A Fold Recognition Approach to Modeling of Structurally Variable Regions
(HS-IKI-MD-04-004)

Christer Levefelt (a94chrle@student.his.se)

School of Humanities and Informatics,
University of Skövde, P.O. Box 408,
SE-541 28 Skövde, Sweden

Master Dissertation at the New Generation Systems study program, 2004.

Supervisor: Dan Lundh

A Fold Recognition Approach to Modeling of Structurally Variable Regions

Submitted by Christer Levefelt to the University of Skövde as a dissertation towards the degree of M.Sc. by examination and dissertation in the School of Humanities and Informatics.

August 31, 2004

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: _____

A Fold Recognition Approach to Modeling of Structurally Variable Regions

Christer Levefelt (a94chrle@student.his.se)

Abstract

A novel approach is proposed for modeling of structurally variable regions in proteins. In this approach, a prerequisite sequence-structure alignment is examined for regions where the target sequence is not covered by the structural template. These regions, extended with a number of residues from adjacent stem regions, are submitted to fold recognition. The alignments produced by fold recognition are integrated into the initial alignment to create a multiple alignment where gaps in the main structural template are covered by local structural templates. This multiple alignment is used to create a protein model by existing protein modeling techniques.

Several alternative parameters are evaluated using a set of ten proteins. One set of parameters is selected and evaluated using another set of 31 proteins. The most promising result is for loop regions not located at the C- or N-terminal of a protein, where the method produces an average RMSD 12% lower than the loop modeling provided with the program MODELLER. This improvement is shown to be statistically significant.

Keywords: protein structure prediction; loop modeling; fold recognition; threading; structurally variable regions

Acknowledgements

I would like to thank my supervisor Dan Lundh for the initial project idea and for his guidance during the work on this dissertation.

Contents

1 Introduction

Current methods for protein structure prediction frequently use an alignment between the target amino acid sequence and a structural template derived from a known protein. The conformation of alignment regions where the sequence is not covered by the template structure must be determined by an alternative approach, termed loop modeling. Together with alignment errors, loop modeling is a major limitation (Fiser et al., 2000) of protein structure prediction methods.

The purpose of this dissertation is to investigate if a fold recognition approach to loop modeling can improve the quality of protein models.

A novel approach to loop modeling is proposed. This approach applies fold recognition to sequences not covered by a structural template. Alignments to local templates identified by fold recognition are integrated into the initial sequence-structure alignment to create a multiple alignment consisting of the target sequence, the main template structure and one local template structure for each region not covered by the main template structure.

The rest of this dissertation is organized as follows. Chapter 2 presents background on protein structure prediction and current methods for loop modeling. Chapter 3 contains the problem description, aim and objectives for the dissertation. Chapter 4 gives a detailed description of the method used in this work. Results of modeling are presented in chapter 5. Chapter 6 discusses and analyzes these results and conclusions are presented in chapter 7.

2 Background

Proteins are macromolecules, consisting of one or more chains of amino acids whose sequence is encoded in the genome of an organism. The chains of a protein fold into a three-dimensional structure that determines its functional properties. Proteins perform many important tasks in organisms (Rost, 1997), such as catalysis of biochemical reactions, transport of nutrients, recognition, and transmission of signals.

Knowledge of protein structures plays a central role in our understanding of the processes of life and has many practical applications. In medicine, for example, knowledge of the structure of a protein involved in harmful biological processes can make it possible to design a drug to target that specific protein. The drug binds specifically to active sites of the protein, thus hitting the problem at its source and minimizing deleterious side effects (Berman et al., 2002).

2.1 Protein Sequence and Structure Relationships

The structure of a protein is determined by its amino acid sequence plus its native solution environment (Rost, 1997). There are three levels of structural relationships (Jones & Hadley, 2000): (i) proteins within the same family have high sequence identity and highly similar structures, (ii) proteins within a superfamily share a common ancestry and may have low to insignificant sequence identity, but still share similar folds, (iii) analogous folds may not be evolutionarily related, but share the same major secondary structure elements with the same arrangement and connectivity.

2.2 Protein Structure Prediction

The Protein Data Bank (PDB)[1] (Berman et al., 2000) is the worldwide archive of structural data of biological macromolecules. These structures are determined experimentally using methods such as X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy (Bourne et al., 2004). Experimental methods for structure determination are, however, time-consuming and limited in their application (Jones & Hadley, 2000) and will therefore not be able to keep up with the increasing number of known protein sequences. At the time of writing, the latest update of PDB (August 24, 2004) held 26,880 experimentally determined structures while the latest release (143) of GenBank[2] (Benson et al., 2003), a comprehensive public database of nucleotide and protein sequences, contained 37,343,937 entries. This huge gap between available sequences and structures has led to great interest in computational methods for prediction of the three-dimensional structure of proteins from their sequence (Schwede et al., 2003).

Traditionally, computational methods for protein structure prediction are divided into three categories (Moult et al., 2003):

- Comparative or homology modeling (Martí-Renom et al., 2000) predicts structure primarily based on similarity between the sequences of the target protein and those of one or more template proteins of known structure. Comparative modeling works because small differences between sequences

[1] Available at URL http://www.rcsb.org/pdb/.

[2] Available at URL http://www.ncbi.nlm.nih.gov.

usually correspond to small differences between structures, but a detectable similarity between sequences is a necessary condition.

- Fold recognition or threading (Jones & Hadley, 2000) is based on the observation that a large percentage of proteins adopt one of a limited number of folds. Where there is a high sequence similarity to a protein of known structure, fold recognition is trivial, but it has the capability to detect structural similarities even in the absence of any detectable sequence similarity.
- New fold methods, also referred to as *ab initio* methods (Moult, 1999; Moult et al., 2003), are intended to construct structural models for a protein sequence from first principles, i.e. without reliance on any direct relationship to a known structure. In practice, though, many of these methods make extensive use of available structural information in devising scoring functions, distinguishing between correct and incorrect predictions and in choosing fragments to incorporate into the model.

There is now general agreement (Moult et al., 2003) that changes in the nature of structure modeling have made these categories outdated, e.g. improved sequence comparison techniques have blurred the boundary between comparative modeling and fold recognition. This work mainly concerns comparative modeling and fold recognition. A more detailed outline of these approaches is stated below.

2.2.1 Comparative Modeling

Current comparative modeling methods consist of four steps (Martí-Renom et al., 2000):

1. Fold assignment and template selection. This identifies all structures related to the target sequence and selects one or more of them to be used as templates.
2. Template–target alignment. Most fold assignment methods produce an alignment between the target sequence and template structures.
3. Model building. The target–template alignment is used to construct a 3D model of the protein.
4. Model evaluation. The quality of a model determines what information can be extracted from it, thus estimating the accuracy of models is essential for interpreting them.

If the model is not satisfactory, the first three steps can be repeated.

Available comparative modeling tools include:

- MODELLER (Šali and Blundell, 1993; Fiser et al., 2000; Martí-Renom et al., 2000) is a program implementing comparative modeling by satisfaction of spatial restraints derived from an alignment between a target sequence and related structures. MODELLER can also perform many additional tasks including loop modeling.
- SWISS-MODEL[3] (Schwede et al., 2003) is an automated comparative modeling server. Three levels of user interaction are provided, first approach mode, requiring only an amino acid sequence as input, alignment mode,

[3] Avialable at URL http://swissmodel.expasy.org.

requiring a sequence-structure alignment and project mode, allowing submission of a manually optimized modeling request.

2.2.2 Fold Recognition

A generic fold recognition method is outlined by Jones and Hadley (2000), requiring:

- A 'fold library' of unique or representative structural templates derived from the database of all known protein structures. Depending on method, the fold library may consist of complete protein chains, structural domains or even conserved protein cores.
- An algorithm for aligning the target protein sequence to a protein structure. This is used to optimally align the target sequence in turn to each fold from the fold library (allowing for insertions and deletions in loop regions). Several different algorithms have been proposed, such as 'double' dynamic programming (Jones et al., 1992) and branch-and-bound searching (Lathrop & Smith, 1996).
- A function to determine the goodness of fit between a sequence and a structure. This is used by the alignment algorithm to optimize the sequence-structure alignment. Most frequently used is some kind of 'pseudo energy' function. A review of scoring functions is provided by Jones and Thornton (1996).

The result of a fold recognition method is a ranking of the fold library according to the 'goodness of fit' of the respective alignments, with the best fitting fold considered the most probable match. When a fold has been selected, the alignment can be passed to an automatic comparative modeling program for modeling of loops and side chains, creating a complete three-dimensional model.

The success of a fold recognition method as described above is dependent on several factors (Jones & Hadley, 2000), namely (i) the quality of the fold library, (ii) the use of an appropriate scoring function, (iii) the algorithm being capable of producing high-quality alignments for sequences onto template folds, and (iv) the post-processing of results.

Available fold recognition tools include:

- THREADER (Jones et al., 1992; Jones et al., 1995; Jones, 1998) uses a double dynamic programming algorithm to align a target sequence to a template structure, taking into account detailed pairwise interactions.
- GenTHREADER[4] (Jones, 1999; McGuffin & Jones, 2003) uses traditional sequence alignment algorithms, which are then evaluated using pairwise potentials and solvation potentials. Potentials and other scores related to different aspects of the sequence-structure alignment are evaluated by a neural network to create a single measure of confidence for alignments.
- 3D-PSSM[5] (Fischer et al., 1999; Kelley et al., 2000) aligns target sequences to structures using a position-specific scoring matrix (PSSM) created from a large multiple alignment based on all members of a superfamily.

[4] Available at URL http://www.psipred.net.

[5] Available at URL http://www.bmm.icnet.uk/servers/3dpssm.

Additional references on fold recognition methods can be found in (Jones & Hadley, 2000) and (McGuffin & Jones, 2003).

2.2.3 Loop Modeling and Structurally Variable Regions

Although the approaches discussed have been shown to be successful, improvements are still necessary to overcome missing structural template regions in the alignment. This is achieved by loop modeling.

The word 'loop' is surrounded by some terminological confusion. At least two different meanings are applied to the term. According to van Vlijmen and Karplus (1997), loops are segments that do not correspond to α-helical or β-strand secondary structure elements.

Moult (1999) defines loops as regions, typically occurring between secondary structure elements, where there are insertions and deletions in the target sequence relative to that of the template(s), or a local loss of sequence similarity. The presence of loops prevents these regions of the backbone to be usefully copied from the template structure.

The term 'structurally variable region' is used by Rohl et al. (2004) for gaps, insertions and regions of low-confidence alignment. This term is better suited for loops in the context of a sequence-structure alignment, as it is not burdened by any alternate meaning. When referring to loops and loop modeling, this dissertation will use the definition of Moult (1999).

In the modeling of an alignment such as the one in Figure 2.1a, some regions are not covered by the structural template. Loop modeling is the process of determining conformations for such regions. Figure 2.1b shows which parts of the structure would have to be determined by alternate means. Here can also be seen that gaps in the template structure do not necessarily occur between secondary structure elements.

Loops often determine functional specificity of a protein, contributing to active and binding sites (Fiser et al., 2000). Consequently, the accuracy of loop modeling is a major factor in creation of useful protein models.

Loop modeling methods can be grouped into three categories (Rohl et al., 2004):

- Knowledge-based methods use known protein structures as a source of loop conformations. Likely conformations are selected based on evaluation using a knowledge-based potential or rule-based filters.
- *De novo* or *ab initio* strategies generate loop conformations by methods such as molecular dynamics, simulated annealing, exhaustive enumeration or heuristic sampling of a discrete set of (φ, ψ) angles, random tweak, or analytical methods.
- Combined approaches. These combine knowledge-based and *de novo* methods in a hybrid approach.

See Fiser et al. (2000) and Rohl et al. (2004) for more references on loop modeling.

a)

```
>P1;T0139
sequence:T0139:::::::
-----------------------------TGISRETSSDVALASHILTALREKQ-------APELSLSSQ-------DLE---LV
>P1;1S3J
structure:1S3J::A::A::::
SADQLMSDIQLSLQALFQKIQPEMLESMEKQGVTPA-------QLFVLASLKKHGSLKVSEIAERMEVKPSAVTLMADRLEQKNLI

>P1;T0139
sequence:T0139:::::::
TKE----DPKALAVALNWDIKKTETVQEACERELALRLQQTQSLHSLR------------------------*
>P1;1S3J
structure:1S3J::A::A::::
ARTHNTKDRRVIDLSL------TDEGDIKFEEVLAGRKAIMARY--LSFLTEEEMLQAAHITAKLAQAAETD*
```

b)

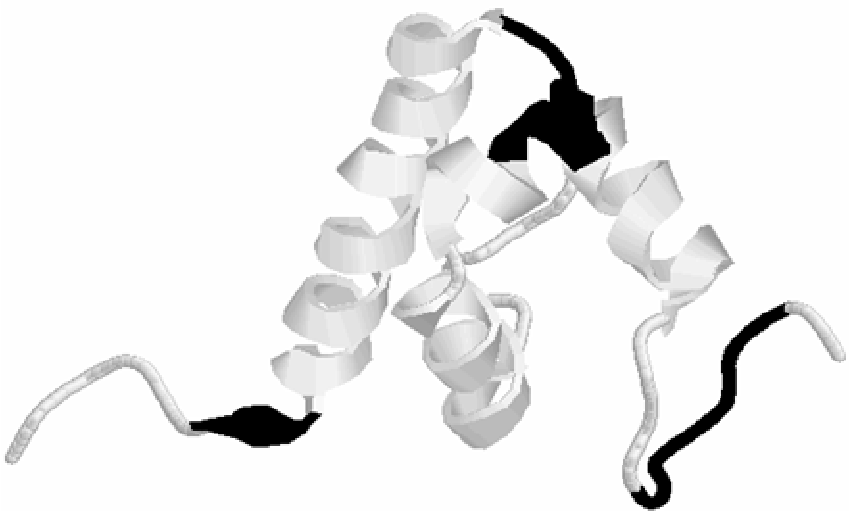

Figure 2.1 a) Alignment between CASP target sequence T0139 (Kinch et al., 2003) and template structure 1S3J. Alignment gaps in the structure are highlighted. b) The native structure of T0139, 1IYR. Residues in its sequence which have been aligned to gaps in the template are highlighted in black. This and other protein figures were created by Protein Explorer[6] (Martz, 2002).

[6] Available at URL http://proteinexplorer.org.

3 Problem Description

As protein function is determined by structure, determination of protein structures is essential for our understanding of the processes of life and is critical to many important areas such as drug design. Loops often determine functional specificity of a protein, contributing to active and binding sites (Fiser et al., 2000). Consequently, the impact of an accurate loop modeling method would be large. Existing methods are reasonably accurate for modeling of short loop regions, but modeling of longer structurally divergent regions is an unsolved problem (Rohl et al., 2004). Fiser et al. (2000) noted that in the first two CASPs, there was no reliable method available for constructing loops longer than five residues, but that recently progress has been made (see Fiser et al. (2000) for references). For example, van Vlijmen and Karplus (1997) suggested an algorithm for loops of nine residues or less. Rohl et al. (2004) presented a promising method for prediction of longer structurally variable regions.

3.1 Hypothesis

Loop modeling can be seen as a mini-protein folding problem where the conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself (Fiser et al., 2000). This indicates that regular structure prediction methods could be applied to loop modeling with some modification. One such method is fold recognition.

Fold recognition is traditionally applied to entire protein chains, finding global folds. However, this work is based on the hypothesis that the conformations of local folds can be analogous to local folds occurring in other known proteins. This would allow a fold recognition method to be applied locally to determine the conformation of regions in a sequence-structure alignment which are not covered by the main template structure.

3.2 The Proposed Method

Such a fold recognition approach to loop modeling or modeling of structurally variable regions is described in Figure 3.1. An initial sequence-structure alignment is used as input. Sequence regions which are aligned to gaps in the template structure are extracted. The sequences are extended with a number of residues from adjacent stem regions to facilitate subsequent modeling. These sequences are submitted to fold recognition and the alignments obtained are integrated into the initial alignment to create a multiple alignment which can be used by modeling software for creation of a protein model.

