Protein Structure Prediction: Knowledge-based Approaches for Loop Prediction and Model Quality Assessment

I n a u g u r a l - D i s s e r t a t i o n
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Pascal Benkert
aus Wetzikon (Schweiz)

Köln 2007

Berichterstatter:	Prof. Dr. D. Schomburg Prof. Dr. R. Schrader
Tag der mündlichen Prüfung:	30. November 2007

für Nina

Acknowledgements

I wish to express my sincere thanks to my supervisor Prof. Schomburg for his support and for giving me the opportunity to join CUBIC for the postgraduate course in bioinformatics and for the PhD thesis.

The Federal Ministry of Education and Research (BMBF) is acknowledged for the financial support.

I also want to express my gratitude to Prof. Schrader for readily accepting the position of the second examiner.

My special thanks are addressed to my cooperation partner and friend Prof. Silvio Tosatto for the fruitful exchange of ideas, many suggestions and his valuable help as a proof-reader.

I also owe my sincere thanks to Dr. Philipp Heuser for prove-reading and for introducing me to the fascinating field of structural biology.

Finally, I thank all my fiends and colleagues I met at CUBIC for their support and the friendly working climate.

Abstract

Knowledge of the three-dimensional structure of proteins is of vital importance for understanding their function and for the rational development of new drugs. Homology modelling is currently the most successful method for the prediction of the structure of a protein from its sequence. A structural model is thereby built by incorporating information from experimentally solved proteins showing an evolutionary relationship to the target protein. The accurate prediction of loop regions which frequently contribute to the functional specificity of proteins as well as the assessment of the quality of the models are major determinants of the applicability of the generated models in order to answer biological questions.

The modelling pipeline established in the course of this work is able to produce very accurate models as shown in a recent community-wide blind test experiment: From 18 processed protein structure prediction test cases, 3 very good models have been submitted (rank 2, 4 and 6 of over 130 participating groups) and the vast majority of the remaining models was above the community average.

The loop modelling routine relies on a comprehensive database of fragments extracted from known protein structures. After the selection of fragments from the database, a variety of filters are applied in order to reduce the number of fragments. In contrast to other knowledge-based loop prediction methods described in the literature, which mostly perform a ranking based on the geometrical fit of the fragments to the anchor groups in the protein, the present method ranks the remaining candidates with an all-atom statistical potential scoring function which investigates the compatibility of the loop including sidechains with its structural environment. On a large test set of over 200 loops, the loop prediction method is able to model loops with median root mean square deviation per loop length below 1 Å for loops up to a length of 7 residues if all fragments, originating from proteins sharing more than 50% sequence identity to the proteins of the test set, are excluded. On the same data basis, the present method outperforms 3 out of 4 commercial loop modelling programs tested in this work.

Furthermore, a composite scoring function consisting of 3 statistical potential terms covering the major aspects of protein stability and two additional terms describing the agreement between prediction features of the sequence and calculated characteristics

of the model is presented. The scoring function performs significantly better than five well-established methods in the discrimination of good from bad models based on a comprehensive test set of 22,420 models and represents a valuable tool for the assessment of the quality of protein models.

Zusammenfassung

Das Wissen über die dreidimensionale Struktur von Proteinen ist von entscheidender Bedeutung für das Verständnis der biologischer Funktion und ist eine wichtige Voraussetzung für die moderne Arzneimittelforschung. Die Vorhersage der Struktur eines Proteins aus deren Sequenz mit Hilfe von computergestützten Methoden wird deutlich erleichtert, wenn Informationen von experimentell gelösten Proteinen benutzt werden können, welche eine evolutionäre Verwandtschaft zum gesuchten Protein aufweisen (Homologiemodellierung). Dabei spielen die präzise Strukturvorhersage von Loopregionen, welche häufig die funktionelle Spezifität von Proteinen ausmachen, sowie die Fähigkeit, die Qualität der erzeugten Modelle zu bewerten, eine wichtige Rolle für die spätere Verwendbarkeit der Modelle zur Beantwortung biologischer Fragestellungen.

Die im Laufe dieser Arbeit entwickelte Modellierungsumgebung wurde kürzlich an einem internationalen Blindversuch zur Proteinstrukturvorhersage getestet und es hat sich gezeigt, dass sehr genaue Vorhersagen erreicht werden können: Von den 18 untersuchten Vorhersagetestfällen wurden 3 sehr gute Modelle eingereicht (Platz 2, 4 und 6 von über 130 teilnehmenden Arbeitsgruppen) und die überwiegende Mehrzahl der restlichen Modelle waren besser als der Durchschnitt.

Die intergrierte Loopmodellierungsroutine basiert auf einer umfangreichen Datenbank von Proteinfragmenten extrahiert aus experimentell gelösten Strukturen. Im Vorhersageprozess werden mehrere Qualitätsfilter verwendet, um die Anzahl der Fragmente zu reduzieren. Im Gegensatz zu anderen beschriebenen wissensbasierten Ansätzen, in welchen das Scoring meist über die Passgenauigkeit der Fragmente zu den Ankergruppen im Protein durchgeführt wird, verwendet die hier vorgestellten Methode eine Scoringfunktion basierend auf statistische Potentialen, welche die Kompatibilität der Loops inklusive Seitenketten mit der strukturellen Umgebung bewertet. Die Methode wurde auf einem Datensatz von über 200 Loops getestet. Der Median des RMSD (Wurzel der mittleren quadratischen Abweichung) pro Looplänge liegt dabei unter 1 Å für Loops bis 7 Residuen. Dabei wurden Fragmente aus Proteinen extrahiert, die weniger als 50% Sequenzidentität zu den Proteinen im Testdatensatz haben. Mit dem gleichen Datensatz liefert dabei die vorliegende Methode genauere Loopstrukturvorhersagen als 3 von 4 untersuchten kommerziellen Loopvorhersage-Programmen.

Zusätzlich wurde eine zusammengesetzte Scoringfunktion entwickelt, bestehend aus fünf Termen: Drei statistischen Potentiale erfassen verschiedene Faktoren der Proteinstabilität und zwei zusätzlich Terme beschreiben die Übereinstimmung zwischen aus der Sequenz vorhergesagten Eigenschaften und gemessenen Eigenschaften des Proteinmodells. Eine statistisch signifikante Verbesserung gegenüber fünf etablierten Energiefunktionen bezüglich der Fähigkeit, zwischen guten und schlechten Modellen zu unerscheiden, wird erreicht, basierend auf einem umfangreichen Testdatensatz von 22'420 Modellen und einer Vielzahl von Qualitätsmaßen. Die hier vorgestellte Scoringfunktion stellt ein wertvolles Hilfsmittel zur Bewertung der Modellqualität dar.

List of Abbreviations

Å	Ångström (1Å= 10^{-10} m)
API	Advanced Programming Interface
B-factor	atomic displacement parameter; temperature factor
BLAST	Basic Local Alignment Search Tool [5]
CATH	Class, Architecture, Topology and Homologous superfamily [153]
CASP	Critical Assessment of techniques for protein Structur Prediction [147]
DSSP	Dictionary of Secondary Structure of Proteins [107]
FM	CASP7 category: (template-)Free Modelling
FSSP	Families of Structurally Similar Proteins [92]
GDT/GDT_TS	Global Distance Test (Tertiary Structure) [244]
HA-TBM	CASP7 category: High Accuracy Template-Based Modelling
HMM	Hidden Markov Model
HOMSTRAD	HOMologous STRucture Alignment Database
LGA	Local/Global Alignment [244]
MD	Molecular Dynamics
MM	Molecular Mechanics
MQAP	Model Quality Assessment Program
NMR spectroscopy	Nuclear Magnetic Resonance spectroscopy
nr	NCBI's non-redundant sequnece database
PDB	Protein Data Bank [18]
pdbaa	sequence database of protein structures from the PDB
PSI-BLAST	Position-Specific Iterative BLAST [6]
PSSM	Position Specific Scoring Matrices
QMEAN	Qualitative Model Energy ANalysis

RMSa	RMSD between terminal fragment residues and anchor groups residues after fitting
RMSD	Root Mean Square Deviation
ROC curves	Receiver Operator Characteristic curves
SCOP	Structural Classification of Proteins [148]
SCWRL	Side Chain placement With a Rotamer Library [31]
SSE	Secondary Structure Element
TBM	CASP7 cathgory: Template-Based Modelling
TXXXX	Targets of CASP7, *e.g.* T0298
X-rays	Röntgen rays
Znat	Z-score of the native structure compared to the ensemble

List of Tables

List of Figures

Contents

1 Introduction

Proteins[a] play a key role in all living organisms. They participate in all processes that characterise life, which are the ability to metabolise nutrients, respond to external stimuli, grow, reproduce and evolve. Proteins are involved in most physiological processes, for example in the immune response, cell cycle, signal transduction, metabolism, catalysis of reactions and transport, and they serve as structural material (*e.g.* actine, collagen, elastin or creatin).

Proteins are composed of 20 different amino acids and the order of the amino acids is determined by the genes. After synthesis, the linear polymer folds in a well-defined 3-dimensional structure [7]. The enormous variety of functions proteins perform can be attributed to a great extent to their ability to specifically and tightly bind other molecules. Binding and function is mediated by the 3-dimensional structure of the protein and the physico-chemical properties of the amino acids sidechains at the active or binding site. Therefore, knowledge about the structure of a protein is of paramount importance in order to understand its function, find explanations for diseases and potentially design drugs against them.

Over the last two decades, large-scale sequencing projects of dozens of genomes (including human) have resulted in a vast amount of sequences. Of these, a considerable fraction has no annotated function or their mechanism of action is virtually unknown. The number of known protein sequences is about two orders of magnitude higher than the number of experimentally solved protein structures. Since experimental methods for the determination of protein structures are time-consuming and fail for some important groups of proteins (*e.g.* membrane proteins), efficient computational methods for the prediction of the protein structure from its sequence are needed.

The prediction of the protein structure from scratch solely based on physical principles (*i.e.* the simulation of the biological process of folding) is, unfortunately, out of reach at present. All current methods for protein structure prediction incorporate to some extent knowledge of experimentally solved structures either by using segments of known

[a]The word "protein" comes from the Greek $\pi\rho\omega\tau\alpha$ ("prota") which means "of primary importance"

protein structures to model the structure of unknown ones or by parametrising energy functions.

In this work, the potential of these so called "knowledge-based" approaches for protein structure prediction is investigated. A method for the modelling of loop regions, as well as a scoring function for the quality assessment of the protein structure models are presented, which both take advantage of the information stored in the set of experimentally solved protein structures. The methods are embedded in a modelling pipeline established in the course of this work.

This chapter starts with a general introduction on proteins and their structure, followed by an overview on methods used in protein structure prediction and ends with the description of the objectives of this thesis.

1.1 Protein structure

1.1.1 General properties of proteins

Proteins are linear polymers consisting of 20 different amino acids. The amino acids are connected by the peptide bond between the carbonyl C of the i^{th} amino acid and the amine N of the $i+1^{th}$ amino acid (Figure 1.1). During the formation of the peptide bond, a water molecule is released. The peptide bond has a shared double bond: the non-bonding electron pair of the nitrogen can be delocalised to form a double bond with the carbonyl C, with the consequence that the π electrons of the $C = O$ bond are moved to the oxygen [2].

As a consequence of the double bond character, the peptide bond is rigid and almost planar which greatly reduces the degrees of freedom. The 6 atoms between two consecutive Cα atoms (including the Cαs) can therefore be considered to be in a plane. The dihedral angle ω (Figure 1.1) is typically very close to 180° for all amino acids (except proline) which is equivalent to the Cα atoms being in *trans* conformation (*i.e.* the Cα's point in opposite directions of the peptide bond). Over 99.9% of all amino acids in proteins (except proline) occur in trans conformation [166]. Proline, as a consequence of the covalent bonding between sidechain and backbone, occurs in

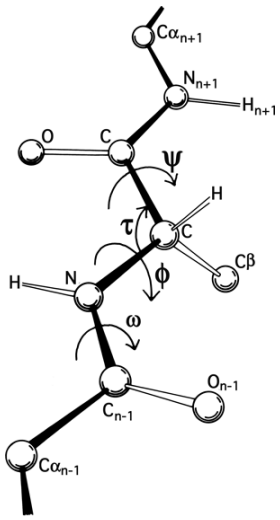


Figure 1.1: Important angles in polypeptides.[b]

approximately 5% of the cases in cis-conformation [54, 130].

Due to the planarity, the conformational degrees of freedom of the protein backbone are mainly reduced on the two torsion angles Φ and Ψ. The dihedral angle Φ describes the angle between the two planes defined by the 4 atoms C_{i-1}, N_i, $C\alpha_i$, C_i and Ψ in analogy is defined by N_i, $C\alpha_i$, C_i, N_{i+1} (i represents any position in the polypeptide chain). Not all Φ/Ψ-angle combinations are energetically favourable as a consequence of steric hindrance between the first sidechain atom and the backbone atoms. This fact can be schematically visualised by the Ramachandran plot [167] (Figure 1.2).

The Ramachandran plot is obtained by treating the atoms as hard spheres and marking the Φ and Ψ angle combinations which do not lead to collisions of the van der Waals spheres. White regions are sterically disallowed, dark regions lead to no van der Waals clashes and the lighter region are possible if the radii are slightly reduced. The distribution of Φ/Ψ-angles observed in experimental structures can sometimes differ substantially from the ideal situation depicted above. The high energy of an unfavourable dihedral angle combination can be compensated for example by other interactions. Glycine and proline show a quite different Ramachandran plot as compared to the other amino acids: Glycine, as a consequence of the missing sidechain (R-group $= -H$), can populate regions which are unfavourable for the other amino

[b]source: http://kinemage.biochem.duke.edu/~jsr/html/anatax.1b.html

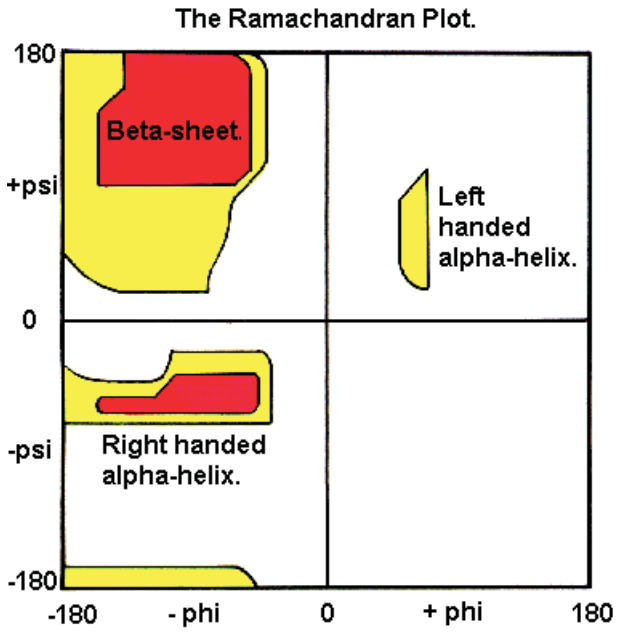


Figure 1.2: The Ramachandran plot.[c]

acids and in proline the Φ dihedral angle is restrained as a consequence of the cyclic nature of this amino acid.

There are four levels of proteins structure: The linear sequence of amino acids, encoded by the nucleotide sequence of the gene, is called the *primary structure*. *Secondary structure* refers to local structural patterns of the protein backbone. The *tertiary structure* is the 3-dimensional conformation of the protein whereas *quaternary structure* describes the arrangement of protein subunits forming complexes.

1.1.2 Amino acids

Amino acids consist of a central carbon atom (the Cα atom) in tetrahedral coordination with four substituents: A hydrogen atom, the amino-group ($-NH_2$), the carboxyl-group ($-COOH$) and an organic sidechain (R-group). The unique physical and chemical properties of the 20 naturally occurring amino acids are therefore a consequence of the difference in the R-group. The properties of the amino acids can be represented

[c]source: http://www.bbk.ac.uk/PPS2/course/section3/rama.html

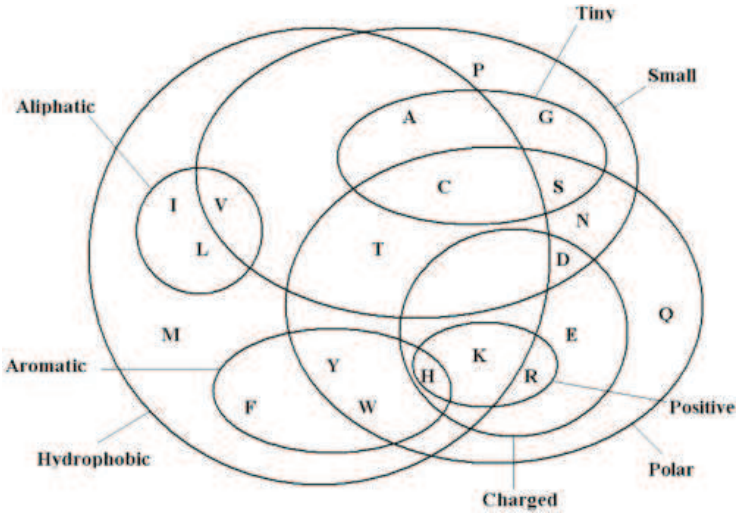


Figure 1.3: Properties of the 20 amino acids [127].

schematically in a Venn diagram [127] (Figure 1.3).

The 20 amino acids are shown below in Figure 1.4. The unique properties of some selected amino acids are described in the following (according to Tramontano [219] and Voet and Voet [231]):

- As a consequence of its missing sidechain, glycine is very flexible and can adopt unusual backbone torsion angles. Glycine is therefore often observed in tight turns.
- Proline is the only imino acid, which means that the sidechain is connected with the backbone forming a nitrogen-containing ring. Proline is often observed in turn structures. Proline is known to be a helix breaker [40]. A conserved proline within a protein family can be an evidence of a specific structural feature and should be taken into account in protein structure prediction and especially in loop modelling.
- Cysteins are the only amino acids able to form inter- and intra-molecular covalent bonds by oxidation of the sulfhydryl groups ($-SH$) of two cysteins to a disulfide bond. These amino acids are therefore of crucial importance in extracellular proteins which are in a reducing environment. The $SH-$group of cysteins is rather reactive and can coordinate metals.

