary to ensure that the χ angle(s) in a given set of conformations are within 30° of the experimental χ angle(s). That is, we generate all possible side chain conformations and select different numbers of conformations per residue based on their negative log conditional probability score in such a way that at least one conformation is within 30° of the experimental χ angle. The rotamer library approximation we use does this automatically for all but two of the side chain positions. This means that the maximum accuracy we can achieve in numbers of χ_1 angles correctly built is 25/27. In one experiment where we build side chains on the experimental structure main chain, we include the exact experimental structure rotamers in the sample space for all 27 residues, as well as rotamers with values from the library (see Table 5.2). For nineteen of the 27 residue positions we sample two side chain conformations per residue, for seven positions we sample three side chain conformations per residue, and for one position we sample four conformations per residue. This is the equivalent of systematically exploring $4^1 \times 3^7 \times 2^{19} \simeq 5 \times 10^9$ possibilities.

5.2.10 Mixing and matching between different parent homolog structures

For one of the targets at CASP1, cellular retinoic acid binding protein I (crabpi), we found after the experiment that certain regions in the closest homolog (muscle fatty acid binding protein; PDB code 2hmb) did not match the experimental structure as well as the next-to-closest homolog did (cellular retinol binding protein II; PDB code 1opa-A). The C_α RMSD between 2hmb, the closest homolog, and the experimental structure is 2.03 Å for the 130 residues that are superimposable. The C_α RMSD between 1opa-A, the next-to-closest homolog, and the experimental structure is 1.87 Å for 130 residues. The C_α RMSD between the final model generated by us at CASP1 (which involved subjectively mixing and matching between 2hmb and 1opa-A) is 1.81 Å for the same 130 residues (we exclude regions that represent insertions in the calculation of this RMSD).

The question then is: given the two parent homolog structures to the crabpi sequence, can the graph-theoretic clique finding method mix and match between the structures and produce a model that is as good as, if not better than, the main chain model built by us at CASP1?

To answer this question, we first define crossover points where mixing between different parent structures can occur. We do this by performing a structural superposition between the 2hmb and 1opa-A structures and determine ranges of main chain where the C_α atoms are less than 1.0 Å to each other. We look for contiguous stretches of the alignment where the residues are all within 1.0 Å of each other to define the crossover points. Exceptions to the 1.0 Å limit are handled in a subjective manner by visual inspection of the 2hmb and 1opa-A structures. We define seven crossover points, leading to eight mix and match regions: 1-20, 21-41, 42-52, 53-73, 74-98, 99-107, 108-122, and 123-140.

Once the two initial models were built, some of the side chains that were built using the minimum perturbation (MP) method were found to clash in each of the models. For these fourteen residues, we sampled multiple conformations per side chain in the two separate initial models and explored all possibilities of mixing and matching the main chains and the side chains. Only three possible conformations per residue with the lowest negative log conditional probabilities for the fourteen positions were chosen due to computational limits. This equates to exploring $3^{14} \times 2^{8} \simeq 10^{9}$ conformations systematically. All other side chains were used as constructed by the MP method.

5.2.11 Building regions of main chains (loops) in an interconnected manner

We apply the CF method to a classic problem in building main chain regions, that of determining the conformation of antibody complementary determining regions (CDRs). In one experiment, we build the four CDRs on the Fv fragment of the D1.3 antibody (PDB code 1vfa) [146] simultaneously, sampling only the one side chain conformation per residue position with the lowest negative log conditional probability (see Table 5.1 for details about the CDRs being built). In another experiment, we build two of the CDRs, H3 and L3, simultaneously sampling the two side chains per residue position with the lowest negative log conditional probabilities for all residues except the proline in L3. In the former

CDR	Residue range	Number of Residues	Sequence
H2	158-166	9	MIWGDGNTD
H3	205-212	8	RERDYRLD
L2	47-55	9	LVYYTTTLA
L3	90-97	8	HFWSTPRT

Table 5.1: Details of the four complimentary determining regions (CDRs) built in the D1.3 antibody (PDB code 1vfa) using the clique finding (CF) method. The name of the CDR, the range of residues built, the number of residues, and the sequence are given.

case, the number of possible choices available is the total product of the number of main chains generated for all the loops using the database method for each of the CDRs. In the latter case, the number of possible choices available is the total product of the number of main chains generated using the database method for the H3 and L3 CDRs times 2^{15} (there is only a single conformation for the proline residue in the L3 CDR). In both these cases, the environment of the experimental structure was used to build the CDRs. The database search found 168, 216, 176, and 166 main chain conformations for the H2, H3, L2 and L3 CDRs (Table 5.4). This is the equivalent of systematically exploring $168 \times 216 \times 176 \times 166 \simeq 10^9$ conformations in the case of building the four CDRs simultaneously with only one side chain per residue and $216 \times 2^8 \times 166 \times 2^7 \simeq 10^9$ conformations in the case of building the H3 and L3 CDRs with two side chains per residue.

5.2.12 Implementation issues

The graphs are stored as edge matrices of size $n \times n$ where n is the number of the nodes. The size of a single element in the matrix, which represents an edge, is one byte, and therefore the weight of an edge is limited by the storage capacity of one byte.

Ideally, the weight of the clique should be equal to the negative log conditional probability of the conformation represented by the clique, as calculated by summing the probabilities of all the atom-atom contacts in the conformation. However, in the summation of the conditional probabilities of the nodes and the edges, contacts between pairs of atoms in main chain of two residues that are within four residues of each other are excluded from the counts. In addition, we also evaluate the conditional probabilities of atomic contacts within a residue and add it to the weights of the nodes. Both these modifications of the residue-specific all-atom conditional probability discriminatory function (RAPDF) as implemented in Chapter 3 have not been evaluated rigourously. We therefore compared the negative log conditional probabilties of consistent conformations obtained by summing up weights of the nodes and edges of the clique representing that conformation, and the negative log conditional probabilities of the same conformations obtained by calculating the conditional probabilities of the interatomic contacts as in Chapter 3 by the RAPDF. The comparison is accomplished by calculating the two types of negative log conditional probabilities of 100,000 conformations for residues 21-32 in the α-lactalbumin structure (PDB code 1alc) and plotting them against each other (Figure 5.3). The fragment in α-lactalbumin is proposed to be an independent folding unit as determined by local hydrophobic burial and experimental evidence [97, 120]. The conformations represent 100,000 cliques with the best weight obtained after exploring up to six residues per residue position with a fixed main chain. That is, each of the conformations represents a different side chain arrangement for the twelve residues in the independent folding unit.

Figure 5.3 shows that even though the correspondence between the two types

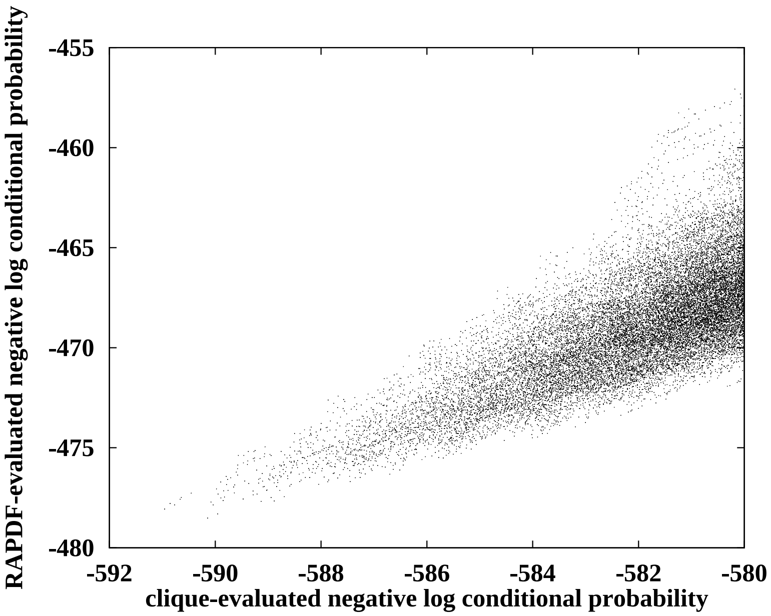


Figure 5.3: Comparison of the total negative log conditional probabilities obtained by summing the weights of nodes and edges in a clique to those obtained by summing up the probabilities of all atomic contacts in the 3D conformation represented by the clique. The negative log conditional probabilities for 100,000 side chain conformations/cliques of an independent folding unit, α-lactalbumin (residues 21-32) are shown [97, 120]. The horizontal axis is the range of conditional probabilities for the 100,000 cliques as evaluated by summing up the probabilities of the nodes and the edges. The vertical axis is the range of conditional probabilities for the 100,000 cliques as evaluated by summing up the probabilities of contacts between atom pairs for each conformation represented by the corresponding clique by the residue-specific all-atom conditional probability discriminatory function (RAPDF) as described in Chapter 3.

of conditional probabilities is not perfect, the lowest negative log conditional probability conformation as evaluated by the RAPDF can be obtained (in this specific case) by taking ten cliques with the lowest clique weights and recalculating the conditional probabilities of the conformations represented by the cliques using the RAPDF. In our implementation, we obtain the top 100 cliques with the lowest weights, and then reevaluate them using the RAPDF and select the conformation with lowest negative log conditional probability.

The storage of the top 100 cliques is accomplished through the aid of a queue data structure of size 100. Initially, the index of the clique with the lowest weight in the queue is calculated (if there are no cliques in the queue, then the index is set to 1). New cliques with lower weights are added to the queue by simply replacing the clique in the queue with the lowest weight with the new clique and recalculating the index of the clique with the lowest weight.

In cases where cliques the size of the protein cannot be found due to exclusion of certain possible residue conformations (nodes) which are inconsistent with the rest of the nodes, then more or different possible conformations for that residue position are generated using the sampling methods described in this work so that the resulting cliques found will be of the size of the protein. Alternately, one can obtain cliques that are smaller in size than the length of the protein, which will not contain atomic coordinates for some residues, and either construct those residue conformations from the set of existing nodes (which will lead to inconsistent protein conformations as per our definition) or by using other methods. In specific cases, pairs of nodes considered inconsistent by the CF method are explicitly (manually) allowed to form an edge. This is generally necessary when *any* consistent conformations the size of the protein cannot be found regardless of the degree of sampling.

Two residues separated by a large distance in the protein that do not physically interact with each other have an edge between them. For proteins that are bigger than, say, 200 residues, this results in a large number of edges per node and increases the running time of the program. In such cases, we consider only a limited subset of the protein and omit all residues beyond a certain cutoff (say 20.0 Å) from the region we are interested in modelling using the CF method.

When the above implementation issues arose during the course of building side chain and main chain conformations using the CF method, they were handled on a case by case basis and are not further discussed in the methods or results sections in this work.

5.3 Results

5.3.1 Building side chains

For the 27 residues that were built using the CF method in the histidine-containing phosphocarrier protein (hpr), the percentage error, i.e., the number of angles that deviated more than 30° after mixing and matching between all the possible side chains on the experimental structure main chain was 29.6% for χ_1 angles, and 42.1% for all χ angles. The percentage errors for the 27 side chains on the model main chain was 40.7% for χ_1 angles and 49.1% for all χ angles.

When the exact experimental structure rotamer was added to the sample space, by removing the library rotamer corresponding to the experimental structure rotamer, the percentage error decreased in the case of the correct main chain to 25.9% for χ_1 angles and 31.6% respectively for all χ angles. For the model main chain, the percentage error decreased to 31.6% for χ_1 angles and 42.1% for all χ angles. Table 5.2 summarises the results of the side chain construction.

Even when we introduce the correct experimental side chain conformation in the sample space, we are unable to build the conformations of 7/27 χ_1 angles and 18/57 all χ angles. We analyse the eighteen χ angles to determine the reason they were built incorrectly. Table 5.3 shows the results of this analysis.

Thirteen of the χ rotamers incorrectly built have an atom with a temperature factor of more than 30.0 Å^2. In twelve cases, the side chains are involved in

	$\chi_1 > 30°$ (%)	all $\chi > 30°$ (%)
Correct main chain with library rotamers	29.6	42.1
Correct main chain with correct rotamers	25.9	31.6
Model main chain with library rotamers	40.7	49.1
Model main chain with correct rotamers	40.7	42.1

Table 5.2: Results of side chain construction for 27 residues using the clique finding (CF) method for the histidine-containing phosphocarrier protein (hpr). All 27 side chains had χ_1 conformations that deviated by more than 30° in the model built by us at the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1) relative to the experimental structure. The percentage errors for χ_1 and all χ angles is given for the cases where the correct and the model main chains were used to the build side chains conmformations using the rotamer library described in Table 4.2. The percentage errors is also given for the cases where the correct and the model main chains were used to build the side chains conmformations including the experimental conformation in the sample space. The total number of χ_1 angles considered is 27 and the total number of χ angles is 57. For all residues, the side chain conformations in the environment when building side chains on the experimental main chain is the same as in the experimental structure of hpr, and the side chain conformations in the environment when building side chains on the model main chain is the same as in the final model of hpr built by us for CASP1.

intermolecular crystallographic contacts of less than 4.0 Å, as determined using the program CONANA [147]. In nine cases, atoms involved in the rotamers are close to water molecules or the sulphate ion in the experimental structure (which are not taken into account by our discriminatory function in a direct manner). All the rotamers built inaccurately may be affected by one or more of these factors.

The total negative log conditional probability of the protein conformation with the eighteen incorrectly built rotamers is lower than for the experimental

χ	Residue	Largest	Discrimination	Number of	Observation
angle		B (Å^2)	ratio	xtal contacts	
χ_2	I7	24.3	1.01	1	residue on surface of protein
χ_1	L14	20.5	0.99	3	deviation of 33°
χ_2	L14	22.7	0.99	3	intermolecular contacts
χ_1	S30	32.6	0.99	4	high B; intermolecular contacts; 2 H_2O molecules within 3.6 Å
χ_2	I36	40.0	0.99	0	high B; residue on surface of protein
χ_2	N38	42.2	1.01	4	high B; 180° rotation
χ_1	E39	30.5	1.01	6	intermolecular contacts
χ_2	E39	35.0	1.01	6	intermolecular contacts
χ_3	E39	52.2	1.01	6	high B
χ_1	I47	40.5	0.99	3	high B; SO_4 ion within 3.8 Å
χ_2	I47	40.5	0.99	3	high B; SO_4 ion within 4.0 Å
χ_3	M48	69.1	1.01	5	high B; 6.2 Å to SO_4 ion
χ_1	D66	41.0	1.01	0	high B
χ_2	D66	45.6	1.01	0	high B; 2 H_2O molecules within 3.8 Å
χ_1	N68	46.3	1.01	0	high B; H_2O molecule within 4.0 Å
χ_3	Q72	32.9	0.99	8	intermolecular contacts
χ_1	I87	22.8	1.01	0	4 H_2O molecules within 3.0-5.0 Å
χ_2	I87	22.8	1.01	0	4 H_2O molecules within 3.0-5.0 Å

Table 5.3: Analysis of χ angles that were incorrectly built for 27 residues in the histidine-containing phosphocarrier protein (hpr) using the clique finding (CF) algorithm. The χ angle, the residue number and name (in one letter code), the largest temperature factor among the atoms defining the χ angle, the ratio of the negative log conditional probabilities between the expermental structure with the incorrect side chain conformation and the experimental structure with the correct side chain conformation (a ratio less than 1.0 indicates successful discrimination in the context of the rest of the structure), the number of intermolecular crystallographic contacts less than 4.0 Å involving the residue, and observations regarding influence of water, the sulphate ion, temperature (B) factors, and number of intermolecular contacts is given.

structure conformation. When single rotamers are considered in the context of the experimental structure, the discriminatory function is unable to distinguish correct rotamer from the incorrect rotamer in 11/18 cases, as determined by the ratios of the negative log conditional probabilities between the experimental structure with the incorrect side chain conformation and the experimental structure with the correct side chain conformation (a ratio less than 1.0 indicates successful discrimination; see Table 5.3). In all but two cases (the two χ angles in isoleucine 87) where the discriminatory function selected the correct rotamer given the exact experimental structure environment, the temperature factors are greater than 30.0 Å^2 and/or there are intermolecular contacts involving the side chain of the residue. This is not to suggest that experimental structure is incorrect or that the discriminatory function is not failing, but that it makes it difficult to assess what the cause of failure is.

We compare specific contacts involving side chain conformations that were built using the CF method and the correct experimental side chain conformations to structurally rationalise why the experimental conformation is preferred over the built conformation. These contacts are thought to be the most influential in determining the side chain conformation of the residue of interest.

The χ_2 rotamer of isoleucine 7 varies by 52.5° between the experimental conformation and the built conformation, one of the two clear non-discrimination cases. As a consequence, the side chain $C_{\gamma 2}$ and $C_{\delta 1}$ atoms, which are the only ones that are different in the two conformations, are 1.1 Å apart. The $C_{\delta 1}$ atom of the isoleucine is in contact with the $C_{\gamma 2}$ atom of threonine 9, and the distance between these two atoms is 3.7 Å for the experimental conformation and 4.3 Å for the built conformation.