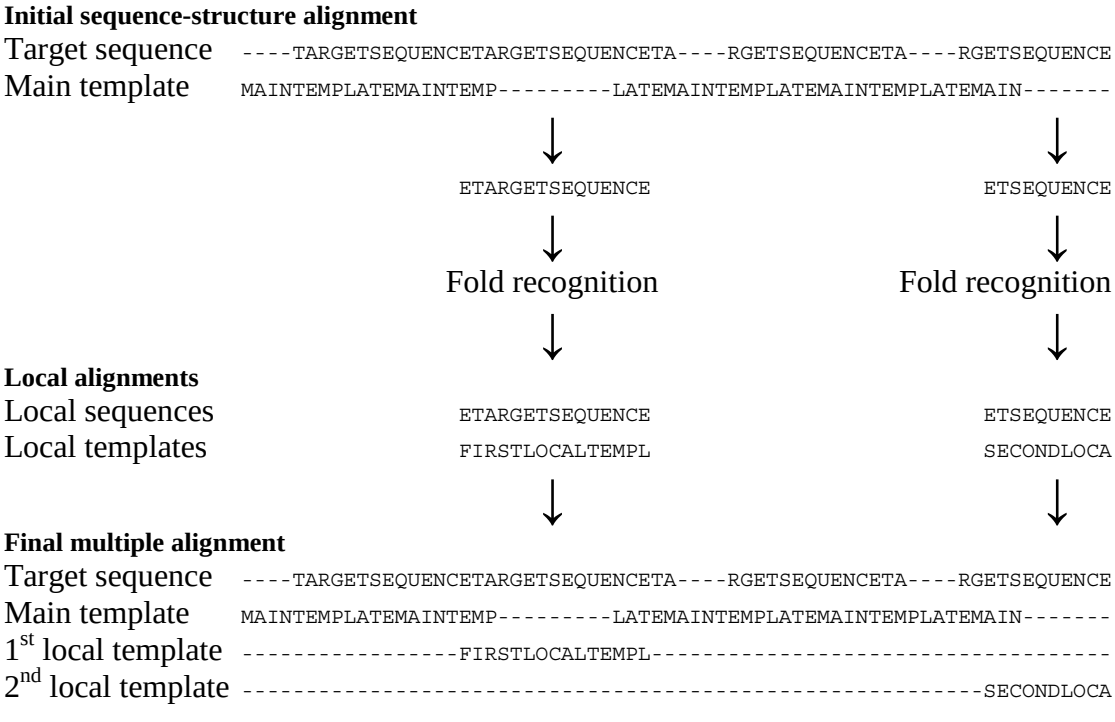

Figure 3.1 The proposed approach to modeling of structurally variable regions by fold recognition.

3.3 Related Work

Svensson et al. (2004) have predicted the protein structure of a putative gene, Flowering Regulating Factor (FRF), in the *Arabidopsis thaliana* genome. Fold recognition using the THREADER tool assigned the fold 2LBD to the sequence. This fold, however, had two long gaps in the alignment, 10 and 23 residues long. Fold recognition was applied to these sequences, assigning them the folds 1GNF and 1KQ1, respectively. The MODELLER tool was then used to build a model with these three folds as templates. The C-terminal region of FRF (117 residues) was not included in the modeling experiment since it did not correspond to 2LBD. Verification of models was performed using Ramachandran plots (Ramachandran et al., 1963) and the software PROCHECK (Laskowski et al., 1993). The best model had an energy profile similar to that of the template and no high-energy regions could be detected. The apparently high quality of the model indicates some promise for a fold recognition approach to loop modeling.

Also of interest could be work on optimizing alignments, but this has not been investigated.

3.4 Aim

The aim of this dissertation is to investigate whether additional structural templates determined by fold recognition of gap regions can improve the model generated from an alignment.

To assess model improvement, models built using existing loop modeling techniques would be compared to models created using a multiple alignment with local structural templates determined by fold recognition.

3.5 Objectives

To achieve the aim the following objectives were derived:

1. Selection and preparation of two sets of proteins, a training set for deriving parameters and a test set for testing. Proteins should be representative of typical targets for structure prediction and must contain a sufficient number of gap regions for conclusions to be drawn with reasonable accuracy. To facilitate comparison to other prediction methods, it would be advantageous to use protein sets for which prediction results of other methods have been published.
2. Identification of factors that could affect the result of a fold recognition approach to loop modeling. Selection of some promising alternatives between the factors for how fold recognition could be applied to a gap region, as well as for how to interpret results from fold recognition. Exhaustive testing of these alternatives by applying every combination of the factors to the training set. Creation of protein models from the multiple alignments produced.
3. Analysis of models built from the training set to select the most promising combination of factors for constructing a prototype method for fold recognition-based modeling of structurally variable regions.
4. Application of the proposed method to the test set to get an indication of the applicability of the derived method. Creation of protein models for evaluation of the method.
5. Analysis of models built from the test set to allow assessment of the quality of the proposed method.

4 Method

This chapter describes the steps, shown in Figure 4.1, that were followed in order to develop a method for loop modeling by fold recognition. Results from the steps are presented in the Results chapter.

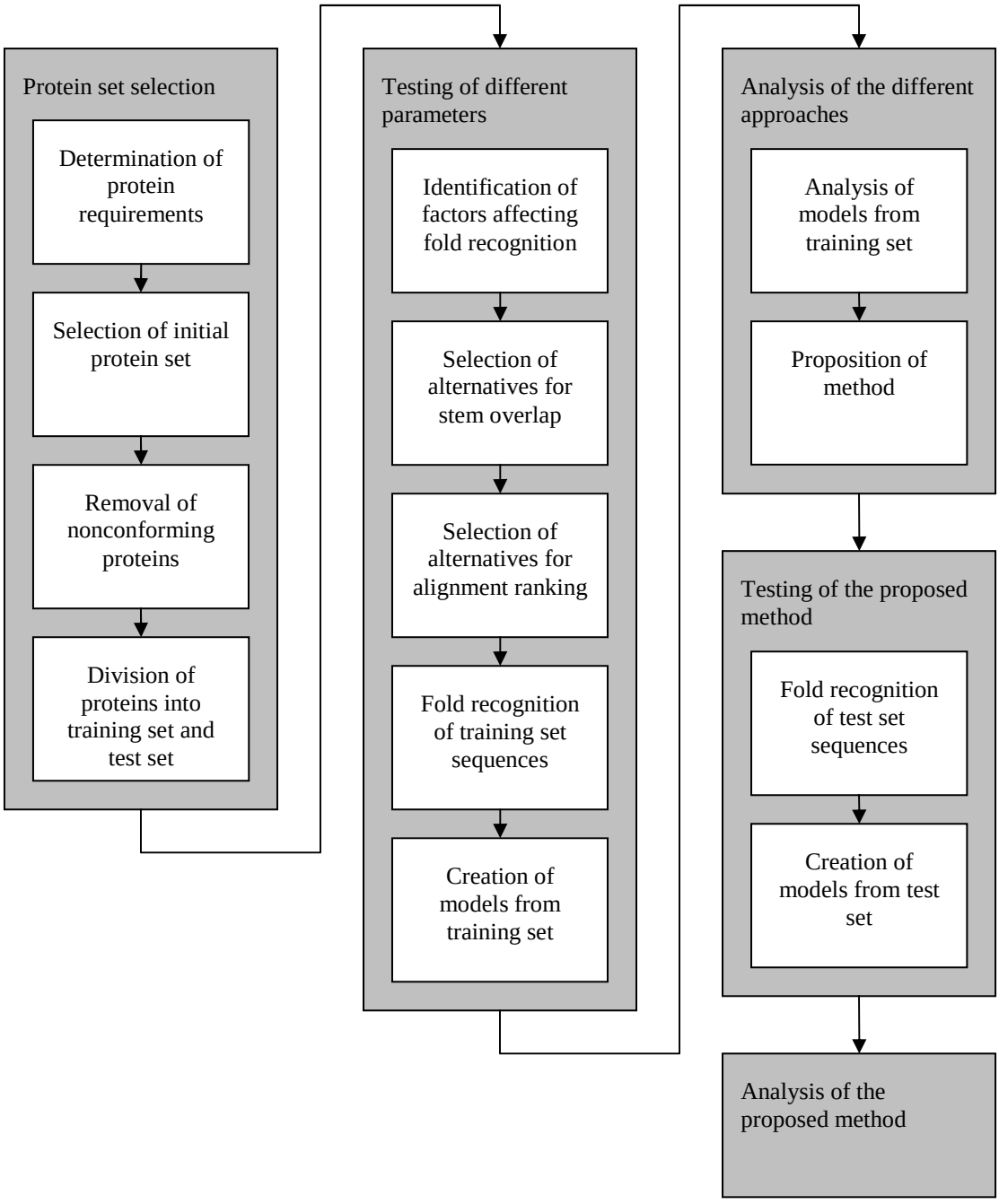

Figure 4.1 The method.

Two sets of proteins were selected, a training set and a test set. Commonly used in deriving artificial intelligence applications, a training set is used for learning (fitting of parameters) and a test set is used only to assess performance (Ripley, 1996, p. 354). Using two independent sets avoided optimistic bias from evaluation of a method on data that was used to create the method. Different parameters were proposed for a fold recognition approach to loop modeling. The training set was used to exhaustively test

all combinations of parameters identified. Based on the results, a method was proposed and was evaluated using the independent test set.

4.1 Protein Set Selection

The first objective was further divided into four sub-objectives:

1. Determination of requirements for proteins to be included in one of the protein sets.
2. Selection of an initial set of proteins.
3. Removal of proteins not conforming to requirements.
4. Division of remaining proteins into a training set and a test set.

A set of proteins were selected for modeling. Using an appropriate set of proteins is important to ensure the relevance of the results. To this end, the following requirements were defined for the proteins to be used:

- Given an alignment between each target protein and a template structure, each alignment should contain at least one gap of appropriate length. Current loop modeling can accurately predict backbone conformations for loops up to nine residues in length (van Vlijmen & Karplus, 1997). Also, fold recognition is not suited for very short sequences since the length of the threaded sequence is inversely proportional to the number of folds identified, i.e. a sequence of one residue in length can be threaded onto any structure. Because of this, only gap regions of length ten or greater will be considered. It would be advantageous if the protein sets contained gap regions of differing length, since this could give an indication of whether the applicability of the approach is dependent on the length of the sequence to which it is applied.
- The proteins should each have experimentally determined structures available. These are needed for evaluation of protein models.
- The protein sets should be large enough for conclusions to be drawn with reasonable confidence. Van Vlijmen and Karplus (1997) used two protein sets of 13 and 8 proteins, defined by Leszczynski and Rose (1986) and Tramontano and Lesk (1992), respectively. Based on the sizes of these sets, the minimum size of each set was set to ten proteins. The test set was allowed to be larger, but the size of the training set was kept at ten proteins, as exhaustive evaluation of the training set would require considerably more computational resources per protein than evaluation of the test set.
- The protein sets should contain prediction targets which are known to have been previously used in protein structure prediction. This will minimize the risk of using proteins that are not representative for structure prediction targets in general or that are otherwise unfit. This will also facilitate performance comparisons of the proposed method to that of other methods.

Given these requirements, the CASP5 targets (Kinch et al., 2003) were initially selected. The CASP4 targets (Murzin & Hubbard, 2001) were later added to increase the number of proteins.

CASP (Critical Assessment of Protein Structure Prediction) is a series of communitywide experiments for evaluation of computational methods for protein structure prediction. To date, there have been five such experiments; CASP1 (Moult

et al., 1995), CASP2 (Moult et al., 1997), CASP3 (Moult et al., 1999), CASP4 (Moult et al., 2001) and CASP5 (Moult et al., 2003). In the experiments, prediction methods are assessed based on the analysis of blind predictions of protein structures. The basic structures of the experiments have been similar (Moult, et al., 2003):

1. Information about structures on the verge of being solved was collected from the experimental community and passed on to the prediction community.
2. Prediction teams deposited predicted models before the experimental results were made public.
3. Predicted models were compared with experiments by using numerical evaluation techniques as well as human assessment.

The CASP targets are well known in protein structure prediction and some of the most successful prediction methods have been applied to them. Murzin and Hubbard (2001) and Kinch et al. (2003) present CASP targets along with PDB identifiers of their structures, where known. Protein sequences were downloaded from the CASP homepages at the Lawrence Livermore National Laboratory Protein Structure Prediction Center[7]. Protein structures were downloaded from the PDB. In order to find any additional native structures not known at publication of the CASP targets, a FASTA search was carried out for each sequence against the PDB. This produced native structures for several more proteins. It also revealed that "native" structures not always had a 100% sequence identity to the CASP target sequences. Typically, the best alignments produced by FASTA searches had a 90%-100% sequence identity, with a sharp drop for successive alignments to around 20%-50% sequence identity. Because of this, it was decided to consider structures with a sequence identity above 85% as "native".

For proteins where several PDB entries were identified as native, one structure and one specific chain within that structure was selected as "true" for purposes of validation of models. In order to select the best structure the following criteria were used in the order specified:

1. Low E-value, favoring structures similar to the target sequence.
2. Structures determined by X-ray diffraction were favored over those determined by NMR. This was to simplify logistics by avoiding the multiple structure versions present in NMR entries.
3. High resolution (for X-ray-determined structures), favoring structures of higher quality.
4. For NMR-determined structures entries containing averaged structures were preferred to those containing multiple structures, again to simplify logistics.

An initial sequence-structure alignment was created for each sequence for which a native structure had been determined. This was done by submitting sequences to the fold recognition tool GenTHREADER (Jones, 1999; McGuffin & Jones, 2003). Filtering options were left at their default settings (masking of low complexity regions). The best scoring alignment for each sequence was selected unless that alignment was to a structure which had been identified as native for that sequence, in which case the next best alignment would be chosen. The reasons for ignoring native structures were to simulate fold recognition of a typical sequence, for which no native

[7] Available at URL http://predictioncenter.llnl.gov.

structure would be available and to not produce perfect alignments since these obviously would not contain any gap regions, making them unsuitable for this work. It should be noted that the use of fold recognition in the creation of the initial alignment is not required by the method; its only purpose is to produce an alignment whose gap regions may be examined. To this end, comparative modeling might be used instead.

Each alignment was examined for regions where the target sequence was aligned to a gap in the template sequence. All such regions of ten or more residues in length were noted and the corresponding proteins included in one of the protein sets. The proteins were randomly divided into a training set of ten proteins and test set containing the rest of the proteins.

4.2 Testing of Different Parameters

The second objective was further divided into five sub-objectives:

1. Identification of factors that could affect the result of a fold recognition approach to loop modeling.
2. Selection of some promising alternatives for how fold recognition could be applied to each gap region.
3. Selection of some promising alternatives for how results from fold recognition could be interpreted and ranked.
4. Application of fold recognition to each sequence in the training set, exhaustively using all combinations of these alternatives.
5. Creation of models from the multiple alignments produced.

According to Martí-Renom et al. (2000), the conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, they note that loops are generally too short to provide sufficient information about their local fold, and thus the conformation of a given segment is also influenced by the core stem regions that span the loop and by the structure of the rest of the protein that cradles the loop.

According to the hypothesis of this work (see section 3.1) fold recognition of the sequence of a loop segment would be able to determine the local fold of the loop. The influence of stem regions could be accounted for by including additional residues on either side of the loop region in the sequence submitted to fold recognition. Influences from the rest of the protein structure are unfortunately not as easy to incorporate in a fold recognition approach.

Based on these influences on loop conformation three possible alternatives were proposed for generating the sequence to submit to fold recognition:

- No stem overlap. The sequence of a gap region is used without modification. This means that structure is determined solely from the sequence of the gap region, ignoring influences from stem regions.
- Three residues stem overlap. Three residues immediately preceding and succeeding the gap region are included in the sequence submitted to fold recognition. This allows for a small influence from stem regions on loop conformation.

- Ten residues stem overlap. Ten residues immediately preceding and succeeding the gap region are included in the sequence submitted to fold recognition. This allows for a more significant influence from stem regions.

Three alternative strategies for ranking of alignments were proposed:

- Low solvation energy. Since loops are usually located at the surface of a protein, corresponding to low solvation energy, this was judged a promising approach.
- High alignment score. This would lead to a homology-like approach.
- A combined quality measure such as the score generated by the GenTHREADER neural network (Jones, 1999). This would allow several aspects of the quality of an alignment to influence ranking, possibly allowing for a more balanced ranking than relying on any single aspect.

The GenTHREADER tool was selected for fold recognition tasks on the basis that it is a representative tool which has been used successfully in a number of fold recognition assignments. It is very fast and reliable (Jones, 1999) and its neural network provides a combined quality measure. This was used for ranking according to the third ranking strategy. For solvation energy, GenTHREADER's solvation energy score was used. When more than one alignment had equal solvation energy, alignments with lower pairwise energy were favored. For ranking by alignment score, GenTHREADER's alignment score was used. When more than one alignment had equal scores, they were ranked first by longest alignment length and then by solvation energy.