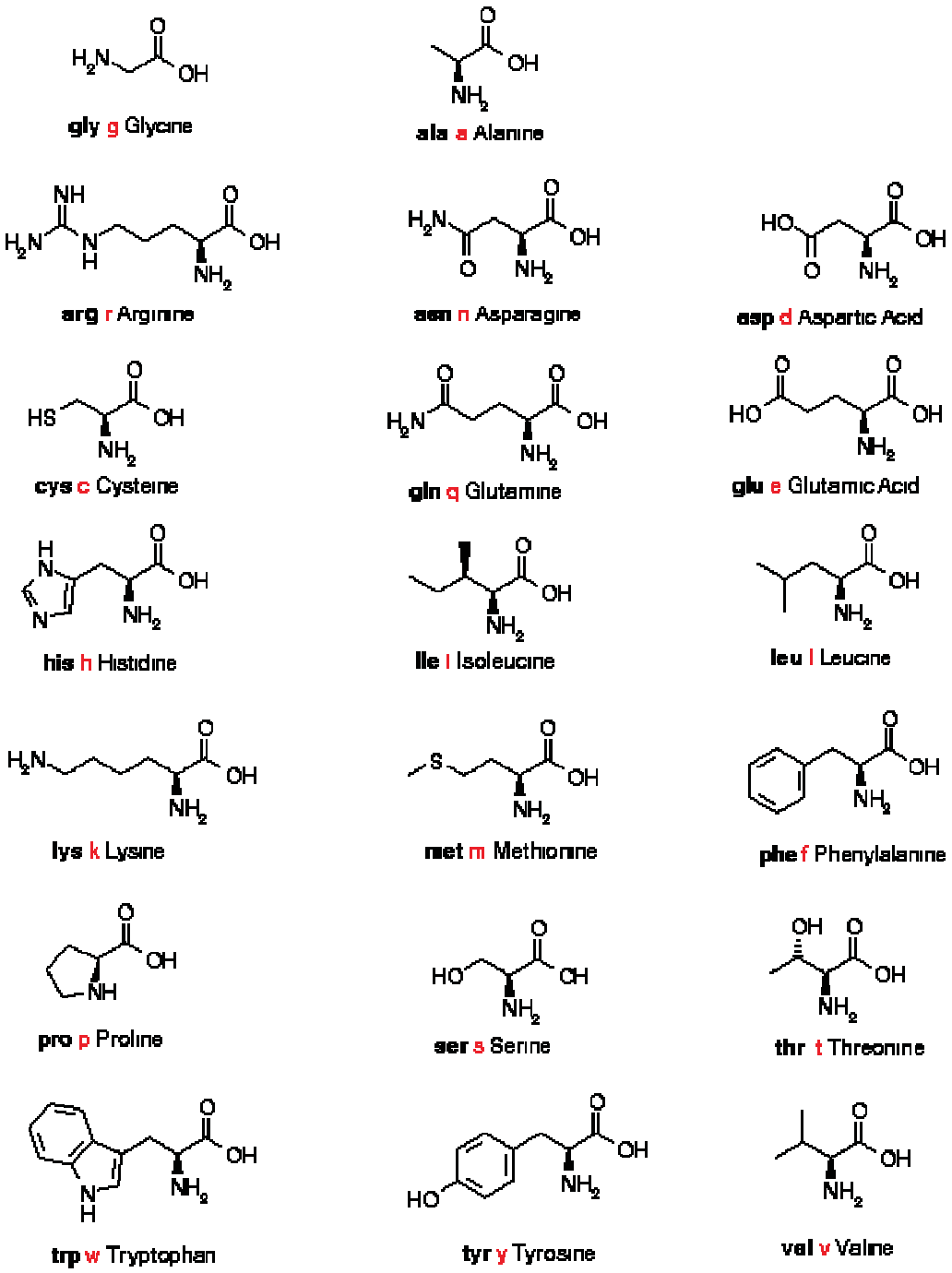


Figure 1.4: The 20 naturally occuring amino acids.

source: http://www.cem.msu.edu/~reusch/OrgPage/aminoacids.htm

- Hydrophobic amino acids such as for example leucine, valine and isoleucine are usually found in the interior of proteins shielded from direct contact with water. Conversely, the hydrophilic amino acids (*e.g.* asparagine and glutamine) are generally encountered on the exterior of proteins as well as in the active centers of enzymes. Charged residues such as the negatively charged asparagate (or aspartic acid) and glutamate (or glutamic acid) as well as lysine and arginine (positively charged) can form salt bridges and are often observed in active sites.

- Another group of amino acids are the aromatic residues (phenylalanine, tryptophane, tyrosine and histidine) which can interact with each other forming π-stacks. Histidine additionally has the important property that it can act both as a base and an acid under physiological pH and therefore plays a central role in active sites (*e.g.* in the catalytic triad in chymotrypsin).

1.1.3 Secondary structure

Secondary structure elements are local structural segments typically stabilised by backbone hydrogen bonds and are the essential building blocks of protein conformation. Secondary structures represent sterically favourable conformations as reflected by the Ramachandran plot in Figure 1.2. The most common secondary structure elements are α-helices and β-sheets. The fact that the amino acids have different propensities to be observed in secondary structure elements was used by Chou and Fasman in the early 1970's to predict secondary structure [40, 41]. For example alanine, glutamate, leucine and methionine were identified as helix formers, while proline and glycine, due to the unique conformational properties, commonly end a helix.

The α-helix is the simplest and most abundant secondary structure element (see Figure 1.5). An α-helix has on average 3.6 amino acids per turn and is stabilised by hydrogen bonds between the amide H at position i and the carbonyl O at position $i-4$. The Φ/Ψ dihedral angles are typically around (-60°, -50°) [219]. The sidechains point outward from the helix. Other, less common, helix types are the 3_{10}-helix and the π-helix.

Another frequently occurring secondary structure element is the β-sheet which is formed by two or more β-strands (*i.e.* polypeptide segments in extended conformation)

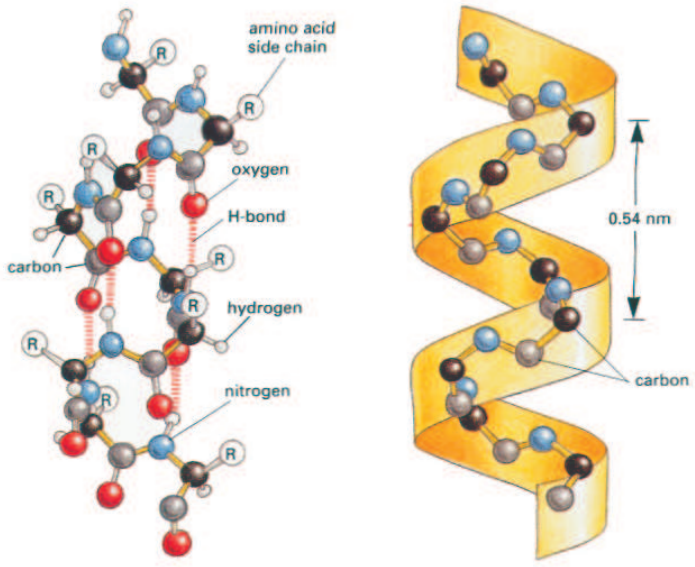


Figure 1.5: The α-helix structure (*source:* [2]).

linked laterally by hydrogen bonds. The sidechains of neighboring residues point into different directions. The strands can be aligned in the same or opposite orientation forming parallel (Φ/Ψ angles around (-119°, 113°) [231]) or anti-parallel β-sheets (Φ/Ψ angles around (-139°, 135°) [231]) which are typically slightly twisted (see Figure 1.6).

Regions without regular structure connecting secondary structure elements are called *loops*. A frequently occurring structural loop motif are reverse turns which are stabilised by a hydrogen bond between cabonyl oxygen at position i and N-H group at position $i + 3$. If a reverse turn is enclosed by β-strands the motif is called β-hairpin. Some turns require a glycine at a certain position as a consequence of the torsion angles falling in the "forbidden" region of the Ramachandran plot for the other amino acids.

1.1.4 Tertiary and quaternary structure

The 3-dimensional arrangement of the secondary structure elements (including the connecting loops) in a single chain is called the tertiary structure. Frequently occurring geometric arrangements of two or three secondary structure elements are also known as motifs or supersecondary structures. Examples are the β-hairpin motif described above

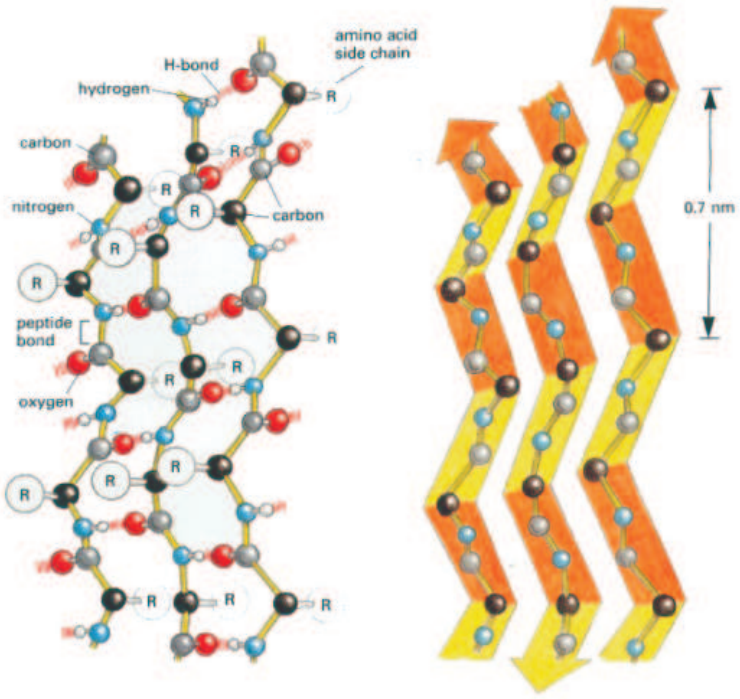


Figure 1.6: An anti-parallel β-sheet (*source:* [2]).

(beta-turn-beta) or the beta-alpha-beta unit. The combination of supersecondary structures is often called domain or fold [219]. An exact definition of the term "domain" is difficult: domains are often described as segments that can independently fold into a stable 3-dimensional structure. In a more evolutionary sight, domains can be seen as evolutionary units which can be duplicated and/or undergo recombination [38]. Two very common arrangements of supersecondary structures are the Rossman fold (beta-alpha-beta-alpha-beta) and the four-helix bundle.

It is commonly assumed that the number of protein folds occurring in nature is limited but there is disagreement about the magnitude of this number (*e.g.* [37, 152, 231]) and whether each fold originated just once (as propagated via divergent evolution) or has been "re-invented" (convergent evolution of structures).

Several hierarchical protein structure classification systems have been developed ranging from entirely manual to fully-automated approaches: SCOP [148], CATH [153] and FSSP [92]. On the highest level, the proteins are typically classified according to their

secondary structure content. For example in CATH, the *Class*-level is organised as follows:

- mainly α-helix
- mainly β-sheet
- α/β proteins
- few secondary structures

The lowest classification level are the protein families in which the members have a clear evolutionary relationship (*i.e.* are homologues).

1.1.4.1 Sequence-structure relationship

Since Anfinsen's pioneering work in 1973 [7] it is known that the primary sequence exclusively determines the 3-dimensional structure of a protein. Anfinsen realised that the driving force for folding is the gradient of free energy and that the native structure of the protein is in its free energy minimum (for a review on folding see [12, 57, 95]).

Folding describes the physical process in which a polypeptide chain folds in its characteristic 3-dimensional structure. The folding process is still not fully understood. In the late 1960's Levinthal [122] demonstrated that the sequential sampling of all possible conformations of the polypeptide chain would take an astronomical amount of time which disagrees with the folding time of microseconds to minutes typically observed in nature. He concluded that proteins fold by a directed process with specific *folding pathways*. This observation was later called the "Levinthal paradox".

In a more modern view, the pathway concept assuming an obligate series of discrete intermediates is replaced by a multiplicity of parallel routes down a *folding funnel* based on the concept of the *energy landscape* [27]. A schematic picture of the funnel-like energy landscape is given in Figure 1.7. The energy landscape in potentially rugged as a consequence of kinetic traps and energy barriers.

Dill illustrates this concept as follows: "water flowing along different routes down mountainsides can ultimately reach the same lake at the bottom" [59]. It is generally

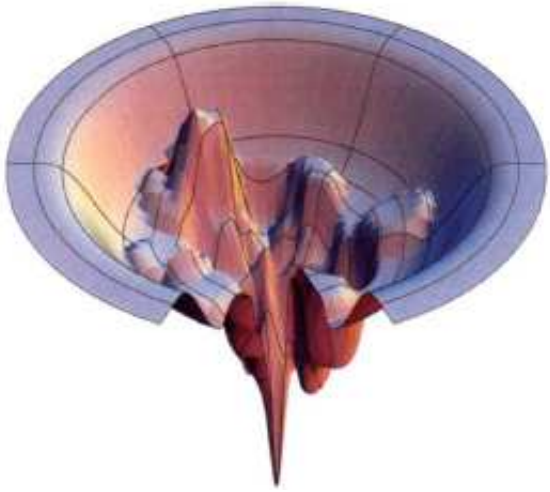


Figure 1.7: Schematic representation of the funnel-like energy landscape [59].

assumed, that the folding process starts with the formation of local secondary structure governed by interactions being close in the polypeptide chain and that the subunits are subsequently assembled further down the folding funnel. Folding involves a balance between loss of conformational entropy and gain in enthalpy. The hydrophobic effect seems to be the driving force and to a certain extent also hydrogen bonding.

Generally, it can be said that sequence determines structure and structure determines the protein function. But unfortunately the prediction of protein structure from scratch solely based on physical principles is at present still out of reach. Most current methods for protein structure prediction incorporate to some extent knowledge of experimentally-solved structures based on the fact that structure is more conserved than sequence.

The relationship between sequence similarity and structural similarity was topic of the seminal work of Chothia and Lesk [39]. The authors showed that the difference in the structure of two proteins increases as the sequence identity decreases (see Figure 1.8).

Sequence similarity is typically expressed as pairwise sequence identity based on an alignment. An alignment is an ordered mapping of the residues of two sequences. A gap (denoted by "-") can be placed when a residue is not aligned with any of the residues of the other sequence. More precisely, sequence identity is defined as the number of positions in the alignment where the residues are identical divided by the length

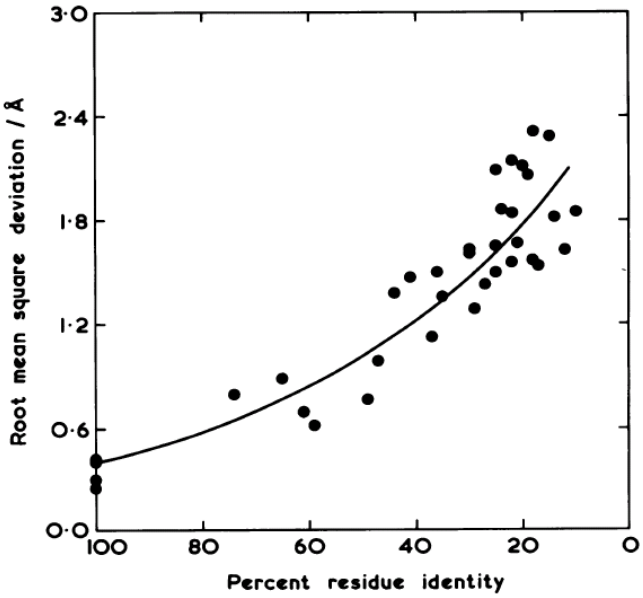


Figure 1.8: Relationship between sequence and structure similarity analysed by Chothia and Lesk [39].

of the shorter sequence. Structural similarity is traditionally expressed by the root mean square deviation (RMSD) between corresponding atoms in an optimal structural superposition (see Formula 2.7 on page 61).

As an example the sequence alignment between myoglobin (PDB code 1mbn, 153 residues) and hemoglobin (PDB code 3hhb, 141 residues) is shown in Figure 1.9. Conserved residues are marked in bold. The structural superposition of the two proteins in given in Figure 1.10. Although the sequence identity is only around 25% ($36 \div 141 \approx 25.5$) the two proteins show a remarkable structural similarity with an RMSD of the backbone atoms below 1.5 Å.

```
myoglobine    VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKF-DRFKHLKTEAEMKASEDLKKHGVTVL
hemoglobine   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPH-FDLSHG------SAQVKGHGKKVA

myoglobine    TALGAILKKKGHHEA-ELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELF
hemoglobine   DALTNAVAHVDDMPNALS-ALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASV

myoglobine    RKDIAAKYKELGYQG
hemoglobine   STVLTSKYR------
```

Figure 1.9: Sequence alignment between myoglobin and hemoglobin.

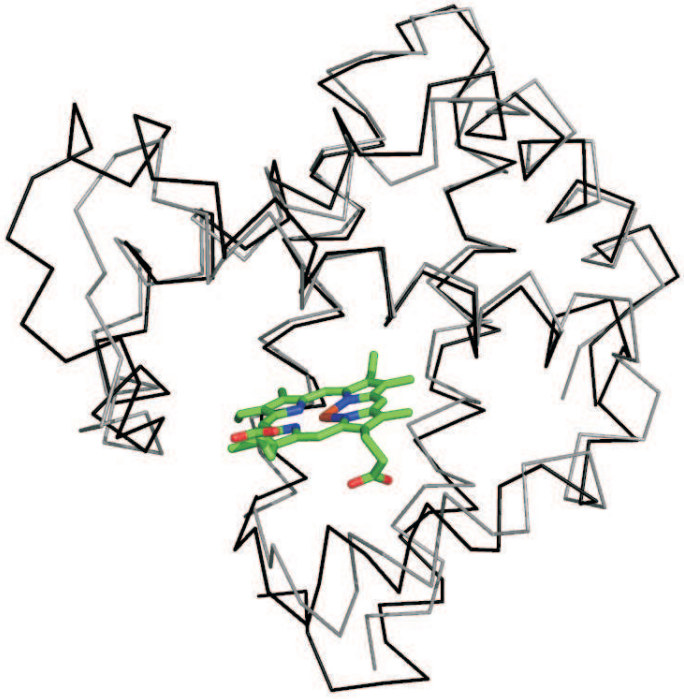


Figure 1.10: Superposition of myoglobin (black) and hemoglobin (light grey) in ribbon representation together with the heme group (sticks representation).

In an extensive evaluation of sequence alignments of protein pairs with similar and dissimilar structure, Rost [175] analysed the minimum sequence identity which is needed to infer structural similarity. The relationship between sequence and structure is dependent on the alignment length, but for long alignments, high sequence identity (>40%) guarantees structural similarity. In the so called "twilight zone" between 20-30% the relationship is uncertain.

1.1.5 Experimental Methods

The two experimental methods able to determine protein structures at atomic resolution are X-ray crystallography and NMR-spectroscopy. More than 85% of the protein structures in the Protein Data Bank *(see next section)* are determined by the former method. Cryo-electron microscopy is also used, but this method can only extract low-resolution information of large protein complexes and is therefore not described here.

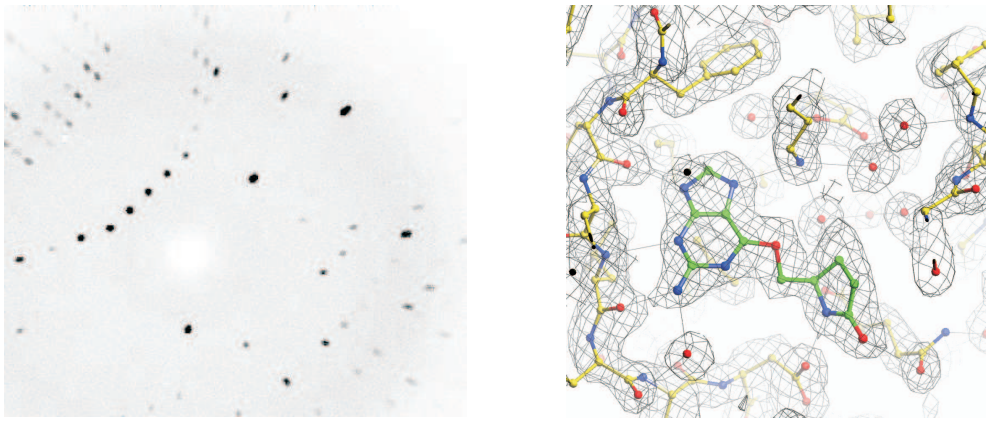


Figure 1.11: Typical images in X-ray crystallography: an expample of a diffraction map (left) and a electron density map (right) derived from it.[d]

In X-ray crystallography, the first and most difficult step is the growth of a well-ordered crystal. The crystal lattice is then irradiated with X-rays leading to a diffraction pattern specific for the given protein structure (see Figure 1.11 left hand side). The X-rays, which have wavelengths in the order of interatomic distances, are dispersed by the electrons in the molecule and interfere with each other resulting in a diffraction pattern reflecting the relative positions of the electrons in the crystal. The electron density is calculated from the amplitudes and the phases of the diffraction waves by a Fourier transform function. Unfortunately, the phase information cannot be measured in this process and additional information is needed in order to estimate the phases (*e.g.* by isomorphous replacement or molecular replacement). After Fourier transform and solving the phase problem, an electron density map can be built as shown in Figure 1.11 right hand side).