The situation with the χ_2 rotamer of isoleucine 36, the other clear non-discrimination case, is very similar. The difference in the χ_2 angle between the experimental and the built conformations is 53.5° and the side chain $C_{\gamma 2}$ and $C_{\delta 1}$ atoms, which are the only ones that are different in the two conformations, are 1.2 Å apart. In this case, the $C_{\gamma 2}$ atom of the isoleucine is in contact with the C_{β} atom of threonine 62 at a distance of 4.6 Å in the experimental conformation whereas the distance between the same two atoms in the built conformation is 3.5 Å .

Given the fact that both the side chains (isoleucine 7 and 36) are on the surface of the molecule it is difficult to rationalise why one conformation is preferred over the other. It could be that the experimental conformation has better packing between the side chain carbon atoms in the isoleucine residues and the side chain carbon atoms in the threonine residues, compared to the built conformation.

In the case of the χ_2 rotamer in asparagine 38, which is also present on the surface, the nitrogen and oxygen atoms in the side chains are in opposite positions to each other (the χ_2 angle is rotated by 180° in the built side chain relative to the experimental conformation). In the experimental structure, the distance between the $O_{\delta 1}$ atom of the asparagine and the N_{ζ} atom of lysine 40 is 3.4 Å, whereas in the built conformation, the $N_{\delta 2}$ atom the of asparagine and the N_{ζ} atom of the lysine are at a distance 3.7 Å, resulting in a contact with bad electrostatics.

In the case of isoleucine 87, where the side chain is partially exposed, the χ_1 angle differs by 225.3° (a complete turn around) in the built conformation relative to the experimental and the χ_2 angle differs by 55°. In this case, the contacts between the $C_{\delta 1}$ and the $C_{\gamma 1}$ atoms in isoleucine 87 and the $C_{\delta 1}$ and the $C_{\gamma 1}$ atoms in isoleucine 82 appear to be more well packed in the case of the experimental conformation compared to the built conformation.

The conformations of threonine 9 and 62, lysine 40, and isoleucine 82 are not among the 27 conformations that were built using the CF method and actually represent the experimental structure conformation.

In all the above cases, the specific contacts mentioned above in experimental conformation have lower negative log conditional probabilities compared to the same contacts in the built conformation. However, the total negative log conditional probability in the built conformation is lower when summed over many other contacts between the atoms in the environment and the side chain atoms relative to the experimental structure in the case of isoleucine 7, asparagine 38 and isoleucine 87.

5.3.2 Mixing and matching between homologs

The C_α RMSD of the structure of the cellular retinoic acid binding protein I (crabpi) we modelled for the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1) relative to the experimental structure for the 130 residues (excluding insertions) was 1.80 Å. The C_α RMSD of the conformation built by mixing and matching between the two templates using the CF method was 1.66 Å for the same 130 residues.

The theoretical limit for the C_α RMSD of mixing and matching between the main chains given the designated crossover points is 1.54 Å. This is determined by considering the C_α RMSDs of the conformations created by mixing and matching between the two crabpi homologs for all the $2^8 = 256$ possibilities relative to the experimental structure.

The optimal conformation (in terms of C_α RMSD) contains clashes because the side chains are built using the MP method, and has a worse (higher) energy than the conformation built by the CF method. This shows that side chain building is also necessary in order to mix and match between template structures before a discriminatory function can accurately identify the lowest RMSD conformations.

The difference between the CF built structure and the optimal conformation is in two segments, residues 1-20 and 21-41. These should have been selected from

1opa-A, the next-to-closest homolog, but are selected from 2hmb, the closest homolog. The C_α RMSDs for these two segments in 2hmb relative to the crabpi experimental structure are 1.2 Å and 2.4 Å respectively. The C_α RMSDs for these two segments 1opa-A relative to the crabpi experimental structure are 0.9 Å and 1.9 Å respectively. There are 21 residues involved in clashing contacts in the optimal conformation, and the multiple side chain conformations for twelve of these residues are not explored by the CF method to build the mix and match model. One of the residues involved in a clash in the optimal conformation is proline 2 which clashes with tryptophan 88 (which is a conserved residue with the correct conformation in the two initial models), and since we do not sample multiple positions for proline residues, the segment containing that residue (1-20) will never be selected from the 1opa-A-based initial model.

5.3.3 Building regions of main chain

Table 5.4 shows the C_α RMSDs for the four CDRs in the situation where they were built simultaneously with only a single side chain conformation sampled per residue and the RMSDs for the H3 and L3 CDRs where they were built simultaneously sampling two side chain conformations per residue. Also listed are main chain (N, C_α, C, O) and all atom RMSDs for the structures so they can be compared to other methods constructing the CDRs. The conformations of the H3 and L3 loops built using the CF method in two different experiments are identical as far as the main chain is concerned. The largest C_α RMSD is 1.33 Å and the largest main chain RMSD is 1.42 Å among the four CDRs between the built conformation and the experimental structure based on a global superposition of the framework upon which the CDRs were built.

CDR	Number of main chain conformations	C_α RMSD range (Å)	C_α RMSD (Å)	Main chain RMSD (Å)	All atom RMSD (Å)
One side chain per residue					
H2	168	0.40 - 6.23	1.33	1.42	1.94
H3	216	0.59 - 5.43	1.01	1.20	2.43
L2	176	0.66 - 5.28	1.10	1.54	2.67
L3	166	0.70 - 5.24	0.86	1.13	2.70
Two side chains per residue					
H3	216	0.59 - 5.43	1.01	1.20	2.65
L3	166	0.70 - 5.24	0.86	1.13	2.78

Table 5.4: Results of simultaneously building complimentary determining regions (CDRs) in the D1.3 antibody structure using the clique finding (CF) method. The name of the CDR, the number of main chain conformations sampled, the C_α RMSD range of the conformations sampled, and the C_α, main chain (N, C_α, C, O), and all atom RMSDs of the conformation selected using the CF method is given. The results are given for two experiments, where four CDRs (H2, H3, L2 and L3) were built simultaneously with one side chain per residue and where two CDRs (H3 and L3) were built simultaneously with two side chains per residue (see Figure 5.4).

The largest all atom RMSD is 2.70 Å for any single loop when the four CDRs are built simultaneously sampling only a single side chain conformation per residue. However, when two side chain conformations per residue are sampled when building the H3 and L3 CDRs, the all-atom RMSD increases slightly. Five side chain conformations are different in the two situations. Analysing the accuracy of each residue individually, three side chain conformations (phenlyalanine 91, arginine 207, and leucine 211) are built with a higher side chain atom RMSDs (relative to the experimental structure) and two side chain conformations (histidine 90 and arginine 210) are built with lower side chain atom RMSDs when multiple side chain conformations are sampled (leading to an overall higher side chain atom RMSD when multiple conformations are sampled). The discriminatory function is unable to distinguish, in the context of the approximate environment, the side chain conformation with the lower RMSD for these three cases. However, when the correct experimental main chain conformation is used to build side chains for all these three residues using only the local main chain information, then both phenylyalanine 90 and leucine 211 are built within 30° of all the experimental χ angles for those residues. In the other case, where the side chain atom RMSD is higher when multiple side chain conformations are sampled (arginine 207), the conformation in the experimental structure may be influenced by the presence of large (> 50.0 Å^2) temperature factors in the side chain atoms.

The C_α and all-atom RMSDs for all the four CDRs (34 residues) relative to the experimental structure as found by the CF method when sampling single side chain conformations per residue position are 1.10 Å and 2.46 Å (see Figure 5.4). The best all-atom atom RMSD that can be obtained, given our rotamer library approximation, and using the database main chain conformations, for all the four CDRs, is 1.71 Å.

The database used to search for main chain conformations contained several antibody conformations but did not include D1.3 antibody structures. The source of the conformations selected by the CF method for each of the four loops in the two different situations are given in Table 5.5. Loops from other antibodies are selected only in the case of the L3 CDR.

CDR	Loop source (PDB code and name of protein)	Residue range in source	Sequence of source	Sequence of CDR
One side chain per residue				
H2	1tgs - Trypsinogen	143-151	NTKSSGTSY	MIWGDGNTD
H3	1npx - NADH peroxidase	329-336	LAVFDYKF	RERDYRLD
L2	1aaz - Glutaredoxin	36-44	IMPEKGVFD	LVYYTTTLA
L3	1rei - Immunoglobulin	90-97	QYQSLPYT	HFWSTPRT
Two side chains per residue				
H3	1npx - NADH peroxidase	329-336	LAVFDYKF	RERDYRLD
L3	1rei - Immunoglobulin	90-97	QYQSLPYT	HFWSTPRT

Table 5.5: Details of the sources for the four complimentary determining regions (CDRs) built using the clique finding (CF) method. The name of the CDR, the source (PDB code and name of protein) of the main chain selected by the CF method, the range of residues in the source that matched the selected conformation, the sequence of residues in the source, and the sequence of the CDR in the D1.3 antibody is given.

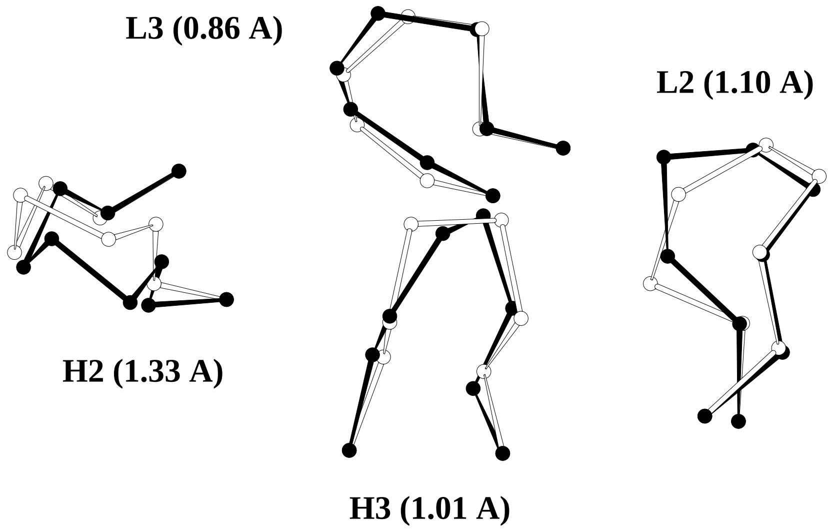


Figure 5.4: Comparison of conformations built (white) using the clique finding (CF) method to the experimental structure (black) for four complimentary determining regions (CDRs) in the D1.3 antibody. Shown are C_α traces of four CDRs, H2 (residues 158-166), H3 (residues 205-212), L2 (residues 47-55) and L3 (90-97) which were built simultaneously using the CF method. The individual C_α RMSDs of the CDRs are 1.42 Å, 1.01 Å, and 1.10 Å and 0.86 Å for each of the CDRs with an overall C_α RMSD of 1.10 Å for all the 34 residues relative to the experimental structure. The C_α RMSDs do not include the root residues and are based on a global superposition.

Experiment	Number of conformations	CPU Time (hh:mm:ss)
Side chain building on hpr	$\simeq 5 \times 10^9$	31:02:14
Mixing and matching crabpi templates	$\simeq 10^9$	18:01:42
Building all four CDRs simultaneously with one conformation per residue	$\simeq 10^9$	33:12:37
Building two CDRs with two conformations per residue	$\simeq 10^9$	37:29:34

Table 5.6: Computation times of the clique finding (CF) method for the three comparative modelling scenarios described in this chapter. The name of the experiment, the number of conformations in the sample space, and the CPU time (hh:mm:ss) as measured using the "time" command on a Silicon Graphics (SGI) Challenge workstation is given.

5.3.4 Computation times

The computation time of this method is proportional to the density (number of edges per node) of the graph. Each graph representing various possibilities for a given structure varies in density. The times, determined using the "time" command in an Unix system for performing the computations of the various experiments are given in Table 5.6. All times were calculated on a multi-user Silicon Graphics (SGI) Challenge workstation utilising a single R10000 processor. In general, finding a consistent structural arrangement of residues in an amino acid sequence with the lowest negative log conditional probability, sampling one billion to ten billion ($10^9 - 10^{10}$) possible conformations, can be accomplished within a 24-48 hour period.

5.4 Discussion

5.4.1 Building side chain conformations

The graph-theoretic clique finding (CF) method described here is a search method that allows the exploration of a large number of possible conformations of a protein. The accuracy of the side chain construction made in this exercise depend on the discriminatory function used. In the case of histidine-containing phosphocarrier protein (hpr), a target for which we constructed a model at the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1), we rebuild the 27 residues with incorrectly built χ_1 angles in the model (Chapter 2). Using the CF method improves the results in terms of percentage error in the χ_1 angles for the 27 residues by at least 60%. Many of these residues represent drastic amino acid substitutions (alanine to the phenylalanine, for example) and since it is in those cases that the conventional minimum perturbation method generally fails, the CF method should be used. The CF method complements existing comparative modelling methods, to build side chain conformations by performing some limited combinatorial searching. While we know from CASP1 that even conformations of conserved amino acid substitions can change between related proteins 25% of the time (Chapter 2), achieving this level of accuracy in a comparative modelling situation is our first goal.

We also build side chain conformations for hpr using the experimental structure main chain. The improvement when sampling multiple (2-4) side chain conformations per residue position compared to using a single side chain conformation with the lowest negative log conditional probabilities for the contacts between atoms within the side chain to the atoms in the local main chain (± four residues) is not significant: the percentage error of χ_1 and all χ rotamers built incorrectly ($> 30°$ relative to the experimental structure) decreases only 3.7% and 5.2% respectively when multiple side chain conformations are sampled.

Modelling side chains on the experimental main chain does not represent a realistic scenario. We know from previous studies that side chain building accuracy in a comparative modelling situation decreases rapidly as the main chain varies [60]. We have shown in Chapter 4 that the side chain building method we employ here in conjunction with the CF method is comparable to other side chain building methods in the literature where the results are described for rebuilding side chains on the experimental structure main chain [50, 51, 52, 53]. Using the CF method to sample multiple side chain conformations using the method in Chapter 4 does not improve side chain building accuracy significantly when rebuilding side chains on the experimental structure main chain. However it does improve side chain construction when an approximate main chain is used.

5.4.2 Mixing and matching

The utility of mixing and matching between template parent structures to construct a model that is better than simply copying the coordinates of a single parent structure has been shown before [49].

At the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1), the closest homolog to the cellular retinoic acid binding protein I (crabpi), the muscle fatty acid binding protein (PDB code 2hmb) was generally used by the different groups participating to model the main chain for all residues (excluding insertions in the alignment between crabpi and 2hmb) [8]. In this case, the main chain of 2hmb can be mixed and matched with another homolog (cellular retinol binding protein II; PDB code 1opa-A) to produce a template model structure with a lower RMSD to the crabpi experimental structure than using either of the homologs by itself. The CF method used to mix and match between the two different templates was done without any manual intervention, and can be used in a comparative modelling scenario where multiple template/parent structures are available for modelling.

Identification of the best model by mixing and matching between regions in the initial model using a discriminatory function is not trivial. In our case, the side chains in the environment were approximate and the initial models were missing atomic coordinates for regions of insertions.

5.4.3 Building main chain regions

The results shown in Tables 5.4 and 5.5 compare favourably with the results for building the antibody complimentary determining regions (CDRs) on a D1.3 antibody structure using the most homologous canonical loops in other antibody structures [58]. For D1.3, Pedersen, et. al report main chain (N, C_α, C, O) RMSDs of 1.41, 0.93, and 1.14 for the H2, L2, and L3 CDRs using canonical construction. However, the only case where the CF method selects a main chain conformation from another antibody (PDB code 1rei) is with L3, where it finds a match of the same CDR with a similar sequence (Table 5.5). All selections were made on the negative log conditional probabilities and no homology information was included.

The main chain RMSDs reported by Pedersen, et. al. [58] show that the CF method is not the only way conformations with low main chain RMSDs can be built for antibody CDRs. However, the example we choose illustrates how the method can be used to build side chains and main chains simultaneously for regions in a protein that are non-local in sequence but interacting with each other in the native conformation. Due to the limited sampling of the side chain conformations, it is not possible to build conformations with an all-atom RMSD of $\simeq 1.0$ Å, and therefore better side chain sampling algorithms need to be developed.

Using the database method to sample main chain conformations in this particular case makes use of a knowledge-based component for building antibody loops.

5.4.4 Sampling side chains and main chains

We have used a side chain sampling method that we developed in Chapter 4 based on selecting the most probable conformation using only the local main chain and using a well-established main chain sampling procedure based on a database search [57, 58]. However, side chain and main chain conformations that need to be sampled can be generated by any means. The number of possibilities that can be handled is limited, and to overcome this, we apply this method in a comparative modelling scenario where distinct structurally context-sensitive units can be built separately.