All alignments from the training set had their gap region sequences extracted and submitted to GenTHREADER. For gap region sequences all filtering options were disabled. Three sequences were submitted for each gap region, one for each stem overlap alternative. From the results returned by GenTHREADER, the ranking alternatives were used to create three alternative local alignments for the corresponding sequence, one for each ranking strategy. Any alignment to a structure labeled as "native" would be ignored and the next best alignment would be used. This was to simulate prediction of a protein with no native structure available, since using a native structure as template for a gap region obviously would make the problem trivial.

The local alignments created from the gap region sequences were integrated into the initial sequence-structure alignment. This way, multiple alignments were created consisting of the target sequence, the main structural template and one local structural template for each gap region. Nine such multiple alignments were created from each initial alignment, one for each combination of stem overlap and alignment ranking.

Since evaluation of a model is more reliable than evaluation of an alignment (Martí-Renom et al., 2000), creation of models from the multiple alignments were necessary. The MODELLER tool (Šali and Blundell, 1993; Fiser et al., 2000; Martí-Renom et al., 2000) was used to build models for all ten alignments for each protein (the initial alignment and the nine multiple alignments). Five models were built from each alignment.

4.3 Analysis of the Different Approaches

The third objective was further divided into two sub-objectives:

1. Analysis of models built from the training set.
2. Proposition of a prototype method for modeling of structurally variable regions by fold recognition.

Models built using initial alignments containing gap regions (initial models) and models built using multiple alignments created by one of the proposed approaches (final models) were evaluated using three different measures of model quality:

- RMSD (Root Mean Square Deviation) between C_α atoms of the created model and the native protein chain. Improvement in models would be indicated by a lower RMSD for the final model than for the initial model.
- RMSD between model and native chain for C_α atoms in each gap region. Improvement in loop modeling would be indicated by a lower RMSD for the final gap region conformation than for the initial conformation.
- Ramachandran plot (Ramachandran et al., 1963) for each model. Improvements in structure quality would be indicated by a higher percentage of residues in most favored regions and a lower number of residues in disallowed regions. According to PROCHECK output, a high-quality model would be expected to have over 90% of its residues in most favored regions.

RMSD values were obtained by fitting using the McLachlan algorithm (McLachlan, 1982) as implemented in the program ProFit[8]. Ramachandran plots were created with the program PROCHECK (Laskowski et al., 1993).

Native structure files were prepared for evaluation by removing all chains but the one designated as "native". Because of problems getting ProFit to accept certain structure files, atoms within non-standard groups ("HETATM" entries) and residues containing multiple alternative locations for atoms were stripped from structure files. ProFit also had problems operating on structure files with missing residues. Because of this, all structure files were examined manually to create an alignment between model and native structure for use with ProFit, replacing missing residues with gaps. Where structure files contained residues that were not present in the sequence used to represent it, these residues were removed from the structure file.

RMSD values were calculated for all models as well as for the template structures used for the initial alignments. Ramachandran plots were created for all models, native structures and template structures. Average values and standard deviations were calculated for model RMSD, gap region RMSD and for region distributions in Ramachandran plots (percentage of residues in most favored regions, additional allowed regions, generously allowed regions and disallowed regions). These average values were evaluated, and based on these a method for loop modeling was proposed using a stem overlap of ten residues and ranking of alignments by the GenTHREADER neural network score.

4.4 Testing of the Proposed Method

The fourth objective was further divided into two sub-objectives:

1. Application of fold recognition to each sequence in the test set.
2. Creation of models from the multiple alignments produced.

[8] Available at URL http://www.bioinf.org.uk/software/profit/.

The proposed method for loop modeling by fold recognition was applied to the test set. All alignments from the test set had their gap region sequences extracted. To each gap region sequence was added the ten immediately preceding and succeeding residues, where possible. The extended region sequence was then submitted to GenTHREADER. All of GenTHREADER's filtering options were disabled. The highest scoring alignment as determined by GenTHREADER's neural network was selected to model the gap region. As before, alignments to structures that were classified as native for the protein were ignored since this would make the problem trivial.

The local alignments produced by GenTHREADER were integrated into the initial alignment between the target sequence and main template structure to create a multiple alignment, each gap region adding one sequence to the alignment.

The MODELLER program was used to build models from the multiple alignments as well as for the initial alignment. Five models were built from each alignment.

4.5 Analysis of the Proposed Method

Models built using initial alignments containing gap regions (initial models) and models built using multiple alignments created by the proposed loop modeling method (final models) were evaluated using the same three measures as above:

- RMSD (Root Mean Square Deviation) between C_α atoms of the created model and the native protein chain.
- RMSD between model and native chain for C_α atoms in each gap region.
- Ramachandran plot for each model.

RMSD values and Ramachandran plots were obtained as above. Average values and standard deviations were calculated for entire structure RMSDs, individual gap region RMSDs and distributions of residues in Ramachandran plots. Average change in RMSD from initial models to final models was also calculated for entire structures and individual gap regions.

Results were evaluated both over all gap regions and over terminal gap regions (located at the C- or N-terminal of the protein chain) and non-terminal gap regions separately. When there appeared to be a significant difference between these two groups, proteins containing only non-terminal gap regions were also evaluated separately. For some of the most interesting results a Student's t-test was performed to determine statistical significance. For this, Microsoft Excel's TTEST function was used to do a paired, one-tailed t-test with initial data and final data as input sets.

5 Results

This chapter presents the results from following the objectives as described in the Method chapter.

5.1 Protein Set Selection

Table 5.1 presents the targets considered for inclusion in the protein sets. They are comprised of all prediction targets from CASP4 (Murzin & Hubbard, 2001) and CASP5 (Kinch et al., 2003). Target names are those used in CASP. Specific chains used as templates (template structures) and selected for model evaluation (native structures) are indicated.

Of the original CASP targets, 110 in total, 15 were removed because no native structure was available and another 54 were removed since the alignments generated for them lacked any gap region of length greater than nine residues. Thus, the final number of targets was 41.

The technique whereby "native" structures were determined resulted in the lowest sequence identity for a "native" structure being 87.2% (1PUG for T0091) and the highest sequence identity for a "non-native" structure being 70.4% (1KKG for T0110).

Table 5.1 CASP targets considered for inclusion in the protein sets.
[a] CASP target id. [b] Length of target sequence. [c] Chain used as reference for model evaluation. [d] Sequence identity of the native chain. [e] Structures not accepted as templates. [f] Main template structure in initial alignment. [g] Sequence identity of main template structure. [h] Number of gap regions longer than nine residues occurring in initial alignment.

Target[a]	Length[b]	Native structure[c]	Native seq. id (%)[d]	Alternative native structures[e]	Template[f]	Template seq. id (%)[g]	Gap regions of length >9[h]
T0086	164	1G1B:A	100.0	1FW9, 1G81, 1JD3	1UAE	20.7	0
T0087	310	1I74:A	98.4		1IR6:A	16.5	2
T0088	156	1OIO:A	100.0	1O9W, 1O9V, 1O9Z	1GQ8:A	16.7	0
T0089	419	1E4F:T	100.0	1E4G	1BA1	15.3	3
T0090	209	1G0S:A	100.0	1G9Q, 1GA7, 1KHZ, 1VIQ	1VIU:A	28.3	2
T0091	109	1PUG:A	87.2	1J8B	1MOJ:A	15.6	0
T0092	241	1IM8:A	97.5		1XVA:A	13.3	0
T0093	160	1MXI:A	100.0	1J85	1IPA:A	21.9	1
T0094	181	1JH6:A	100.0	1JH7, 1FSI	1REC	11.0	0
T0095	244	1H6G:A	97.9	1L7C	1VHN:A	13.1	0
T0096	239	1HW1:A	100.0	1E2X, 1H9G, 1H9T, 1HW2	1J5Y:A	13.4	2
T0097	105	1G7D:A	100.0		1K6K:A	12.4	0
T0098	121	1FC3:A	100.0	1LQ1	1VI0:A	9.9	0
T0099	56	none	n/a		n/a	n/a	n/a
T0100	342	1QJV:A	100.0		1GQ8:A	31.7	2
T0101	400	1RU4:A	100.0		1RMG	9.8	1
T0102	70	1O82:A	100.0	1E68, 1O83, 1O84	1KV8:A	7.1	0
T0103	372	1GA6:A	100.0	1GA1, 1GA4, 1KDV, 1KDY, 1KDZ, 1KE1, 1KE2, 1NLU	1SIO:A	30.3	0
T0104	158	1HTW:A	100.0	1FL9	1RZ3:A	12.7	2
T0105	94	1H5P:A	100.0		1OQJ:A	26.7	0
T0106	128	1IJX:A	100.0		1GVF:A	14.8	0
T0107	188	1I82:A	99.5	1I8A, 1I8U	1ATG	6.4	0

Target[a]	Length[b]	Native structure[c]	Native seq. id (%)[d]	Alternative native structures[e]	Template[f]	Template seq. id (%)[g]	Gap regions of length >9[h]
T0108	206	1J84:A	98.9	1J83	1QEX:A	20.4	0
T0109	182	1J9A:A	97.8		1UOC:A	13.2	0
T0110	128	1JOS:A	100.0		1PA4:A	17.7	1
T0111	431	1E9I:A	100.0		4ENL	50.1	0
T0112	352	1E3J:A	100.0		1LLU:A	24.0	0
T0113	261	1E3W:B	100.0	1E3S, 1E6W, 1SO8	1H5Q:A	23.1	0
T0114	87	1GH5:A	100.0	1G6E	1QCS:A	9.2	0
T0115	300	1H72:C	100.0	1FWK, 1FWL, 1H73, 1H74	1S4E:B	19.7	1
T0116	811	1NNE:A	100.0	1EWQ, 1EWR, 1FW6	1TAQ	15.5	6
T0117	250	1OT3:A	100.0	1J90, 1OE0	1QHI:A	13.6	1
T0118	149	1M0D:A	100.0	1FZR, 1M0I	1KNY:A	26.2	1
T0119	338	1KRH:A	100.0		1CQX:A	19.8	0
T0120	336	1IK9:A	98.1	1FU1	1O5Z:A	11.3	2
T0121	372	1G29:1	98.7		1B0U:A	26.7	1
T0122	248	1GEQ:A	100.0		2TYS:A	31.0	1
T0123	160	1EXS:A	100.0		1BEB:A	65.4	0
T0124	242	1JAD:A	97.1		1CUN:A	16.0	2
T0125	141	1GAK:A	100.0		1LIS	16.0	0
T0126	163	1JOB:A	100.0	1F35, 1JOD, 1JYT	1CBY	24.5	0
T0127	350	1G8P:A	100.0		1FNN:A	11.1	2
T0128	222	1P7G:A	98.6		1AVM:A	49.3	1
T0129	182	1IZM:A	98.9		1AOX:A	13.2	0
T0130	114	1NO5:A	100.0		1J0L:A	28.6	1
T0131	100	none	n/a		n/a	n/a	n/a
T0132	154	1NNG:A	100.0		1NJK:A	14.3	1
T0133	312	none	n/a		n/a	n/a	n/a
T0134	251	none	n/a		n/a	n/a	n/a
T0135	108	none	n/a		n/a	n/a	n/a
T0136	523	1ON3:A	100.0	1ON9	1UYR:A	17.4	0
T0137	133	1O8V:A	99.2		1MDC	22.5	0
T0138	135	1M2E:A	100.0	1M2F, 1R8J	1PEY:A	18.5	0
T0139	83	1IYR:A	100.0	1KOY	1S3J:A	18.1	0
T0140	103	1MJC	100.0	3MEF	1LCL	15.5	0
T0141	187	1J3G:A	100.0		1LBA	23.3	2
T0142	282	1NZH:A	100.0	1NTF	1I9Y:A	24.5	0
T0143	216	1WCZ:A	99.5	1QY6	1P3C:A	22.3	0
T0144	172	none	n/a		n/a	n/a	n/a
T0145	216	none	n/a		n/a	n/a	n/a
T0146	325	1NRK:A	97.2		1PJ5:A	11.7	1
T0147	245	1M65:A	100.0	1M68, 1PB0	1J6O:A	13.9	0
T0148	163	1IN0:A	100.0		1DD5:A	15.3	0
T0149	318	1NIJ:A	100.0		1O5Z:A	9.1	4
T0150	102	1H7M:A	100.0	1GO0, 1GO1	1CK9:A	34.0	0
T0151	164	1UE1:A	100.0	1UE5, 1UE6, 1UE7	1QVC:A	29.0	1
T0152	210	none	n/a		n/a	n/a	n/a
T0153	154	1MQ7:A	100.0	1SIX, 1SJN, 1SLH, 1SM8, 1SMC, 1SNF	1DUP:A	31.6	1
T0154	309	1MOP:A	99.3	1N2B, 1N2E, 1N2G, 1N2H, 1N2I, 1N2J, 1N2O	1IHO:A	42.9	2
T0155	133	1NBU:A	100.0		1DHN	33.1	1
T0156	157	1NXJ:A	99.4		1VI4:A	45.9	0
T0157	138	1NMN:A	100.0	1NU0, 1OVQ	1VHX:A	32.6	0
T0158	319	none	n/a		n/a	n/a	n/a
T0159	309	1R9L:A	99.4	1R9Q	4MBP	12.6	0
T0160	128	none	n/a		n/a	n/a	n/a

Target[a]	Length[b]	Native structure[c]	Native seq. id (%)[d]	Alternative native structures[e]	Template[f]	Template seq. id (%)[g]	Gap regions of length >9[h]
T0161	156	1MW5:A	98.1		1QSP:A	15.4	1
T0162	286	1IZN:A	100.0		1WER	12.2	1
T0163	369	1NG4:A	99.7	1NG3	1B3M:A	19.0	0
T0164	166	1IO0:A	100.0		1UW4:B	13.9	1
T0165	318	1ODT:C	99.4	1L7A, 1ODS	1EVQ:A	17.9	2
T0166	150	1LJ9:A	100.0		1S3J:A	16.1	1
T0167	185	1M3S:A	100.0	1VIV	1JEO:A	35.6	0
T0168	327	1MKI:A	96.3		1BTL	11.8	3
T0169	156	1MK4:A	100.0		1GHE:B	16.0	0
T0170	69	1UZC:A	98.6		1J7N:A	8.7	0
T0171	256	1M33:A	97.3		1MT3:A	15.6	0
T0172	299	1M6Y:A	98.7	1N2X	1B74:A	11.5	3
T0173	303	1Q74:A	100.0	1Q7T	1UAE	17.8	0
T0174	417	1MG7:A	100.0		1UAE	12.2	8
T0175	248	1NKV:A	97.6		1KPH:A	10.1	0
T0176	100	1N91:A	100.0		1BX4:A	8.0	0
T0177	240	1MW7:A	100.0		1LFP:A	32.1	0
T0178	219	1MZH:A	100.0		1JCL:B	26.9	0
T0179	276	1IY9:A	100.0		1JQ3:C	43.1	0
T0180	53	none	n/a		n/a	n/a	n/a
T0181	111	1NYN:A	100.0		1KC6:A	26.1	0
T0182	250	1O0X:A	100.0		1QXW:A	33.3	0
T0183	248	1O0Y:A	100.0		1MZH:A	41.3	1
T0184	240	1O0W:A	100.0		1TGO:A	17.5	1
T0185	457	1J6U:A	98.5		1GQQ:A	29.2	2
T0186	364	1O12:A	97.8		1GKP:A	15.1	0
T0187	417	1O0U:A	100.0		1UAE	16.8	4
T0188	124	1O13:A	99.2		1EO1:A	27.4	0
T0189	319	1O14:A	100.0		1RKD	15.4	0
T0190	114	none	n/a		n/a	n/a	n/a
T0191	282	1NVT:A	100.0		1NPD:A	33.0	1
T0192	171	none	n/a		n/a	n/a	n/a
T0193	211	1R72:A	100.0		1R9L:A	12.3	0
T0194	237	none	n/a		n/a	n/a	n/a
T0195	299	none	n/a		n/a	n/a	n/a

The 41 targets with one or more gap regions, randomly divided into a training set of 10 proteins and a test set of 31 proteins are shown in Table 5.2 (training set) and Table 5.3 (test set) together with length and position of their gap regions.