In the refinement process a model of the protein structure is fitted in the electron density map using information about standard geometries for bond lengths and angles. The accuracy of the electron density map and the corresponding model of the protein structure depend on quality and amount of available data compared to the number of unknowns (atoms in the protein) and is expressed by the term "resolution" (in Ångstrom). From the model of the structure it is possible to recompute the diffraction map and compare it with the original one. The difference is reflected by the R factor.

[d] http://en.wikipedia.org/wiki/Portal:Xray_Crystallography,
http://biop.ox.ac.uk/www/lab_journal_1998/Endicott.html

A good structure should have an R value of less than resolution divided by 10.

Nuclear magnetic resonance (NMR) spectroscopy is a method which allows to determine the structure of a protein in solution. The solution is exposed to a powerful magnetic field which causes the spin of the nuclei to be oriented in direction of the external field. An additional magnetic field is used in order to measure the frequency at which the different atom nuclei switch the spin orientation (called resonance frequency). The resonance frequency of an atom depends on its type but also on the environment. The magnetic interaction of the spins of two atoms close in space can be measured and its intensity depends on the distance, which allows to derive a set of distance constraints. Given a sufficient number of constraints a finite set of models can be built. The more constraints are given and the closer the models become. For highly flexible regions the derivation of distance constraints is hindered and therefore the models in these segments are less similar.

1.1.6 The Protein Data Bank

The experimentally determined structures of proteins (but also other macromolecules) are deposited in the publicly accessible Protein Data Bank (PDB) [18]. Each structure in the PDB has a unique identifier composed of four letter. At the date of this work (September 2007) the PDB holds 45,506 structures, most of which are proteins determined by X-ray crystallography. The PDB contains a considerable amount of redundancy (*e.g.* because some proteins involved in diseases have been solved with different bound ligands). A non-redundant subset of the PDB composed of structures with less than 90% sequence identity and resolution better than 3 Å yields in approximately 12,000 structures. The size of the PDB has grown exponentially over the last years as it can be seen from Figure 1.12.

Regardless of the exponential growth of the PDB, the number of new folds (based on the SCOP classification) entering the PDB has decreased over the last years. Virtually no new fold were solved over the last two years. This can be attributed to the fact that on one hand some proteins (especially membrane proteins) are very difficult or impossible the determine with current methods. On the other hand, structural genomics initiatives have solved many of the missing folds over the last years.

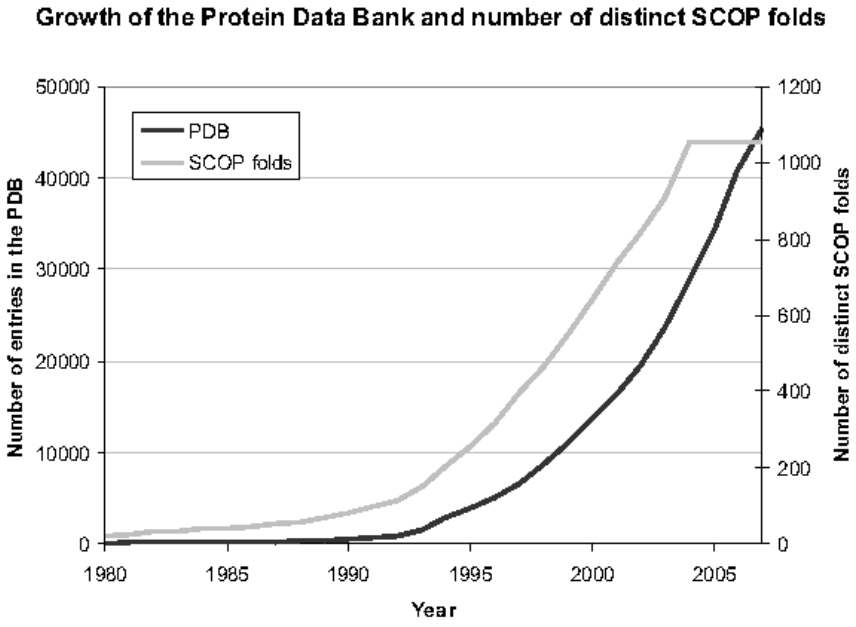


Figure 1.12: Growth of the Protein Data Bank from 1972-2007 *(data source: www.pdb.org)*.

1.1.7 Structural genomics

The goal of the worldwide structural genomics initiatives is to provide structural information for most of the known protein sequences through a combination of experimental and computational methods [33].

The structural genomics effort started around the year 2000 and can be split in three main groups: the Protein Structure Initiative (PSI) by the US National Institute of Health, the Japan-based program led by the RIKEN research foundation and the european effort with the Structural Genomics Consortium (SGC) and SPINE.

One aspect of structural genomics initiatives is the emphasis on high throughput protein structure determination, which allows to solve structures faster and with lower costs. In the last seven years, more than 5,000 new protein structures from the structural genomics centers have been deposited in the PDB (see Figure 1.13).

[e]source: http://sg.pdb.org/

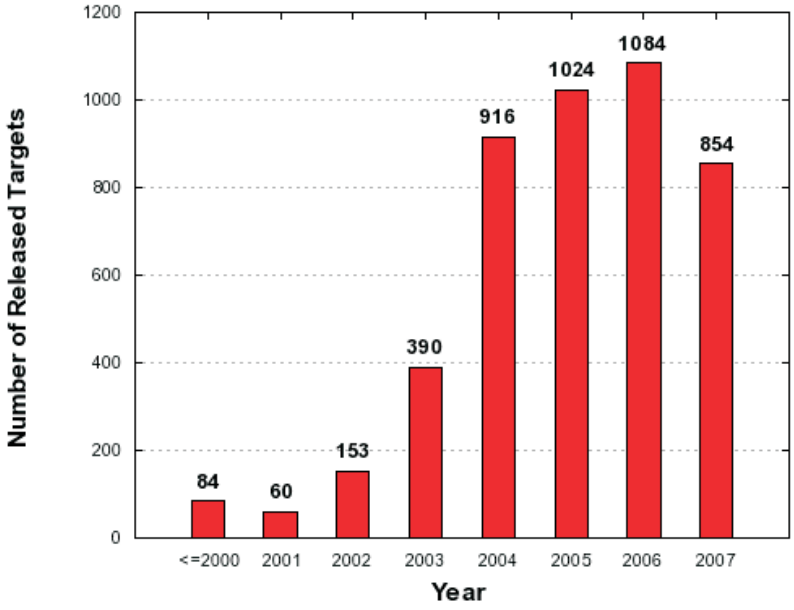


Figure 1.13: New structures solved by the structural genomics centers (As of: September 2007).[e]

The structural genomics projects attempt to deliver structural templates for members of all protein families in which they were very successful until now (a review on the expectations and outcomes of the structural genomics initiatives can be found in [34]). Targets for structural genomics are proteins with less than 30% sequence identity to any structure in the PDB. Protein sequences above this cutoff typically have a similar structure as mentioned above and can therefore be solved by homology modelling *(see next section)*. At the beginning of the year 2005, about 36% of the Pfam families (Pfam is a manually curated database of protein families) contained at least one member with known structure. This allows to model the other family members [34]. It has been estimated in 2004 [33] that around 57% of the domains of all sequences can be modelled with the current PDB. An estimated number of 10,000-16,000 structures have to be determined experimentally in order to model most of the current sequences [33, 230].

1.2 Protein structure prediction

The functional characterisation of a protein sequence is strongly facilitated by the knowledge of its 3-dimensional structure. Structural information can be used to ask new biological questions and efficiently design experiments. To close the gap between the number of known sequences (approximately 4.8 million in UniProt/TrEMBL[f]) and the fraction for which the structure is solved (approximately 45,500 in the PDB), efficient methods for protein structure prediction are needed that complement current efforts in structural genomics (see Chapter 1.1.7).

Protein structure prediction refers to the prediction of the tertiary structure of a protein given its sequence by means of computational methods. Two fundamental principles are acting on proteins that guide their 3-dimensional structure: the laws of physics and the theory of evolution. Accordingly, there are two different classes of protein structure prediction methods: *ab initio* methods and template-based methods.

Ab initio or *de novo* methods try to predict the structure of a protein from the sequence alone based on the laws of physics and chemistry assuming that the native structure is in the global free energy minimum. In contrast, template-based methods take into account structural information from experimentally solved protein structures ("the templates") to build a model of the target sequence relying on the fact that structure is more evolutionarly conserved than sequence [39] and that proteins adopt a limited number of folds [37, 152, 231]. Traditionally template-based modelling has been split into the two fields of fold recognition and comparative (homology) modelling, depending on the approach used for template identification. A constantly increasing overlap between the three fields can be observed over the last years making the boundaries increasingly blurred. An overview on the different methods is given below.

The accuracy of models generated by template-based modelling techniques is highly dependent on the sequence identity between the target sequence and the template of known structure. It based on the relationship between sequence and structure of a protein described in Chapter 1.1.4.1. The application of protein structure models is determined by their accuracy [11]. High to medium accuracy models generated by comparative modelling, based on a template with more than 30% sequence identity to

[f]source: `http://www.ebi.ac.uk/swissprot/sptr_stats/index.html`

the target can for instance be used for structure-based drug design, the investigation of the shape and volume of the binding site or for refining function prediction based on sequence [98, 161].

1.2.1 CASP

Critical Assessment of techniques for protein Structure Prediction (CASP) is a community-wide experiment taking place every two years with the aim of assessing the progress in this field [143, 147]. CASP is a blind test experiment where the predictors receive a set of protein sequences for which the structure is about to be experimentally solved. During the prediction season, of approximately 3 months, the native structures remain unknown to the predictors. Afterwards the quality of the submitted models is analysed by independent assessors and the results are presented at the CASP conference and in a special issue of the journal *Proteins* (*e.g.* [145], [144]).

The number of prediction targets steadily increased over the years from 33 at the beginning of CASP in the year 1994 to 95 accepted targets at the seventh round of CASP in summer 2006. The targets are categorised according to modelling difficulty in comparative modelling, fold recognition (homologues and analogues, respectively) and new folds. For the last CASP round, the categories have been redefined to reflect developments in methods in template-based modelling and (template-)free modelling.

1.2.2 Overview of methods

1.2.2.1 Ab initio

Ab initio or *de novo* methods try to predict the native structure of the protein by simulating the biological folding process. Folding simulations using molecular mechanics force-fields and molecular dynamics simulations are not discussed here since these applications are limited to very small polypeptides and require an enormous amount of computational time.

In practice, most of the *ab initio* methods incorporate to some extent available structural information either through the use of fragments from known protein structures

or in devising scoring functions. This is the reason why the term "new folds" and "free modelling" have been used to describe this field in the last rounds of CASP.

The two major problems in *ab initio* structure prediction are the vast number of conformations that have to be sampled and the inaccuracies of the scoring functions. The combinatorial explosion can be approached by using reduced representation of conformations and by efficient sampling strategies. Successful approaches include methods which build structures from short protein fragments (so called fragment assembly methods) such as ROSETTA [21, 196] and lattice-based simulations [154, 246]. A combination of both is implemented in TASSER (Threading/ASSEmbly/Refinement) [245] which assembles the model from structural fragments of templates identified by threading, if possible, and uses a lattice-based approach for the remaining parts. Usually, a vast amount of conformations is generated from which the final model is selected by clustering the solutions and applying a composite scoring function.

1.2.2.2 Fold recognition

Fold recognition is based on the notion that protein structure is much more evolutionarly conserved than sequence and that the number of adopted protein folds is limited. Two proteins can share the same fold even if the sequence similarity is either very low or does not exist. In previous CASP rounds (until CASP7), the fold recognition targets have been divided in homologous and analogous folds. Homologues are evolutionarly related and diverged from a common ancestor. Analogues have no evolutionary relationship and are a result of convergent evolution, meaning that nature has independently "re-invented" the fold. The definition of analogues is rather vague and strongly depends on our ability to detect remote evolutionary relationships: as a result of advances in sequence comparison methods such as PSI-BLAST [6], proteins which have been originally regarded as analogues have been later confirmed to be homologues.

The traditional division in homology (comparative) modelling and fold recognition was based on the difficulty to detect a suitable template. Whereas in homology modelling the template could be more or less easily identified (*e.g.* by a simple BLAST run), more advanced methods were used in fold recognition. Nowadays, fold recognition

methods are not only standard in the field of protein structure prediction and part of virtually all comparative modelling pipelines but also of *ab initio* methods (*e.g.* some fragment assembly methods). In the following, approaches for template identification which arose from the fold recognition field are briefly described.

Historically, fold recognition can be divided into threading methods and sequence similarity-based methods. Threading methods were developed in the hope to detect analogous folds with no evolutionary relationship. They take their name from the conceptual threading of the sequence of a protein through a library of folds with the intention to identify the fold that fits the given sequence best. The fitness of each residue is assessed separately by analysing its compatibility with the given local conformation and the structural environment. This has led to the development of contact potentials [104, 197, 200, 209] and 3D-profiles which encode the structural environment of the residues [24]. Dynamic programming is usually applied in order to align the sequence to the template structure. By this stepwise mapping of the target sequence onto the structure of the template, the structural environment changes accordingly. This problem divides the threading methods into those using the "frozen approximation" leaving the structural environment as in the template and those using the "defrosted approximation" in which the surrounding amino acids are updated [85, 201]. The models of the query protein, based on the alignment to the different template folds are often further evaluated by contact potentials and other statistical potentials. The application of these methods is not restricted to fold recognition and similar methods are used in model quality assessment in general (see Chapter 1.2.4).

Sequence similarity-based methods try to identify templates which are evolutionarly related to the target sequence. Sequence-sequence comparison methods such as FASTA [160] and BLAST [5] are the most simple methods to assign a fold of a protein (*e.g.* by a BLAST search of the query protein sequence against the sequences of all experimentally solved proteins). BLAST, which stands for Basic Local Alignment Search Tool, has become one of the standard tools in the bioinformatics community and beyond it. The algorithm basically consists of three steps: First, the sequence database is scanned for exact matches of sequence fragments of fixed length contained in the query sequence (the "seeds"). In the second stage, the seeds are extended in both directions. Finally, high scoring ungapped alignments are collected and gapped alignments of

the query sequence with the corresponding database sequences are generated using a modified version of the Smith-Waterman algorithm for local alignments [203]. The statistical significance of the hits is reported as an E-value which reflects the number of different alignments with equivalent or better score that are expected to occur in a database search by chance. Basic ingredients of an alignment algorithm based on dynamic programming such as Smith-Waterman and Needleman-Wunsch [150] (for global alignments covering the entire length of both sequences) are a substitution matrix which defines the similarity between two amino acids [89] and the penalty of setting a gap (usually a separate gap open and a gap extension penalty are used).

A new generation of alignment algorithms came up in the mid 1990's based on the assumption that conserved sequence motifs should have a stronger influence on the alignment than variable regions resulting in the development of position-specific scoring matrices (PSSMs) [22]. As opposed to the ordinary substitution matrices (20 x 20 amino acids), PSSMs or profiles are composed of 20 x N entries (where N is the length of the sequence) and are generated by analysing the amino acid variability in a multiple sequence alignment of the family of the query protein. A profile describes a family of homologous proteins and not a single sequence. As a consequence, profile-sequence comparison methods have been developed with PSI-BLAST [6] as the most prominent representative. PSI-BLAST (Position-Specific Iterative-BLAST) uses the same heuristics as the original BLAST (explaining its speed) and additionally an iterative generation of multiple sequence alignments and profiles in order to increase the search sensitivity. In a closely related approach the family-specific information is stored in hidden Markov models (HMMs) [63, 108].

The sensitivity in detecting weak evolutionary relationships as well as the accuracy of the alignment has been further increased by the use of profile-profile (or HMM-HMM) comparison methods [155, 179, 180, 232, 243]. In these approaches the query profile is aligned to the profile of the template protein using a scoring function which calculates the compatibility of two columns in the profiles. Several alternative column-column scoring functions have been proposed in the literature as well as alternative ways to generate the profiles and to build the alignments (a review can be found in [140, 235]).

A clear trend to combine sequence and structure information is observable in the field over the last years, either by incorporation of structural information in sequence profiles

directly [1, 156, 210] or by integrating sequence information in threading [60, 157, 188]. A variety of approaches to integrate structural information from the templates in the sequence profiles have been proposed. Structural information can be integrated using predicted structural profiles in terms of secondary structure and sometimes solvent accessibility [65, 165, 178, 250]. Secondary structure information for example is used by comparing observed secondary structures in the template and predicted states in the target.

1.2.2.3 Comparative modelling

As mentioned in Chapter 1.1.4.1, a sequence identity of roughly 30% is generally sufficient to infer structural similarity between two proteins. This is the fundamental idea behind homology or comparative modelling. With the growing number of experimentally solved protein structures, this concept has become a powerful method to predict the structure of a large fraction of the known protein sequences (see Chapter 1.1.7).

Homology modelling basically consists of six steps: template identification and selection, target-template alignment, initial model building, loop prediction, sidechain prediction and, finally, refinement and quality assessment (see Figure 2.1 in Methods for an overview). A short description of all steps is given below. Loop prediction as well as model quality assessment are picked out as central themes of this thesis in the next two sections.

The first two steps (template identification and alignment building) have been described in detail in the previous section. Usually, more than one template is identified and it is necessary to select the best candidate(s) for a given modelling problem. In this context, sequence identity between target and template is the most important argument but there are other factors which should be taken into account in template selection:

- A phylogenetic tree based on a multiple sequence alignment of the protein family can help to identify the template closest to the target sequence.
- The "environment" of the template should be analysed and compared to the situation in the target, *e.g.* quaternary interactions (Is the template part of a

complex and the target not?), protein-ligand interactions or chemical conditions (solvent, pH etc.).

- The quality of the experimental structure should be considered as well, *e.g.* resolution and R-factor of X-ray structures.

Multiple templates can be used as well, either by building alternative models based on the single templates and subsequently selecting the best one, or by combining parts of multiple templates. The simple rule that combining multiple templates instead of using a single best template results in better models does not hold, as it has been shown by Venclovas and Margelevicius in the CASP6 evaluation [227]. However, as identifying the best template among several is not always a trivial task, using multiple templates increases the chance of selecting the best template.

The alignment produced in the fold recognition step is often not the optimal one (*e.g.* BLAST typically produces local alignments covering only a part of the target). Specialised methods should be used in order to align the target sequence to the template structure.