5.4.5 Tractability and complexity of clique finding

Clique finding in a graph is an NP-hard problem with a worst-case estimate of $O(3^{n/3})$, where n is the number of nodes in the graph [148, 145, 149]. The big-O estimate indicates that even the best algorithm for finding all the cliques in a graph will take at least $k \times 2^{n/3}$ time, where k is some constant, in the worst case. For building the main chains and side chains at CASP2, a typical graph had around 5000 nodes and we were able to search graphs with 30,000 nodes using an SGI Challenge R10000 workstation within a 24-hour period. None of the graphs we have encountered represent worst-case scenarios, i.e., they do not take time in the order of 3^n, where n is the number of nodes. This is presumably due to the nature of the representation and its relation to protein structure, and illustrates that big-O and NP-hard estimates, which apply in the worst case scenarios, are not necessarily relevant to biological problems.

5.4.6 Choice of the Bron and Kerbosch algorithm for clique finding

There is no rigourous proof of the time taken for the Bron and Kerbosch algorithm in the average case scenario, but plots given in [145] show that it works well in practice. In one test case, the authors generate a number of random graphs and the computing time per clique remains linear in the size of the graph. In a second test case where special graphs of size $3 \times n$, which contain the largest number of cliques per node, are used, the computing time is proportional to 3.14^n ms, where 3^n ms would be the theoretical limit for these graphs [148, 145, 149].

In the case of a practical application involving graph-theoretical techniques to compare protein structures, this algorithm is reported to produce the best performance among several different clique finding algorithms [144]. Also, as we will demonstrate in the next chapter, this algorithm seems to perform well in the case of realistic homology modelling problems.

5.4.7 Sizes of problems that can be handled in reasonable amounts of times

Tables 5.6 and 6.6 give an indication of the sizes of problems that can be handled using the CF method. The size of problem generally depends on the number of residues being built, and the number of main chain and side chain possibilities considered. In general, the CF method can handle problems that are the equivalent of exhaustively sampling $10^9 - 10^{10}$ possible conformations of a protein in a 24-48 hour period on an SGI Challenge (R10000 processor) workstation.

5.4.8 Advantages of this method compared to conventional search methods

There are three primary advantages of this method to traditional methods that search conformational space in proteins: First, the calculation of the fitness of a conformation is extremely fast. Since the weights are precalculated in advance for the nodes and the edges, the weight of a clique is calculated simply by summing the edges of the nodes. For this to be useful, the discriminatory function that allows the calculation of weights must work in a pairwise manner and must be additive [101, 102, 103, 104]. Second, inconsistent conformations are never evaluated for their weights and are rejected in advance—i.e., they are never found as cliques in the graphs. This has the advantage in that the density of the graph, and consequently the speed of the algorithm, can be controlled by applying filters to eliminate edges before clique finding occurs. Third, the conformations represented by the cliques are found in a discrete manner (depending on the main chain and side chain sampling) and this allows the method to "jump through" conformational space without regard to energetic barriers and ensures it does not get stuck in local minima. Finally, in a comparative modelling situation, there is the advantage of this method in that it takes into account the context of the environment when building main chains or side chains.

5.4.9 Limitations of this method

The foremost limitation of this method is in the fact that clique finding itself is an intractable problem computationally. Even though the worst-case big-O estimate does not apply in the cases we encounter, the size of problems that can be solved with current computing abilities is limited to the equivalent of exploring 10^{10} conformations. Also, as mentioned earlier, the discriminatory function used must be able to represent weights of nodes and edges independently of other nodes and edges [101, 102, 103, 104, 105, 106, 107, 108, 109, 111]. This eliminates discriminatory functions that base their calculation of strengths of interatomic contacts in a context-sensitive manner, such as discriminatory functions that use the accessible surface area of atoms as a measure of solvation preference [102, 103, 104, 119, 120].

5.5 Summary

The interconnected nature of interactions in protein structures appears to be the major hurdle preventing the construction of accurate homology models. We present an algorithm that uses graph theory to handle this problem. Each possible conformation of a residue in an amino acid sequence is represented using the notion of a node in a graph. Each node is given a weight based on the degree of the interaction between its side chain atoms and the local main chain atoms. Edges are then drawn between pairs of residue conformations/nodes that are consistent with each other (i.e., clash-free and satisfying geometrical constraints). The edges are also weighted based on the interactions between the atoms of the two nodes. Once the entire graph is constructed, all the maximal sets of completely connected nodes (cliques) are found using a clique finding algorithm. The cliques with the best weights represent the optimal combinations of the various main chain and side chain possibilities, taking the respective environments into account. The algorithm is used in a comparative modelling scenario to build side chains, regions of main chain, and mix and match between different homologs in a context-sensitive manner.

In the next chapter, we assess the predictive power of this approach by applying it to blind tests where the experimental structure is not known in advance.

Chapter 6

Handling context-sensitivity in protein structures using graph theory: bona fide prediction

6.1 Introduction

Comparative models of five structures, polyribonucleotide nucleotidyl s-transferase (pns1/target 4; 76 residues [150]) from *E. coli*, neurocalcin delta (ncd/-target 7; 193 residues) from *B. taurus*, cucumber stellacyanin (csc/target 9; 109 residues [151]) from *C. sativus*, ubiquitin conjugating enzyme (ubc9/target 24; 158 residues [152]) from *M. musculus*, and endoglucanase I (egi/target 28; 371 residues [153]) from *T. reesei*, were built. We used the graph-theoretic clique finding (CF) method described in Chapter 5, in conjunction with the discriminatory function and side chain sampling method described in Chapters 3 and 4 to build some side chains and main chain segments after constructing an initial model by copying a subset of the atomic coordinates from the parent structure(s).

6.2 Methods

6.2.1 Search for parent sequences with known structure

Target protein sequences were obtained from the web page provided by the CASP2 organisers [154]. A basic BLAST search [155], using the program blastp and the default BLOSUM62 scoring matrix, was performed on the PDB [113] to identify parent sequences with known structures that are related to the target sequence.

In one case, pns1/t4, where no apparent homology could be detected by conventional sequence searches, distantly related sequences with known structure were found using the Hidden Markov Model (HMM) package HMMER [156]. A maximum discrimination HMM was first constructed from a multiple sequence alignment of seventeen sequences related to the pns1/t4 family. The multiple sequence alignment was obtained from the PredictProtein server [157], which uses the MaxHom program [158] for performing sequence alignments by extracting all sequences in the Swissprot sequence database [159] that have a percentage identity of 30% or more to the sequence submitted. The HMM, which is a statistical model of the sequence variability at each position in the pns1/t4 multiple sequence alignment, was then aligned using the Smith/Waterman algorithm [160] to the set of sequences in the PDB using HMMER. The two highest scoring sequences based on this alignment were considered to be distant homologs.

6.2.2 Sequence and structure alignment

Multiple sequence alignments were generated with the AMPS package [70, 71]. The AMPS-derived alignment was used to identify regions of sequence variability within the target sequence family. AMPS pairwise alignments were also used to determine the degree of sequence identity between the target sequences and the parent sequences with known structure (see Table 2.1). The default PAM250 mutation matrix and a length independent gap penalty of 8.0 were used. In the case of target sequences with multiple parent structures, structural alignments between the parent structures were generated using the G program [72]. These structural alignments were used to examine the structural variation at a given residue position to determine regions that are structurally conserved and regions that are not. The structure and sequence conservation for each residue was examined to identify main chain regions that might require rebuilding.

Visual inspection of the initial AMPS alignments revealed regions in two cases (pns1/t4 and egi/t28) where we thought the alignment was dubious. The alignment in these regions was adjusted manually.

In the case of pns1/t4, an insertion of two residues in pns1/t4 relative to 1csp in the sequence alignment was moved from residues 9-10 to residues 17-18, because the AMPS alignment placed the insertion in the middle of a β-strand. The single residue insertion at residue 21 was moved to residue 26 for the same reason (see Figure 6.1a).

For egi/t28, we noticed that aligning an identical stretch of four residues with sequence QNGV (residues 275-278 in egi/t28; see Figure 6.1e) between the target sequence and the parent sequence led to a higher degree of percentage sequence identity for the entire alignment. We therefore made this correction by introducing an insertion and a deletion as shown in Figure 6.1e.

6.2.3 Construction of an initial model

Following the sequence alignment, for each parent structure, an initial model was generated by copying atomic coordinates for the entire main chain (excluding any insertions) and for side chains where the residues that represent sequence identities from the parent structure. Residues that differ in sequence were constructed by mutating the residues using a minimum perturbation (MP) technique implemented by the program MUTATE [74]. The MP method changes a given amino acid to the target amino acid preserving the values of equivalent χ angles between the two side chains. The χ angles not present in the model are constructed by MUTATE using an internally developed library based on residue type.

6.2.4 General description of the graph-theoretic clique finding approach

Each possible conformation of a residue represents a node in the graph. Residues can have different main chain and side chain conformations. The nodes are weighted based on the strength of the interaction between pairs of atoms within the residue side chain and between the side chain and the local main chain atoms.

Edges are drawn between every pair of residue conformations if there are no clashes between atoms of the interacting residues and if the interaction between the two residues is covalently acceptable. A clash is said to occur if there are two non-hydrogen atoms, belonging to two different residues, with a contact of less than 2.0 Å. Contacts between pairs of atoms in the main chain of neighbouring residues are not evaluated for clashes. If the interaction weight of a side chain with the local main chain is extremely positive (> 10.0), then an edge is not drawn between the nodes. If two residue positions are within one main chain region being built, then both their conformations must be connected by a single covalently linked main chain conformation before an edge can be drawn between them. Edges are also not drawn between different possible conformations of the same residue.

Once a graph representing the various possible side chains and main chains is constructed, we search for maximal completely connected graphs (cliques). Cliques the size of the target structure (which are largest sized cliques that can be found in this representation) represent self-consistent arrangements of the individual amino acid conformations. That is, they represent possible candidates for the final structure. The clique with the best weight is taken to represent the correct conformation. In practice, only a subset of residues can be constructed because of computational limitations. A full description of the method is given in Chapter 5.

There are four main components to the implementation of the CF method for structure prediction. They are:

- A discriminatory function for assigning weights to nodes and edges.
- A method for sampling side chain conformations.
- A method for sampling main chain conformations.
- A method for finding cliques.

In a comparative modelling scenario, only those main chain and side chain conformations that are thought to vary from the parent structure are built using this method.

6.2.5 Description of discriminatory function

Our objective here is to assign weights to nodes and edges by determining the strength of the interaction of a side chain in a node to the local main chain and by determining the strength of interaction between two nodes/residues that form an edge. To do this, we use an all-atom distance dependent conditional probability-based discriminatory function which is used to calculate the probability of contacts of a given distance between pairs of atom types in a protein conformation. A detailed description of this formalism is given in Chapter 3.

The weight of an edge or a node is assigned by summing over the conditional probabilities of the appropriate atomic contacts. In the case of a node, it is the set of interatomic distances between the side chain conformation and the local main chain, and distances between atom types within a residue. In the case of an edge between nodes, it is the set of distances between pairs of atoms in the two residues.

6.2.6 Building side chain conformations

Multiple side chain conformations for a given residue position were generated by exploring all the possible side chain conformations given the rotamer library approximation and selecting the most probable conformations based on the interactions of a given conformation with the local main chain. For each χ angle in a side chain conformation, up to three rotamers were considered using the rotamer library in Table 4.2. For each possible side chain conformation, the interactions between the atoms within the side chain and between the side chain and the local main chain (± four residues, total of nine, where available) were evaluated using the conditional probability discriminatory function described above. The side

Name of target	Number of side chains	Number of conformations
egi/t28	18	$6^3 \times 4^3 \times 3^7 \times 2^5 \simeq 10^9$
ubc9/t24	18	$6^2 \times 5^2 \times 4^2 \times 3^7 \times 2^5 \simeq 10^9$
csc/t9	15	$6^4 \times 5^2 \times 3^9 \simeq 6 \times 10^8$

Table 6.1: Details of side chain sampling for three CASP2 targets. For each target (egi/t28, ubc9/t24, csc/t9), the number of side chains and the number of conformations explored is given. All side chains were built on main chain that was copied from the parent structure. The total number of side chain conformations explored is calculated by taking the product of the number of side chain conformations explored per residue (specified by the mantissas in column 3) for all residues whose side chains were built using the CF method (specified by the sum of the exponents in column 3) in the context of the rest of the model. For example, in the case of egi/t28 three residues with six conformations each, three residues with four conformations, seven residues with three conformations and two residues with five conformations each were used to construct the graph which was handed over to the CF method.

chain conformations with the lowest negative log conditional probability were taken to represent the most probable conformations. A detailed description of this side chain sampling method is given in Chapter 4.

Fifteen (in the case of csc/t9) to eighteen side chains (in the case of ubc9/t24 and egi/t28) were identified by a preliminary environmental analysis of the initial model as positions for sampling. The environmental analysis was performed visually using interactive computer graphics, identifying side chains with implausible packing, clashes, and unfavourable electrostatic interactions (hydrogen bonding, salt bridges) with other side chains and/or main chain. Between two to six different most probable side chain conformations were built for each such residue position. The optimal arrangement of the fifteen to eighteen side chain conformations sampled was determined using the CF method in the context of the rest of the initial model. Table 6.1 gives the details of side chain sampling for each target.

6.2.7 Building main chain conformations

In two cases (csc/t9 and egi/t29), the initial model with the CF built side chains was used as a template for building regions of insertions, deletions, and regions of suspected main chain uncertainty. In one case (ubc9/t24), two initial models were created from the two different parent structures (PDB codes 1aak and 2uce). Main chain regions selected for rebuilding were deleted from the initial models. The side chains for these models were built using the MP method for all but ten side chains, where two side chain conformations per residue position were built using the side chain sampling method described in the previous section. A total of thirteen main chain regions from these two models were mixed and matched using the CF method. A graph was constructed based on these possible main chain and side chain conformations and was searched for maximal completely connected subgraphs representing plausible conformations of the model given the side chain and main chain choices per residue position. The conformation represented by the clique with the lowest negative log conditional probability was used as a template for further building of main chain regions.

For three regions (csc/t9 residues 1-2, 106-108; ubc9/t24 residues 164-166), main chains were sampled using a simple combinatorial main chain grid search, with a 60 degree grid. Since only the terminal residues were built in this manner, there is only one root residue flanking the region being built in the initial model. The conformations obtained from the grid search were fitted on to the N, C_α, C, O, and C_β atoms of the root residue. In the case of csc/t9 residues 1-2, residue two was built manually in an extended conformation on the initial model using the program QUANTA [76] and the conformations from the grid search were fitted using N, C_α, C, O, and C_β atoms of the manually built residue position.

The ninteen other main chain regions were built by searching a database of main chain regions using distance constraints from the parent structure [58]. The matching main regions were positioned in the model structure using the method of Martin, *et. al.* [57].

Once the rebuilt main chain regions were sampled, side chain conformations within the main chain and 2-10 side chain conformations that were believed to be in contact with the main chain being built were also sampled using the methods described in the previous section. In five cases, multiple regions of insertions and deletions were built simultaneously. The optimal arrangement of the possible side chains and main chains was determined using the CF method by selecting the conformation corresponding to the clique with the lowest negative log conditional probability. Further detail on the main chain sampling for the 22 main chain regions is given in in Tables 6.6 and 6.5, along with the results of the prediction.

6.2.8 Clique finding

Clique finding was accomplished using the Bron and Kerbosch algorithm [145] as implemented in Chapter 5.

6.2.9 Model refinement

The final models produced by the cliques were energy minimised for 100 steps using the steepest descent method and either the CHARMM or Discover potentials without electrostatics [76, 75]. This procedure was intended to remove steric clashes and to produce acceptable bond lengths and angles rather than change the conformation significantly.

Structure (PDB code)	Source	Function	Sequence identity (%)	Resolution (Å)	Reference
Neurocalcin delta - ncd/t7					
1rec	*B. taurus*	calcium sensor	51.3	1.9	[161]
Endoglucanase - egi/t28					
1cel	*T. reesei*	cellobiohydrolase	49.0	1.8	[162]
Ubiquitin conjugating enzyme - ubc9/t24					
1aak	*A. thaliana*	ubiquitin conjugating enzyme	40.4	2.4	[163]
2uce	*S. cerevisiae*	ubiquitin conjugating enzyme	37.8	2.7	[164]
Cucumber stellacyanin - csc/t9					
2cbp	*C. sativus*	cucumber basic protein	33.6	2.5	[165]
Polyribonucleotide nucleotidyltransferase - pns1/t4					
1csp	*B. subtilis*	cold shock protein	27.2	2.5	[166]
1mjc	*E. coli*	cold shock protein	23.1	2.0	[167]

Table 6.2: Percentage sequence identity between the target sequence and other homologous sequences with known structures for CASP2 targets, based on the alignment used for building the comparative models (the percentage identity based on the correct structure-based alignment is given in Figure 6.1). For each structure, the target name and numbers are given, along with details of the known homologs.