The 10 proteins in the training set contained a total of 23 gap regions ranging in length from 10 to 54 residues, with the number of gap regions per chain varying from one to eight. Eight gap regions (35%) were located at the C- or N-terminal of the chain.

The 41 proteins in the test set contained a total of 54 gap regions ranging in length from 10 to 125 residues. The number of gap regions per chain varied from one to six. 22 gap regions (41%) were located at the C- or N-terminal of the chain.

Also shown in Table 5.2 and Table 5.3 is the absence of gap region residues from native structure files, which affected model evaluation. As shown, in the training set 10 gap regions (43%) had at least one missing residue and four (17%) had less than three residues present in the native structure file, making these gap regions totally useless for evaluation purposes. In the test set, 22 gap regions (41%) had at least one missing residue and four (7%) had less than three residues present. Of the total

number of residues in the training set, 32% were missing. For the test set, this number was 22%.

Table 5.2 Gap regions in training set.
[a] CASP target id. [b] Residue numbers included in gap region. [c] Length of gap region. [d] Number of residues in the gap region that were missing from the native structure file. • Terminal gap region.

Target[a]	Region[b]	Length[c]	Missing Residues[d]	Target[a]	Region[b]	Length[c]	Missing Residues[d]
T0087	62-71	10	0	T0174	27-36	10	8
T0087	152-167	16	0	T0174	75-89	15	0
T0093	149-160•	12	4	T0174	198-211	14	0
T0120	205-215	11	4	T0174	228-250	23	0
T0120	283-336•	54	54	T0174	272-281	10	0
T0127	95-112	18	0	T0174	303-314	12	0
T0127	245-271	27	0	T0174	353-372	20	0
T0130	102-114•	13	9	T0174	403-417•	15	15
T0153	131-154•	24	19	T0183	1-27•	27	0
T0168	1-32•	32	8	T0185	130-145	16	1
T0168	131-146	16	0	T0185	445-457•	13	11
T0168	153-165	13	0				

Table 5.3 Gap regions in test set.
[a] CASP target id. [b] Residue numbers included in gap region. [c] Length of gap region. [d] Number of residues in the gap region that were missing from the native structure file. • Terminal gap region.

Target[a]	Region[b]	Length[c]	Missing Residues[d]	Target[a]	Region[b]	Length[c]	Missing Residues[d]
T0089	107-121	15	0	T0132	1-15•	15	10
T0089	357-369	13	0	T0141	1-20•	20	0
T0089	394-419•	26	26	T0141	48-58	11	0
T0090	1-13•	13	0	T0146	310-325•	16	1
T0090	150-159	10	5	T0149	37-53	17	0
T0096	3-12	10	2	T0149	81-90	10	0
T0096	203-239•	37	9	T0149	114-125	12	0
T0100	115-128	14	0	T0149	287-296	10	0
T0100	249-262	14	0	T0151	144-164•	21	21
T0101	168-197	30	0	T0154	1-10•	10	2
T0104	62-73	12	0	T0154	288-309•	22	19
T0104	130-139	10	0	T0155	121-133•	13	13
T0110	107-128•	22	22	T0161	147-156•	10	1
T0115	1-10•	10	4	T0162	246-255	10	0
T0116	55-71	17	0	T0164	72-82	11	0
T0116	549-563	15	0	T0165	30-40	11	0
T0116	571-580	10	0	T0165	92-103	12	0
T0116	606-621	16	0	T0166	139-150•	12	5
T0116	673-683	11	0	T0172	105-123	19	0
T0116	687-811•	125	46	T0172	235-244	10	0
T0117	1-16•	16	11	T0172	257-299•	43	5
T0118	140-149•	10	4	T0184	14-24	11	0
T0121	248-372•	125	0	T0187	48-59	12	0
T0122	165-177	13	7	T0187	84-95	12	0
T0124	1-15•	15	1	T0187	253-281	29	0
T0124	227-242•	16	2	T0187	289-304	16	0
T0128	1-15•	15	11	T0191	158-167	10	0

5.2 Analysis of the Different Approaches

Table 5.4 shows average RMSDs obtained for the training set from modeling using multiple alignments created by all combinations of stem overlap and alignment ranking, as well as for each of the stem overlap and ranking alternatives, all other things being equal. Shown for comparison are average RMSDs for initial models, built from the initial alignment with gap regions modeled using MODELLER's own loop modeling, and the main template structures, i.e. the structure used in the initial alignment and for all non-gap regions in the multiple alignment. Averages shown are calculated over all proteins in the training set, and for models, over all five models created per protein. Also shown are standard deviations for all averages. Lower standard deviations mean more confidence can be placed in obtaining similar performance for another set of proteins. RMSDs are presented for entire structures and for gap regions (an average over RMSDs calculated for all individual gap regions).

When comparing the three stem overlap options over all different ranking alternatives, Figure 5.1 shows that using no overlap produced the worst models in terms of average structure RMSD, while ten residues overlap resulted in a slightly better RMSD than an overlap of three residues. Using no overlap also resulted in a higher standard deviation among the model RMSDs, providing further indication that some amount of stem overlap would need to be used. For stem overlaps of three and ten residues, average RMSDs of entire structures were higher than those of the main templates, but lower than those of the initial models, i.e. final models showed improvement over initial models.

Table 5.4 Average RMSDs for training set.
[a] Category for which averages are shown. [b] Average RMSD from native structures over all proteins in training set. [c] Standard deviation of RMSD from native structures. [d] Average RMSD from native conformations of gap regions over all proteins in training set. [e] Standard deviation of RMSD from native conformations of gap regions.

	Structure RMSD		Gap Region RMSD	
Category of structures [a]	Average (Å)[b]	St. dev.[c]	Average (Å)[d]	St. dev.[e]
Main template structures	11.73	9.7	n/a	n/a
Initial models	14.75	8.3	6.21	3.6
Stem overlap:				
No stem overlap	20.14	16.2	8.69	8.6
3 residues overlap	13.64	8.2	5.09	3.1
10 residues overlap	12.94	8.2	4.97	3.2
Alignment ranking by:				
Solvation energy	14.15	8.8	6.07	3.8
Alignment score	16.55	13.2	6.07	6.8
GenTHREADER score	16.02	13.2	6.60	6.6
Combinations of stem overlap and alignment ranking:				
Solvation energy, no overlap	16.14	9.5	7.38	4.4
Solvation energy, overlap 3	13.39	8.4	5.68	3.4
Solvation energy, overlap 10	12.93	8.1	5.16	3.2
Alignment score, no overlap	22.51	18.4	8.98	10.2
Alignment score, overlap 3	13.83	8.2	4.61	3.1
Alignment score, overlap 10	13.32	8.3	4.63	3.5
GenTHREADER score, no overlap	21.77	18.6	9.70	10.0
GenTHREADER score, overlap 3	13.70	8.1	4.97	2.7
GenTHREADER score, overlap 10	12.57	8.4	5.13	3.0

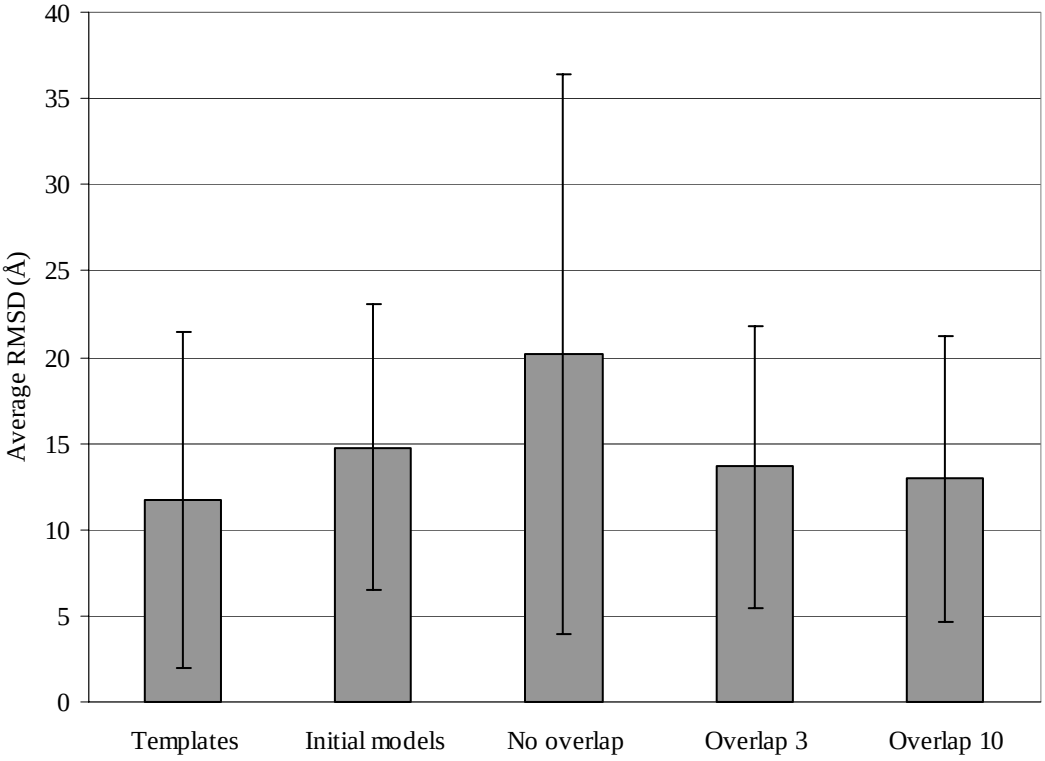

Figure 5.1 Average structure RMSDs for different amounts of overlap. Error bars show standard deviation.

Comparing ranking strategies (see Figure 5.2), solvation energy produced models of a quality similar to that of initial models, while ranking by alignment score and GenTHREADER score produced worse models; the difference in average RMSD was not great, but the difference in standard deviation was more pronounced.

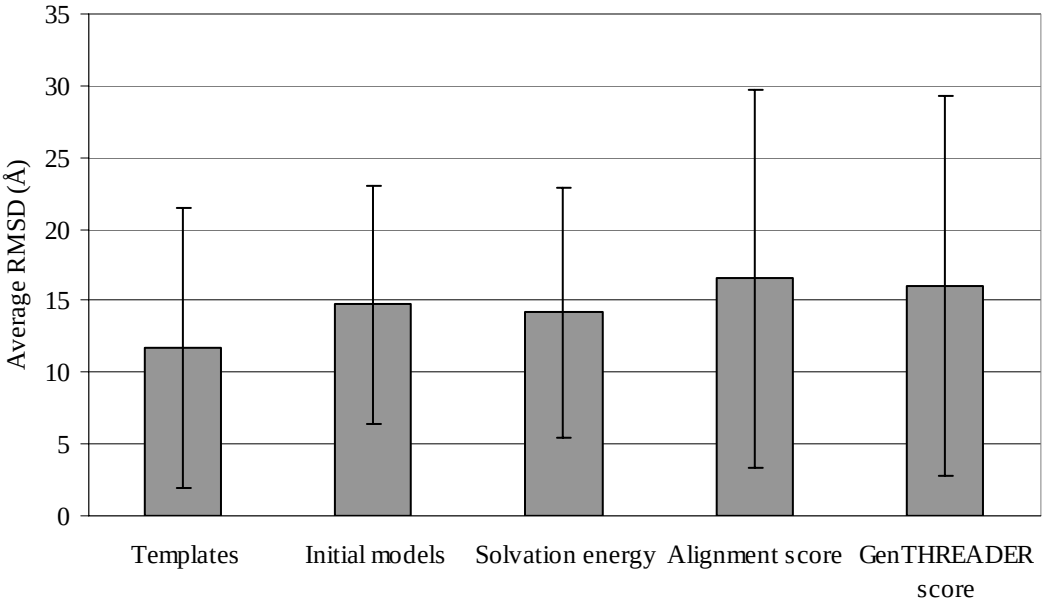

Figure 5.2 Average structure RMSDs for different alignment rankings. Error bars show standard deviation.

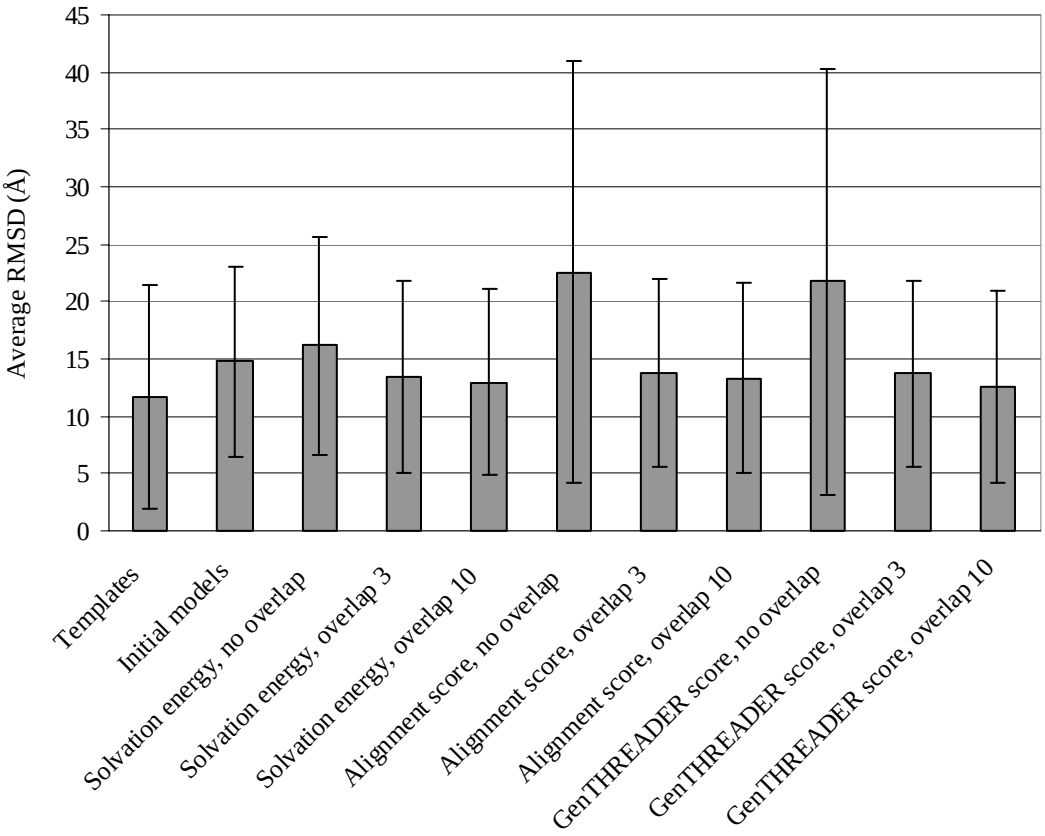

Figure 5.3 Average structure RMSDs for different combinations of stem overlap and alignment rankings. Error bars show standard deviation.

Figure 5.3 shows entire structure RMSDs and Figure 5.4 shows gap region RMSDs for all combinations of stem overlap and ranking strategies. They confirm that using no overlap residues resulted in models of lower average quality for all the different ranking strategies, both in terms of entire structure RMSD and gap region RMSD. One interesting discrepancy is that the combination of no stem overlap with ranking by solvation energy produced only slightly worse RMSD than initial models. Especially interesting is the standard deviation of this combination, which was much lower than that of the other two combinations involving no stem overlap.

Overlaps of three and ten residues produced similar results for all ranking strategies, with averages between those of the main templates and initial models. It should be noted that all averages for stem overlaps of three and ten residues were better than those of the initial models. In terms of structure RMSD, a stem overlap of ten residues and ranking by GenTHREADER score produced best results, while in terms of gap region RMSD a stem overlap of three residues and ranking by alignment score proved best. However, there were no great differences in average quality between any of the combinations using a stem overlap of three or ten residues.