In terms of fold recognition sensitivity and specificity as well as in terms of accuracy of the resulting alignments, profile-profile methods have been shown to outperform sequence-sequence and profile-sequence methods [100, 132, 179, 187, 243]. In general, integrating structural information (*e.g.* based on multiple structural alignments of templates [1, 110] or environment-specific gap penalties [191, 210]) tend to improve the alignment accuracy but most probably not the fold recognition sensitivity. With decreasing sequence identity between target and template (especially below 30%), the alignment accuracy drops rapidly and alignment errors become the major source of errors in homology models.

The alignment produced by a dynamic programming algorithm using a specific gap penalty is not necessarily the best alignment to generate the model. Thus, using sub-optimal alignments, representing alternative paths in the alignment matrix, may identify more suitable alignments [45, 138, 149, 186, 228]. Additionally, a set of sub-optimal alignments can be used to predict the local alignments reliability. Local alignment paths used by a higher number of sub-optimal alignments can be regarded as more reliable. An alternative way to assess the local alignment reliability has recently

been proposed by Tress *et al.* [223, 224]: the local alignment quality is predict based on the information about the observed frequencies in the sequence profiles.

There is no alignment protocol that is clearly superior over other protocols for every protein family and similarity level. Elofsson [65] for example pointed out that, for proteins related to the family level, purely sequence-based methods tend to produce better models, whereas at fold level, sequence-based methods including predicted secondary structure outperform purely sequence-based approaches. Thus, many groups produce several alignments based on different protocols, parameters and sometimes sub-optimal alignments. The final model is then selected based on a scoring function (see Chapter 1.2.4).

Building a model based on the alignment between target and template is fairly straightforward. A variety of methods can be used which can be roughly divided into three groups [133]:

- modelling by assembly of rigid bodies [20, 88]
- modelling by segment matching or coordinate reconstruction [105, 123]
- modelling by satisfaction of spatial restraints [8, 181]

Assembly of rigid bodies relies on the fact that the structure of proteins belonging to the same family can be roughly divided into structurally conserved regions (SCRs), or the structural "core" and structurally variable regions (SVRs). The model is built by assembling the core segments from one or several templates and modelling of the structurally conserved regions (loop prediction).

In the second approach, a model is constructed by using a subset of the coordinates of the template (typically Cα atoms of conserved residues) as guiding positions on which short all-atom segments are fitted. These segments can either be extracted from experimentally-solved structures [43, 93] or obtained by a conformational search guided by the Cα-trace [15, 55].

In modelling by satisfaction of spatial restraints, a model for the target sequence is derived by minimising the violations of all restraints on the target. The restraints are obtained from the alignment to the templates (*e.g.* distances and angles) and are

usually supplemented by other stereochemical restraints (*e.g.* bond lengths and angles, torsion angles and non-bonded contacts).

The accuracy of models generated by the different approaches does not differ much since other factors such as template selection and target-template alignment have a much stronger impact on the quality of the final model.

In a next step, the backbone of regions which cannot be directly obtained from the templates (*i.e.*, the structurally variable regions) have to be modelled. These regions often correspond to loop regions at the protein surface which connect regular secondary structure elements and are the location where mutations (amino acid substitutions, insertions and deletions) tend to accumulate. Since loops often define the functional specificity of proteins and contribute to the binding site, an accurate prediction of loop structures finally determines the usefulness of the homology model (*e.g.* for protein-ligand docking). A detailed introduction to loop prediction is given in the next section.

Sidechain modelling represents the last step toward a first all-atom model of the target. It has been shown that the principal factor determining the sidechain conformation, beside packing in the structural core, is the local backbone conformation [23, 183]. The observation that sidechains show a strong preference for specific conformations led to the development of rotamer libraries [163].

Most methods use as starting point the most frequent rotamer for each amino acid and subsequently optimise the conformations. Since the sidechain conformation of conserved residues in homologous structures are often identical, they are usually copied

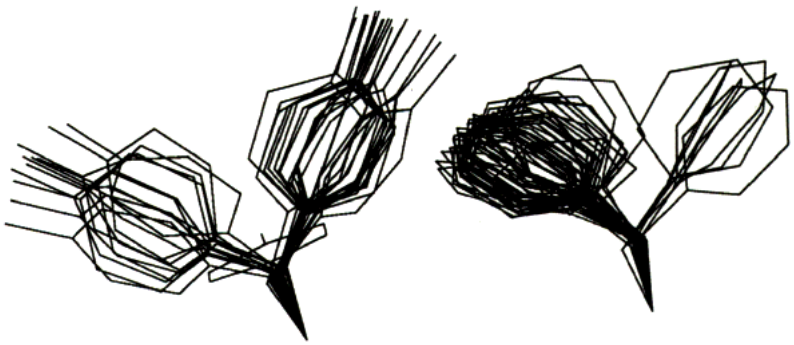


Figure 1.14: Some sidechain conformations observed for tyrosine and phenylalanine [36].

from the template instead of using a rotamer library. A frequently used program for sidechain modelling is *SCWRL* [31], which uses a heuristic search strategy based on backbone-dependent rotamer libraries extracted from a set of known structures. As a consequence of the relationship between backbone and sidechain conformation, the limiting factor on sidechain accuracy is backbone accuracy [42].

Refinement refers to the attempt to bring an approximate model of the target protein closer to the experimental structure. The most frequent sources of errors in comparative modelling are: alignment errors, incorrect templates, wrong loop modelling, distorsions or shifts in correctly aligned regions and errors in sidechain packing. As observed at CASP, predicted models are still rarely closer to the native structure than the best template [222]. The CASP experiment also revealed that refinement is problematic and no method is currently able to improve consistently over the initial model [116].

Estimating the accuracy of a model is an essential step in comparative modelling since the quality of a model determines its usefulness. The stereochemistry of a model can be analysed with standard tools such as PROCHECK [117] or WHATCHECK [96]. Scoring functions used to identify the best model among a set of alternative conformations or to identify regions of structural errors fall into two broad categories: physics-based energy functions and knowledge-based scoring functions based on 3D profiles (*e.g.* VERIFY3D [129]) or statistical potentials (*e.g.* PROSA [199] or ANOLEA [136]). A comprehensive introduction in model quality assessment is given in Chapter 1.2.4.

1.2.3 Loop modelling

As the sequence identity between target and template decreases, an increasing number of insertions and deletions as well a local loss of sequence similarity is observed, typically in solvent-exposed regions between secondary structure elements. These regions, often referred to as loops, have to be remodelled since the backbone of the template cannot be used. As mentioned above, loops often determine the functional specificity of proteins belonging to the same family (*e.g.* the hypervariable region in antibodies) and therefore the accuracy of loop modelling (or loop prediction) strongly influences the usefulness of a model for function annotation or structure-based drug design [91, 98].

Loop prediction can be seen as a constrained "mini-folding" problem [77] in which a polypeptide segment with a given sequence is modelled using geometric constraints imposed by the backbone atoms on both sides of the loop that anchor it to the remainder of the protein (called anchor groups or loop stems). It has been shown that segments of up to nine residues with identical sequence can have entirely unrelated conformations [46, 185]. Thus, the conformation of a loop is determined not only by its sequence but also by the geometry of the anchor region and the structural environment.

Many loop modelling procedures have been described in the literature and they can be generally grouped into *ab initio* methods and database search techniques (knowledge-based loop prediction) as well as combinations of both. Loop modelling basically consists of two steps: sampling (the conformational space) and scoring, optionally with an intermediate filtering step. *Ab initio* loop prediction methods are based on a conformational search in the given structural environment usually guided by an energy function. Algorithms used in conformational search include discrete sampling of energetically favourable main chain dihedral angles [52, 56, 146, 251], random tweak methods [190, 207, 241], analytical methods [86, 218], molecular dynamics simulations [26, 77], Monte Carlo with simulated annealing [32, 47] and many more. Usually, the loop is incrementally built up from one anchor and a *loop closure* algorithm [30, 112, 190] is used in order to generate closed conformations. There are also approaches which build the loop from both the N-terminal and C-terminal anchor group and connect the fragments in the middle [99, 171, 251]. The conformations generated by *ab initio* methods are often evaluated using a scoring function based on terms from molecular mechanics force fields [80, 99, 168, 171, 241] sometimes in combination with statistical potentials [77, 207].

On the other hand, knowledge-based or database search methods extract the loop conformations from experimentally solved protein structures from the PDB [28, 53, 61, 70, 71, 105, 120, 134, 139, 151, 208, 220, 226]. In contrast to *ab initio* methods, the local loop geometries predicted by knowledge-based approaches represent physically reasonable conformations since they are observed in native protein structures. In knowledge-based approaches, protein structure fragments of the desired length are selected from the database which approximately fits to the geometry imposed by the anchor groups. The fragments are usually scored according to the "goodness of fit" of

the fragment to the anchor region and other criteria such as sequence similarity between the database fragment and the loop to be modelled [70], the use of environmentally constrained substitution tables [53, 214] or the energy of the fragments based on a distance-dependent statistical potentials [52]. A subsequent optimisation and ranking of database loops with a molecular mechanics force field has also been suggested [226].

The accuracy of knowledge-based approaches is limited by the completeness of the PDB concerning structural fragments of a given length. In 1994, Fidelis *et al.* [74] estimated that fragments of up to 7 residues can be accurately modelled (RMSD $<$ 1 Å) with the PDB. Lessel and Schomburg [120] confirmed these results and showed that the coverage is even lower if stricter and more realistic cutoffs are used. *I.e.* fragments are not fitted on each other but on the terminal anchor residues and a RMSD cutoff of 0.8 Å was used. As a result of the exponential growth of the PDB over the last years the coverage of loop conformations has increased dramatically and recent publications report a much higher coverage even for longer loops [62]. Fernandez-Fuentes and Fiser [69] calculated a coverage of $>$95% for fragments up to 10 residues.

Several loop classification methods have been described in the literature [29, 71, 72, 126, 151, 239]. The most common classification criteria are geometry of the surrounding secondary structure elements, loop length, loop sequence, torsion angles and solvent accessibility.

Beside alignment accuracy, loop prediction is still a major source of errors in comparative modelling [221] and only short and medium loops (less than approximately 8 residues) can be modelled with acceptable accuracy [174]. The prediction accuracy for longer loops rapidly drops in all current methods although remarkable progress has been reported recently, if in addition to an extensive conformational sampling strategy, crystal contacts are taken into account in loop ranking [99, 251]. This also demonstrates the limits of loop prediction: beside the fact that many loops are highly flexible, the conformation of a loop in a crystal structure may be determined in part by packing constraints and does not present the native conformation of the loop in solution.

Loop prediction methods are usually tested in "self prediction" experiments which means that the loop is cut out from the protein and rebuilt with the given method in the fixed structural environment. This does not represent a realistic modelling situation

in which the geometry of the anchor region, as well as the structural environment, are only approximately correct. Furthermore, in knowledge-based loop prediction, often different sequence similarity thresholds are used in order to remove trivial results. *I.e.* loops from close homologues of the query protein which are usually not present in the application case. Because loops from homologous protein structures are often the best available fragments in the database, the sequence identity cutoff used in the evaluation of the method strongly influences the prediction accuracy.

Another problem, which makes a fair comparison of current loop modelling protocols difficult, is the fact that no standard benchmark set for loop prediction exists. Most methods are tested on their own test sets and the performance is often compared to other methods based on only a few examples. In a recent benchmarking by Rossi *et al.* [174], four commercial loop modelling programs have been tested on a comprehensive test set covering loops of 4 to 12 residues based on a the work of Jacobson *et al.* [99]. The results were rather disillusioning in that only short loops (4 to 7 residues in length) could be modelled with acceptable accuracy for structure-based drug design and all methods have considerable problems in loop ranking (*i.e.* the top-scoring loop was rarely the loop with minimal RMSD compared to the native conformation). These results underline the general problem in loop prediction: the bottleneck in loop modelling seems to be no longer the sampling step (as a consequence of advances in sampling algorithms and the growth of the PDB) but the subsequent scoring of the conformations.

1.2.4 Model quality assessment

Particularly *ab initio* methods, but increasingly also template-based approaches, usually produce a considerable amount of alternative models. Selecting the model being closest to the native conformation of a given protein out of an ensemble of models, independent of being produced during conformational search in a template-free approach [172, 248] or on the basis of alternative alignments or different templates [48, 101, 206], is a crucial step in protein structure prediction in general. This section provides an overview on the topic and the methods used in the assessment of model quality. An in-depth introduction to the theoretical background of statsitical potential

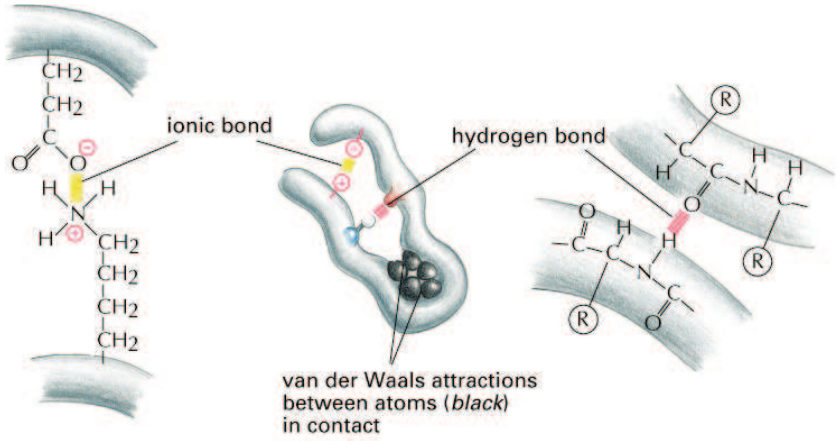


Figure 1.15: Schematic representation of physical forces occuring in proteins (*source:* [2]).

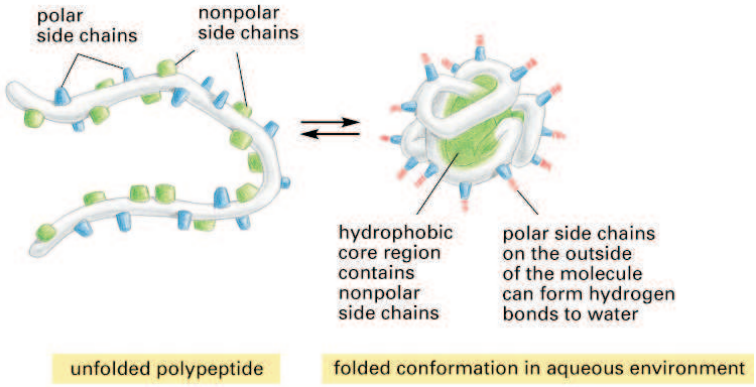


Figure 1.16: Schematic representation of hydrophobicity (*source:* [2]).

scoring functions is given in Methods (Chapter 2.4.1.1).

Scoring functions rely on the thermodynamic hypothesis stating that the native state of a protein lies in the free energy minimum under physiological conditions [119]. There are basically two categories of scoring functions: physics-based energy functions and knowledge-based statistical potentials. The former are true effective energy functions describing interactions observed in proteins and their parametrisation is performed either by fitting experimental data or based on quantum chemical calculations [25, 79, 118]. A schematic representation of some important forces in proteins is given in Figure 1.15 and 1.16.

Statistical potential energy functions are derived from data of known protein structures and are usually formalised as either distance-dependent or -independent pairwise potentials of mean force [9, 128, 135, 184, 189, 197, 198, 213, 249]. Alternatively, statistical potentials have been derived for other structural features such as torsion angles [3, 16, 19, 111, 193, 215] and solvent accessibility [94, 104].

Statistical potentials are based on the inverse Boltzmann equation, which relates frequencies of observed structural features to their energy. A detailed description of the theoretical background of statistical potentials is given in Methods on page 55. They have the advantage of being fast and simple to construct and they are widely used for various purposes among which are fold recognition [102, 141, 170, 200, 202], identification of the native structure among decoys[g] [158, 225], model quality assessment [16, 66, 215, 233] or prediction of thermo stability [83, 84, 97, 159].

Combining several statistical potential terms covering different aspects of protein structures or models is a popular strategy and the combined potentials have been shown to outperform any single potential [16, 66, 111, 135, 198, 215, 233]. Model quality assessment programs are used to assess models generated by various methods and the quality of the models range from very coarse *ab initio* models often having a wrong fold to very accurate template-based models. Therefore, scoring functions consisting of several terms and being optimised on a diverse set of models will be more suitable for the task of discriminating good from bad models or for the identification of the most native-like structure. Model quality assessment programs have been tested the first time in a community-wide experiment in 2004 during CASP6 as part of CAFASP (Critical Assessment of Fully Automated Structure Prediction) [76] and only recently at CASP7 [49].

[g]Decoys are computer generated conformations of protein sequences that possess some characteristics of native protein structures, but are not biologically real.

1.3 Objectives

Homology modelling is currently the most successful approach for the prediction of the 3-dimensional structure of a protein from its sequence. A model of the protein is thereby built by using information from experimentally solved protein structures (the templates) showing an evolutionary relationship to the target protein, relying on the fact that the structure of a protein is more evolutionarly conserved than its sequence.

The objectives of this thesis are to optimally take advantage of the information contained in the database of known protein structures especially for the prediction of loop regions and for the assessment of the quality of the generated models. Both tasks are of crucial importance for the final application of the models.

Both loop prediction as well as the scoring functions used for the quality assessment loops and entire models can benefit from the steadily growing number of known protein structures. In knowledge-based loop prediction, the coverage of the conformational space by fragments extacted from known structures increases with the number of known proteins. A comprehensive and up-to-date fragment database will be established in the course of this work. Furthermore, scoring functions based on the statistical analysis of structural features observed in experimentally solved proteins are potentially more accurate and wider applicable as the number of folds increases. These statistical potentials can be used for the assessment of entire models but also for the ranking of candidate fragments in loop prediction. In this work, it shall be investigated whether a statistical interaction potential on atomic level can be used for the ranking of complete loops after sidechain modelling. The knowledge-based loop prediction algorithms described in the literature typically take into account only the loop backbone in the scoring step and mostly rank the loops according to the geometrical fit of the fragments on the anchor groups of the protein. This approach is problematic since the anchor region is typically distorted with respect to the native structure.

For the assessment of the quality of protein models, a scoring function shall be implemented being able to identify good models among a set of alternatives. It will be investigated whether the combination of multiple terms can improve the prediction of the model accuracy. In order to be able to cope with loop prediction and model quality assessment, a comparative modelling pipeline needs to be implemented.

2 Methods

This chapter is structured according to the typical modelling workflow shown in Figure 2.1. Establishing a complete comparative modelling pipeline was a basic prerequisite for dealing with loop prediction and model quality assessment which are described later in this chapter. The modelling pipeline has been implemented in C++. A description of the most important classes can be found in the last section on page 74.