6.3 Results

6.3.1 Sequence alignment

Table 6.2 shows the parent structures that were selected for each family and the percentage identity to the target sequence as determined by the alignment used for constructing the initial models. In one case, ncd/t7, the experimental coordinates are not available to us at this time; the accuracy of model building for that target will be evaluated at a later date.

To judge the accuracy of the alignments, we compare the alignment generated by a structural superposition of the parent structure and the target experimental structure to the sequence alignment used in the modelling exercise.

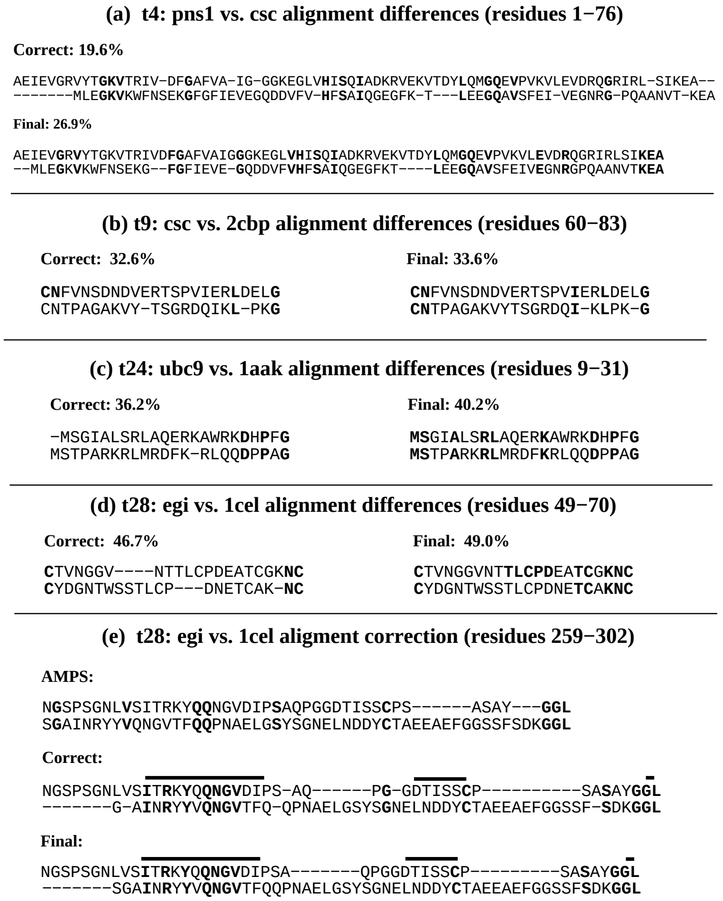
For three of the proteins (pns1/t4, csc/t9, and ubc9/t24), neither the final alignments nor the initial AMPS alignments (which are identical in the case of csc/t9 and ubc9/t24) agree with those produced by structural superposition of the target experimental structures with the respective parent structures. A comparison of the alignment differences in non-loop regions identified by the comparative modelling evaluation program [168] is shown in Figure 6.1a-d. Figure 6.1e shows an example of a hand-corrected AMPS alignment that is correct.

In the case of pns1/t4 (Figure 6.1a) the alignment used for model building is incorrect for more than 50% of the residues, even though the proteins are related (the structural alignment between the parent and target structures results in a C_α RMSD of 2.52 Å for 64/67 residue positions that are aligned; see Figure A.7). Given such an alignment error, the rest of the model building process is doomed to failure. The results of main chain and side chain building for pns1/t4 is thus not discussed in detail.

In two other cases (csc/t9 and ubc9/t24; Figure 6.1b and 6.1c), the AMPS-generated alignments were incorrect for one region in each structure.

The "alignment difference" in egi/t28 (Figure 6.1d), residues 49-70, illustrates that structure-based alignments are not necessarily meaningful. What is identified as an alignment error by the comparative modelling evaluation program is not really an error, but rather an example of a large main chain shift (with a C_α RMSD of 4.85 Å for the 21 residues). The structural alignment between the parent and the target experimental structures is meaningless in this region.

The alignment correction in egi/t24 (Figure 6.1e) underscores the importance of visual inspection. The C_α RMSD between the model constructed using the AMPS alignment and the target experimental structure is 4.24 Å for the 292 main chain positions that were copied from the parent. The C_α RMSD between the model constructed using the hand-corrected alignment and the target exper-



(a) t4: pns1 vs. csc alignment differences (residues 1−76)

Correct: 19.6%

```
AEIEVGRVYTGKVTRIV-DFGAFVA-IG-GGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRL-SIKEA--
-------MLEGKVKWFNSEKGFGFIEVEGQDDVFV-HFSAIQGEGFK-T---LEEGQAVSFEI-VEGNRG-PQAANVT-KEA
```

Final: 26.9%

```
AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRLSIKEA
--MLEGKVKWFNSEKG--FGFIEVE-GQDDVFVHFSAIQGEGFKT----LEEGQAVSFEIVEGNRGPQAANVTKEA
```

(b) t9: csc vs. 2cbp alignment differences (residues 60−83)

```
Correct:  32.6%                 Final: 33.6%

CNFVNSDNDVERTSPVIERLDELG        CNFVNSDNDVERTSPVIERLDELG
CNTPAGAKVY-TSGRDQIKL-PKG        CNTPAGAKVYTSGRDQI-KLPK-G
```

(c) t24: ubc9 vs. 1aak alignment differences (residues 9−31)

```
Correct: 36.2%                  Final: 40.2%

-MSGIALSRLAQERKAWRKDHPFG        MSGIALSRLAQERKAWRKDHPFG
MSTPARKRLMRDFK-RLQQDPPAG        MSTPARKRLMRDFKRLQQDPPAG
```

(d) t28: egi vs. 1cel alignment differences (residues 49−70)

```
Correct:  46.7%                 Final:  49.0%

CTVNGGV----NTTLCPDEATCGKNC      CTVNGGVNTTLCPDEATCGKNC
CYDGNTWSSTLCP---DNETCAK-NC      CYDGNTWSSTLCPDNETCAKNC
```

(e) t28: egi vs. 1cel aligment correction (residues 259−302)

AMPS:

```
NGSPSGNLVSITRKYQQNGVDIPSAQPGGDTISSCPS------ASAY---GGL
SGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGL
```

Correct:

```
NGSPSGNLVSITRKYQQNGVDIPS-AQ------PG-GDTISSCP----------SASAYGGL
-------G-AINRYYVQNGVTFQ-QPNAELGSYSGNELNDDYCTAEEAEFGGSSF-SDKGGL
```

Final:

```
NGSPSGNLVSITRKYQQNGVDIPSA-------QPGGDTISSCP---------SASAYGGL
-------SGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGL
```

Figure 6.1: Differences between the alignment used for the modelling exercise (labelled "Final") and the correct alignment based on a structural superposition (labelled "Correct") for CASP2 targets, and an example of an alignment correction. In (a-c), the final sequence-based alignment used to build the model is incorrect in comparison to the correct structure-based alignment. In (d), the main chain region in egi/t28 residues 49-70 varies by more than 4.0 Å between the parent and the target structures, and a structural alignment in that region is not meaningful. In (e) an example of an hand-modified alignment that is correct is shown. The model constructed using the modified alignment (labelled "Final") is lower in C_α RMSD by more than 2.0 Å to the experimental structure compared to the model constructed using the AMPS-generated alignment, considering only main chain regions that were copied from the parent. These regions are indicated by a thick black line for part of the correct and final alignments in (e).

Name of target	All MC Built SC	All MC Copied SC	Built MC Built SC	Built MC Copied SC	Copied MC Built SC	Copied MC Copied SC
egi/t28	46.5% (71)	52.3% (65)	49.0% (53)	53.2% (47)	38.9% (18)	50.0% (18)
ubc9/t24	45.2% (43)	46.0% (37)	56.0% (25)	63.2% (19)	33.3% (18)	33.3% (18)
csc/t9	47.4% (38)	40.0% (28)	69.6% (23)	46.2% (13)	13.3% (15)	33.3% (15)

Table 6.3: Analysis of side chain residues that were built using the clique finding (CF) method for CASP2 targets. For each target (egi/t28, ubc9/t24, csc/t9), the percentage of χ_1 angles that deviate more than 30 ° for side chains that were built using the CF method (labelled "Built SC") is shown. For comparison, the percentage error that would have resulted had those side chains been built using the minimum perturbation (MP) method (labelled "Copied SC") is shown. The second and third columns (labelled "All MC") make this comparison for all side chains that were built on any main chain region, built or copied, the fourth and fifth columns make this comparison for side chains that were built on main chain regions not copied from a parent structure (labelled "Built MC"), and the last two columns make this comparison for side chains that were built on main chain regions that were copied from a parent structure (labelled "Copied MC"). Numbers in parenthesis show the total number of χ_1 angles that were considered for the percentage error calculation.

imental structure is 1.92 Å for the same number of residues. The hand-corrected alignment matches the structural one exactly for these residues.

6.3.2 Side chain building

Table 6.3 shows the details of side chain construction for the various targets. The percentage of χ_1 angles that deviate more than 30° from the experimental structure for side chains that were built using the CF method is shown. For comparison, the percentage of χ_1 angles that deviate more than 30° from the experimental structure if the MP method had been used to build side chains for those residues is also shown.

Table 6.3 shows that in cases were the parent main chain was copied, the percentage error in the χ_1 angles is significantly reduced in egi/t28 and csc/t9 by 11% and 20% respectively by building those side chains with the CF method.

In the case of ubc9/t24, the percentage error is similar regardless of the method used and the source of the main chain. However, when we consider the columns labelled "All MC" in Table 6.3, we see that the percentage error in the case of csc/t9 has risen (by 7%) upon using the CF method. This presumably reflects the fact that the insertions in egi/t28 and ubc9/t24 were built relatively accurately, leading to better predictions with the side chains, whereas the insertions in csc/t9 had large errors (Cα RMSDs greater than 3.0 Å) leading to inaccurate side chain predictions. These observations are supported by the data under the columns labelled "Built MC" in Table 6.3.

Table 6.4 shows an analysis of side chains that were built on main chains copied from the parent experimental structure using the CF method that had an error in the χ_1 angles of more than 30° relative to the target experimental structure. Figures 6.2 and 6.3 show specific examples of side chain construction using the CF method in csc/t9 and ubc9/t24 respectively.

There were seven side chains with incorrect χ_1 angles in egi/t28, six in ubc9/t24, and two in csc/t9. The C_α-C_α distance between the model and the experimental structure for the residue position, and the largest temperature factor in any atom in the χ_1 rotamer for the side chain is shown. Out of the fifteen side chains containing errors in the χ_1 angles, eight of the errors are associated with the presence of high (> 30.0 Å^2) temperature factors in the side chain atoms or a main chain shift in the residue C_α (> 1.0 Å) position in the model relative to the experimental structure. In two cases (W36 and Y94 in egi/t28), the experimental conformation was not acceptable in the model because of clashes. The clashes are directly attributable to main chain shifts between the target and parent structures (which was used to construct the model) in either the residue

Residue	C_α-C_α distance (Å)	Largest B (Å^2)	Problem
egi/t28			
W36	0.87	20.7	experimental conformation clashes with model I72 built using the minimum perturbation method
E73	0.90	23.6	experimental conformation $O\epsilon_1$ and model G4 oxygen at 2.7 Å (unfavourable electrostatics); G4 B is 55.8 Å^2
Y94	0.30	16.0	experimental conformation clashes with model L349 built using the minimum perturbation method; L349 has a main chain shift of 2.62 Å in the target relative to the parent structure and B for side chain atoms in L349 is 39.7 Å^2
V119	0.46	49.0	high B
Q149	0.31	47.0	high B
E342	2.40	73.6	main chain shift; high B
T355	0.55	16.3	discriminatory function fails
ubc9/t24			
R21	4.36	19.3	region of alignment error in model
R25	1.60	31.0	main chain shift in target relative to parent; high B
C51	0.22	17.5	discriminatory function fails
L89	0.80	23.0	peptide flip over/shift in surrounding main chain in target relative to the parent structure
K118	1.14	27.2	main cain shift in target relative to parent
Y142	1.33	21.3	main chain shift in target relative to parent
csc/t9			
T11	0.93	20.1	interacts with region 14-24 (C_α RMSD 5.23)
D66	4.72	95.9	main chain shift; high B

Table 6.4: Analysis of side chains with an error of more than 30° in the χ_1 angle built by using the clique finding method on main chains that were copied from the parent experimental structure for CASP2 targets. For each residue with an error in the χ_1 angle, The distance between the C_α atoms of the corresponding residues in the experimental structure and the model, the largest temperature factor (B) of any of the atoms determining the χ_1 rotamer, and a brief comment about the nature of the error is shown.

with the incorrect rotamer or in the residue with which it is clashing. In one case (E73 in egi/t29), unfavourable contacts between the side chain atoms of E73 to the main chain atoms of another residue (G4) might be responsible for the incorrect prediction. The atoms in G4 have high temperature factors (<B> is 55.8 Å^2), and the position of the carbonyl in G4 is different by 2.65 Å in

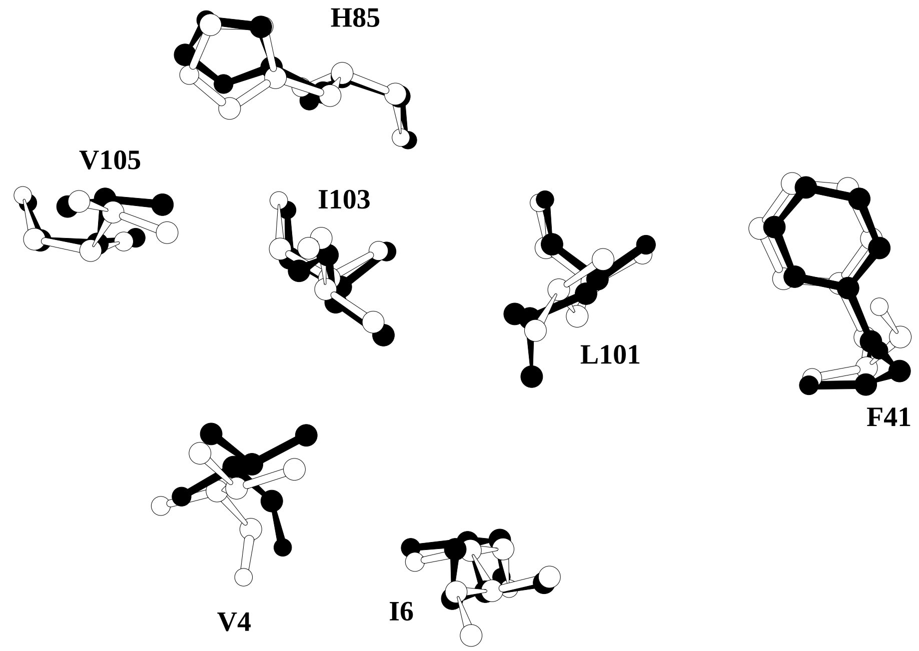


Figure 6.2: Comparison of some side chain conformations predicted (white) using the clique finding (CF) method to the experimental structure (black) for cucumber stellacyanin (csc/t9). All the side chains shown were built on main chain that was copied from the parent structure, where the CF method generally performs well (see Table 6.3). For csc/t9, the percentage of side chains accurately predicted in the case of copied main chains is 86.7%.

the model relative to the experimental structure. In two cases (T355 in egi/t28 and C51 in ubc9/t24), it appears as if the discriminatory function is unable to select the correct rotamer and in two other cases (T11 in ubc9/t24 and L89 in csc/t9) the side chain built interacts with a main chain region that was predicted incorrectly.

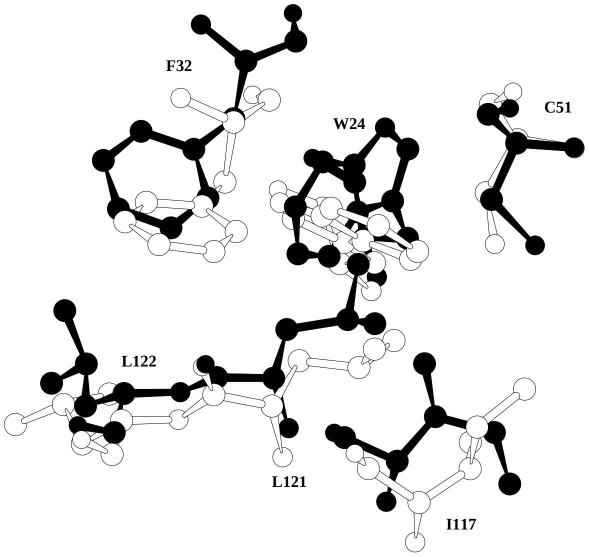


Figure 6.3: Comparison of some side chain conformations predicted (white) using the clique finding (CF) method to the experimental structure (black) for the ubiquitin conjungating enzyme (ubc9/t24). All the side chains show (except for L121 and L122, which are identities) were built on main chain that was copied from the parent. However, for these residues, there is a main chain shift (> 1.0 Å) in the target relative to the parent structure which generally results in inaccurate side chain prediction.