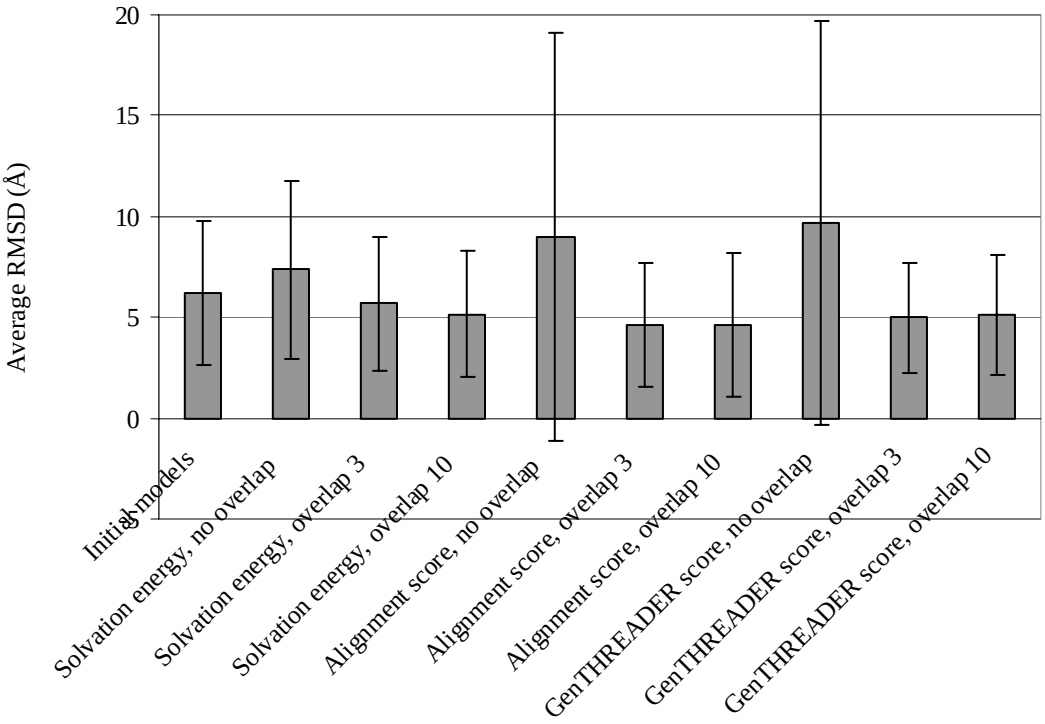

Figure 5.4 Average gap region RMSDs for different combinations of stem overlap and alignment rankings. Error bars show standard deviation.

Table 5.5 Average distribution of residues in most favored regions and disallowed regions of the Ramachandran plots.
[a] Category for which averages are shown. [b] Average percentage of residues in most favored regions over all proteins in training set. [c] Standard deviation of percentage of residues in most favored regions. [d] Average percentage of residues in disallowed regions over all proteins in the training set. [e] Standard deviation of percentage of residues in disallowed regions.

	Most favored regions		Disallowed regions	
Category of structures [a]	Average (%)[b]	St. dev.[c]	Average (%)[d]	St. dev.[e]
Native structures	91.2	2.1	0.1	0.2
Main template structures	91.8	3.2	0.3	0.6
Initial models	84.8	5.9	1.4	1.1
Stem overlap:				
No stem overlap	72.9	17.2	4.0	4.0
3 residues overlap	83.4	6.3	1.4	1.0
10 residues overlap	82.8	7.9	1.5	1.2
Alignment ranking by:				
Solvation energy	79.7	12.7	2.1	2.7
Alignment score	79.9	11.7	2.3	2.7
GenTHREADER score	79.4	13.1	2.4	2.9
Combinations of stem overlap and alignment ranking:				
Solvation energy, no overlap	74.4	17.3	3.5	4.1
Solvation energy, overlap 3	83.2	6.9	1.4	1.1
Solvation energy, overlap 10	81.6	9.9	1.6	1.1
Alignment score, no overlap	72.9	16.0	4.0	3.8
Alignment score, overlap 3	83.5	5.8	1.4	1.0
Alignment score, overlap 10	83.4	7.2	1.5	1.3
GenTHREADER score, no overlap	71.5	18.5	4.4	4.2
GenTHREADER score, overlap 3	83.4	6.3	1.4	1.0
GenTHREADER score, overlap 10	83.3	6.3	1.5	1.1

Table 5.5 shows average distributions of residues in most favored regions and disallowed regions of Ramachandran plots for models built using multiple alignments created by all combinations of stem overlap and alignment ranking, as well as for each of the stem overlap and ranking alternatives, all other things being equal. Shown for comparison are averages for native structures, main template structures and initial models. Also shown are standard deviations for these averages.

The results of the Ramachandran plots are in line with RMSD values, in that the alternatives using no stem overlap produce the worst values (lower percentage of residues in most favored regions and higher percentage of residues in disallowed regions). Standard deviations followed the same pattern. There were no significant differences in average Ramachandran plot values between different alignment rankings. Models for stem overlaps of three and ten residues produced average Ramachandran values similar to those of initial models, but no averages approached the quality of native structures and main template structures.

On the basis of the training set results, a stem overlap of ten residues and alignment ranking by the GenTHREADER score was proposed as the method to apply to the test set since this combination produced the lowest average RMSD for entire structures.

5.3 Analysis of the Proposed Method

Modeling using the prepared multiple alignments failed for targets T0101 and T0187 and thus no final models were available for these targets.

Table 5.6 shows average RMSDs for initial models (built from initial sequence-structure alignments) and final models (built from multiple alignments generated by the proposed method) with main template RMSD for reference. Averages are over all five models built from each alignment.

Also shown is the change in final model RMSD, ranging from 85% lower to 34% higher than initial models. For 21 of the 31 targets (68%), final models exhibited a better average RMSD than initial models. Of the remaining targets, two failed modeling and eight produced final models of lower quality than initial models. The average change was 7% lower RMSD than for initial models with a standard deviation of 19. The t-test probability for this improvement in entire structure RMSD was 0.089. With a conventional significance level of 0.05, this change was not statistically significant.

Table 5.6 Average RMSD from native structures.
[a] CASP target id. [b] RMSD for main template structure. [c] Average RMSDs for initial models. [d] Standard deviation for initial model RMSDs. [e] Average RMSDs for final models. [f] Standard deviation for final model RMSD. [g] Change in average RMSD from initial models to final models. Negative values indicate an improvement. [h] Modeling from multiple alignment failed.

		Structure RMSD				
		Initial models		Final models		
Target[a]	Template (Å)[b]	Average (Å)[c]	St. dev.[d]	Average (Å)[e]	St. dev.[f]	Change (%)[g]
T0089	17.38	17.74	0.2	17.49	0.1	-1
T0090	1.90	5.10	0.6	6.83	0.7	34
T0096	15.47	23.69	0.7	18.24	0.3	-23
T0100	3.00	6.86	0.4	7.02	0.3	2
T0101	14.62	16.46	0.2	n/a[h]	n/a[h]	n/a[h]
T0104	12.96	15.27	0.2	14.85	0.4	-3
T0110	6.94	6.58	0.3	7.23	0.3	10
T0115	7.72	9.54	0.2	9.57	0.1	0
T0116	38.1	46.34	0.5	38.99	0.2	-16
T0117	6.81	7.57	0.2	7.46	0.2	-1
T0118	23.44	20.38	0.8	19.8	0.6	-3
T0121	4.99	97.18	1.3	14.11	0.1	-85
T0122	2.77	3.60	0.2	3.00	0.1	-17
T0124	31.57	44.32	1.1	42.42	0.9	-4
T0128	1.24	4.48	0.3	4.41	0.4	-2
T0132	3.93	5.60	0.4	5.15	0.6	-8
T0141	8.56	15.52	1.1	13.09	0.3	-16
T0146	9.80	12.95	0.6	10.94	0.1	-16
T0149	21.84	21.23	0.3	20.69	0.2	-3
T0151	5.38	5.58	0.5	5.38	0.3	-4
T0154	3.72	4.64	0.4	4.44	0.3	-4
T0155	0.91	0.85	0.0	0.86	0.0	1
T0161	18.11	19.02	0.4	18.27	0.1	-4
T0162	21.67	22.50	0.6	22.64	0.5	1
T0164	14.61	14.90	0.3	14.51	0.4	-3
T0165	14.26	16.29	0.2	15.73	0.2	-3
T0166	3.61	4.89	0.6	4.98	0.5	2
T0172	21.28	31.89	0.7	22.53	0.2	-29
T0184	19.85	16.97	0.3	17.02	0.4	0
T0187	20.48	21.43	0.2	n/a[h]	n/a[h]	n/a[h]
T0191	6.00	7.24	0.1	6.51	0.1	-10

Figure 5.5 shows average initial and final model RMSDs together with RMSD of the main template used for each protein. As can be seen, there was a strong correlation between RMSD of the main template and that of both initial and final models. Exceptions were proteins containing long terminal gap regions, which were modeled as straight stretches of amino acids by MODELLER, resulting in very bad RMSD values for initial models.

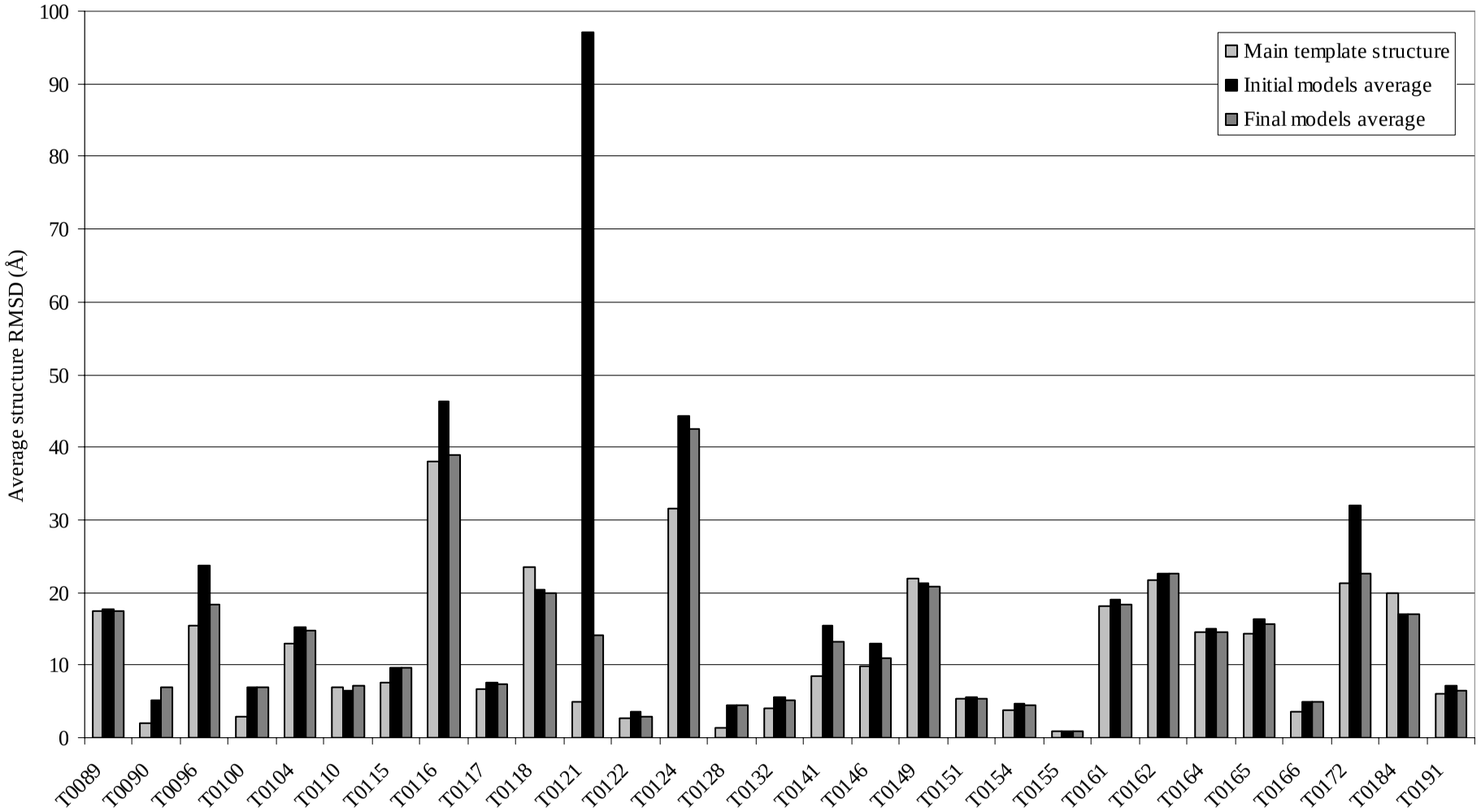

Figure 5.5 Average structure RMSDs for initial and final models.

Table 5.7 presents all 54 gap regions in the test set together with the local structural templates used for modeling them and average RMSD for gap region residues in initial and final models. The change in RMSD from initial to final models is shown, ranging from 84% lower RMSD to 160% higher RMSD. 24 of the 54 gap regions (44%) had an improved RMSD in final models. Five gap regions were in proteins which failed modeling and 21 gap regions were modeled with less accuracy than in initial models. Four gap regions could not be evaluated because of too many missing residues in native structure files. If not counting these four regions, 48% of gap regions had an improved RMSD. For some gap regions which could be evaluated the reliability of RMSD values are affected by the number of residues missing in their native structure files, as shown in Table 5.3 and Table 5.4. Average change in quality was a 1% higher RMSD with a standard deviation of 50.

There was a great difference between the 22 terminal and the 32 non-terminal gap regions. Changes for terminal gap regions ranged from 84% lower to 160% higher RMSD, covering both extremes from the entire test set. 32% of terminal gap regions showed an improvement in RMSD and the average change was a 22% higher RMSD with a standard deviation of 67. Changes for non-terminal gap regions ranged from 78% lower to 34% higher RMSD. 53% of terminal gap regions showed an improvement in RMSD with the average change being 12% lower RMSD (standard deviation 28).

The t-test probabilities for gap region RMSD changes in the entire test set was 0.061, and for terminal and non-terminal gap regions 0.089 and 0.0098, respectively. With a conventional significance level of 0.05, the change for non-terminal gap regions is statistically significant, while the other two changes are not.

If only counting proteins containing no terminal gap regions (9 proteins), the average change in entire structure RMSD was a decrease of 4% with a standard deviation of 6 and a t-test probability of 0.012, making these numbers statistically significant.

Table 5.7 Average RMSD from native conformation of gap regions.
[a] CASP target id. [b] Residue numbers included in gap region. [c] Local template for gap region. [d] Average gap region RMSD for initial models. [e] Standard deviation for initial model gap region RMSD. [f] Average gap region RMSD for final models. [g] Standard deviation for final model gap region RMSD. [h] Change in average gap region RMSD from initial models to final models. Negative values indicate an improvement. [i] Modeling from multiple alignment failed. [k] Too many residues were missing in native structure file for RMSD to be calculated. • Terminal gap region.