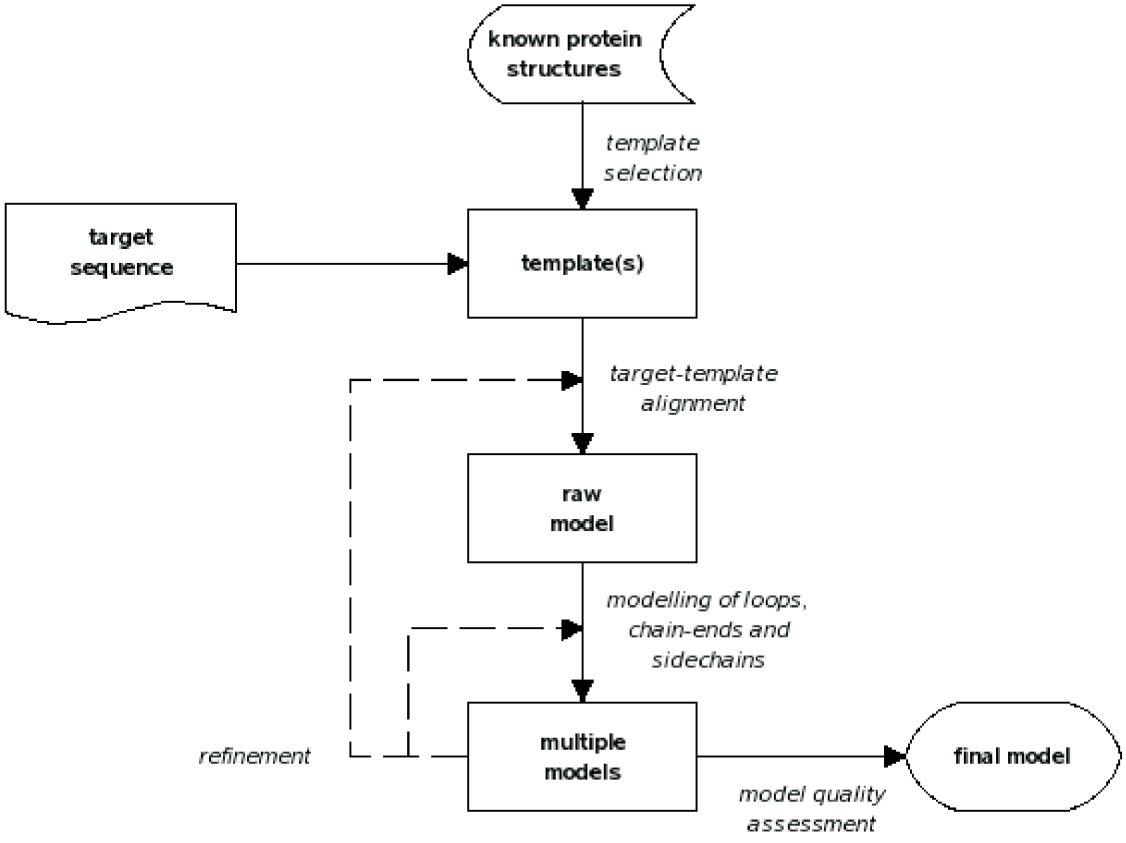


Figure 2.1: Basic steps in homology modelling.

2.1 Template selection and alignment

2.1.1 Databases

The non-redundant sequence database (nr) from the National Center for Biotechnology Information (NCBI) has been downloaded from the official ftp-server[a]. The *nr* database contains all publicly available sequences from a variety of sources (*e.g.* translations from GenBank [17] and RefSeq [164] as well as sequences from Swissprot [10], PIR [13] and the PDB [18]). In order to further reduce the redundancy (*e.g.* because of protein families being over-represented), NCBI's non-redundant sequence database was clustered at color 90% sequence identity using the tool CD-HIT [125]. The resulting database (*nr90*) was subsequently used to generate the profiles used for template identification and target-template alignment.

The database containing the sequences of all known protein structures from the Protein Data Bank (PDB) [18], frequently called *pdbaa*, has been obtained from the Dunbrack Lab[b]. In comparison to the *pdbaa* sequence database from NCBI, the version from Dunbrack Lab contains additional information such as resolution, R value, R free value and sequence length in the header of each entry. These information are crucial for template selection.

2.1.2 Template identification and selection

The template structures are identified using a variation of the PDB-BLAST protocol. The term PDB-BLAST was introduced in a work of Rychlewski and co-workers [179] in which several strategies of using sequence profiles for fold recognition have been compared. In PDB-BLAST, the profile generated by PSI-BLAST [6] is stored and used to scan the database of known protein structures. In the implementation used in this work, the profile generated after 4 PSI-BLAST iterations on the *nr90* sequence database is subsequently used for a final iteration on the *pdbaa*. After each PSI-BLAST iteration only sequences with an E-value ≤ 0.001 are retained. The maximum number

[a] `ftp://ftp.ncbi.nih.gov/blast/db`
[b] `http://dunbrack.fccc.edu/Guoli/pisces_download.php`

of sequences in the alignment was set to 1000.

One or several templates are selected manually based on the observed sequence identity to the target and their quality (*i.e.* resolution, target coverage, completeness). The sequence identity is calculated based on the alignments provided by PSI-BLAST.

2.1.3 Target-template alignment

The target-template alignments are built based on a profile-profile alignment protocol (see section 1.2.2.2 in the Introduction). The profiles for both target and template are calculated by PSI-BLAST with 5 iterations on the *nr90* data bank using an E-value < 0.001. The alignments are generated using a modified version of the profile-profile alignment functionality included in the Align-package, a C++ library provided by the Tosatto group [216]. The library has been extended and benchmarked as part of the CUBIC-project of Oscar Bortolami under the author's supervision.

A total number of 20 alternative alignments is generated by applying different gap open and gap extension penalties and by applying a global (Needleman-Wunsch [150] and a local (Smith-Waterman [203]) alignment algorithm.

The following strategy was used in order to optimise the gap penalties. The quality of sequence alignments is assessed by comparing them with structural alignments as gold standard. Therefore a representative set of structural alignments has been built as described by Marti-Renom *et al.* [132]. The final data set consists of 300 structural alignments of pairs of proteins sharing less than 40% sequence identity and belonging to the same homologous superfamily as defined by CATH [153], a hierarchical classification system for protein domain structures.

100 structural alignments have been used for training (optimising the gap penalties) and the rest for testing. The structural alignments were generated with CE [192]. An exhaustive search over a reasonable range for the gap penalties was performed in order to identify gap open and gap extension penalties which lead to a maximum overlap of the sequence alignments with the corresponding structural alignments. The quality of the resulting alignments was assessed based on the fraction of identically

Table 2.1: Optimised gap open (g_o) and gap extension (g_e) penalties used for local and global alignments, respectively.

global		**local**	
g_o	g_e	g_o	g_e
8	0.2	6.5	0.5
4.5	0.1	6.5	0.7
7	0.1	6.5	0.3
5.5	0.2	7	0.3
4.5	0.15	7.5	0.5
6	0.1	7.5	0.3
7.5	0.2	7.5	0.25
7	0.2	8	0.3
7	0.08	8	0.25
8	0.15	8.5	0.3

aligned residues. The final penalties are shown in Table 2.1. The optimal gap open and gap extension penalties, *i.e.* those values that produce the most similar alignments compared to the structural alignments, are shown in the first row and the sub-optimal penalties below.

In analogy to the scores for aligning two residues in a sequence alignment, profile-profile alignment algorithms need a scoring function which quantifies the degree of similarity of two profile columns being aligned. Several different implementations have been investigated and a column-column scoring function, as proposed by Panchenko in 2003 [155], has been used (formula 2.1). The score of aligning position i of the target with position j of the template is given by:

$$S_{i,j} = \frac{n_i(\vec{F}_i * \vec{W}_j) + n_j(\vec{F}_j * \vec{W}_i)}{n_i + n_j} \tag{2.1}$$

where n_i and n_j are the number of independent observations of different amino acid types in columns i and j representing a measure of the diversity within the columns. $\vec{F}_i$ and $\vec{F}_j$ are the vectors of observed frequencies in column i and j, respectively, in the profile. $\vec{W}_i$ and $\vec{W}_j$ represent the corresponding columns in the profiles or PSSMs (Position Specific Scoring Matrices).

2.2 Model building

2.2.1 Building the raw model

In a first step, the target sequence is mapped on the template structure according to the alignment, *i.e.* the sidechains of all non-conserved residues are removed and the amino acid type of the template is "mutated" to the one of the target. The sidechain conformation of conserved residues are inherited directly from the template, which turned out to be a good strategy (see section 3.1.5.4 in Results and Discussion). The sidechain conformations of the remaining residues are calculated with SCWRL [31]. Deletions (*i.e.* residues of the template not present in the target) are automatically removed from the structure. For insertions, "dummy residues" with the corresponding amino acid type of the target residue and consisting only of a Cα atom are added at the appropriate position in the structure. At all time, the mapping between the position in the alignment and the corresponding position in the model has to be guaranteed and is checked after each modification. Additionally, while loading a protein structure file, information from the program DSSP [107] (such as secondary structure assignment, solvent accessibility, torsion angles) is mapped to each residue and the integrity is checked. The resulting structure is called here the "raw model" since it is starting point of all subsequent modelling steps.

2.2.2 Defining the structural core and structurally variable regions

The structural core consists of those regions of the template which have preserved their structure during evolution and whose backbone conformation can be directly copied from the template. In order to illustrate the situation, the sequence alignment and the structural superposition of the two homologous proteins papain (PDB identifier 1ppn) and actinidin (PDB identifier 2act) are shown in Figure 2.2. The sequence identity between the two proteins is 47%. The structures are coloured according to the structural deviation between corresponding residues of the alignment. The region coloured in blue represents the structural core with low deviation between target and

template. As it can be seen, the structurally variable regions are mainly located around insertions and deletions.

The identification of the structural core, is facilitated by the use of the following information:

1. the sequence conservation in a multiple sequence alignment of the protein family of the target
2. the agreement between the secondary structure in target and template
3. the analysis of the local model energy profile (see section 2.4.5 on page 67)

The multiple sequence alignment of the target protein family is automatically produced based on the PSI-BLAST search used to generate the targets profile. The conservation within the protein family is visually inspected with JalView [44]. The multiple sequence alignments can be further refined by using MUSCLE [64], a highly accurate algorithm for multiple sequence alignments. A web service for MUSCLE is implemented in JalView and therefore, the PSI-BLAST based alignments can be directly refined in this environment.

The agreement between the secondary structure of the template and the target is investigated by comparing the calculated secondary structure of the template as derived from DSSP [107] with the predicted secondary structure of the target sequence. A consensus secondary structure prediction of PSIPRED [103], SSpro [35] and ProfSec/PHD [177] is built by simple majority voting [4], *i.e.* by assigning to each amino acid the secondary structure state predicted by at least two of the three methods (otherwise the residue is defined as being in coil state).

Regions of the model not belonging to the structural core (*i.e.* structurally variable regions) usually have to be remodelled. The structurally variable regions are mainly composed of protein surface loops containing insertions and deletions as well as the chain ends. Often, loops without insertions and deletions need to be remodelled as well, depending on the degree of sequence conservation between target and template. Highly non-conserved loops are likely to adopt different local folds as compared to the template loops. On the other hand, loop prediction is only possible with a certain

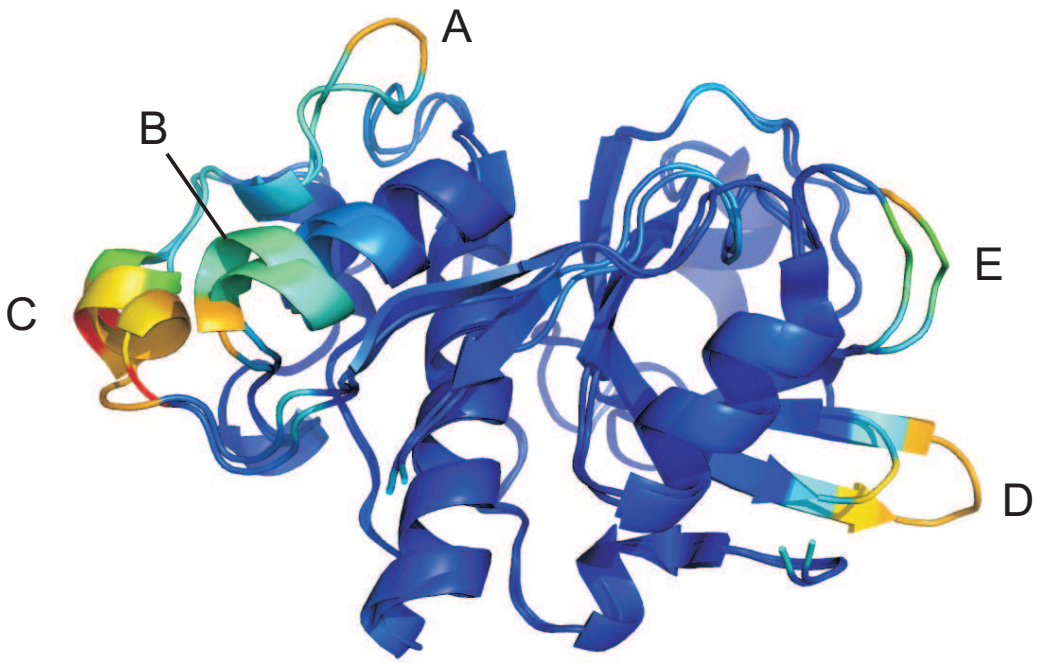
```
          1        10        20        30        40        50
1ppn      IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNEYSEQELLDCDRR-
2act      LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ
conserv   :* *****: ***  :*:** **:******::*:*** ** :*:*   *****:** * D
dssp_1ppn CCCCEEHHHHCCCCCCCECCCCCCHHHHHHHHHHHHHHHHHHHCCCCCECHHHHHHHCCEC

            60        70        80        90        100       110
1ppn      -SYGCNGGYPWSALQLVAQY-GIHYRNTYPYEGVQRYCRSREKGP-YAAKTDGVRQVQPY
2act      NTRGCDGGYITDGFQFIINDGGINTEENYPYTAQDGDCDV-ALQDQKYVTIDTYENVPYN
conserv   D: ** ***:   :*:::: D**::  :***      *  I  : D  :  *:  :*  :
dssp_1ppn CECHHHCCCHHHHHHHHHHHCCECECCCCCCCCCCCCCCH HHHHCCCECCCEEEECCCC

             120       130       140       150       160       170
1ppn      NEGALLYSIANQPVSVVLEAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGP----NYIL
2act      NEWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWI
conserv   ** ** : : :*****:*:***  *  * :*** **** ****:::**** DDDD *::
dssp_1ppn CHHHHHHHHHHCCEEEEECCCCHHHHHCCCCEECCCCCCCCCEEEEEEEEEEECCCEEEEE

                180       190       200       210
1ppn      IKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
2act      VKNSWDTTWGEEGYMRILRNVG-GAGTCGIATMPSYPVKY
conserv   :**** *:*** **:** *: *I: * **: *   ****:
dssp_1ppn EECCECCCCCECCEEEEECCCC CCCHHHCCCCCEEEECC
```

(a) Structure-based sequence alignment with insertions and deletions highlighted. The last line shows the secondary structure composition of the second protein.



(b) Superposition of two homologues coloured according to the local structural deviation.

Figure 2.2: Structural core and structurally variable regions: Alignment and superposition of the two homologous proteins papain (1ppn) and actinidin (2act).

accuracy, typically depending on the length of the fragment to be modelled. Therefore, deciding whether to re-model a loop or not remains a difficult task. These regions which would benefit from an accurate loop modelling are still difficult to identify and the prediction of these regions is an active field of research [68, 73, 77]. In order to investigate the tendency of a loop to adopt a different fold, a local statistical potential scoring function has been implemented investigating the local sequence to structure fitness. In other words, the scoring function assesses the likelihood that a given region of the target sequence adopts the structure provided by the template. High local energies suggest that the sequence does not "feel comfortable" with the given structure provided by the template and therefore a local refolding is rather likely. The local scoring function is described in Chapter 2.4.5.

Suitable start and end points of the loop modelling process, the so called anchor groups, have to be identified. The anchor groups are located in the transition of the structural core and the structurally variable region. Usually, in loop prediction the anchor groups are set near the end points of the surrounding secondary structure elements which are rather likely to be structurally conserved. As mentioned above, investigating the sequence conservation in these regions further provides evidence for the positioning.

For the models submitted to CASP, the position of the anchor groups has been defined manually by investigating the agreement between the position of the secondary structure end points between target and template and by looking at the sequence conservation. In order to combine all information needed to accomplish this task, a condensed "model information" output file is generated as shown in Figure 2.3. The following information is provided (in the same order as in the data lines):

- The alignment between target (in the example above CASP7 target T0379) and template (PDB identifier 2b0c, chain A) is shown in the first two data lines.
- The sequence conservation (denoted as "conserv") is described by an asterisk for identical residues and a colon for similar residues according to the definition used in CLUSTALW [212].
- The line "conf" shows the average confidence of the secondary structure predictions calculated by PSIPRED and ProfSec. Both methods provide a measure

```
alignment type:          global

Target sequence:
name:                    T0379
length:                  208

Template sequence:
name:                    2b0cA
length:                  199


Sequence identity:       0.201923
Energy of the raw model: -16.383


          1        10        20        30        40        50        6
T0379     MIRNIVFDLGGVLIHLNREESIRRFKAIGVADIEEMLDPYLQKGLFLDLESGRKSEEEFR
2b0cA     AKMLYIFDLGNVIVDIDFNRVLGAWSDLTRIPLASLKKSFHMGEAFHQHERGEISDEAFA
conserv   :    :****:*:: :     :  :  :: : :  :        :*   * *  *:* *.
conf      920056505760450070777788862455448887506740340000004788788888
consensus CCEEEEEECCCCEEEECCHHHHHHHHHHCCCCCHHHHHHHHHCCCHHHHHHCCCCCHHHHH
psipred   CCEEEEECCCCCEEECCCHHHHHHHHHHCCCCCHHHHHHHHHCCCHHHHHHCCCCCHHHHH
SSpro     CCEEEEEECCCCEEEECCHHHHHHHHHHCCCCCHHHHHHHHHCCCHHHHHHCCCCCHHHHH
phd       CCCCEEEECCCCEEEEECCHHHHHHHHHCCCCCHHHHHCHHHHCCCEECCCCCCCCCHHHHH
dssp      CCCEEEECCECCCEEEECHHHHHHHHHHHCCCHHHHHHHCCCCHHHHHHCCCCCHHHHH

          0        70        80        90        100       110.......
T0379     TELSRYIGKELTYQQVYDALLGFLEEISAEKFDYIDSLRP-DYRLFLLSNTNPYVLDLAM
2b0cA     EALCHEMALPLSYEQFSHGWQAVFVALRPEVIAIMHKLREQGHRVVVLSNTN-----RLH
conserv     *:: :   *:* *::  :  ::  : :* :  :  ** D :*:::*****IIIII :.
conf      888887340004678888887567640077640565400 3 643555007870456450 0
consensus HHHHHHCCCCCHHHHHHHHHHHCCHHHHHHHHHHC CCEEEEEECCCHHHHHHH
psipred   HHHHHHCCCCCHHHHHHHHHHHHCCHHHHHHHHHC CCEEEEEECCCHHHHHHH
SSpro     HHHHHHCCCCCHHHHHHHHHHHHCCHHHHHHHHHC CCCEEEEECCCHHHHHHH
phd       HHHHHHCHHHCHHHHHHHHHHHHHHHHHHCHHHCCC CCEEEECCCCCCHHHHCC
dssp      HHHHHHCCCCCHHHHHHHHCCEEEECHHHHHHHHHHHCCCEEEEEECCC     CCC

          120       130       140       150       160       170.......
T0379     SPRFLPSGRTLDSFFDKVYASCQMGKYKPNEDIFLEMIADSGMKPEETLFIDDGPANVAT
2b0cA     TTFWPEEYPEIRDAADHIYLSQDLGMRKPEARIYQHVLQAEGFSPSDTVFFDDNADNIEG
conserv   :  ::  :  :  ::*::*:*: :*  **   *   ::   *: * :*:*:**:: *: :
conf      000000035524442120000110677887677787777607870005553278000567
consensus CCCHHHHCCCHHHHHHHEEEHHHHCCCCCCCHHHHHHHHHHCCCCHCCEEEECCCCHHHHHH
psipred   HHHHHHHCCCHHHHHHHEEEHHHHCCCCCCHHHHHHHHHHHCCCHHHEEEECCCHHHHHH
SSpro     CCCHHHHHCCHHHHHHHEEEHHHCCCCCCCHHHHHHHHHHCCCCHCCEEEECCCHHHHHH
phd       CCCCCCCCCCHHHHHHHHHHHHCCCCCCCHHHHHHHHHHCCCCCCEEEEECCCCCCHHH
dssp      CCCCHHHCHHHHHHCCEEEEHHHHCCCCCCHHHHHHHHHHHCCCHHHEEEEECCHHHHH

          180       190       200
T0379     AERLGFHTYCPDNGENWIPAITRLLREQK
2b0cA     ANQLGITSILVKDKTTIPDYFAK-----V
conserv   *  **:::  :    :::  : :IIIII
conf      875075005006800036898888874295
consensus HHHCCCEEEECCCHHHHHHHHHHHHCC
psipred   HHHHCCEEEEECCHHHHHHHHHHHHCC
SSpro     HHHCCCEEEECCCHHHHHHHHHHHHCCC
phd       HHHCCCCCECCCCCCCHHHHHHHHHHCC
dssp      HHCCCCEEEECCCCCHHHHHHHC     C


Gaps in the alignment:.
-----------------------
nr of deletions:     1
nr of insertions:    2
```

Figure 2.3: Example of a “model information" output file used for the positioning anchor groups serving as starting points of the loop prediction process.

of confidence ranging from 0 (*i.e.* no reliable assignment of secondary structure possible) to 9 (*i.e.* high confidence).