6.3.3 Main chain building

Table 6.5 shows the details for the 22 main chain regions that were built using the CF method, where the main chains were sampled by the database or grid search methods described in the METHODS section. Table 6.6 gives the details

	Region built	#	Sequence	Parent RMSD (Å)	Type	Root RMSD (Å)	Sample range (Å)	**Region RMSD (Å)**	Problem
egi/t28									
	42-48	7	HDANYNS	2.14	(D)	2.53	1.76-7.43	**3.12**	roots
*	78-81	4	AASG	1.15	(D)	0.60	0.68-2.49	**0.77**	
	96-103	8	PSSSGGYS	6.60	(2)	2.86	6.20-8.26	**7.43**	sampling/roots
	155-161	7	GANQYNT	2.16	(D)	0.95	1.29-5.55	**3.57**	context
	177-190	14	VQTWRNGTLNTSHQ	5.63	(D)	2.31	10.36-16.30	**11.39**	sampling/roots
*	214-219	6	CTATAC	2.76	(D)	1.02	1.02-3.54	**1.14**	
+	240-244	5	GDTVD	1.15	(D)	0.77	1.78-3.70	**2.23**	
	256-268	13	NTDNGSPSGNLVS	1.85	(7)	0.46	4.07-13.58	**5.36**	sampling
	282-287	6	SAQPGG	6.23	(D)	5.64	3.28-7.29	**5.02**	sampling/roots
	293-301	9	CPSASAYGG	2.31	(D)	2.82	3.66-10.50	**8.70**	sampling/roots
ubc9/t24									
*	37-46	10	TKNPDGTMNL	2.32	(5)	0.85	1.72-9.20	**2.64**	
+	56-62	7	KKGTPWE	0.53	(0)	0.57	0.60-5.45	**0.60**	
+	73-79	7	KDDYPSS	1.20	(0)	0.83	1.13-4.78	**1.18**	
*	106-111	6	EEDKDW	1.44	(2)	0.66	1.32-4.77	**2.38**	
	164-166	3	APS	6.05	(1)	4.19	4.57-6.47	**6.29**	sampling/roots
csc/t9									
	1-2	2	GS	-	(2)	0.68	1.46-5.20	**4.53**	fitting error
	14-24	11	SVPSSPNFYSQ	2.45	(4)	1.15	4.07-9.40	**5.23**	sampling
+	42-45	4	PANA	1.92	(0)	0.45	1.33-2.64	**1.90**	
*	51-57	7	METKQSF	1.55	(1)	0.50	1.07-5.18	**1.57**	
	77-83	7	ERLDELG	1.45	(1)	2.71	2.62-3.82	**3.56**	roots/alignment
+	90-93	4	TVGT	0.82	(0)	0.43	0.66-2.41	**0.83**	
	106-108	3	VAA	0.46	(2)	0.67	3.07-6.90	**5.49**	fitting error

Table 6.5: Analysis of the predictions of 22 main chain regions that were built using the clique finding (CF) method for CASP2 targets. All RMSDs shown are C_α RMSDs in Å and are based on a global superposition of the structures being compared. For each target (egi/t28, ubc9/t24, and csc/t9), the range of residues in the built region, the number of residues in the built region, the sequence of the region being built, the C_α RMSD of the two root residues, the C_α RMSD of the built region (not including the roots) between the model and the target experimental structure, the C_α RMSD for equivalent residues ('-' if there were no equivalent residues) between the parent structure and the target experimental structure, the region type in parenthesis (a number greater than 0 indicates there was an insertion of that many residues, a 'D' signifies a deletion, and a 0 signifies a region that is neither an insertion or a deletion but was built because we thought the main chain conformation would differ from the parent), the range of C_α RMSDs that were sampled, and a brief comment about the nature of the problem in building the region accurately (if there was one). *Bona fide* successful predictions where copying the parent would not have sufficed are indicated by '*' and cases where the CF method works well (even though copying the main chain from the parent would have sufficed) are indicated by '+'.

of main chain region building process, including the interconnected manner in which they were built (i.e., combining main chain and side chain possibilities simultaneously).

Comparing the C_α RMSD between the target experimental structure and the model and the C_α RMSD between the target experimental structure and the

	Regions built	Number of main chain conformations	Number of side chain conformations per main chain	Overall C_α RMSD (Å)
egi/t28				
	42-48	364	$6^2 \times 5^2 \times 3^6 \times 1^1 \simeq 6 \times 10^6$	3.12
*	78-81,98-103	$1013 \times 586 \simeq 6 \times 10^6$	$3^1 \times 2^8 \times 1^6 = 768$	6.08
	155-161,177-190	$591 \times 96 \simeq 5 \times 10^5$	$2^{13} \times 1^8 = 8192$	9.5
*	214-219	468	$4^3 \times 3^6 \times 2^4 \times 2^1 \simeq 15 \times 10^5$	1.14
+	240-244	497	$6^4 \times 3^4 \times 1^1 \simeq 10^5$	2.23
	256-268	294	$3^{12} \times 1^3 \simeq 5 \times 10^5$	5.36
	282-287	973	$6^5 \times 3^1 \times 1^4 \simeq 2 \times 10^4$	5.02
	293-301	991	$6^2 \times 3^8 \times 5^1 \simeq 10^6$	8.70
ubc9/t24				
*	37-46,73-79,106-111	$517 \times 451 \times 461 \simeq 10^8$	$2^5 \times 1^{23} = 32$	2.22
+	56-62	461	$6^6 \times 3^2 \times 1^2 \simeq 4 \times 10^5$	0.60
	164-166	78	$6^4 \times 3^1 \times 1^2 = 3888$	6.29
csc/t9				
+	42-45,90-93	$456 \times 311 \simeq 14 \times 10^4$	$3^6 \times 1^4 = 729$	1.47
	77-83,106-108	$205 \times 102 \simeq 2 \times 10^4$	$3^{10} \times 1^3 \simeq 6 \times 10^4$	4.24
	1-2	1256	$6^6 \times 3^1 \simeq 14 \times 10^4$	4.53
	14-24	595	$3^{13} \times 1^2 \simeq 10^8$	5.23
*	51-57	133	$6^2 \times 3^{10} \simeq 2 \times 10^6$	1.57

Table 6.6: Computational details of 22 main chains that were built using the clique finding (CF) method for CASP2 targets. For each target (egi/t28, ubc9/t24, and csc/t9) the regions that were built (in the case of multiple regions built simultaneously, the regions are all indicated in a single row), the number of possible main chain conformations (in the case of multiple regions, this is the product of the number of possible conformations of each of the regions), the number of side chain conformations per main chain, and the overall C_α RMSD between the experimental structure and the model with the lowest negative log conditional probability based on a global superposition is given. The total number of conformations explored is the product of the number of side chain conformations times the number of side chain conformations per main chain. The sum of the exponents for the side chain conformations represent the number of side chains for which conformations were varied. The mantissa indicates the number of side chain conformations for each of the residues in the exponent. A mantissa of one indicates that only one side chain conformation was considered per residue (an alanine, glycine, or proline residue except in the case of ubc9/t24 residues 37-46, 73-79 and 106-111). In cases where the sum of the exponents exceeds the number of residues in the region being built, the excess number indicates the number of side chain positions in the environment where multiple side chain conformations were explored. *Bona fide* successful predictions for at least one of the regions being built where copying the parent would not have sufficed are indicated by '*' and cases where the CF method works well (even though copying the main chain from the parent would have sufficed) are indicated by '+'. Residues 78-81 and 96-103, in egi/t28, have an overall combined C_α RMSD of 6.08 Å but the individual RMSDs are 0.77 and 7.43 respectively. The total number of conformations explored, considering both side chain and main chain conformations simultaneously, is generally in the order of 10^9-10^{10} conformations.

parent for each built region, we see that in ten of built regions, the C_α RMSDs are similar to, and in some cases better than, the C_α RMSD had we simply copied the parent main chain.

There are five regions corresponding to insertions that represent accurate and *bona fide* blind predictions where simply copying the parent would not have sufficed (these rows are prefixed by an '*' in Table 6.5). The sizes of these regions range from four to ten residues (with sizes of the insertions ranging from one to five residues) with C_α RMSDs ranging from 0.77 Å (for a four residue region involving deletion) to 2.64 Å (for a ten residue region involving a five residue insertion). One of the more dramatic predictions include the construction of three regions in ubc9/t24 (residues 37-46, 73-79, and 106-111) simultaneously with an overall C_α RMSD of 2.22 Å for the 23 residues (Figure 6.4).

There are another five regions where copying the parent would have generally sufficed for building these regions (rows are prefixed by a '+' in Table 6.5) but were built using the CF method because we thought these regions would vary. However, these cases illustrate that the CF method works well and the C_α RMSDs range from 0.60 Å to 2.23 Å.

The last column in Table 6.5 makes a brief comment about the nature of problem for each main chain that had a C_α RMSD greater than 3.0 Å between the model and the experimental structure. Out of the twelve regions that had large C_α RMSDs, nine of them were predicted incorrectly due to either lack of adequate sampling (no conformation with a C_α RMSD lesser than 3.0 Å), large C_α RMSDs for the two root residues (greater than 2.0 Å), or both. In two of the cases (csc/t9 residues 1-2 and residues 106-108), a technical error in which the main chains returned by the grid search method were fitted incorrectly

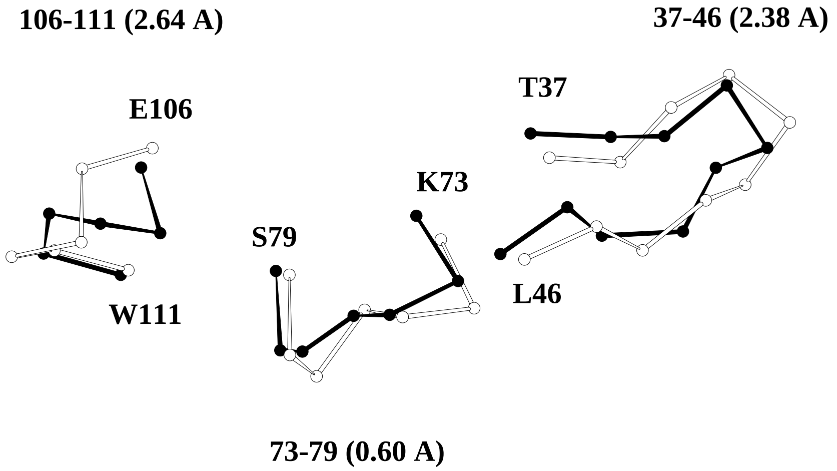


Figure 6.4: Comparison of conformations predicted (white) using the clique finding (CF) method to the experimental structure (black) for three context-sensitive regions in the ubiquitin conjugating enzyme. Shown are C_α traces of three regions, residues 37-46 (five residue insertion), 73-79, and 106-111 (two residue insertion), which were built simultaneously using the CF method with individual C_α RMSDs of 2.64 Å, 0.60 Å, and 2.38 Å for each region respectively, and an overall C_α RMSD of 2.22 Å for all the 23 residues relative to the experimental structure. The C_α RMSDs do not include the root residues and are based on a global superposition.

to the framework led to incorrect predictions. In one case (egi/t28 residues 155-161), we sample main chains with C_α RMSDs between 1.29 Å and 5.55 Å, have a C_α RMSD of 0.95 Å in the root positions, but the predicted region has a C$_\alpha$ RMSD of 3.57 Å. This error is due to the fact that this region in egi/t28 interacts with residues 177-190, which could not have been predicted accurately due to inadequate sampling. These two regions are interconnected and cannot be built separately, and if the main chain in one region cannot be sampled adequately, then the other region is likely be predicted incorrectly. This example illustrates the importance of handling context-sensitivity when building comparative models.

Figure 6.5 compares the C$_\alpha$ trace of the ubc9/t24 complete model to its corresponding experimental structure.

6.3.4 Model refinement

After energy minimization, the C$_\alpha$ RMSD between the model and experimental structure increased slightly, as at CASP1 (see Chapter 2).

6.3.5 Overall accuracies of the model compared to the experimental structure

Table 6.7 shows the overall C$_\alpha$ and all-atom RMSDs, and the percentage errors for χ_1 and all χ angles, between the model and the experimental structure for each of the four targets.

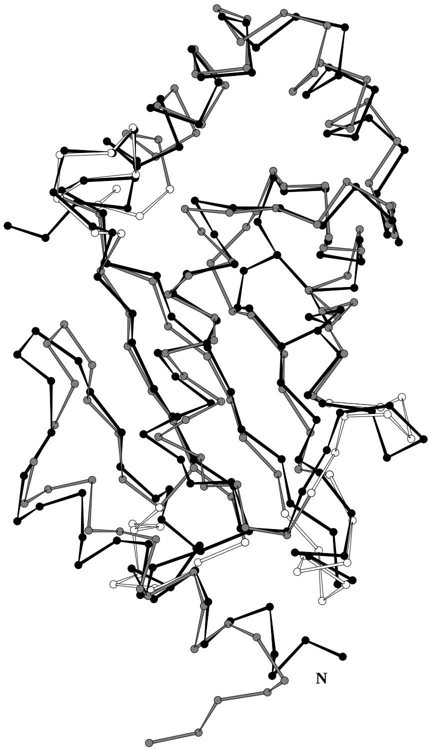


Figure 6.5: Visual comparison between the model and the experimental structure (black) for the ubiquitin conjugating enzyme illustrating regions of main chain that were copied from the parent structure (grey) and regions that were built using the clique finding (CF) method (white). This protein was modelled by us for CASP2 with a C_α RMSD of 2.47 Å. The RMSDs for the built regions are available in Table 6.5.

6.4 Discussion

6.4.1 Alignment

At the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1), we learned that automated sequence alignment methods are

Target (number)	C_α RMSD (Å)	All-atom RMSD (Å)	$\chi_1 > 30°$ (%)	All $\chi > 30°$ (%)
egi/t28	3.37 (371)	3.76 (2735)	33.8 (287)	40.2 (524)
ubc9/t24	2.47 (158)	3.29 (1266)	47.5 (120)	48.2 (284)
csc/t9	2.75 (108)	4.09 (837)	38.9 (90)	49.7 (171)
pns1/t4	7.38 (76)	8.13 (588)	54.1 (61)	61.0 (141)

Table 6.7: Accuracy of the models that were built compared to the experimental structures for CASP2 targets. The C_α and all-atom RMSDs, and the percentage errors in χ_1 and all χ torsions between the model and the experimental structures are given. The numbers listed in parenthesis are the number of atoms/χ angles considered for all residues.

inadequate and that a visual inspection is necessary to optimise the alignment. However, at CASP1, we were lucky that all optimisations by hand based on sequence identity proved to be correct. Here, only one such optimisation in egi/t28 produced the correct alignment (Figure 6.1e). The other hand corrected alignment in pns1/t24 was wrong (Figure 6.1a). This particular error could be attributed to the low level of global sequence identity in the target. However, in ubc9/t24 (Figure 6.1c), the structural alignment differs significantly from the sequence based one, and visual inspection of the alignment would have yielded no clues about the shift in the helix in that region. In fact, in all cases the correct structure-based alignments have a lower percentage sequence identity than the sequence alignments that were used (Figure 6.1). This indicates that a sequence alignment that relies on percentage identity or homology alone cannot effectively produce the correct alignment, and that visual inspection and hand-optimisation of alignments has its limits. As we suggest in Chapter 2, better alignment methods that take structural information into account need to be developed.

6.4.2 Side chains

The percentage error in all the χ angles built using the minimum perturbation (MP) method at CASP1 for three targets was around 50.0% (Table 2.2) and the overall accuracy in χ_1 angles was around 45% (Table 5 in [8]). These results taken together with the data in Tables 6.3 and 6.7 and suggest that there is utility to building side chains taking interconnectedness into account. However, we were unable to handle more than eighteen residues simultaneously due to tractability issues—the clique finding algorithm can handle problems of sizes that correspond to systematically exploring between 10^9 and 10^{10} conformations.