			Gap region RMSD				
			Initial models		Final models		
Target[a]	Region[b]	Template[c]	Average (Å)[d]	St. dev.[e]	Average (Å)[f]	St. dev.[g]	Change (%)[h]
T0089	107-121	1AUA	7.41	0.5	4.73	0.1	-36
T0089	357-369	7ODC:A	6.37	1.5	6.33	0.2	-1
T0089	394-419•	1AUK	n/a[k]	n/a[k]	n/a[k]	n/a[k]	n/a[k]
T0090	1-13•	1F82:A	3.48	0.7	4.03	0.9	16
T0090	150-159	1CM3:A	1.70	0.5	1.81	0.1	6
T0096	3-12	1VJT:A	3.25	0.4	3.21	0.3	-1
T0096	203-239•	1JCU:A	10.82	0.8	10.48	0.4	-3
T0100	115-128	1V86:A	3.77	0.5	4.35	0.2	15
T0100	249-262	1OAC:A	7.21	0.5	4.78	0.3	-34
T0101	168-197	1PFO	n/a[i]	n/a[i]	n/a[i]	n/a[i]	n/a[i]
T0104	62-73	1PFO	3.60	0.2	4.02	0.3	12
T0104	130-139	1TBM:A	3.02	0.5	3.24	0.2	7
T0110	107-128•	1FUI:A	n/a[k]	n/a[k]	n/a[k]	n/a[k]	n/a[k]
T0115	1-10•	1PIE:A	1.90	0.9	1.85	0.4	-3

			Gap region RMSD				
			Initial models		Final models		
Target[a]	Region[b]	Template[c]	Average (Å)[d]	St. dev.[e]	Average (Å)[f]	St. dev.[g]	Change (%)[h]
T0116	55-71	1K9D:A	5.03	0.7	5.92	0.1	18
T0116	549-563	1GKR:A	7.73	0.2	7.28	0.1	-6
T0116	571-580	1E9S:E	2.81	0.1	2.64	0.2	-6
T0116	606-621	1AYL	6.02	1.0	6.58	0.4	9
T0116	673-683	16PK	5.47	0.2	1.91	0.7	-65
T0116	687-811•	1MOJ:A	58.95	2.2	16.61	0.1	-72
T0117	1-16•	1HZ4:A	1.61	0.6	2.68	0.0	67
T0118	140-149•	1PFO	1.82	0.3	3.20	0.1	76
T0121	248-372•	1GVF:A	102.68	1.8	16.14	0.1	-84
T0122	165-177	1IWG:A	2.75	0.4	2.79	0.3	1
T0124	1-15•	1GM5:A	5.31	1.0	7.54	0.1	42
T0124	227-242•	1BE3:B	5.65	0.3	3.97	0.9	-30
T0128	1-15•	1JB0:B	0.46	0.4	1.20	0.6	160
T0132	1-15•	1BYB	0.99	0.4	1.34	0.5	35
T0141	1-20•	1DN1:A	6.50	0.6	6.74	0.4	4
T0141	48-58	1TVF.A	3.69	0.4	4.14	0.1	12
T0146	310-325•	1SF9:A	5.43	1.8	7.57	0.1	39
T0149	37-53	1ACC	4.78	1.0	6.40	0.2	34
T0149	81-90	1LDJ:A	3.92	0.2	2.74	1.1	-30
T0149	114-125	1LRW:A	5.08	0.4	5.58	0.2	10
T0149	287-296	1DF0:A	4.16	0.9	3.75	0.2	-10
T0151	144-164•	1FIQ:C	n/a[k]	n/a[k]	n/a[k]	n/a[k]	n/a[k]
T0154	1-10•	1OYG:A	2.35	0.5	2.37	0.7	1
T0154	288-309•	1H80:A	0.10	0.1	0.25	0.2	157
T0155	121-133•	1L5J:A	n/a[k]	n/a[k]	n/a[k]	n/a[k]	n/a[k]
T0161	147-156•	1YGE	3.28	0.7	5.08	0.0	55
T0162	246-255	1AV1:A	4.18	0.4	3.17	0.2	-24
T0164	72-82	1TVF:A	4.90	0.4	4.76	0.2	-3
T0165	30-40	1VK3:A	5.13	0.3	4.40	0.5	-14
T0165	92-103	1I3Q:A	5.42	0.3	5.20	0.2	-4
T0166	139-150•	1BYB	3.57	0.2	2.90	0.6	-19
T0172	105-123	1GPR	6.75	0.2	4.84	0.4	-28
T0172	235-244	1UFK:A	5.81	0.9	1.65	0.3	-72
T0172	257-299•	1C3C:A	25.85	1.3	13.93	1.0	-46
T0184	14-24	1I4S:A	3.69	0.6	0.81	0.1	-78
T0187	48-59	1DQR:A	n/a[i]	n/a[i]	n/a[i]	n/a[i]	n/a[i]
T0187	84-95	1IWG:A	n/a[i]	n/a[i]	n/a[i]	n/a[i]	n/a[i]
T0187	253-281	2ACY	n/a[i]	n/a[i]	n/a[i]	n/a[i]	n/a[i]
T0187	289-304	1FNO:A	n/a[i]	n/a[i]	n/a[i]	n/a[i]	n/a[i]
T0191	158-167	1TGO:A	4.56	0.2	2.41	0.1	-47

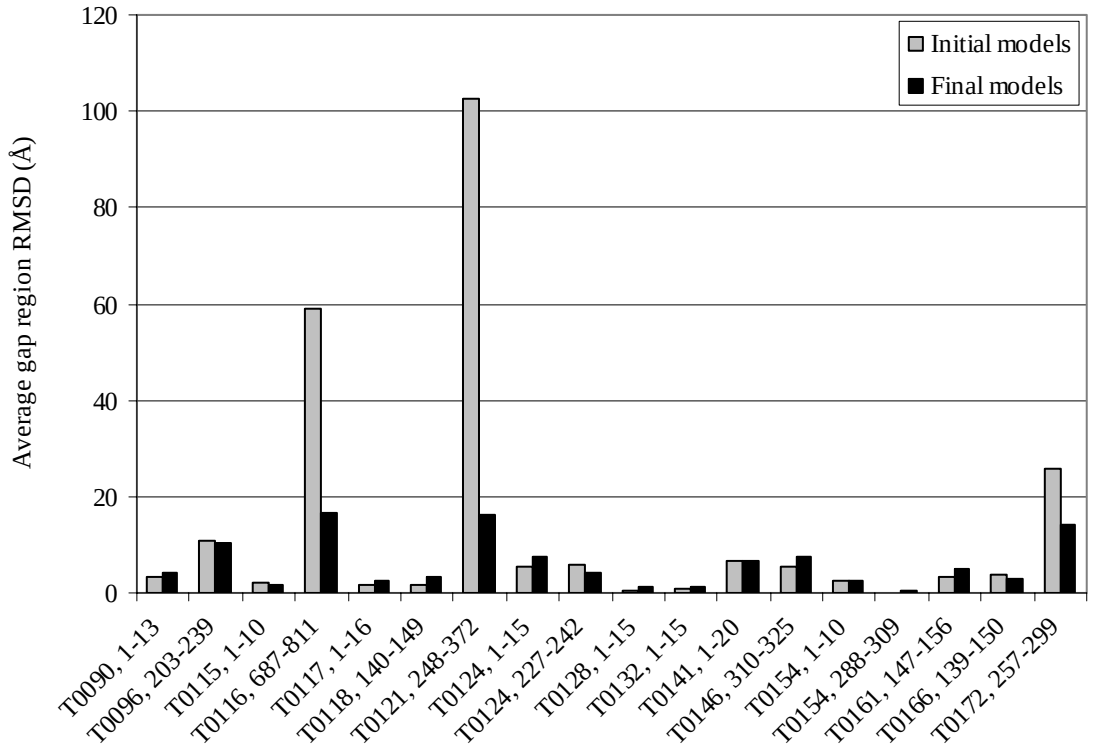

Figure 5.6 Average RMSDs for terminal gap regions in initial and final models.

Figure 5.6 shows average RMSD for all terminal gap regions in initial and final models of the proteins in the test set. Most final models had higher RMSDs than initial models. Again, exceptions were long terminal gap regions, which were modeled as straight stretches by MODELLER, resulting in very bad RMSDs for initial models. Figure 5.7 shows corresponding values for non-terminal gap regions. These are more promising, with the majority of gap regions having a better average RMSD in final models than in initial models.

Figure 5.8 shows the relation of change in gap region RMSD to the length of the region for both terminal regions and non-terminal regions. Again they show a great improvement for long terminal gap regions. Two non-terminal gap regions of 29 and 30 residues were in proteins for which creation of final models failed. Without them, the longest non-terminal gap region was 19 residues long. With only gap regions between 10 and 19 residues in length available, it is hazardous to make any statements about relation between gap region length and quality of models.

Figure 5.9, Figure 5.10 and Figure 5.11 show changes in gap region quality by solvation energy, alignment score and GenTHREADER score, respectively. A high correlation between the change in RMSD and the respective score indicates a good choice for ranking of alignments. As expected there is a general tendency for gap regions with low RMSD to have low solvation energy, a high alignment score and a high GenTHREADER score. However, the GenTHREADER score which was selected for use in the proposed method does not seem to show a higher correlation to the change in RMSD of gap regions.

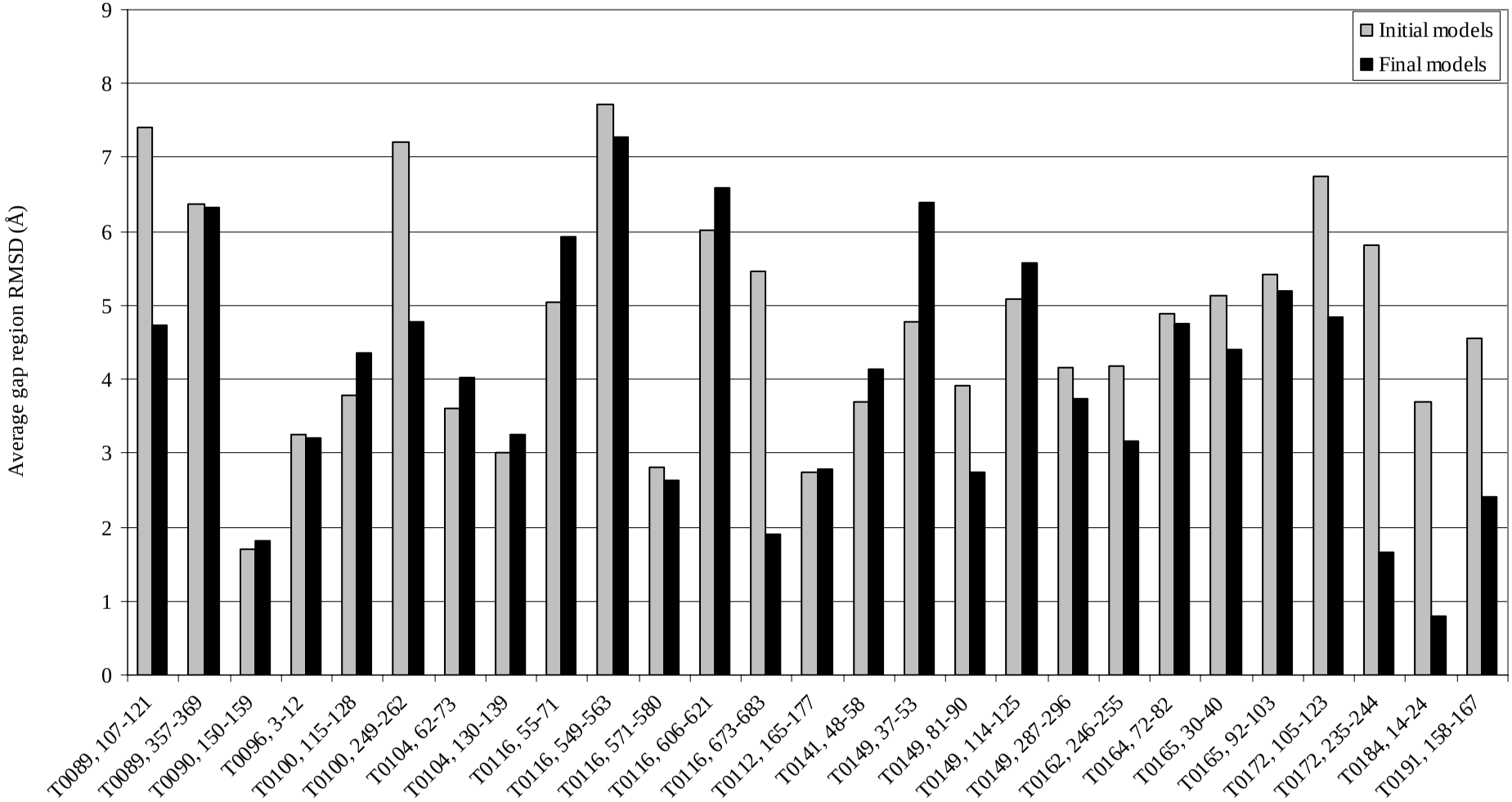

Figure 5.7 Average RMSDs for non-terminal gap regions in initial and final models.

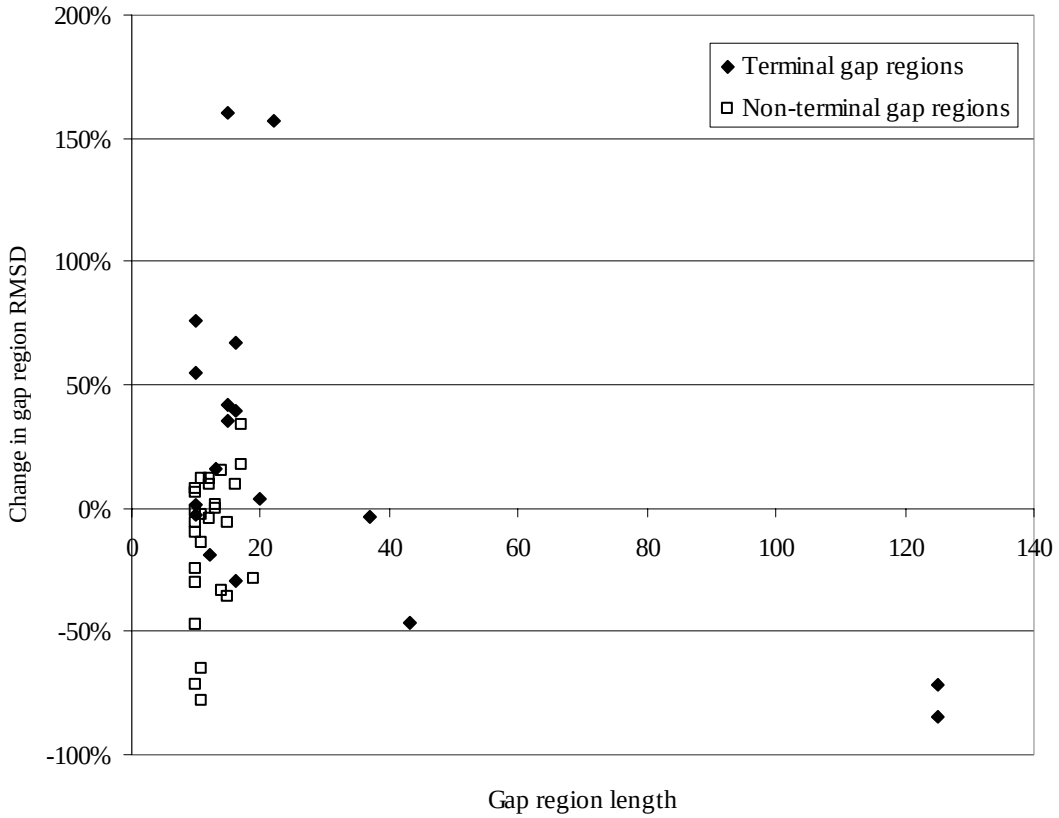

Figure 5.8 Change in gap region RMSD by gap region length and terminalness.

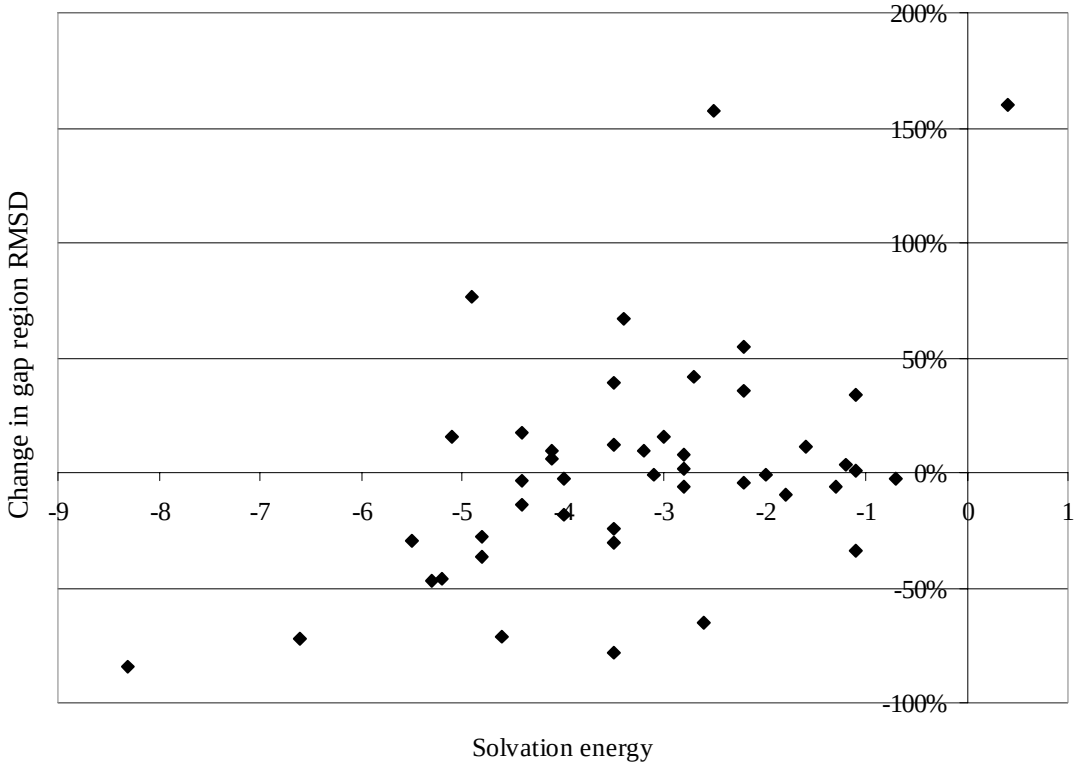

Figure 5.9 Change in gap region RMSD by solvation energy of gap region alignment.