- "consensus" is the consensus of the three secondary structure predictions shown on the subsequent lines based on majority voting as described above.
- The last data line ("dssp") shows the calculated secondary structure of the model derived from DSSP.

2.3 Loop prediction

As mentioned in the introduction, there are basically two approaches to the loop prediction problem: knowledge-based and *ab initio*. We follow a knowledge-based strategy by scanning a database of fragments (extracted from the PDB) for suitable backbone conformations. A schematic representation of the loop prediction routine is shown in Figure 2.4. A detailed description of all steps is given below.

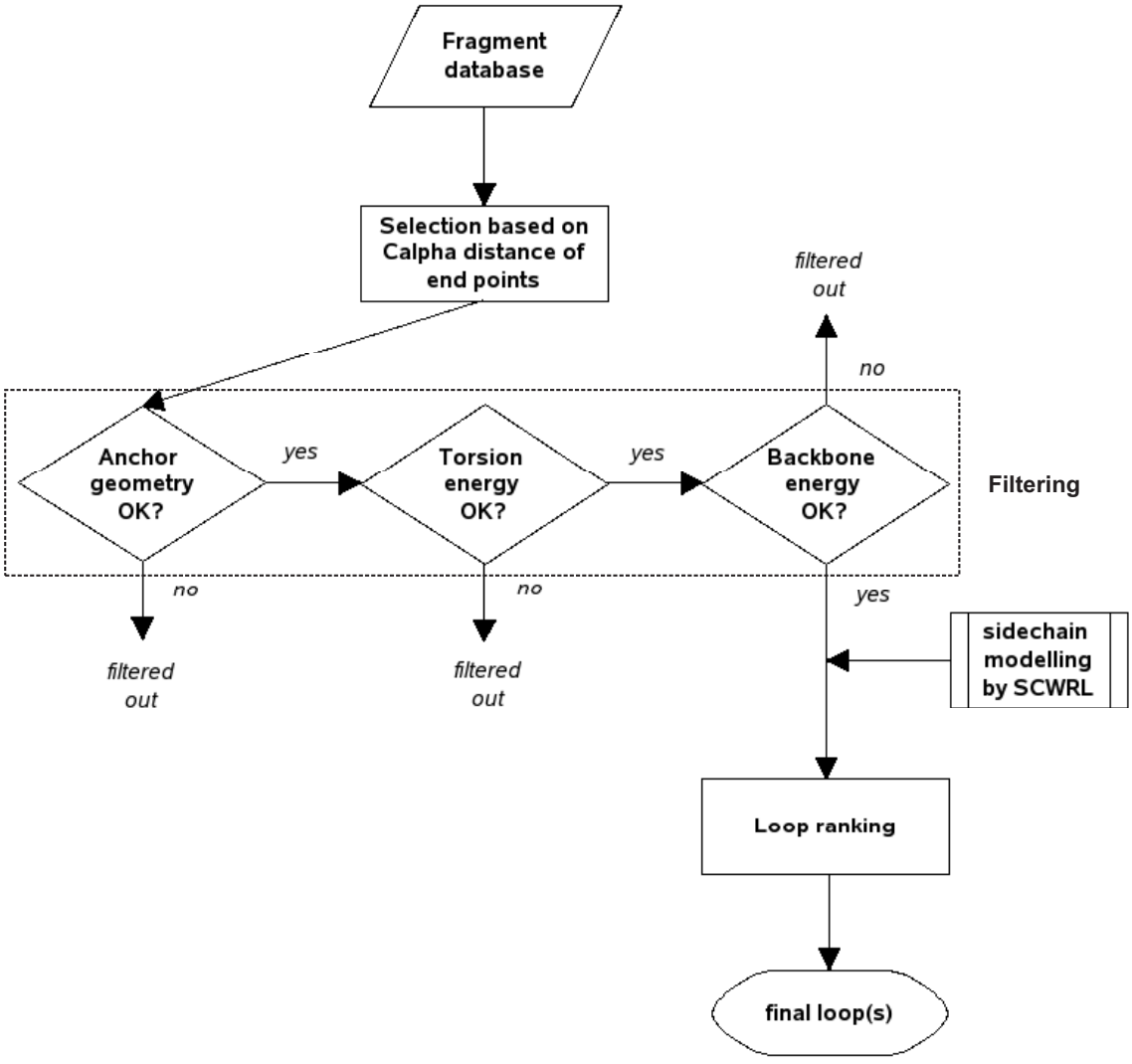


Figure 2.4: Schematic representation of the loop prediction routine.

2.3.1 Fragment database

The fragment database is based on a non-redundant subset of protein structures from the PDB [18]. The selection is generated using the PISCES server [236] which allows to extract sets of protein structures. The following selection criteria are used:

- pairwise sequence identity $< 95\%$
- resolution < 3.0 Å
- R-value < 0.3
- only structures determined by X-ray crystallography

The selection criteria represent a trade-off between quality of the structures and quantity of the fragments in order to increase the coverage of the conformational space. Since only protein backbone coordinates are stored in the database, a resolution cutoff of 3 Å represents a reasonable compromise since at this resolution the backbone is usually well-defined in proteins solved by X-ray crystallography.

The resulting data set contains 12,376 protein chains which are cut into fragments of length 3-20 by the class `Fragmentor` (see section Implementation, page 74). In a first step, the chain is inspected concerning chain breaks and missing residues. Structurally continuous substructures are then defined which are subsequently fragmented using sliding windows of length 3 to 20 residues. Only complete fragments containing all 4 backbone atoms per residue are accepted and stored in a MySQL database. The structure of the fragment database is shown in Table 2.2. Since in the application case only queries on fragments of the same length are performed, specific fragment tables for each length are generated in order to enhance query speed. The fragment tables contain approximately 2.5 to 2.9 million fragments each.

The table `structure` contains information about all protein structures used to generate the fragments (*e.g.* PDB identifier, chain identifier, resolution R-value etc.). The table `fragment2structure` stores begin and end position of the fragment in the corresponding structure (starting from 0) and additionally the two corresponding

Table 2.2: Name and number of entries of the tables in the fragment database. The fragments of the length 3-20 amino acids are stored in separate tables.

Table name	Number of entries
fragment3	2,907,542
fragment4	2,879,976
fragment5	2,853,117
fragment6	2,826,819
fragment7	2,801,064
fragment8	2,775,811
fragment9	2,751,095
fragment10	2,726,790
fragment11	2,702,933
fragment12	2,679,522
fragment13	2,656,505
fragment14	2,633,917
fragment15	2,611,692
fragment16	2,589,817
fragment17	2,568,295
fragment18	2,547,087
fragment19	2,526,182
fragment20	2,505,539
fragment2structure	48,543,703
structure	12,376

primary keys of the tables `fragment` and `structure`. The primary keys of the `fragment` tables are unique over all tables. An alternative, relational database structure has been investigated using an atom, a residue and fragment tables including the corresponding connection tables. But this approach resulted in an explosion of the query time most probably as a consequence of the multitude of joining operations on huge tables. Therefore, the database was denormalised and all necessary data was condensed in one table (the `fragment` tables). An overview on the fields of the `fragment` tables is shown in Table 2.3. The query speed was further increased by sorting the table according to the fragment end-distance since this represents the primary selection criteria used. In addition to the selection by fragment end-distance, an advanced selection using sequence or secondary structure constraints is possible. Therefore an index has been put on these three columns in order to increase the query speed.

Table 2.3: Structure of the table `fragment3` containing fragments of the length of 3 residues.

Field name	Datatype	Description
ID	int(11)	primary key
dist_bin	smallint(5)	fragment end-distance (rounded)
end_distance	float	fragment end-distance
anchor_coordinates	tinytext	backbone coordinates of the anchor residues
loop_coordinates	text	backbone coordinates of all loop residues
torsion_angles	text	torsion angles of all loop residues
sequence	char(3)	sequence of the fragment
SSE_pattern	char(3)	secondary structure of the fragment
chain_end_ID	char(1)	identifier for chain-end fragments: N,C
SSE_N_flank [a]	char(1)	type of the left flanking secondary structure
SSE_C_flank [a]	char(1)	type of the right flanking secondary structure
N_flank_length [a]	int(2)	length of the left secondary structure
C_flank_length [a]	int(2)	length of the right secondary structure
solvation_avg	float	average solvation of the fragment
solvation_pattern	varchar(3)	solvation pattern: 0=buried, 1=exposed
pdb_ID	varchar(4)	PDB identifier of the original structure
chain_ID	char(1)	chain identifier of the original structure

[a] These fields are only used for "real" loops, *i.e.* fragments which only consist of residues with the secondary structure type *coil* and are immediately enclosed by secondary structure elements.

Since the sequence and the secondary structure composition of the fragments are stored in the database as text entries, queries with regular expressions on these fields are possible. This can be especially useful when constraints derived from the analysis of the sequence conservation in the protein family or knowledge about the position of the surrounding secondary structure elements should be used as described in section 2.2.2.

Below, a virtual example of a constraint query on the fragment database is provided:

```
SELECT * FROM fragment10 WHERE (end_distance BETWEEN 10 AND 14) AND
(SSE_pattern LIKE 'HH____CC__') AND (sequence LIKE '__G_______');
```

2.3.2 Loop test sets

A parameterisation test set consisting of 50 loops of length 3-15 residues was used in order to optimise all loop prediction parameters described in the next section. The

same parameterisation as described by Michalsky *et al.* in the LIP program is used [139].

The performance of the loop prediction routine described in this work is compared to 4 commercial loop modelling programs which have been recently benchmarked by Rossi *et al.* [174] with a test set covering loops from 4-12 residues (a filtered test set based on the work of Jacobson and *et al.* [99]). The test set as well as the results of the 4 commercial programs were obtained from the author (Karen Rossi). Additionally, a test set of 14 loops of length 4-9 is used in order to compare the performance to seven other programs. Although being small and probably not representative, this test set is frequently used in the literature and is applied here as well for the sake of completeness. The results of the other loop prediction programs are obtained from two publications [53, 139] and from the LIP website[c].

2.3.3 Selecting, filtering, ranking of fragments

The loop prediction protocol involves basically 3 steps as shown in Figure 2.4: Selection of fragments from the database, filtering in order to reduce the set of candidates and finally ranking of the remaining loops based on a scoring function.

2.3.3.1 Loop selection from the fragment database

In the first step, fragments are selected from the database based on a simple geometric criterion comparing the distance between the terminal Cα atoms of the fragment with the corresponding Cα distance of the anchor groups (*i.e.* the framework in which the fragment is inserted). Upper and lower bounds for the difference between these two distances are defined for each loop length. The bounds have been manually adjusted so that less than 5 percent of the all Top10 fragments per loop length are rejected in the parametrisation set. The thresholds for different loop lengths are summarised in Table 2.4. Adjusting the bounds represents another trade-off between speed and accuracy. Retrieving more fragments by less restrictive cutoffs slows down the whole loop prediction process since more data (especially the coordinates) have to

[c] `http://www.drug-redesign.de/LIP/LIP_WebseiteTestsets.html`

be transferred from the database and processed by the filters and the scoring function described below. On the other hand, the presence of more candidates makes the task of identifying the best fragment among others more difficult. All selected fragments are subsequently fitted on the the anchor groups by least squares fitting over the coordinates of the 4 backbone atoms N, Cα, C and O of both end points.

2.3.3.2 Loop filtering steps

In the next step, four quality filters are applied in order to remove unsuitable fragments, thereby reducing the candidate set for the final ranking step. The first filter analyses the "goodness of fit" *i.e.* how well the backbone of the terminal fragment residues matches the anchor backbone geometry provided by the protein framework. The root mean square deviation between anchor residues and terminal fragment residues is calculated (called RMSa). Fragments with a RMSa above a loop length dependent threshold are rejected. In analogy to the strategy used in the selection process, the cutoff values for the RMSa filter were set such that not more than 5 percent of all Top10 fragments are filtered out (Table 2.4).

The second filter rejects fragments having serious clashes with the environment after fitting into the framework. Two atoms are defined as clashing if the distance between them is less than 70% of the sum of their van der Waals radii. The van der Waals radii have been taken from a work of Li and Nussinov [124]. A similar threshold has

Table 2.4: Threshold used in loop selection and for the anchor group RMSD filter.

	difference between Cα-distances[a]		
loop length	**lower bound**	**upper bound**	**RMSa cutoff**[b]
L $\leq$6	-1.15	0.85	1
6 < L $\leq$8	-1.5	1.75	1.35
8 < L $\leq$12	-2.25	2.5	1.5
L > 12	-2.75	2.75	1.75

[a] Cα-distance of the fragment end points compared to the Cα-distance of the anchor groups.
[b] RMSD between the terminal fragment residues and the anchor group residues after fitting.

been used in a recent publication on loop prediction [70]. The fitting process based on least squares fitting results in only an approximately correct orientation of the fragment in the protein framework and therefore loops with accurate local geometry compared to the native loops can still have considerable clashes. In an earlier work from our lab, Heuser *et al.* [90] approached the problem by accepting one clash with the environment. Furthermore, "soft" clashes can be expected to be removed in a subsequent energy minimisation step.

The two initial filtering steps (*i.e.*, RMSa filter and clash filter) are performed during the retrieval process of the fragments from the MySQL database and, depending on the modelling situation, the filters remove a large fraction of the selected fragments. This approach allows to restrict the number of candidate fragments to be stored simultaneously and therefore reduces the main memory consumption. The remaining loop objects (see section Implementation, page 74) are stored in a vector for further processing.

The third filter analyses the torsion energy of the remaining fragments. As described in the Introduction, the 20 amino acids show, as a consequence of the steric restrictions imposed by their side chains, preferences for certain torsion angles. The fragments of the database originate from structures having completely different amino acid compositions and therefore analysing the torsion energy can be used to estimate how well the given loop sequence matches the dihedral angles of the fragment. The torsion angle potential is especially valuable for filtering since it relies only on the backbone atoms and does not need the sidechains which have not been modelled yet. Z-scores of the torsion energies of all fragments are calculated by subtracting the mean and dividing by the standard deviation of the whole set. Loops with torsion energy Z-scores above 1 standard deviation are removed. If a maximum number of 20,000 fragments is exceeded after the first round, the threshold is gradually lowered with a step size of 0.2 standard deviations.

In the last filtering step the compatibility of the loop backbone with its framework is investigated before the actual scoring is performed. This step was necessary since side chain modelling is the rate limiting process in the whole modelling pipeline typically taking a fraction of a second (maximum 1 second) per loop. Sidechain modelling is performed by SCWRL [31] but since an external program is used, the protein structure

including the loop has to be temporary saved, the program executed and the output has to be reloaded.A combination of the following 3 terms is used in the backbone scoring step:

- A pairwise distance-dependent statistical potential based on Cα atoms in order to analyse the interactions of the loop with its environment.
- A solvation potential based on Cα atoms investigating the propensity of the loop residues for the given degree of solvent exposure.
- The "goodness of fit" of the terminal loop residues to the anchor groups as expressed by the RMSa.

The theoretical background of statistical potentials of mean force and how they are extracted is described in detail in the next section. For all 3 terms, Z-scores are calculated and the Z-scores for each loop are simply summed up. An inspection of the distribution of the scores revealed that the values are at least approximately normal distributed which is a prerequisite for the derivation of Z-scores. The use of Z-scores enables the combination of statistical potential terms with the RMSa distance measure. Such a combination would be difficult if the raw energies are used directly since, depending on the structural environment which determines the number of contacts between loop and framework, the amplitude of the energies potentially differs significantly between different modelling situations, which complicates the combination with the distance measure. Z-scores reflect how well a certain fragment fits in the given environment (sterically and energetically) compared to all other fragments in the set. A good, near-native fragment should have reasonable scores for all three terms. Based on the combined backbone score, the best 3,000 loops are retained. The number of loops passing the torsion energy filter (20,000) and the backbone filter have been optimised based on the parametrisation set.

2.3.3.3 Loop scoring

In the next step, sidechains are added to the loop residues by executing SCWRL. Since the loop is now complete in terms of its atomic composition, a more fine-grained, all-atom energy function can be applied in order to rank the remaining fragments. A

variety of different terms and parameters for the statistical potential terms has been investigated. The performance of some selected combinations are shown in Results and Discussion, page 135ff. In the final scoring function only the all-atom interaction potential has been used. Among other alternative implementations, a combined scoring function consisting of 4 terms has been investigated using a torsion angle potential over 3 residues, an all-atom solvation potential, an all-atom pairwise interaction potential as well as the RMSa. All terms are combined by summing of the individual Z-scores.

The torsion angle potential reflects the propensity of the loop sequence to adopt the local geometry described by the torsion angles of the fragment. The same bin sizes for the Φ and Ψ angles have been applied as for model quality assessment (see section 2.4.1.5). The short-range pairwise interaction potential assesses the direct interactions with the structural environment. The upper limit of 10 Å has been set manually after inspection of the interaction curves. At an atomic distance of approximately 10 Å the energy curves reach a pseudo energy of zero. The solvation potential describes the propensity of a certain atom for the observed degree of solvent exposure as approximated by the number of atoms within a sphere of 6 Å around the central atom. A threshold of 6 Å has been chosen in order to assure that no water molecule fits between the two atoms. The solvation potential to some extend favours loops forming contacts with the protein surface instead of pointing into the solvent. A variety of additional functionalities are provieded by the loop prediction routine which are briefly described here:

- A clustering library implemented in C by Michiel de Hoon (originally developed for the analysis of gene expression data) is integrated [51]. In order to remove redundancies, the set of loops can be clustered based on a given RMSD value cutoff using various clustering strategies (*e.g.* single-linkage, complete-linkage *(default)*, centroid-linkage and average-linkage clustering).