Table 6.3 also shows that the error in building side chains on insertions, deletions, and regions of main chain variation is significantly higher than when building side chains where the main chain is copied from another parent structures. This reflects the inaccuracies in the main chain building process, and reflects the fact that side chain prediction is limited by the accuracy of the main chain predicted [60]. As at CASP1, the problem in analysing the accuracy of predicted side chains is hampered by the fact that experimental structure itself contains atoms in the side chain with large temperature factors (Table 6.4).

6.4.3 Main chains

Building main chains in an interconnected manner (i.e., building multiple main chains and side chains in the environment simultaneously) has improved the predictability of insertions and deletions. At CASP1, none of the insertions and deletions were predicted accurately—in the case of models that were built by us, none of the insertions and residues flanking deletions had a C_α RMSD less than 3.0 Å (Chapter 2). At CASP2, five of the insertions and residues flanking deletions have C_α RMSDs less than 3.0 Å.

Even though regions that did not correspond to insertions or deletions were built by the CF method and predicted accurately with C_α RMSDs less than 2.0 Å to the experimental structure, the C_α RMSD had we just copied the parent would have produced similar results (see Table 6.5). However, we could not predict in advance whether these regions would vary or remain conserved between the target and the parent structures and all the predictions made by CF method for these regions were reasonably accurate.

There are other regions of main chain variation that were not built by us where the C_α RMSD does vary significantly between the parent and the experimental structure. This was partly due to the fact that we were not accurate at predicting exactly which main chain regions would vary, and in cases where we suspected a main chain variation would occur, we did not have the computational resources to build the entire region.

6.4.4 Bona fide prediction

Traditional side chain and main chain building methods are tested by building (i.e., *reproducing*) side chains and main chains given the context of the original experimental structure [50, 51, 52, 53, 54, 55, 56, 57]. The stark contrast between the results in Chapters 3, 4 and 5 and the results in Tables 6.5 and 6.3 for main chain and side chain construction highlight the importance and difficulty of *bona fide* prediction, when the correct answer is not known in advance.

While the results at CASP2 are encouraging in the sense that they are better than the results at CASP1, more work lies ahead in improving the method described in this work so it is robust at prediction in approximate environments.

6.5 Summary

We constructed five comparative models in a blind manner for the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2). The method used is based on a novel graph-theoretic clique finding approach, and attempts to address the problem of interconnected structural changes in the comparative modelling of protein structures. We discuss briefly how the method is used for protein structure prediction, and detail how it performs in the blind tests. We find that compared to CASP1, significant improvements in building insertions and deletions and side chain conformations have been achieved.

The final conclusions chapter compares the progress of our comparative modelling approach from CASP1 to CASP2 and discusses the prospects of this approach for handling the protein structure prediction problem.

Chapter 7

Conclusion

7.1 Progress of comparative modelling

Given the differences in the difficulty of targets predicted at the first and second experiments on the Critical Assessment of protein Structure Prediction methods (CASP1 and CASP2), it is hard to compare progress from one experiment to the next. Figure 7.1 illustrates an attempt to do this, by measuring the difficulty of building a model by taking the product of the fraction of non-identical residues and the fraction of residues in insertions and deletions, and plotting it against the C_α RMSD of the models relative to the experimental structure. The results indicate that even though the models at CASP2 were more difficult to build, the C_α RMSD has gotten better. Taken in conjunction with the results presented in Chapter 6, this suggests that progress indeed has been made in comparative model building in the years between the two experiments by taking into account the context-sensitivity of interactions seen in protein structures.

In terms of methodology, we have moved forward from CASP1 in the following ways: we can now sample relatively large numbers of side chain and main

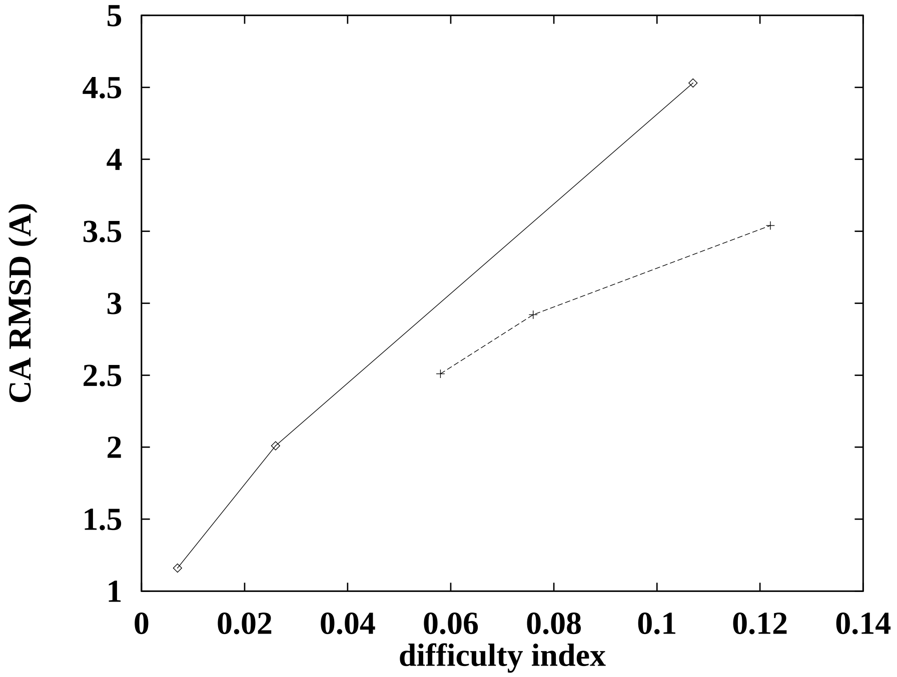


Figure 7.1: Plot of C$_\alpha$ RMSD vs. the difficulty index for models of targets from the first and second experiments on the Critical Assessment of protein Structure Prediction methods (CASP 1 and 2). The thick line represents CASP1 targets and the dashed line represents CASP2 targets. The difficulty index is the the product of the fraction of non-identical residues and the fraction of residues in insertions and deletions. The larger the index, the more difficult it is to build the model for that target.

chain conformations using the clique finding method. We have also developed a discriminatory function that is fairly accurate and allows for fast evaluation of a conformation when used by the clique finding method. The side chain and main chain approximation methods, while not completely adequate, have helped us build regions in a model in an interconnected manner that we were unable to do at CASP1.

7.2 The road ahead

The first implementation of the graph-theoretic clique finding approach shows a great deal of promise. However, while we see some improvements in building side chains and regions of main chain in an interconnected manner, the results at CASP2 show that we still have a long way to go before we can build models that rival experiment in accuracy.

We see room for further improvement in the methodology described here: from the results shown in Chapters 3 and s6, it appears as if the discriminatory function is generally able to select the correct conformation if it is present in the sample space. A transformation of the conditional probabilitities for discrete distance values into a continuous function may enable us to refine the final model produced by the graph-theoretic clique finding method.

Main chain sampling algorithms which generate conformations that are closer to the experimental structure need to be developed. These could involve systematic searching of all main chain conformations for a given region and using filters and clustering to reduce the number of conformations considered [59].

Improved side chain sampling methods that further narrow down the choices for a residue conformation will help in reducing the number of nodes in a graph. This can be accomplished through the aid of a main chain dependent rotamer library [52, 131]

The clique finding method is still limited in terms of the numbers of conformations it can handle. Improvement in clique finding algorithms by using approximation algorithms, and filtering based on weights of nodes and edges, will enable us to build a greater numbers of side chains and larger main chain regions simultaneously, compared to the size of problems we handle in this work.

7.3 Final remarks

The protein structure prediction problems remains one of the fundamental unsolved problems in molecular biology. However, comparative modelling methods provide hope in at least predicting structures where the sequences are related. This should prove to be extremely useful in deducing protein structure-function relationships in whole organisms given the huge number of sequences being produced by the genome sequencing projects, where at least a quarter of the currently known protein sequences belong to protein families for which there are structures in the Protein Data Bank [169].

The concept of a clique offers a way to elegantly represent the web of interactions seen in proteins, and also to cluster them flexibly and efficiently. The general properties of cliques, which readily describe the nature of protein structures, lead us to believe that the approaches outlined here can be extended to other categories of structure prediction, such as fold recognition and *ab initio* modelling.

Appendix A

Visual overview

A.1 Visual comparison between the model and the corresponding experimental structure for CASP1 and CASP2 targets

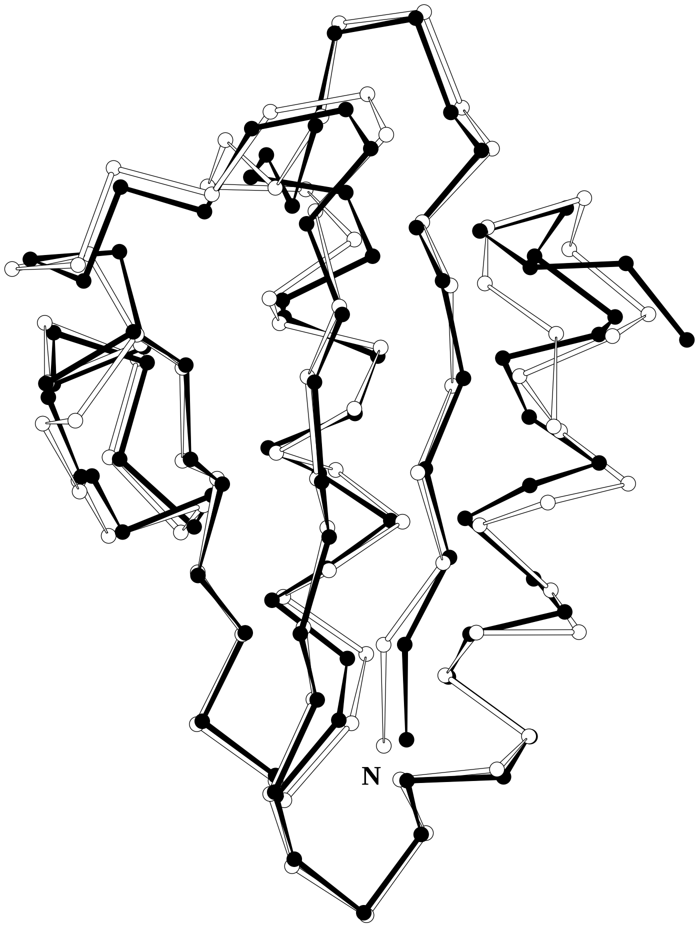


Figure A.1: Visual comparison between the model (white) and the experimental structure (black) for the histidine-containing phosphocarrier protein [63]. This protein was modelled by us for the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1) with a C_α RMSD of 1.18 Å.

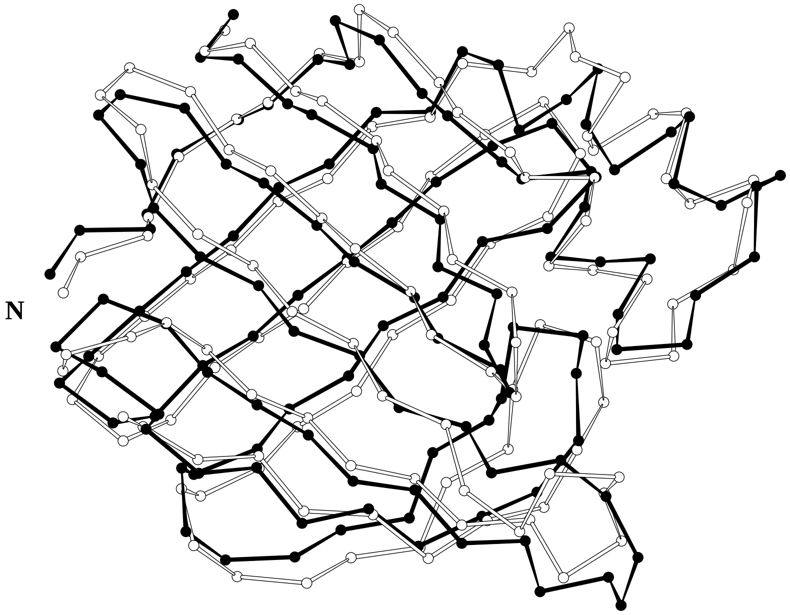


Figure A.2: Visual comparison between the model (white) and the experimental structure (black) for the cellular retinoic acid binding protein I [65]. This protein was modelled by us for CASP1 with a C_α RMSD of 2.01 Å.

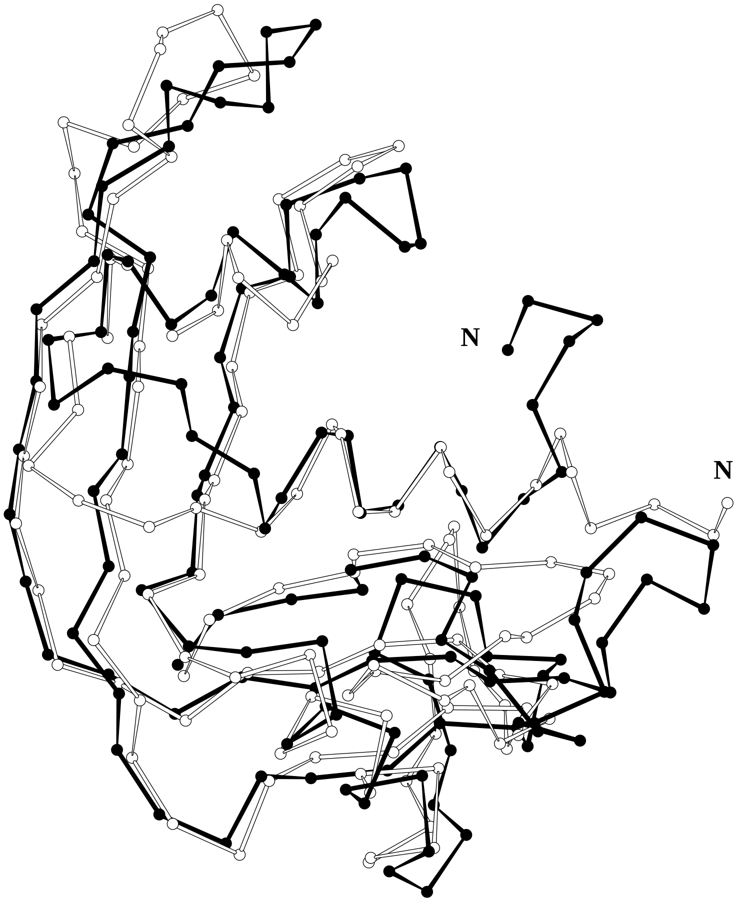


Figure A.3: Visual comparison between the model (white) and the experimental structure (black) for the eosinophil derived neurotoxin [66]. This protein was modelled by us for CASP1 with a C_α RMSD of 4.55 Å.

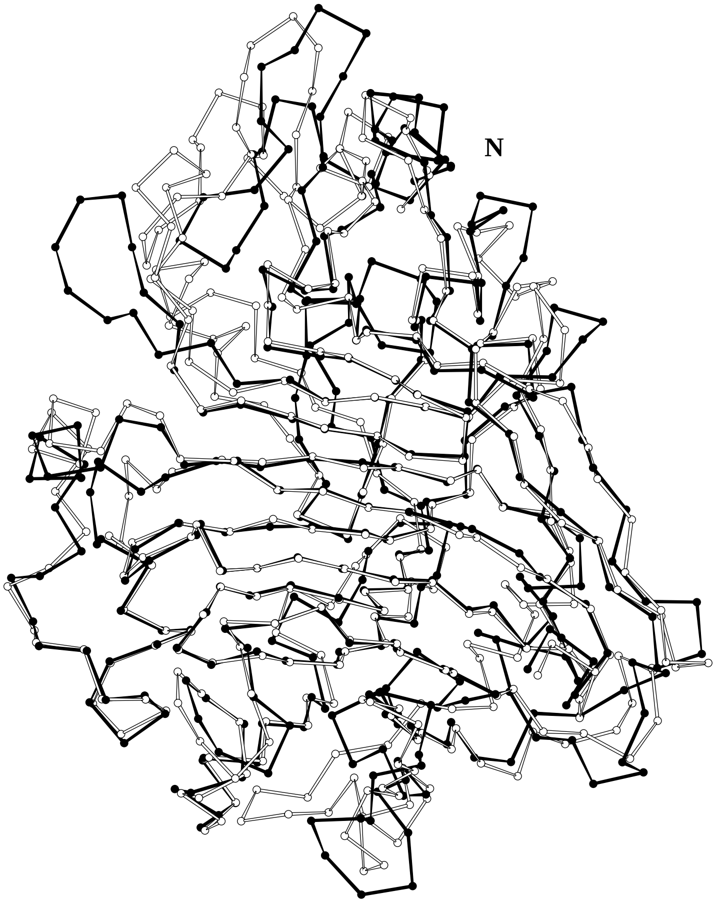


Figure A.4: Visual comparison between the model (white) and the experimental structure (black) for endoglucanase I [153]. This protein was modelled by us for the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2) with a C_α RMSD of 3.37 Å.