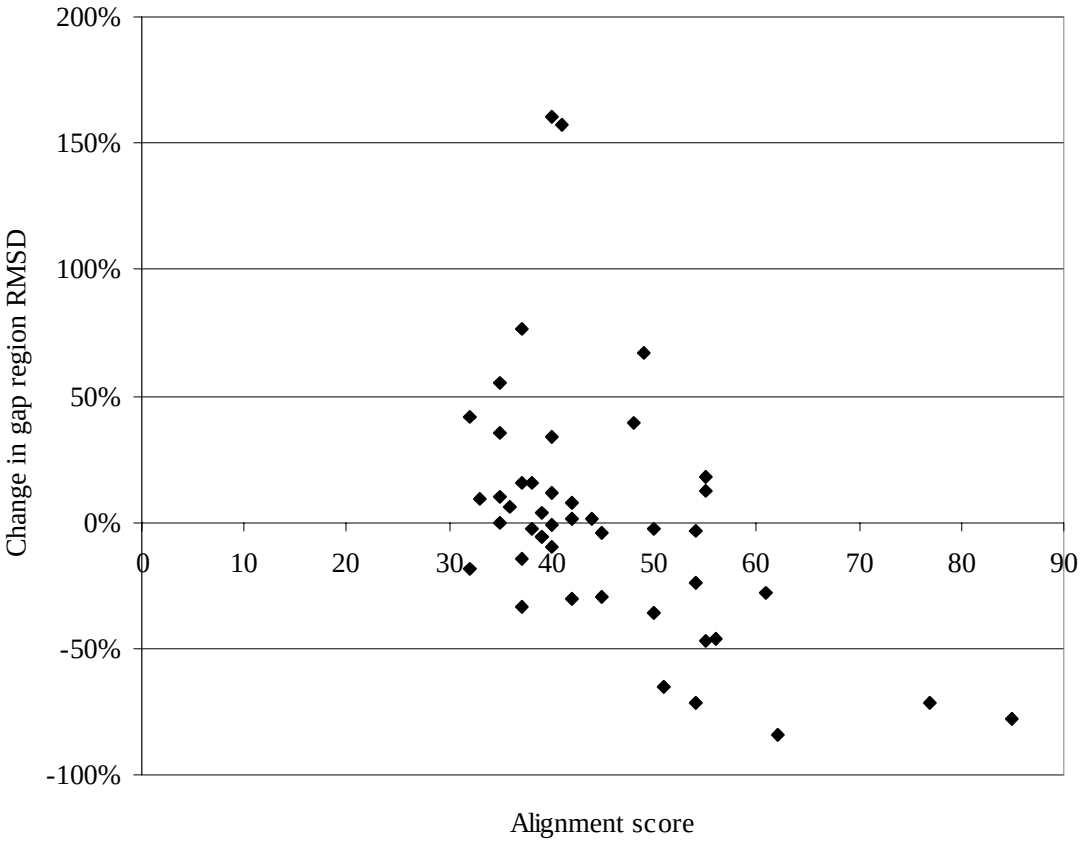

Figure 5.10 Change in gap region RMSD by alignment score of gap region alignment.

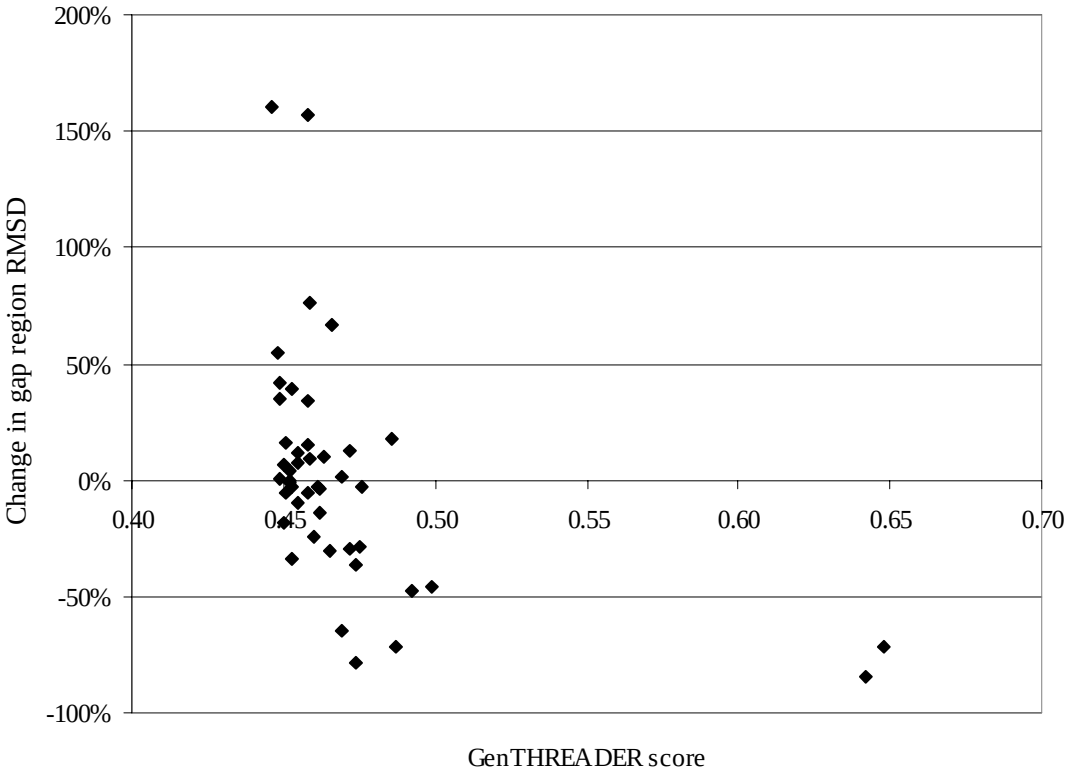

Figure 5.11 Change in gap region RMSD by GenTHREADER score of gap region alignment.

Table 5.8 shows average distributions of residues in most favored regions of Ramachandran plots for initial and final models in the test set. There did not seem to be any great correlation between these and RMSD values. The change in average number of residues in most favored regions ranged from a decrease of 7.5 percentage units to an increase of 1.8 percentage units. Average was a decrease of 1.3 percentage units.

Table 5.9 shows average distributions of residues in disallowed regions of Ramachandran plots for initial and final models in the test set. Again these values did not seem to be correlated to RMSD values. The change in average number of residues in disallowed regions ranged from a decrease of 0.7 percentage units to an increase of 0.7 percentage units. Average was a change of 0.0 percentage units.

According to the number of residues in most favored regions of the Ramachandran plots, there was on average some degradation in final models compared to initial models. However, these residues were still located in additional allowed or generously allowed areas of the plots, since the average number of residues in disallowed regions did not increase.

Table 5.8 Average distribution of residues in most favored regions of Ramachandran plots.
[a] CASP target id. [b] Residues in most favored regions for native structure. [c] Residues in most favored regions for main template structure. [d] Average percentage of residues in most favored regions for initial models. [e] Standard deviation of residues in most favored regions for initial models. [f] Average percentage of residues in most favored regions for final models. [g] Standard deviation of residues in most favored regions for final models.

			Most favored regions			
			Initial models		Final models	
Target[a]	Native (%)[b]	Template (%)[c]	Average (%)[d]	St. dev.[e]	Average (%)[f]	St. dev.[g]
T0089	90.0	92.3	79.2	1.5	77.7	1.2
T0090	92.7	86.9	88.9	1.6	87.9	0.9
T0096	95.5	92.8	84.0	2.0	84.4	2.5
T0100	88.9	86.8	77.6	2.1	75.7	2.3
T0104	89.9	92.5	76.6	4.1	77.6	3.8
T0110	92.9	54.9	74.7	2.8	73.6	3.6
T0115	89.1	89.9	82.7	2.0	82.2	1.4
T0116	88.1	80.8	82.1	1.3	80.1	2.4
T0117	86.7	94.0	81.7	1.9	81.6	2.3
T0118	93.7	87.1	75.2	3.8	74.2	3.2
T0121	92.8	92.9	90.5	0.8	85.1	0.9
T0122	97.1	95.0	94.0	0.6	94.6	0.6
T0124	95.3	98.0	88.3	2.4	87.7	0.7
T0128	88.1	91.5	87.6	1.4	85.6	2.2
T0132	91.9	87.0	87.2	1.6	88.8	0.8
T0141	72.0	86.7	80.9	2.0	77.7	1.8
T0146	80.0	92.1	82.0	1.7	80.5	0.8
T0149	88.7	91.7	83.3	2.2	75.8	2.3
T0151	76.5	70.9	85.0	2.9	83.3	1.5
T0154	90.2	90.8	91.3	0.5	89.3	1.1
T0155	93.3	94.3	92.6	2.2	89.7	1.3
T0161	96.2	92.9	79.1	4.0	78.0	3.0
T0162	85.6	91.4	76.2	2.7	77.9	3.1
T0164	88.2	94.2	78.9	3.2	78.7	1.8
T0165	89.8	86.0	80.4	2.8	78.6	1.1
T0166	93.3	98.2	96.0	1.3	95.5	1.8
T0172	92.7	86.7	83.1	2.1	79.9	2.2
T0184	95.8	88.1	80.0	2.0	80.9	3.6
T0191	86.7	87.3	88.7	0.8	88.1	0.8

Table 5.9 Average distribution of residues in disallowed regions of Ramachandran plots.
[a] CASP target id. [b] Residues in disallowed regions for native structure. [c] Residues in disallowed regions for main template structure. [d] Average percentage of residues in disallowed regions for initial models. [e] Standard deviation of residues in disallowed regions for initial models. [f] Average percentage of residues in disallowed regions for final models. [g] Standard deviation of residues in disallowed regions for final models.

			Disallowed regions			
			Initial models		Final models	
Target[a]	Native (%)[b]	Template (%)[c]	Average (%)[d]	St. dev.[e]	Average (%)[f]	St. dev.[g]
T0089	0.0	0.0	2.1	0.9	1.9	1.0
T0090	0.0	0.0	0.3	0.3	1.0	1.0
T0096	0.5	0.0	1.9	1.1	1.3	0.7
T0100	0.0	0.0	1.9	0.5	1.8	0.6
T0104	0.7	0.6	2.3	1.2	2.1	1.5
T0110	0.0	3.3	2.0	1.4	1.5	0.4
T0115	0.4	0.0	0.8	0.6	1.1	0.8
T0116	0.0	2.4	1.8	0.2	2.2	0.2
T0117	0.0	0.4	1.7	0.2	2.1	1.4
T0118	0.9	0.0	2.1	1.3	2.2	1.0
T0121	0.0	0.0	0.9	0.5	0.9	0.9
T0122	0.0	0.0	0.5	0.5	0.1	0.2
T0124	0.0	0.0	0.8	0.4	1.0	1.2
T0128	0.0	0.6	1.8	0.8	1.7	1.3
T0132	0.0	0.0	0.9	1.2	0.7	0.5
T0141	0.6	0.0	1.5	0.6	2.0	0.3
T0146	0.4	0.0	1.7	0.6	1.6	0.8
T0149	0.4	0.0	0.9	0.4	1.6	0.5
T0151	3.1	1.8	0.6	0.6	0.3	0.4
T0154	0.4	0.0	0.5	0.2	0.7	0.2
T0155	0.0	0.0	0.2	0.4	0.5	0.8
T0161	0.0	0.0	2.8	1.1	2.1	1.0
T0162	0.0	0.7	1.8	0.7	1.4	0.8
T0164	0.0	0.0	2.0	1.3	1.4	0.3
T0165	0.7	0.8	3.0	0.7	2.6	0.5
T0166	0.0	0.0	0.1	0.3	0.6	0.6
T0172	0.0	0.0	2.3	0.8	2.8	0.9
T0184	0.0	0.1	1.7	0.7	1.2	0.5
T0191	0.0	0.0	0.8	0.3	0.6	0.6

6 Discussion

This chapter discusses and analyzes the method and results presented in previous chapters.

6.1 Protein Set Selection

The use of prediction targets from CASP provided a well known and well defined body of data, ensuring that the proposed method was applied to data representative of what is used in the protein structure prediction community. Although no attempt has been made to compare results of this work to any method used in CASP, the use of CASP targets should also facilitate comparisons to published results of existing methods using the same targets. However, direct comparisons to existing results are hazardous for a method such as fold recognition, which is dependent on a changing body of data, i.e. known protein structures.

As noted in the Method chapter, creation of the initial alignment could have been done by an alternative method, but fold recognition by GenTHREADER was chosen as a representative method to find a sequence-structure alignment of appropriate quality.

An unforeseen problem was discrepancies between structure files and their associated sequences. Most often these were structure files missing residues which were present in the sequence. Unfortunately, as shown in Table 5.2 and Table 5.3, missing residues often corresponded more or less directly to gap regions in the sequence-structure alignment, especially for terminal gap regions. These probably reflect difficulties in experimental methods to determine conformation of structurally variable regions and certainly affected both actual models created (for template structures) and evaluation of created models (for native structures). Some gap regions were affected to the degree that calculation of RMSD was not possible (ProFit would not calculate a RMSD for less than three residues). Even for other gap regions, it can be assumed that quality and/or reliability of results decreased with increasing number of missing residues.

As an alternative approach, sequences could have been extracted from PDB structure files to ensure sequences and structures would match. However, this would not reflect a realistic process where the protein structure is unknown.

6.2 Testing of Different Parameters

The choice of stem overlap length is an important variable in the method. A shorter overlap would be more challenging to integrate with the rest of the model, but too long an overlap would also mean difficulties for modeling, since this would mean having two possibly completely different template structures for one sequence of residues. Different modeling algorithms could well require different amounts of stem overlap for optimal results. Of course, the other aspect of varying stem overlap length is that it changes what sequence is submitted for fold recognition. This could change what template structure is selected for a gap region, which could dramatically affect model quality.

Unsurprisingly the alternatives using no stem overlap produced the worst results, confirming that stem regions do have an influence on the conformation of gap regions. Stem overlaps of three and ten residues produced results of similar quality to

each other. It is possible that better results could be achieved for some length between these two.

The use of GenTHREADER output for ranking by solvation energy and alignment score may not have given entirely accurate results. Since GenTHREADER reports only the ten best alignments according to its neural network score, alignments that would have ranked highly by solvation energy or alignment score may not have been included among GenTHREADER's top ten alignments. Thus it is possible that there are alignments that would rank higher according to these strategies than those which are used here.

6.3 Analysis of the Different Approaches

There were no dramatic differences between the best performing approaches. However, the best were a stem overlap of ten residues and ranking by GenTHREADER score (for entire structures) and a stem overlap of three residues and ranking by alignment score (for gap regions). When deciding which approach to use, it was decided to favor good results for entire structure RMSD, and thus ten residues overlap and ranking by GenTHREADER score was selected. Results from the Ramachandran plots were in line with RMSDs, e.g. they indicated lower quality when using no stem overlap.

6.4 Analysis of the Proposed Method

Model building failed for two targets, T0101 and T0187, both with the error message "STDEV < 0". Messages from Professor Andrej Sali in the modeler_usage email list archive[9] state that this error occurs when using more than one template structure and is caused by alignment of the target sequence to two or more very different structures.

As can be seen in Figure 6.1, the alignment for T0187 aligns three structures to each other because of two nearby gap regions. With a stem overlap of ten residues, such situations will occur when two gap regions are closer than 20 residues from each other.

The alignment for T0101 (Figure 6.2) has only two aligned structures, but has four residues from the gap region aligned to the local template another gap region of six residues (shorter than the minimum length of ten residues required for addition of a template of its own). One possible solution for the modeling problem of these targets could be to treat gap regions separated by a small number of residues as one. This would replace the template structures 2ACY and 1FNO in the alignment of T0187 with one template covering both gap regions. Similarly, the template 1PFO in the alignment of T0101 would be replaced with a longer template, better covering the short adjacent gap region. These templates, obtained by a longer alignment with the target sequence, would hopefully be more similar to the main template and thus easier to integrate. They would also remove the additional risk of having two local templates for one set of residues in addition to the main template, as shown in Figure 6.1.