- The colony energy approach as introduced by Xiang *et al.* [241] has been implemented. In this approach, the energy of a loop decreases with the presence of other loops with similar conformation and low energy assuming that the conformational space around global energy minimum is more populated than the rest of the energy landscape.

- Four different loop fitting strategies have been implemented. Fitting on the 4 backbone atoms on both sides *(default)*, fitting on backbone without the oxygen atom (since it is defined by the other 3 atoms), fitting on the backbone of two consecutive residues on both sides, fitting on 3 consecutive Cα-atoms on both sides.

- Both loops and chain ends can be modelled (in the later case only one anchor group given).

- After building the all-atom loop model, the sidechains of the loop together with the sidechains of surrounding residues within a given distance cutoff can be rebuilt.

- A user-defined number of protein structures from the top ranking loop predictions can be saved as PDB files.

- A variety of rankings and quality measures are calculated for benchmarking purposes.

2.4 Model quality assessment

In protein structure prediction, a considerable number of alternative models are usually produced from which subsequently the final model has to be selected. Thus, a scoring function for the identification of the best model within an ensemble of alternative models is a key component of most protein structure prediction. Model quality assessment includes the global assessment of the quality of the entire model but also the local quality assessment analysing the reliability of different regions of a specific model. This section will focus on the first task but in the last section an extension for local quality assessment is described. QMEAN [16], which stands for Qualitative Model Energy ANalysis, is a composite scoring function describing the major geometrical aspects of protein structures. Five different structural descriptors are used. The local geometry is analysed by a new kind of torsion angle potential over 3 consecutive amino acids. A secondary structure-specific distance-dependent pairwise residue-level potential is used to assess long-range interactions. A solvation potential describes the burial status of the residues. Two simple terms describing the agreement of predicted and calculated secondary structure and solvent accessibility, respectively, are also included.

A variety of different implementations are investigated and several approaches to combine and optimise them are described. Only the parameters used in the final implementation of the statistical potentials are shown here together with the description of the optimisation strategy. The rest of the data can be found in the Results section, page 108ff. QMEAN was tested on several data sets as described below and was compared to five well-established model quality assessment programs.

2.4.1 Statistical potentials

2.4.1.1 Theoretical background

The analysis of experimentally solved protein strutures reveals obvious regularities such as the tendency of hydrophobic residues to be buried, the pairing of oppositely charged atoms or the interaction of aromatic rings [87]. Statistics about these empirical or

knowledge-based parameters can help understanding the interactions which contribute to the stability of protein structures and their analysis has a long history going back to the work of Tanaka and Scheraga in 1976 [209].

In the early 1990's Sippl introduced a statistical mechanics formalism based on the inverse Boltzmann principle in order to derive a potential of mean force [197, 198, 200]. The Boltzmann principle relates the energy of a conformational state c_i to its probability of occurrence at the thermodynamic equilibrium:

$$p(c_i) = \frac{e^{\frac{-E(c_i)}{kT}}}{\sum_j e^{\frac{-E(c_j)}{kT}}} \tag{2.2}$$

where k is the Boltzmann's constant and T is the absolute temperature. The summation j over all allowed states of the system is called the partition function or Boltzmann sum (denoted as $Z(C)$). In analogy, the inverse Boltzmann principle relates the probability density function $p(c_i)$ to the energy of a given state:

$$E(c_i) = -kTln(p(c_i)) + kTln(Z(C)) \tag{2.3}$$

In a similar way, the *net potential of mean force* [198] can be derived for a specific subsystem (*i.e.* specific interaction) s_k by subtracting the mean force of reference thereby removing all energies which are common to all subsystems. This can be described as conditional probabilities [205] reflecting the probability of a conformational state c_i in the presence of a specific interaction s_k:

$$\Delta E(c_i|s_k) = E(c_i|s_k) - E(c_i) = -kTln(\frac{p(c_i|s_k)}{p(c_i)}) + kTln(\frac{Z(C|S)}{Z(C)}) \tag{2.4}$$

For example in a distance-dependent pairwise potential c_i refers to the distance and s_k to the identities of the two atoms. In torsion potentials c_i stands for a given pair of Φ/Ψ dihedral angles and s_k for the amino acid type. According to Sippl [198], $Z(C|S) = Z(C)$ can be assumed which results in the following equation:

$$\Delta E(c_i|s_k) = -kTln(\frac{p(c_i|s_k)}{p(c_i)}) \tag{2.5}$$

The numerator is the observed probability of a specific interaction whereas the denominator reflects the expected probability if there where no interactions (*i.e.*, the reference state). The observed probabilities can be directely estimated based on statistics on a representative set of protein structures from the PDB [18]. Different approaches have been described for the estimation of the reference distribution [184, 189, 198, 249]. The majority of statistical potentials relies on the "uniform density" reference state used by Sippl [197] in which it is assumed that the distribution in the reference state is the same as in folded proteins. Therefore the probability distribution of the reference state is an average over all amino acids in the dataset. This distribution can be directly obtained from database statistics as well. An alternative implementation of the reference state has been used by Zhou and Zhou in the DFIRE potential [249]. In their work the reference state is approximated by using uniformly distributed non-interacting points in finite spheres. For the potentials of mean force described in this work, the reference state as proposed by Sippl is used and all potentials are derivations from the following general form:

$$\Delta E(c_i|s_k) = -kTln\left(\frac{\frac{f(c_i|s_k)}{f(s_k)}}{\sum_k \frac{f(c_i|s_k)}{f(s_k)}}\right) \tag{2.6}$$

Typical features investigated by statistical potentials are backbone torsion angles, solvent accessibility and pairwise interactions between non-bonded atoms. As done in this work, different statistical potential terms cen be combined to a single scoring function (see Introduction page 30).

The physical basis of statistical potentials has been questioned [75, 147, 173, 211]. The Boltzmann equation describes a particular system in its thermodynamic equilibrium, whereas statistical potentials assume the system to be a database of protein structures in the free energy minimum. According to this assumption, structural elements such as pairwise distances or torsion angles obey a Boltzmann-type distribution based on a hypothetical reaction at equilibrium in which a unique structure consisting of averaged amino acids "mutates" to a unique sequence [75, 194].

The pseudo energy of the entire protein is calculated by summing up the energies of the individual amino acids. In both cases (summing up different energy terms

and summing up residue energies) thermodynamic additivity is assumed, *i.e.* the components contribute independently to the total energy. This is a fundamental principle used in all energy functions both knowledge-based and physics-based but it only represents a simplification (probably as a consequence of missing alternatives). A critical discussion of the additivity principles in biochemistry can be found in a good review of Dill from 1997 [58].

The non-redundant set of protein structures used to derive the potentials is described in the next section. The different statistical potentials (*i.e.* distance-dependent pairwise potential, torsion angle potential and solvation potential) are introduced in the subsequent sections.

2.4.1.2 Extraction of the statistical potentials

All statistical potentials were extracted from a non-redundant set of high-resolution protein structures from the December 2006 version of the PDB [18]. The PISCES server [236] was used in order to select a subset of the experimentally solved protein structures. The following selection criteria were used:

- pairwise sequence identity $< 30\%$
- resolution < 1.8 Å
- R-value < 0.2
- only structures determined by X-RAY crystallography

This resulted in an initial selection of 1,801 protein chains. To reduce over-training of the potentials for structures subsequently used for training and testing, all target sequences of CASP6 and CASP7 were blasted against the 1,801 chains. All detectable hits were removed resulting in 1,688 structures. The following three filters were applied in order to further increase the quality of the set of protein structures used for the subsequent statistical analysis:

- Proteins having less than 90% of the amino acids resolved in structure (with respect to the sequence) were not included (171 chains removed).

- Structures with a substantial part being flexible (*i.e.* more than 20% of the residues having an residue-averaged B-factor above two standard deviations) were removed (25 chains).

- Structures with missing backbone atoms (21 chains removed).

For each of the remaining 1,471 structures, DSSP [107] was executed in order to assign secondary structure, solvent accessibility and the torsion angles.

2.4.1.3 Distance-dependent pairwise potential

The distance-dependent contact frequencies were extracted from the protein data set described above. The radial distribution of atoms around the central atom is investigated as schematically represented in Figure 2.5. In order to reduce the bias introduced by sequentially local interactions (the contacting atoms are assumed to be free particles), only atom pairs separated by at least 4 residues were included. Alternatively, a sequential separation cutoff of 7 and an implementation without any cutoff has been investigated but resulted in worse performance *(data not shown)*.

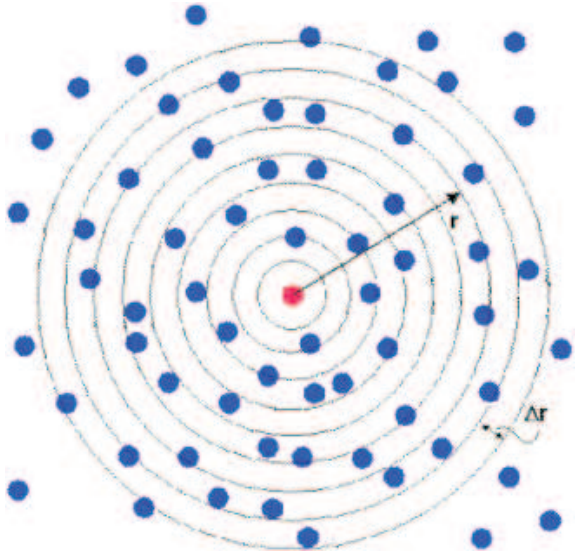


Figure 2.5: Radial distribution of atoms investigated for the derivation of the distance-dependent interaction potential.

Cα and Cβ atoms, respectively, have been investigated as possible interaction centers. Additionally, an all-atom version using all 167 atom types occurring in proteins was implemented and is used in loop ranking. In the secondary structure specific implementation of the residue-level pairwise potential the potentials are calculated based on frequency counts extracted from residues of the same secondary structure state while ignoring the secondary structure state of the contacting residues. A distance range of 3 to 25 Å (Δr = 0.5 Å) turned out to produce the best results. The final potential integrated in QMEAN is based on Cβ atoms and uses the secondary structure specific implementation. The calculation of the residue-level pairwise potentials has been carried out as described by Sippl (see Chapter 2.4.1.1).

2.4.1.4 Solvation potential

The degree of residue burial was approximated by counting the number of interaction centers (Cβ atoms for QMEAN) within a sphere of 9 Å around the given amino acid in a similar way as described by Jones [102] and in FRST [215]. The cutoff of 9 Å used in this work resulted in a slightly better performance of the potential than other cutoffs tested (see Results and Discussion, page 3.2.1ff). The relative accessibility was then calculated by dividing the counts by the maximum number of counts observed for the given amino acid type in the protein data set. The solvation potential reflects the propensity of a certain residue for a given solvent accessibility compared to any other residue. The potential has been implemented as described in section 2.4.1.1.

2.4.1.5 Torsion angle potential

The single residue torsion angle potential reflects the propensity of a certain residue for a given torsion compared to any other residue. The torsion angles were discretised in 10 degree bins. The 3-residue torsion angle potential described here is a further development of the single residue torsion angle potential by others [3, 111, 193, 215]. The description of the local geometry for a certain residue was extended by including the torsion of the adjacent residues. The coarseness was increased by using 45 degree bins for the center residue and a bin size of 90 degree for the dihedral angles of the neighboring residues. Several alternative bin sizes have been investigated ranging from

30 degrees to 90 degrees (see Results and Discussion, page 3.2.1ff). The identity of the neighbours was not taken into account.

2.4.1.6 Agreement terms

A term describing the agreement between the predicted secondary structure of the target sequence and the observed secondary structure of the model as calculated by DSSP was built. The DSSP output was converted into the 3-state format (helix, sheet, coil) as used in EVA [67] an automatic evaluation pipeline for protein structure prediction. A consensus secondary structure prediction approach was investigated in the attempt to increase prediction accuracy. A consensus between PSIPRED [103], SSpro [35] and ProfSec [177] was built based on simple majority voting [4]. The fraction of residues with identical predicted and observed secondary structure states was used as a simple quality measure. In the final implementation of QMEAN, only PSIPRED was used since the consensus of the methods currently included did not lead to an improved performance. A similar measure describing the agreement between the predicted binary burial status (buried/exposed) as provided by ACCpro [35] and observed solvent accessibility based on DSSP was implemented. The relative solvent accessibility was calculated by dividing the solvent accessibility extracted from DSSP by the maximum solvent accessibility for the given amino acid type observed in the training set. Afterwards, the relative solvent accessibility was transformed into the binary classification based on a cutoff of 25%. No consensus scheme was tested in this case.

2.4.2 Measures for the structural similarity between model and target

The traditional measure of expressing the similarity of two protein structures is the RMSD (Root Mean Square Deviation), calculated after a rigid-body superposition:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2} \tag{2.7}$$

where δ is the distance between two corresponding atoms among N pairs of equivalent atoms (usually either Cα atoms, backbone atoms or all atoms).

In order to evaluate the quality of the models in the two CASP test sets described below the GDT_TS score was used as an objective measure for the structural similarity between model and target. The GDT_TS score was calculated using the TMscore software from Zhang and Skolnick [247]. GDT_TS is a well-established score used in the evaluation process of the last CASP rounds having the advantage of being less sensitive to local errors in models as compared to the traditional RMSD. GDT (Global Distance Test) describes the maximum percentage of residues which can be structurally aligned within a defined distance cutoff. In GDT_TS 4 increasing distance cutoffs are used (x = 1, 2, 4 and 8 Å) and the average of the percentage aligned residues px is calculated:

$$GDT\_TS = \frac{p_1 + p_2 + p_4 + p_8}{4} \tag{2.8}$$

For the decoy sets from the Decoys 'R' us website *(see below)*, the RMSD values as provided in the sets have been used directly.

2.4.3 Data sets

In this section, the data sets used for training (*i.e.* optimising parameters and weighting factors) and testing (*i.e.* comparison with other methods) are described.

2.4.3.1 CASP6 decoy set for training

Parameter optimisation as well as the evaluation of weighting factors for the combined energy function was performed on the CASP6 set (a description of the CASP experiment is given in the Introduction, page 19). This set consists of all the models submitted to the 64 accepted targets of CASP6. In order to increase the quality of the data set and to reduce the influence of random predictions or very difficult targets, all models having a GDT_TS score of less than 0.2 were removed for training (11,475 models). The final data set consists of 15,893 models.

2.4.3.2 Standard decoy sets from Decoys 'R' us

The ability of a scoring function to identify the native structure among various decoy structures was investigated and compared to other state-of-the-art tools with the help of the following three frequently used decoy sets from the Decoys 'R' us website[d] [182]: *4state_reduced* [158], *lattice_ssfit* [240] and *LMDS* [109] (a short description of the decoy sets can be found in Wallner *et al.* [233]). The performance of the other methods on these decoy sets has not been recalculated here, but the corresponding data were taken directly from a recent publication [215]. The two quality measures *Znat* and *rank1* used in the results section describe the Z-score of the native structure compared to the ensemble of decoys and the number of cases in which the native structure was ranked first in a given decoy set, respectively.

2.4.3.3 Molecular dynamics decoy set

The decoy set generated by Fogolari and co-workers [81] was used to estimate the performance on near-native structures. It consists of over 6,000 snapshots from five independent molecular dynamics simulations. One simulation started from the native structure and the other four from minimised conformations of the thermo-stable subdomain from the chicken villin headpiece consisting of 36 residues (PDB identifier 1vii). The decoy set can also be downloaded from the Decoys 'R' us website and covers RMSD values from 2 to 12 Å. In contrast to the three test sets described above, this set contains several near-native conformations.

2.4.3.4 CASP7 decoy set: testing model quality assessment

The CASP7 server models for all 95 accepted targets were downloaded from the CASP website[e]. This is the same data basis used in the blind test for model quality assessment programs which was part of CASP7. Although all quality predictions submitted for the quality assessment category of other groups were available on the CASP website, this data were not used here. Rather, predictions were recalculated with

[d] `http://dd.compbio.washington.edu/`
[e] `http://predictioncenter.org/casp7/`

some well-established model quality assessment programs (MQAPs) downloadable from the CAFASP4 website[f]. This has the following reasons. First, many of the MQAPs joining CASP7 have not been published yet and from the abstracts submitted it was mostly impossible to understand how they work. Second, the top performing MQAPs all integrated consensus information in their calculation, which is not in the scope of this work. In consensus methods the quality of a certain model is assessed by taking into account information contained in the ensemble of models. These methods are unable to assess the quality of a single model (as the methods described here). Third, the data is sometimes difficult to compare. Some MQAPs fail to predict the model quality for many servers or have not submitted any predictions for some targets.

The following model quality assessment programs were used: FRST [215], Modcheck [162], ProQ [233], DFIRE [249] and RAPDF [184]. Only server models for which all of the five MQAPs were able to return a prediction were evaluated resulting in a total number of 22,420 models over all 95 targets. ProQ has been executed in two different modes either using secondary structure information (provided as a PSIPRED prediction) or not.

The 95 targets were divided into the two categories free-modelling (FM) and template-based modelling (TBM) as introduced in the seventh round of CASP[g]. Since several targets are multi-domain structures and the domains can sometimes be assigned to different categories, multi-domain targets were assigned to the category of the most difficult domain they include (*i.e.* a target consisting of a FM domain and a TBM domain was assigned to the FM category). The final division is shown in Table 5.1 in the Appendix.

2.4.4 Evaluation criteria

A variety of quality measures have been used in order to compare the performance of the different methods. $logP_{B1}$ and $logP_{B10}$ are the log probabilities of selecting the highest GDT_TS model as the best model or among the ten best-scoring models, respectively. Suppose the best scoring conformation xi has the GDT_TS rank of Ri

[f] `http://www.cs.bgu.ac.il/~dfischer/CAFASP4/`
[g] `http://predictioncenter.org/casp7/meeting_docs/difficulty.html`

in n decoy conformations, then the log probability is given by:

$$logP_{B1} = log_{10}(\frac{R_i}{n}) \quad for\, logP_{B10}: \;\; R_i = min[R_1, \ldots, R_{10}] \tag{2.9}$$

Fraction enrichment (*F.E.*) is the percentage of top 10% lowest RMSD conformations or highest GDT_TS models among the top 10% best-scoring structures. In the fraction enrichment curves variable cutoffs are used ranging from 5% to 50%. The enrichment as defined in Tsai *et al.* ($E_{15\%}$) is calculated by dividing the number of top 15% highest GDT_TS models found among the top 15% best predicted models by the number obtained in a random selection (15% * 15% * number of structures in the decoy set). Z_{nat} is the Z-score of the native structure as compared to the ensemble of models. $rank1$ and $rank10$ are the number of targets in which the native structure (or the best model based on GDT_TS, excluding the native structure) was found on the first rank or among the Top10 predictions, respectively. *GDT_TS loss* is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set. Two kinds of regression coefficients have been used: Pearson's correlation coefficient r^2 and Spearman's rank correlation coefficient rho.

Parameter optimisation for the statistical potentials (such as distance range, bin size, resolution and interaction center) was performed on the CASP6 set. In order to measure the ability of the statistical potential to predict the model quality, the Pearson correlation coefficient between the predicted model energy and the measured quality in terms of GDT_TS was used. A variety of alternative implementations of the statistical potentials were investigated and the best performing torsion angle potential, solvation potential and pairwise potential are selected based on the correlation coefficients.