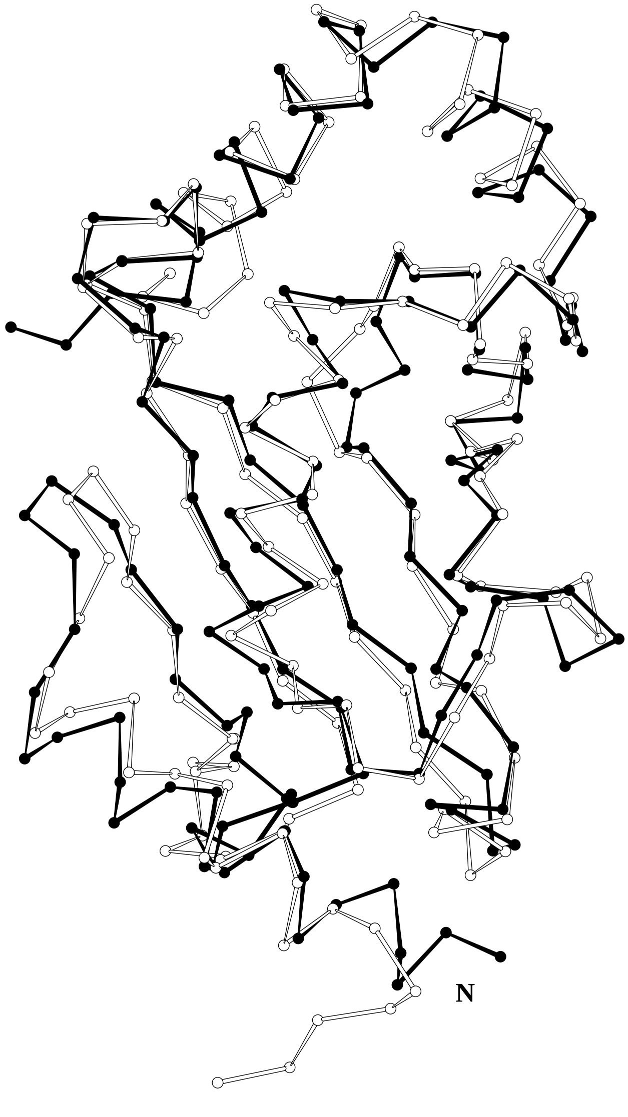


Figure A.5: Visual comparison between the model (white) and the experimental structure (black) for the ubiquitin conjugating enzyme [152]. This protein was modelled by us for CASP2 with a C_α RMSD of 2.47 Å.

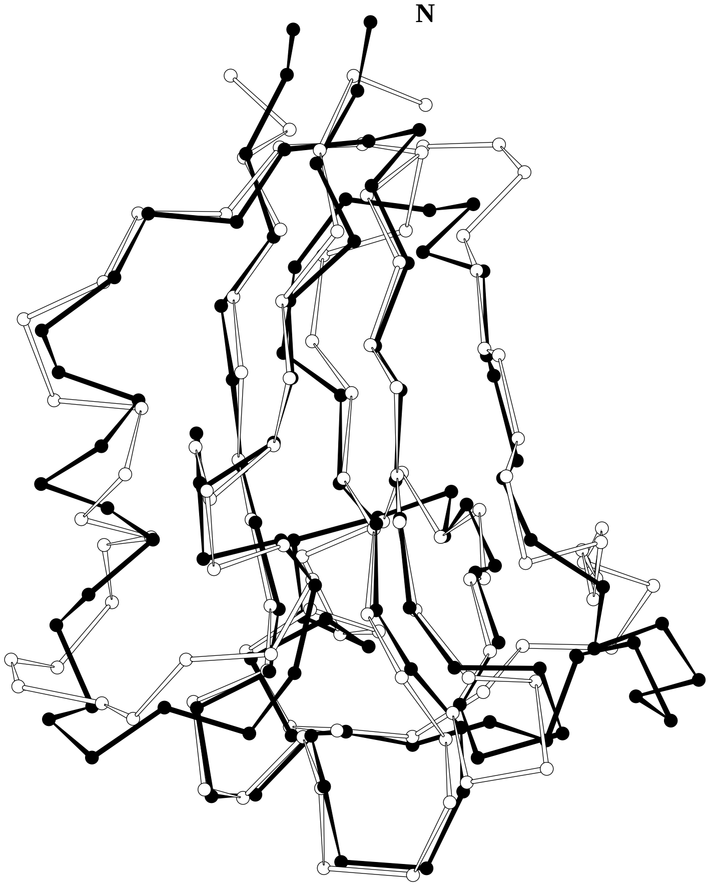


Figure A.6: Visual comparison between the model (white) and the experimental structure (black) for cucumber stellacyanin [151]. This protein was modelled by us for CASP2 with a C_α RMSD of 2.75 Å.

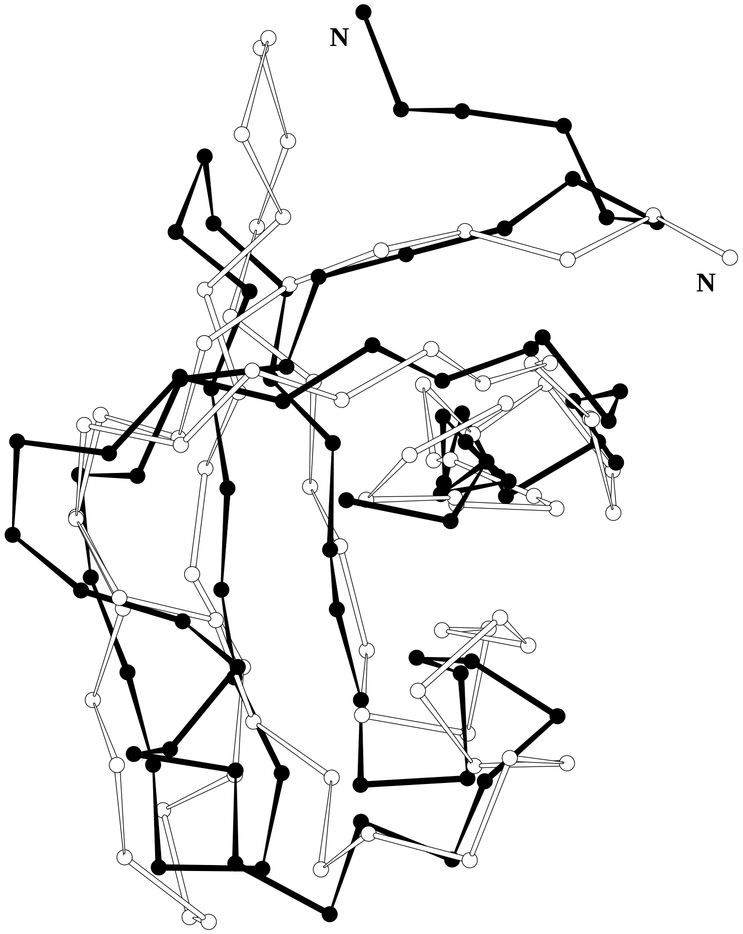


Figure A.7: Visual comparison between the model (white) and the experimental structure (black) for the polyribonucleotide nucleotidyl s-transferase [150]. This protein was modelled by us for CASP2 with a C_α RMSD of 7.38 Å.

BIBLIOGRAPHY

[1] Blundell, T., Johnson, L. *Protein crystallography.* Academic Press (London), 1976.

[2] Kaptein, R., Boelens, R., Scheek, R., van Gunsteren, W. Protein structures from NMR. Biochemistry 27:5389–5395, 1988.

[3] Crowther, R. Probing biological structure. Nature (London) 339:426–427, 1989.

[4] Branden, C., Tooze, J. *Introduction to Protein Structure.* Addison Wesley Publishing Inc., 1991.

[5] Clore, M., Gronenborn, A. Applications of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination. Progress in NMR Spectroscopy 23:43–92, 1991.

[6] Anfinsen, C. Principles that govern the folding of protein chains. Science 181:223–230, 1973.

[7] Moult, J., Judson, R., Fidelis, K., Pedersen, J. A large-scale experiment to assess protein structure prediction methods. Proteins: Struct., Funct., Genet. 23:ii–iv, 1995.

[8] Mosimann, S., Meleshko, R., James, M. A critical assessment of comparative molecular modeling of tertiary structures in proteins. Proteins: Struct., Funct., Genet. 23:301–317, 1995.

[9] Lemer, C. M.-R., Rooman, M., Wodak, S. Protein structure prediction by threading methods: evaluation of current techniques. Proteins: Struct., Funct., Genet. 23:337–355, 1995.

[10] Defay, T., Cohen, F. Evaluation of current techniques for ab initio protein structure prediction. Proteins: Struct., Funct., Genet. 23:431–445, 1995.

[11] Luecke, H., Quiocho, F. High specificity of a phosphate transport protein determined by hydrogen bonds. Nature (London) 347:402–406, 1990.

[12] Hughes, R., Hatfull, G., Rice, P., Steitz, T., Grindley, N. Cooperativity mutants of the gamma delta resolvase identify an essential interdimer interaction. Cell 63:1331–1338, 1990.

[13] Herzberg, O. An atomic model for protein-protein phosphoryl group transfer. J. Biol. Chem. 267:24819–24823, 1992.

[14] Liu, X., Zhu, H., Huang, B., Rogers, J., Yu, B., Kumar, A., Jain, M., Sundaralingam, M., Tsai, M. Phospholipase A2 engineering. Probing the structural and functional roles of N-terminal residues with site-directed mutagenesis, X-ray, and NMR. Biochemistry 34:7322–7334, 1995.

[15] Brick, P., Bhat, T., Blow, D. Structure of tyrosyl-tRNA synthetase refined at 2.3 Å resolution. Interaction of the enzyme with tyrosyl adenylate intermediate. J. Mol. Biol. 208:83–98, 1988.

[16] Rould, M., Perona, J., Soll, D., Steitz, T. Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science 246:1135–1142, 1989.

[17] Schulz, G., Muller, C., Diederichs, K. Induced-fit movements in adenylate kinases. J. Mol. Biol. 213:627–630, 1990.

[18] Cahoon, E., Lindqvist, Y., Schneider, G., Shanklin, J. Redesign of soluble fatty acid desaturases from plants for altered substrate specificity and double bond position. Proc. Natl. Acad. Sci. USA 94:4872–4877, 1997.

[19] Ulmner, K. Protein engineering. Science 219:666–671, 1983.

[20] Pantoliano, M., Whitlow, M., Wood, J., Dodd, S., Hardman, K., Rollence, M., Bryan, P. Large increases in general stability for subtilising BPN' through incremental changes in the free energy of unfolding. Biochemistry 28:7205–7213, 1988.

[21] Hua, Q., Hu, S., Frank, B., Jia, W., Chu, Y., Wang, S., Burke, G., Katsoyannis, P., Weiss, M. Mapping the functional surface of insulin by design: structure and function of a novel A-chain analogue. J. Mol. Biol. 264:390–403, 1996.

[22] Rezaie, A., Olson, S. Contribution of lysine 60f to S1' specificity of thrombin. Biochemistry 36:1026–1033, 1997.

[23] Hunter, W., Bailey, S., Habash, J., Harrop, S., Helliwell, J., Aboagye-Kwarteng, T., Smith, K., Fairlamb. Active site of trypanothione reductase. A target for rational drug design. J. Mol. Biol. 227:322–333, 1992.

[24] Blundell, T. Structure-based drug design. Nature (London) 384:23–26, 1996.

[25] Bryson, J., Betz, S., Lu, H., Suich, D., Zhou, H., O'Neil, K., DeGrado, W. Protein design: a hierarchic approach. Science 270:935–941, 1995.

[26] Smith, C., Regan, L. Guidelines for protein design: the energetics of beta sheet side chain interactions. Science 270:980–982, 1995.

[27] Cordes, M., Davidson, A., Sauer, R. Sequence space, folding and protein design. Curr. Opin. Struct. Biol. 6:3–10, 1996.

[28] Clegg, M., Gaut, B., Learn G.H., J., Morton, B. Rates and patterns of chloroplast DNA evolution. Proc. Natl. Acad. Sci. USA 91:6795–6801, 1994.

[29] Holm, L., Sander, C. Mapping the protein universe. Science 273:595–603, 1996.

[30] Hubbard, T., Murzin, A., Brenner, S., Chothia, C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 25:236–239, 1997.

[31] May, A., Johnson, M., Rufino, S., Wako, H., Zhu, Z., Sowdhamini, R., Srinivasan, N., Rodionov, M., Blundell, T. The recognition of protein structure and function from sequence: adding value to genome data. Phil. Trans. Roy. Soc. Lond. 344:373–381, 1994.

[32] Browne, W., North, A., Phillips, D. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol. 42:65–86, 1969.

[33] Taylor, W. Protein structure modelling from remote sequence similarity. J. Biotechnol. 35:281–291, 1994.

[34] Chothia, C., Lesk, A. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.

[35] Holm, L., Sander, C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res. 22:3600–3609, 1994.

[36] Holm, L., Sander, C. The FSSP database. <`http://www2.ebi.ac.uk/-dali/fssp/TABLE2.html`>, 1997.

[37] Lesk, A., Fordham, W. Conservation and variability in the structures of serine proteases of the chymotrypsin family. J. Mol. Biol. 258:501–537, 1996.

[38] Pawlowski, K., Bierzynski, A., Godzik, A. Structural diversity in a family of homologous proteins. J. Mol. Biol. 258:348–366, 1996.

[39] Hartley, B. Homologies in serine proteinases. Phil. Trans. Roy. Soc. Lond. 257:813, 1970.

[40] Pearl, L., Taylor, W. A structural model for the retroviral proteases. Nature (London) 329:351–354, 1987.

[41] Barnes, A., Wynn, C. Homology of lysozomol enzymes and related proteins: Prediction of posttranslational modification sites including phosphorylation of mannose and potential epitopic and substrate binding sites in the $\alpha-$ and $\beta-$submits of hexosaminidases, $\alpha-$glucosidase and rabbit and human isomaltase. Proteins: Struct., Funct., Genet. 4:182–189, 1988.

[42] Ring, C., Sun, E., McKerrow, J., Lee, G., Rosenthal, P., Kuntz, I., Cohen, F. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proc. Natl. Acad. Sci. USA 90:3583–3587, 1993.

[43] Cohen, F., Gregoret, L., Amiri, P., Aldape, K., Railey, J., McKerrow, J. Arresting tissue invasion of a parasite by protease inhibitors chosen with the aid of computer modeling. Biochemistry 30:11221–11229, 1991.

[44] Carson, M., Bugg, C., DeLucas, L., Narayana, S. Comparison of homology models with the experimental structure of a novel serine protease. Acta Cryst. D50:889–899, 1994.

[45] Turkenburg, J., Dodson, E. Modern developments in molecular replacement. Curr. Opin. Struct. Biol. 6:604–610, 1996.

[46] Mollison, K., Mandecki, W., Zuiderweg, E., Fayer, L., Fey, T., Krause, R., Conway, R., Miller, L., Edalji, R., Shallcross, M., Lane, B., Fox, J., Greer, J., Carter, W. Identification of receptor-binding residues in the inflammatory complement protein C5a by site-directed mutagenesis. Proc. Natl. Acad. Sci. USA 86:292–296, 1989.

[47] Arcoleo, J., Greer, J. Hemoglobin binding site and its relationship to the serine protease-like active site of haptoglobin. J. Biol. Chem. 257:10063–10068, 1982.

[48] Lichtarge, O., Bourne, H., Cohen, F. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257:342–358, 1996.

[49] Greer, J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins: Struct., Funct., Genet. 7:317–334, 1990.

[50] Lee, C., Subbiah, S. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217:373–388, 1991.

[51] Holm, L., Sander, C. Fast and simple monte carlo algorithm for side chain optimization in proteins: application to model building by homology. Proteins: Struct., Funct., Genet. 14:213–223, 1992.

[52] Dunbrack, R., Karplus, M. Backbone-dependent rotamer library for proteins: application to side chain prediction. J. Mol. Biol. 230:543–574, 1993.

[53] Laughton, C. Prediction of protein side chain conformations from local three-dimensional homology relationships. J. Mol. Biol. 235:1088–1097, 1994.

[54] Bruccoleri, R. E., Karplus, M. Chain closure with bond angle variation. Macromolecules 18:2767–2773, 1985.

[55] Moult, J., James, M. N. G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins: Struct., Funct., Genet. 2:146–163, 1986.

[56] Bruccoleri, R. E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.

[57] Martin, A., Cheetham, J., Rees, A. Modelling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. USA 86:9268–9272, 1989.

[58] Pedersen, J., Searle, S., Henry, A., Rees, A. Antibody modelling: Beyond homology. Immunomethods 1:126–136, 1992.