[9] Available at URL http://salilab.org/modeller/discussion_forum.shtml.

```
>P1;T0187
sequence:T0187:::::::
HLSNVEIHLIGNVQKVC--DEAKSLAKEKGFNAEIITTS-------LDCEAREAGRFIASIMKEVKFKDRPLKKPAALIF--------GGETVVHVK
>P1;1UAE
structure:1UAE:::::::
RLGGGVYRVLPDRI-------------------------------------ETGTFLV----------------AAAIS--------RGKIIC---
>P1;1DQR
structure:1DQR::A::A::::
----------------------------------------------------------------------------------------------------
>P1;1IWG
structure:1IWG::A::A::::
----------------------------------------------------------------------------------------------------
>P1;2ACY
structure:2ACY:::::::
----VDYEIFGKVQGVFFRKYTQAEGKKLGLVGWVQNTDQGTVQGQLQGPASKVRHMQEWLE---------------------------------
>P1;1FNO
structure:1FNO::A::A::::
-----------------------------------------------GQWKLLRLLKQQLEEMGLVNITLSEKGTLMATLPANVEGDIPAI----
```

Figure 6.1 Multiple alignment for T0187, positions 296–392.

```
>P1;T0101
sequence:T0101:::::::
NRN-TGLEINNGGSYNTVINSDAYRNYDPKK----NGSMADGFGPKQKQGPGNRFVGCRAWENSDDGFDLFDSP
>P1;1RMG
structure:1RMG:::::::
GNEGGLDGIDVWGSNIWVHD---------------------------------VEVT------NKDECVTVKSPA
>P1;1PFO
structure:1PFO:::::::
----------KNQSIDSGISS---LSYNRNEVLASNGDKIESFVPKEGKKAGNKFIVVERQKRS----------
```

Figure 6.2 Multiple alignment for T0101, positions 186–259.

The longest non-terminal gap region in the test which was successfully modeled was 19 residues long. T0101 and T0187, which failed modeling contained the two longest non-terminal gap regions in the test set (30 and 29 residues, respectively), which may indicate that the method is not suited for longer gap regions. In the training set, however, non-terminal gap regions of 20, 23 and 27 residues were successfully modeled.

One potential problem was the presence of alignment gaps in local alignments, creating a new gap region as a miniature version of the original problem. These were not specifically dealt with, but were left for MODELLER's loop modeling function to handle. It is possible that not using local alignments with too many gaps could lead to better models.

In the integration of local alignments into the main sequence-structure alignment, a wholly mechanistic approach was taken. The multiple alignments could probably have been improved by manual tweaking. However, as the method is currently presented, a fully automated application could be implemented.

These first results for the proposed loop modeling approach, with 68% of the targets achieving a better RMSD than using MODELLER's built-in loop modeling cannot be regarded as a total success, but is a strong indication that a fold recognition approach to loop modeling is worthy of further investigation. The average decrease in RMSD was 7% (standard deviation 19).

Less impressive is a result of 44% percent of gap regions with improved RMSD. However, there was a great difference in performance between terminal and non-terminal regions. For 53% of non-terminal regions RMSD improved, but even where quality decreased, the decrease was not as severe as for terminal regions. None of the terminal regions had more than a 34% increase in RMSD, while the RMSD of two terminal regions increased by over 150%. Averaged over all structures, the proposed method increased gap region RMSD by 1% with a standard deviation of 50 and for

terminal gap regions by 22% (standard deviation of 67). For non-terminal gap regions, however, RMSD decreased with 12% on average, with standard deviation 28. This change was found to be statistically significant as was the average decrease in RMSD for entire structures not containing any terminal gap regions. This was 4% with a standard deviation of 6.

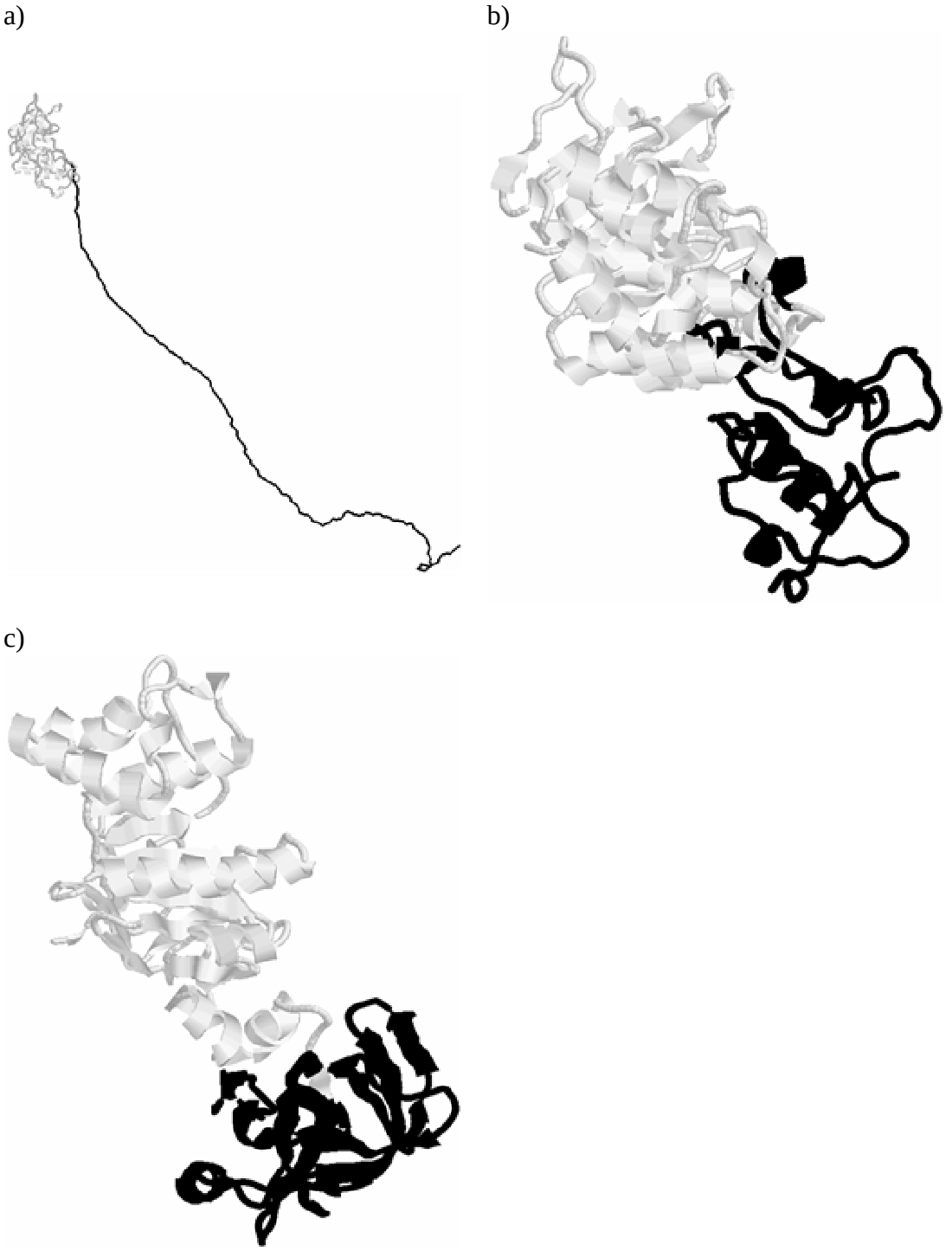

Figure 6.3 a) One of the initial models for T0121. b) One of the final models for T0121. c) Chain 1 of 1G29, the native structure for T0121. The gap region is highlighted in black in all structures.

For the two very long terminal gap regions (125 residues each), the fold recognition approach showed great improvement. However, this is rather because of deficiencies in MODELLER's loop modeling facility than because of merits of the fold

recognition approach. Since MODELLER tend to model these regions as straight stretches (see Figure 6.3a), RMSDs for these regions were extremely high and any method which produced a somewhat folded conformation would result in a better RMSD, regardless of whether that conformation was in any way related to the native one, and indeed even the new conformations for these regions had high RMSD values. Disregarding these artificially positive results, performance for terminal gap regions appears even worse.

The scatterplots of change in gap region RMSD values by different alignment rankings for the test set (Figure 5.9, Figure 5.10 and Figure 5.11) do not suggest any obvious advantage of using the GenTHREADER score over solvation energy or alignment score for that data set.

Results from Ramachandran plots showed no clear correlation to RMSD values but results for final models had a general tendency to deteriorate slightly from initial models.

7 Conclusions

A novel approach to modeling of structurally variable regions has been proposed. This approach applies fold recognition to sequence regions in a sequence-structure alignment which are not covered by the structural template. The result is a multiple alignment created from the initial sequence-structure alignment through the addition of local template structures.

A set of ten CASP prediction targets was used to test several options for what number of residues to include from adjacent stem regions and how to rank local alignments obtained through fold recognition.

Based on the results, a method was proposed where gap region sequences were extended using ten residues from each adjacent stem region. Sequences were then submitted to fold recognition and the output alignments ranked according to the score generated by the GenTHREADER neural network. The highest ranking alignment for each gap region was integrated into the initial sequence-structure alignment to produce a multiple sequence alignment which was used for modeling of the protein.

The proposed method was tested on a set of 31 additional prediction targets from CASP, independent from the first set used. Models were built using the modeling program MODELLER and evaluated in terms of RMSD for entire chains, RMSD for individual gap regions and Ramachandran plots for the chains.

Results indicate that modeling of structurally variable regions by fold recognition is a promising approach. While the method did not perform well on gap regions located at the C- or N-terminal of a chain, non-terminal gap regions were more accurately modeled.

While average results were better than those of MODELLER's loop modeling functionality, results were not consistently better. No comparisons to other loop modeling methods have been made. Because of this, the proposed method is currently not suitable for replacement of those techniques, but at the very least provides a complement to them that in some cases may produce better conformations for structurally variable regions. A serious drawback is that the method may result in models of lesser quality than that of initial models. It is hoped that further work may improve the method or indicate for which proteins the method can be expected to produce improved results.

As the only input needed is a sequence-structure alignment, the method can be applied to alignments created through either comparative modeling or fold recognition. The method can be implemented as an automated server. This is, however left for future work.

It is hoped that these initial results can be improved by further adjustment of parameters such as stem overlap or development of a better strategy for ranking of alignments. Other possibilities for improvement are to what degree gaps in local alignments should be allowed and whether to use one local structural template for several adjacent gap regions. Also, as with regular fold recognition, performance will improve with time as more experimentally determined structures are made available.

While the proposed method has been compared to the loop modeling in the program MODELLER, it would be of great interest to perform a more thorough comparison to other loop modeling methods, such as those used in CASP. Also, performance for longer gap regions should be investigated.

Conclusions

The aim of this dissertation, to investigate whether additional structural templates determined by fold recognition of gap regions can improve the model generated from an alignment, has been achieved. Models created by the proposed method were not universally improved; however, it was shown that for non-terminal gap regions in the test set and for the proteins which contained them, average RMSD was improved. This improvement was shown to be statistically significant.

References

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acids Research*, 31(1), pp. 23–27.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp. 235–242.

Berman, H.M., Goodsell, D.S. and Bourne, P.E. (2002) Protein Structures: From Famine to Feast. *American Scientist*, 90(4), pp. 350–359.

Bourne, P.E., Addess, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., Weissig, H., Westbrook, J. and Berman, H.M. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Research*, 32(Database Issue), pp. D223–D225.

Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K.J., Kelley, L.A., MacCallum, R.M., Pawowski, K., Rost, B., Rychlewski, L. and Sternberg, M. (1999) CAFASP-1: Critical Assessment of Fully Automated Structure Prediction Methods. *Proteins: Structure, Function, and Genetics*, 37(S3), pp. 209-217.

Fiser, A., Do, R.K.G. and Šali, A. (2000) Modeling of loops in protein structures. *Protein Science*, 9(9), pp. 1753–1773.

Jones, D.T. (1998) THREADER: protein sequence threading by double dynamic programming. In: S.L. Salzberg, D.B. Searls and S. Kasif, editors. *Computational Methods in Molecular Biology*. Elsevier Science, Amsterdam, chapter 13, pp. 285–311.

Jones, D.T. (1999) GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences. *Journal of Molecular Biology*, 287(4), pp. 797–815.

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

Jones, D.T., Miller, R.T. and Thornton, J.M. (1995) Successful Protein Fold Recognition by Optimal Sequence Threading Validated by Rigorous Blind Testing. *Proteins: Structure, Function, and Genetics*, 23(3), pp. 387–397.

Jones, D.T. and Thornton, J.M. (1996) Potential energy functions for threading, *Current Opinion in Structural Biology*, 6(2), pp. 210–216.

Jones, D. and Hadley, C. (2000) Threading methods for protein structure prediction. In: D. Higgins and W. Taylor, editors. *Bioinformatics: Sequence, structure and databanks*, Oxford University Press, Oxford, chapter 1, pp. 1–13.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J.E. (2000) Enhanced Genome Annotation Using Structural Profiles in the Program 3D-PSSM. *Journal of Molecular Biology*, 299(2), pp. 499–520.

Kinch, L.N., Qi, Y., Hubbard, T.J.P. and Grishin, N.V. (2003) CASP5 Target Classification. *Proteins: Structure, Function, and Genetics*, 53(S6), pp. 340–351.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), pp. 283–291.

Lathrop, R.H. and Smith, T.F. (1996) Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions. *Journal of Molecular Biology*, 255(4), pp. 641–665.

Leszczynski, J.F. and Rose, G.D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science*, 234(4778), pp. 849–855.

Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F. and Šali, A. (2000) Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29, pp. 291–325.

Martz, E. (2002) Protein Explorer: easy yet powerful macromolecular visualization, *Trends in Biochemical Sciences*, 27(2), pp. 107–109.

McGuffin, L.J. and Jones, D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, 19(7), pp. 874–881.

McLachlan, A.D. (1982) Rapid comparison of protein structures. *Acta Crystallographica Section A*, A38, pp. 871-873.

Moult, J. (1999) Predicting protein three-dimensional structure. *Current Opinion in Biotechnology*, 10(6), pp. 583–588.

Moult J., Pedersen, J.T., Judson, R. and Fidelis, K. (1995) A Large-Scale Experiment to Assess Protein Structure Prediction Methods. *Proteins: Structure, Function, and Genetics*, 23(3), pp. ii–iv.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., and Pedersen, J.T. (1997) Critical Assessment of Methods of Protein Structure Prediction (CASP): Round II. *Proteins: Structure, Function, and Genetics*, 29(S1), pp. 2–6.

Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J.T. (1999) Critical Assessment of Methods of Protein Structure Prediction (CASP): Round III. *Proteins: Structure, Function, and Genetics*, 37(S3), pp. 2–6.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2001) Critical Assessment of Methods of Protein Structure Prediction (CASP): Round IV. *Proteins: Structure, Function, and Genetics*, 45(S5), pp. 2–7.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2003) Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round V. *Proteins: Structure, Function, and Genetics*, 53(S6), pp. 334–339.

Murzin, A. and Hubbard, T.J.P. (2001) Prediction Targets of CASP4. *Proteins: Structure, Function, and Genetics*, 45(S5), pp. 8–12.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1), pp. 95–99.

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rohl, C.A., Strauss, C.E.M., Chivian, D., and Baker, D. (2004) Modeling Structurally Variable Regions in Homologous Proteins With Rosetta. *Proteins: Structure, Function, and Bioinformatics*, 55(3), pp. 656–677.

Rost, B. (1997) Learning From Evolution To Predict Protein Structure, In: D. Lundh, B. Olsson and A. Narayanan, editors. *Bio-Computing and Emergent Computation: Proceedings of BCEC97*, Skövde, Sweden, September 1st–2nd, World Scientific, pp. 87–101.

Šali, A. and Blundell, T.L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3), pp. 779–815.

Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, 31(13), pp. 3381–3385.

Svensson, M., Lundh, D., Ejdebäck, M. and Mandal, A. (2004) Functional prediction of a T-DNA tagged gene of *Arabidopsis thaliana* by *in silico* analysis. *Journal of Molecular Modeling*, 10(2), pp. 130–138.

Tramontano, A. and Lesk, A.M. (1992) Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins: Structure, Function, and Genetics*, 13(3), pp. 231–245.

van Vlijmen, H.W.T. and Karplus, M. (1997) PDB-based Protein Loop Prediction: Parameters for Selection and Methods for Optimization. *Journal of Molecular Biology*, 267(4), pp. 975–1001.