[59] Fidelis, K., Stern, P., Bacon, D., Moult, J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng. 7:953–960, 1994.

[60] Chung, S., Subbiah, S. The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. Protein Sci. 4:2300–2309, 1995.

[61] Šali, A., Potterton, L., Yuan, F., van Vlijmen, H., Karplus, M. Evaluation of comparative protein modeling by MODELLER. Proteins: Struct., Funct., Genet. 23:318–326, 1995.

[62] Samudrala, R., Pedersen, J., Zhou, H., Luo, R., Fidelis, K., Moult, J. Confronting the problem of interconnected structural changes in the comparative modelling of proteins. Proteins: Struct., Funct., Genet. 23:327–336, 1995.

[63] Pieper, U., Kapadia, G., Zhu, P., Peterkofsky, A., Herzberg, O. Structural evidence for the evolutionary divergence of Mycoplasma from gram-positive bacteria: the histidine-containing phosphocarrier protein. Structure 3:781–790, 1995.

[64] Read, J., Brayer, G., Jurášek, L., James, M. Critical evaluation of comparative model building of *Streptomyces griseus* trypsin. Biochemistry 23:6570–6575, 1984.

[65] Kleywegt, G., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., Jones, T. Crystal structure of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure 2:1241–1258, 1994.

[66] Mosimann, S., James, M. X-ray crystallographic structure of recombinant eosinophil-derived neurotoxin at 1.83 Å resolution. J. Mol. Biol. 260:540–552, 1996.

[67] Bleasby, A., Wootton, J. Construction of validated, non-redundant composite protein sequence databases. Protein Eng. 3:153–159, 1990.

[68] Lipman, D., Pearson, W. Rapid and sensitive protein similarity searches. Science 227:1435–1441, 1985.

[69] Murzin, A., Brenner, S. E., Hubbard, T. J. P., Chothia, C. SCOP: Structural Classification of Proteins. <http://www.bio.cam.ac.uk/scop/>, 1997.

[70] Barton, G. J. Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol 183:403–428, 1990.

[71] Barton, G. J., Sternberg, M. J. E. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.

[72] Pedersen, J. T. G, a molecular modelling program available upon request to the author.

[73] McLachlan, A. Gene duplication in the structural evolution of Chymotrypsin. J. Mol. Biol. 128:49–79, 1979.

[74] Read, R. MUTATE, a program that implements the minimum perturbation method for building comparative models.

[75] MSI. DISCOVER and INSIGHT are trademarks of Biosym Technologies, San diego, California, USA.

[76] MSI. QUANTA is a trademark of MSI Technologies.

[77] Fidelis, K., Moult, J. Unpublished.

[78] Toner, M., Moult, J. Unpublished.

[79] Baumann, U., Huber, R., Bode, W., Grosse, D., Lesjak, M., Laurell, C. Crystal structure of cleaved human alpha 1-antichymotrypsin at 2.7 A resolution and its comparison with other serpins. J. Mol. Biol. 218:595–606, 1991.

[80] Pedersen, J. T., Moult, J. Ab initio structure Prediction for small polypeptides and protein fragments using genetic algorithms. Proteins: Struct., Funct., Genet. 23:454–460, 1995.

[81] Liao, D., Herzberg, O. Refined structures of the active Ser83-Cys and impaired Ser46-Asp histidine-containing phosphocarrier proteins. Structure 2:1203–1216, 1994.

[82] Jia, Z., Quail, J., M., V., Hengstenberg, W., Delbaere, L. The 1.6 Å structure of the histidine-containing phosphotransfer protein HPr from *Streptococcus faecalis*. J. Mol. Biol. 236:1341–1355, 1994.

[83] Jia, Z., Quail, J., Waygood, E. B., Delbaere, L. T. J. The 2.0 Å resolution structure of *Escherichia coli* histidine-containing phosphocarrier protein HPr: A redetermination. J. Biol. Chem. 268:22490–22501, 1993.

[84] Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J., Sacchettini, J. Three-dimensional structure of recombinant human muscle fatty acid-binding protein. J. Biol. Chem. 367:18541–18550, 1992.

[85] Winter, N., Bratt, J., Banaszak, L. The crystal structures of holo- and apo-cellular retinol binding protein II. J. Mol. Biol. 230:1247–1259, 1993.

[86] Lalonde, J., Bernlohr, D., Banaszak, L. X-ray crystallographic structures of adipocyte lipid binding protein complexed with palmitate and hexadecanesulfonic acid. Properties of cavity binding sites. Biochemistry 33:4885–4895, 1994.

[87] Sacchettini, J., Gordon, J., Banaszak, L. Crystal structure of rat intestinal fatty acid-bidning protein. Refinement and analysis of the E. coli-derived protein with bound palmitate. J. Mol. Biol. 208:327–339, 1989.

[88] Benning, M., Smith, A., Wells, M., H., H. Crystallisation, structure determination, and least-squares refinement to 1.75Å resolution of the fatty acid-binding protein isolated from *Manduca sexta*. J. Mol. Biol. 228:208–219, 1992.

[89] Wlodawer, A., Svensson, L., Sjolin, L., Gilliland, G. Structure of phosphate-free ribonuclease-A refined at 1.26 Å. Biochemistry 27:2705–2709, 1988.

[90] Mazzarella, L., Capasso, S., Demasi, D., Di Lorenzo, G. Bovine seminal ribonuclease structure at 1.9 Å resolution. Acta Cryst. 49:389, 1993.

[91] Mosimann, S., Ardelt, W., James, M. Refined 1.7Å x-ray crystallographic structure of P-30, an amphibian ribonuclease with anti-tumor activity. J. Mol. Biol. 236:1141–1153, 1994.

[92] Herzberg, O., Moult, J. Analysis of steric strain in the polypeptide backbone of protein molecules. Proteins: Struct., Funct., Genet. 11:223–229, 1991.

[93] Risler, J., Delorme, M., Delacroix, H., Mevarech, M. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. 204:1019–1029, 1988.

[94] Ponder, J., Richards, F. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.

[95] Holm, L., Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C_α trace: application to model building and detection of co-ordinate errors. J. Mol. Biol. 218:183–194, 1991.

[96] Wilson, C., Gregoret, L., Agard, D. Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. J. Mol. Biol. 229:996–1006, 1993.

[97] Unger, R., Moult, J. An analysis of protein folding pathways. Biochemistry 30:3816–3823, 1991.

[98] Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.

[99] Weiner, S., Kollman, P., Nguyen, D., Case, D. An all atom force field for simulations of proteins and nucleic acids. J. Comp. Chem. 7:230–252, 1986.

[100] Jorgensen, W., Tirado-Rives, J. The OPLS potential function for proteins. Energy minimisations for crystals of cyclic peptides and crambin. J. Amer. Chem. Soc. 110:1657–1666, 1988.

[101] Sippl, M. 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. 213:859–883, 1990.

[102] Bowie, J., Lüthy, R., Eisenberg, D. Method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.

[103] Jones, D., Taylor, W., Thornton, J. A new approach to protein fold recognition. Nature 258:86–89, 1992.

[104] Bryant, S., Lawrence, C. An empirical energy function for threading protein sequence through the folding motif. Proteins: Struct., Funct., Genet. 16:92–112, 1993.

[105] Sippl, M. Recognition of errors in three-dimensional structures of proteins. Proteins: Struct., Funct., Genet. 17:355–362, 1993.

[106] MacArthur, M., Laskowski, R., Thornton, J. Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. Curr. Opin. Struct. Biol. 4:731–737, 1994.

[107] Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., Sippl, M. Progress in fold recognition. Proteins: Struct., Funct., Genet. 23:376–386, 1995.

[108] Jones, D., Miller, R., Thornton, J. Successful protein fold recognition by optimal sequence threading validated by rigourous blind testing. Proteins: Struct., Funct., Genet. 23:387–397, 1995.

[109] Madej, T., Gibrat, J., Bryant, S. Threading a database of protein cores. Proteins: Struct., Funct., Genet. 23:356–369, 1995.

[110] Skolnick, J., Kolinksi, A. Simulations of the folding of a globular protein. Science 250:1121–1125, 1990.

[111] Sun, S. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. Protein Sci. 2:762–785, 1993.

[112] Subramaniam, S., Tcheng, D. K., Fenton, J. A knowledge-based method for protein structure refinement and prediction. In States, D., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R., editors, *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology*, pages 218–229, 1996.

[113] Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., Tsumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.

[114] Mosteller, F., Rourke, R., Thomas Jr., G. *Probability with statistical applications.* Addison-Wesley Publishing Company, 1970.

[115] Bahar, I., Jernigan, R. Inter-residue potentials in globular proteins: dominance of highly specific hydrophilic interactions at close separation. J. Mol. Biol. 266:195–214, 1997.

[116] Head-Gordon, T., Brooks, C. Virtual rigid body dynamics. Biopolymers 31:77–100, 1991.

[117] Orengo, C., Michie, A., Jones, S., Swindells, M., Jones, D., Thorton, J. Protein Structure Classification. <`http://www.biochem.ucl.ac.uk/-bsm/cath/`>, 1993.

[118] Braxenthaler, M., Samudrala, R., Pedersen, J., Luo, R., Milash, B., Moult, J. PROSTAR: The protein potential test site. <`http://prostar.-carb.nist.gov`>, 1997.

[119] Holm, L., Sander, C. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225:93–105, 1992.

[120] Avbelj, F., Moult, J. Determination of the conformation of folding initiation sites in proteins by computer simulation. Proteins: Struct., Funct., Genet. 23:129–141, 1995.

[121] Pedersen, J. T., Moult, J. Folding simulation with genetic algorithms and a detailed molecular description. J. Mol. Biol. 269:240–259, 1997.

[122] Avbelj, F., Moult, J., Kitson, H., James, M., Hagler, A. Molecular dynamics study of the structure and dynamics of a protein molecule in crystalline ionic environment, *Streptomyces griseus* Protease A. Biochemistry 29:8658–8676, 1990.

[123] Huang, E., Subbiah, S., Levitt, M. Recognising native folds by the arrangement of hydrophobic and polar residues. J. Mol. Biol. 252:709–720, 1995.

[124] Sippl, M., Ortner, M., Jaritz, M., Lackner, P., Flöckner, H. Helmholtz free energies of atom pair interactions in proteins. Folding and Design 1:289–298, 1996.

[125] Kabsch, W., Mannherz, H., Suck, D., Pai, E., Holmes, K. Atomic structure of the actin:DNase 1 complex. Nature (London) 347:37–44, 1990.

[126] Wendolski, J., Salemme, F. PROBIT: a statistical approach to modeling proteins from partial coordinate data using substructure libraries. J. Mol. Graph. 10:124–127, 1992.

[127] Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci. 1:409–417, 1992.

[128] Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. A new approach to the rapid determination of protein side chain conformations. J. Mol. Biol. 217:373–388, 1991.

[129] Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.

[130] McGregor, M., Islam, S., Sternberg, M. Analysis of the relationship between side chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198:295–310, 1987.

[131] Dunbrack, R., Karplus, M. Conformational analysis of the backbone-dependent rotamer preferences for protein sidechains. Nature: Struct.Biol. 1:334–340, 1994.

[132] Creamer, T., Rose, G. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. Proc. Natl. Acad. Sci. USA 13:5937–5941, 1992.

[133] Creamer, T., Rose, G. Alpha-helix-forming propensities in peptides and proteins. Proteins: Struct., Funct., Genet. 19:85–97, 1994.

[134] Fraenkel, A. Complexity of protein folding. Bull. Math. Biol. 55:1199–1210, 1993.

[135] Unger, R., Moult, J. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. Bull. Math. Biol. 55:1183–1198, 1993.

[136] Chen, R. Monte carlo simulations for the study of hemoglobin-fragment conformations. J. Comp. Chem. 10:488–494, 1989.

[137] Wilson, S., Cui, W. Applicatons of simulated annealing to peptides. Biopolymers 29:225–235, 1990.

[138] Okamoto, Y., Fukugita, M., Nakazawa, T., Kawai, H. α-helix folding by monte carlo simulated annealing in isolated C-peptide of Ribonuclease A. Protein Eng. 4:639–647, 1991.

[139] Abagyan, R., Totrov, M. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235:983–1002, 1994.

[140] Unger, R., Moult, J. Genetic algorithms for protein folding simulations. J. Mol. Biol. 231:75–81, 1993.

[141] Harel, D. *Algorithmics.* Garland Publishing Inc., 1992.

[142] National Research Council. *Mathematical challenges from theoretical/computational chemistry.* National Academy Press, 1995.

[143] Grindley, H., Artymiuk, P., Rice, D., Willet, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J. Mol. Biol. 229:707–721, 1993.

[144] Artymiuk, P., Poirrette, A., Rice, D., Willett, P. Comparison of protein folds and sidechain clusters using algorithms from graph theory. In *Proceedings of the CCP4 Study Weekend*, June 1995.

[145] Bron, C., Kerbosch, J. Algorithm 457: Finding all cliques of an undirected graph. Comm. ACM 16:575–577, 1973.

[146] Bhat, T., Bentley, G., Boulot, G., Green, M., Tello, D., Dallacqua, W., Souchon, H., Schwarz, F., Mariuzza, R., Poljak, R. Bound water molecules and conformational stabilisation help mediate an antigen-antibody association. Proc. Natl. Acad. Sci. USA 91:1089–1093, 1994.

[147] Moult, J. CONANA, a program to analyse intermolecular crystallographic contacts.

[148] Moon, J., Moser, L. On cliques in graphs. Israel J. Math. 3:23–28, 1965.

[149] Tarjan, R., Trojanowski, A. Finding a Maximum Independent Set. SIAM J. Comput. 6:537–546, 1977.

[150] Bycroft, M., Hubbard, T., Proctor, M., Freund, S., A.G., M. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. Cell 88:235–242, 1997.

[151] Hart, P., Nersissian, A., Herrmann, R., Nalbandyan, R., Valentine, J., Eisenberg, D. A missing link in cupredoxins: Crystal structure of cucumber stellacyanin at 1.6 Å resolution. Protein Sci. 5:2175–2183, 1996.

[152] Tong, H., Sixma, T. Ubiquitin conjugatin enzyme structure. Unpublished, 1997.

[153] Kleywegt, G., Zou, J., Divne, C., Davies, G., Sinning, I., Stahlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T., Jones, T. The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 Å with related enzymes. Unpublished results, 1997.

[154] Bryant, S., Hubbard, T., Moult, J. Critical Assessment of protein Structure Prediction methods (2) web page. <`http://iris4.carb.nist.-gov-casp2/`>, 1997.

[155] Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.

[156] Eddy, S., Mitchison, G., Durbin, R. Maximum Discrimination Hidden Markov Models of Sequence Consensus. J. Comp. Biol. 2:9–23, 1995.

[157] Rost, B., Schneider, R., Sander, C. The PredictProtein server. <`http://-www.embl-heidelberg.de/predictprotein/`>, 1996.

[158] Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Struct., Funct., Genet. 9:56–68, 1991.

[159] Rodrigues-Tome, P., Stoehr, P., Cameron, G., Flores, T. The European Bioinformatics Institute (EBI) databases. Nucleic Acids Res. 24:6–12, 1996.

[160] Smith, T., Waterman, M. Identification of common molecular subsequences. J. Mol. Biol. 147:197–197, 1981.

[161] Flaherty, K., Zozulya, S., Stryer, L., McKay, D. Three-dimensional structure of recoverin, a calcium sensor in vision. Cell 75:709–716, 1993.

[162] Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J., Teeri, T., Jones, T. The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei. Science 265:524–528, 1994.

[163] Cook, W., Jeffrey, L., Sullivan, M., Vierstra, R. Three-dimensional structure of a ubiquitin-conjugating enzyme (E2). J. Biol. Chem. 267:15116–15121, 1992.

[164] Cook, W., Jeffrey, L., Xu, Y., Chau, V. Tertiary structures of class I ubiquitin-conjugating enzymes are highly conserved: crystal structure of yeast Ubc4. Biochemistry 32:13809–13817, 1993.

[165] Guss, J., Merritt, E., Phizackerley, R., H.C., F. The structure of a phytocyanin, the basic blue protein from cucumber, refined at 1.8 A resolution. J. Mol. Biol. 262:686–705, 1996.

[166] Schindelin, H., M.A., M., Heinemann, U. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. Nature (London) 364:164–168, 1993.

[167] Schindelin, H., Jiang, W., Inouye, M., Heinemann, U. Crystal structure of CspA, the major cold shock protein of Escherichia coli. Proc. Natl. Acad. Sci. USA 91:5119–5123, 1994.

[168] Zemla, A., Venclovas, C., Fidelis, K., Moult, J. Ab initio protein structure prediction and comparative modeling evaluator. <`http://predictioncenter.llnl.gov/`>, 1997.

[169] Chothia, C. One thousand families for the molecular biologist. Nature (London) 357:543–544, 1992.