The weighting factors for the combined scoring function are evaluated by an exhaustive search strategy over reasonable ranges for the different weighting factors. The final combination is selected based on the maximum correlation coefficient. Several alternative optimisation strategies were investigated. Pearson's correlation coefficient vs Spearman's rank correlation, energy vs Z-scores compared to sequence-shuffled models. Parameters were optimised on a target-specific basis (*i.e.* regressions for all models of each target separately) or on a global basis by maximising the regression over all models from all targets simultaneously.

The target-specific optimisation was accomplished by averaging the Pearson's correlation coefficient over all targets provided that at least a suitable fraction (*i.e.* 150 models which is around 30%) have a GDT_TS higher than 0.2. In this way, 12 of the 64 accepted targets of CASP6 set were excluded from the target-specific evaluation. All but one belong to the novel fold or fold recognition category. The following targets were excluded in the target-specific optimisation process (in brackets the number of models with GDT_TS $>$ 20): T0202 (118), T0206 (94), T0228 (23), T0238 (129), T0242 (139), T0248 (5), T0262 (70), T0272 (4), T0273 (88), T0197 (51), T0198 (104), T0199 (12). This approach was used with the intent to reduce the influence of very difficult free modelling targets in which most of the groups failed to build a reasonable model. These targets are expected to add no value in the optimisation process. In contrast to the Pearson correlation, the Spearman rank correlation allows to investigate a relationship which does not have to be necessarily linear. As described in Pettitt *et al.* [162] Z-scores were built comparing the score of the model with the scores of models after sequence shuffling (1000 times in this work).

2.4.4.1 Statistical significance

In the target-specific assessment, the performance of the methods is evaluated by averaging the results over all targets using a variety of evaluation criteria. The difference in the performance of two methods on the individual targets is investigated using Student's t-test on paired samples. For the quality measures used in this work, a Shapiro-Wilk test (using the Gnu R package) was used in order to analyse whether the scores are approximately normally distributed which is a prerequisite for the t-test. For five of the quality measures (Pearson correlation coefficient, Spearman rank correlation coefficient, the two enrichment measures and *Znat*) the analysis confirmed that the vast majority of the data sets can be regarded as normally distributed (p-value $>$ 0.05). In a related work [131], which was part of the assessment of model quality in CASP4, it has been shown that Student's t-test and the Wilcox signed rank test (which does not rely on a normal distribution of the data) reached the same conclusions.

In Student's t-test, the two-sided upper and lower confidence limits are given by the following equiation:

$$\mu_{u,l} = D \pm \frac{t_{crit}(df, c)S}{\sqrt{n}} \tag{2.10}$$

where D is the average performance difference of the two methods on the targets investigated and S the standard deviation. $t_{crit}(n-1, c)$ is the critical value from the t-distribution, df is the degrees of freedom which is equal to the number of targets minus 1 and c is the confidence level which is 1 minus the significance α. A confidence level of 95% was used in the two-tailed t-test. A schematic representation is given in Figure 2.6.

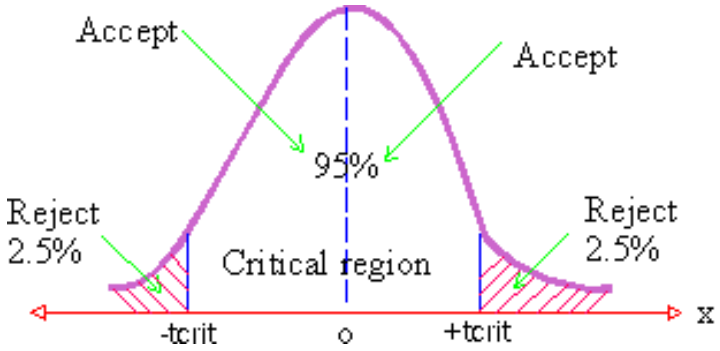


Figure 2.6: Two-tailed t-test on the 95% confidence level.

The null hypothesis states that the two methods perform equally good on the set of targets based on the given evaluation criteria. This hypothesis is rejected according to the Student's t-test if either the upper confidence limit μ_u is below zero or the or the lower confidence limit μ_l is positive. In this case one method performs significantly better than the other.

2.4.5 Local model quality assessment

In comparison to the approach used to analyse the quality of entire models, the scoring function for loop ranking and for local model quality assessment has been especially adapted by using a more short-range implementation of the interaction potential and by using all-atom instead of residue-level solvation and interaction potentials in order to capture more details.

Table 2.5: Differences in the implementation of the local and global energy function.

scoring function term	parameter	local	global
interaction potential	range	2-10 Å	3-25 Å
	bin size	0.5 Å	1 Å
	number of atom types	167 (all-atom)	1 (Cβ)
solvation potential	radius of sphere	6 Å	9 Å
	number of atom types	167 (all-atom)	1 (Cβ)
torsion angle potential	# of residues	3	3

As it can be seen from Table 2.4.5, only contacts within 10 Å are captured in order to assess to interactions with the structural environment. For the task of assessing the quality of entire models (see Chapter **??**) best results are obtained if "interactions" between the Cβ atoms separated up to 25 Å are taken into account. In analogy, a more short-range and fine-grained implementation (compared to the model quality assessment case) has been used for the solvation potential.

The difference in the implementation of the global and the local scoring function can be attributed to the difference of the problems they investigate. In model quality assessment sometimes very rough models are investigated (*e.g.* models from *ab initio* structure prediction or fold recognition) and therefore a coarse-grained implementation (*i.e.* a bin size of 1 Å and a residue-level interaction potentials) seems to be more appropriate. Since only Cβ atoms are used, longer atomic distances have to be considered in order to capture all direct interactions (*e.g.* of two long sidechains pointing toward each other). Furthermore, a global scoring function attempts to assess the fitness of every residue in the sequence to the fold provided by the model. Therefore a pairwise long-range statistical potential should describe not only direct interactions to surrounding atoms but to some extent also "mediated interaction" to atoms being further away in space. In other word, typical distances between pairs of atoms observed in frequently occurring, structurally conserved folds or supersecondary structure elements are likely to influence the energy function and this signal seems to be useful for assessing the quality of models. On the other hand, local energy functions, should only take into account close, direct contacts and therefore the potentials were restricted on short-range interactions.

For the prediction the local model quality, the energy of each residue is calculated using the three statistical potentials described above. In order to smooth the energy profile not only the central residue but also neighbouring residues in a sliding window are taken into account. Different window sizes have been investigated ranging from 1 (*i.e.* only the central amino acid) to 11 (*i.e* five residues on both sides). For the anchor group prediction task in which it is tried to identify the region where the target structure begins to differ from the structure of the template, also asymmetric sliding windows have been investigated. (*E.g.* for the identification of the N-terminal anchor groups, the sliding window covers the central residue and some residues in N-terminal direction (away from the location of the gap). If the sliding window contains structurally undefined positions, the following workaround is used. For gaps (*i.e.* insertions) the average energy of the preceding and the following residue is used and at the chain end the energy of the last residue is taken.

A simple strategy was used in order to combine the three statistical potential terms in a final score. For each of the three terms, the local energies are normalised by calculating Z-scores over the entire model. A combined local score is then built by summing up the three Z-scores for each position in the model. The Z-scores are built in order to cope with the different magnitudes of the three terms and to allow a combination with other features such as sequence conservation, secondary structure content, hydrophobicity etc. It should be mentioned here that this approach only represents a first approximation and that more advanced strategies (*e.g.* machine learning algorithms) should be used in order to optimise the combination of the terms. A comprehensive test set should be used for the evaluation which was not in the scope of this work. The aim was to investigate whether the statistical potentials developed for the quality assessment of entire models and for loops ranking can be used for the analysis of the local model accuracy.

2.4.6 Analysis of gaps and the location of anchor groups

A non-redundant set of homologous pairs of proteins from the HOMSTRAD database [142] is used for the analysis of the distribution of gap lengths (*i.e.* the size of insertions and deletions) occuring in typical modelling situations. A filtered test set of

insertions and deletions (see below) has been built in order to investigate the structural environment on both sides of the gaps for the location of suitable anchor groups and several approaches for the prediction of anchor groups based on the analysis of the local model energy are described.

2.4.6.1 HOMSTRAD test set

HOMSTRAD (HOMologous STRucture Alignment Database) [142] is a curated database storing structural alignments of members of the same homologous protein family. The version from May 2007 containing 1032 protein families was used in order to generate a non-redundant set of pairs of homologous proteins representing realistic modelling situations (*i.e.* target-template pairs with a maximum sequence identity of 40%). A similar procedure has been used in our lab in the past in order to build a test set for anchor group evaluation [121, 238]. Beside other reasons (*e.g.* high sequence cutoff of 50%, presence of very fragmented alignments, no information about resolution of the proteins and chain identifier), this test set was not used here because it is based on the PDB release 8/96. Since then, the size of the PDB has grown by roughly a factor 10 whereas the number of different SCOP superfamilies [148] increased by about a factor 4 (information from the PDB website[h]). The following quality filters were applied in order to build the test set:

- Only families containing exactly 2 members are used (alignments of families with more members are based on multiple structural alignments, which often differ from the pairwise ones).
- A maximum pairwise sequence identity of 40% is used, representing a realistic modelling situation.
- Both sequences need to be at least 80 residues long.
- Both structures need to be resolved by a resolution < 3.0 Å.
- Only structures determined by X-ray crystallography are used.

[h] `http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html&`

This resulted in a final non-redundant set of 257 homologous pair of proteins superimposed on each other. Based on the structure-based sequence alignment all gaps (*i.e.* insertions and deletions) are identified. In order to build a realistic test set (called "anchor group test set" in the following) for the analysis of the structural consequences of insertions and deletion as well as for the analysis of the location of suitable anchor groups, the following rules are applied:

- Gaps close to the chain ends (15 residues apart) are not used since in this case one of the anchors is missing (*i.e.* can most probably not be placed in a structurally conserved region.
- In order to investigate the structural effect of a single gap, no further gap within 10 residues along the sequence is allowed. In the modelling case, two close (separated by a few residues) gaps would be merged to a single (longer) gap.
- Gaps within secondary structure elements are not considered.
- Only gaps in loop regions having secondary structure elements within 10 residues on both sides are taken into account. This reflects a typical loop modelling situation. Usually predictors place the anchor groups close to the ends of the secondary structure elements. The following definition for secondary structure elements is used for the anchor group test set: helix consist of at least 2 residues in helix conformation (according to DSSP) and strands need to have a minimal length of 3 residues.
- The region should be identified where target and template structure begin to differ. Therefore, at least three consecutive residue pairs with backbone RMSD below 1.8 Å need to be present on both sides (in analogy to Lessel und Schomburg [121]).
- Only gaps smaller than 5 residues are included in the final test set. Approximately three-quarter of the insertions and deletions occuring in typical modelling sutuations are below 5 residues (see Results and discussion on page 156).

The final anchor group test set contains 105 insertions and 124 deletions.

2.4.6.2 Anchor group prediction

Based on the anchor group test set described above, the regions on both sides of the gaps (*i.e.* 10 residues in N- and C-terminal direction) are analysed concerning the location of suitable anchor groups. The following set of simple rules has been used for the prediction of the anchor groups and the RMSD between target and template at the given positions as well as the resulting gap length are derived:

- fix distance (1-4 residues) from gap on both sides.
- fix depth in the surrounding secondary structure element (1-3 residues inside the SSE),
- as reference for the "optimal" anchor groups, the location of minimal RMSD between target and template is used as well as the first position (starting from the gap) where the RMSD drops below 2 Å or 1.5 Å, respectively.

The anchor group prediction based on these simple criteria is compared to a prediction which takes into account the local model energy around the gaps. For this purpose, raw models are generated based all alignment used in the anchor group test set (*i.e.* by replacing the sidechains and by removing residues in the case of deletions). Several possible approaches for the prediction of optimal anchor groups based on the inspection of the local energy profile are investigated (see Results and Discussion, Chapter 3.4.2).

In order to analyse the correlation between local structural deviation (between target and template) and local energy of the raw model, the S-score has been used as in several related publications [68, 195, 234, 247]. In contrast to the RMSD, the S-score has an upper limit for the contribution of individual atoms. This makes sense in the given application, since two residues with an 5 Å are as inaccurate as a pair being 10 Å apart.

The S-score is given by the following formula:

$$S - score \quad = \quad \frac{1}{1 + (\frac{d_i}{d_0})^2} \tag{2.11}$$

where d_i is the distance between two atoms (here the Cα atoms) and d_0 is the distance threshold which has been set to $\sqrt{5}$ as in the other approaches. The S-score ranges from 1 (for a perfect agreement between target and template) to 0 (infinite distance).

2.5 Implementation

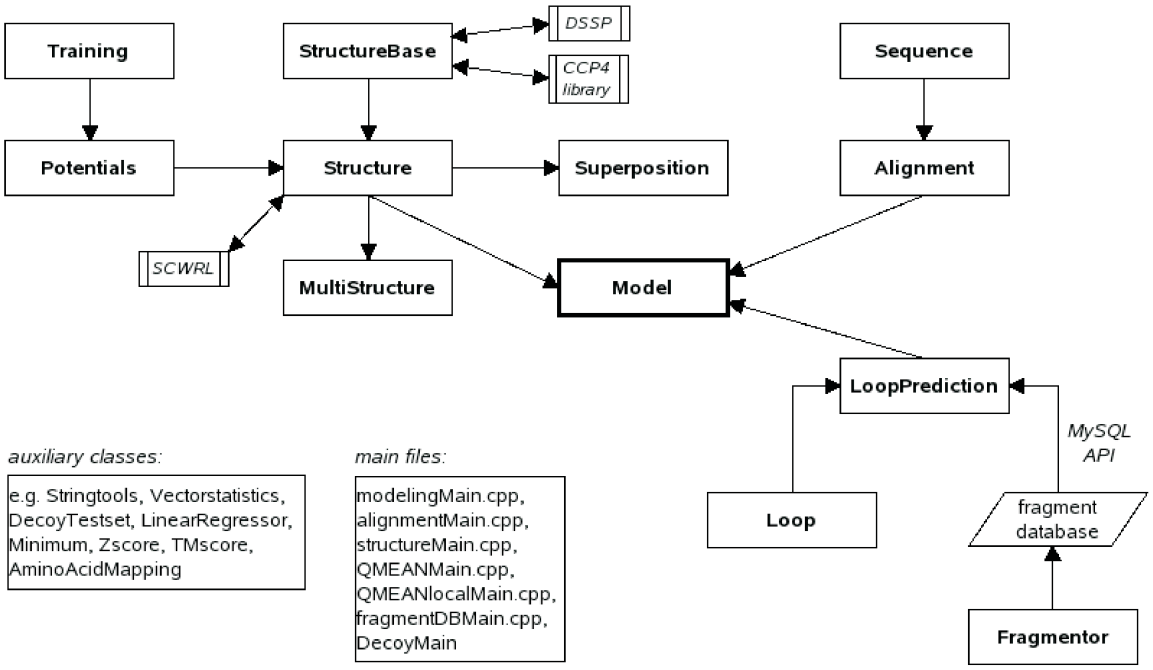


Figure 2.7: Most important C++ classes of the modelling pipeline.

The modelling pipeline presented in this work has been implemented in C++. The most important classes and their interconnections are shown schematicaly in Figure 2.7.

The central class `Model` combines an instance of the classes `Alignment` and `Structure` and is connected to the loop modelling class `LoopPrediction`. At any time of the modelling process `Model` ensures the correct mapping of amino acid positions in the alignment, the structure of the template and the resulting model and guides the initial model building process based on the given template structure and the alignment (*i.e.* the mapping of the target sequence on the template backbone).

The actual changes of the template structure in the modelling process are performed solely in the class `Structure`. These changes include: mutations (change the identity of a residue and remove its sidechain), protection of residues (mark residues such that their

conformation is not changed in the sidechain building routine), deletion and insertion of residues in the template structure, sidechain modelling (using SCWRL [31]), etc.

The class `Structure` itself inherits from `StructureBase` which is responsible for loading and saving PDB-files, for the correct assignment of properties such as torsion angles, secondary structure and solvent accessibility information from DSSP [107]. It provides methods for the selection of atoms using the CCP4 Coordinate Library [114]. The selection of atoms within a sphere is used in the derivation and application of the pairwise statistical potential and the solvation potential as well as in the clash check routine in loop modelling. As compared to `StructureBase`, the class `Structure` additionally contains all methods for the energy calculation of single residues, segments (as needed in loop prediction) and whole structures based on the statistical potentials described in this work.

The class `Training` is used to derive the frequencies of structural features from a set of protein structures and converts them in potentials of mean force as described in Chapter 2.4.1. The data is stored in text files. All classes using the statistical potentials need to include the class `Potentials` which loads the data from the text files and stores them in internal datatyps.

The class `LoopPrediction` contains the loop modelling routine with an interface to the fragment database using the MySQL C-API[i] based on the `mysqlclient` library. The `Fragmentor` class performs the fragmentation of a given non-redundant set of protein structures and the storage of the data in the MySQL database. The fragmentation process is described in Chapter 2.3.1.

The class `Superposition` allows to superimpose two structures either in a sequence-dependent manner by parsing the output of the program TMscore [247] or in a sequence-independent manner by using the algorithm of Lessel and Schomburg [121]. In both cases, the distances of the corresponding residues is calculated. For the later approach, additionally a web server with the name *Protein3Dfit* has been implemented[j].

Multiple main-files have been implemented resulting in different executables which provide access to different functionalities of the modelling pipeline such as modelling

[i] `http://dev.mysql.com/doc/refman/5.1/en/c.html`
[j] `http://www.protein3dfit.uni-koeln.de`

as a whole, loop prediction and model quality assessment (global or local). In all cases, the "-h" option displays an overview on the functionality of the given executable.

The modelling pipeline itself requires an alignment and a template structure as input (optionally an output directory and the path to the secondary structure and solvent accessibility prediction files can be provided). After execution, the user is guided through the modelling process in an interactive manner. The initial modelling steps (template detection and alignment building) are performed with separate Python and Perl scripts.

3 Results and Discussion

The Results and Discussion chapter is structured as follows: In the first section, the results from CASP7 experiment are used as a basis to analyse and discuss the performance of the modelling pipeline established in this work. The section starts with a brief recapitulation of the steps involved in homology modelling. In the next section (page 108ff), the scoring function used for model quality assessment is described in detail since the two subsequent sections both rely on the energy function terms introduced there. Afterwards, the general performance of the loop prediction routine is investigated and compared to several other loop prediction methods (page 135ff). The last section deals with the local analysis of model quality and a statistical analysis of the regions around gaps serving as potential anchor groups for the loop modelling process is presented (page 153ff).

3.1 CASP7 results

3.1.1 The comparative modelling pipeline

The basic steps in homology modelling or comparative modelling are template identification and selection, target-template alignment and model building including loop and side chain prediction. A schematic representation of a typical comparative modelling workflow is given in Figure 2.1 at the beginning of the Methods. Usually multiple models are built from which the final model is selected using some kind of energy or scoring function (typically called *model quality assessment program*). In an optional refinement step it can be tried to remove local errors in the model in order to come closer to the target.

The modelling pipeline as well as an early version of the QMEAN scoring function [16] for model quality assessement (see Chapter 3.2) have been recently tested at the seventh round of community-wide CASP experiment. The goal of CASP is to objectivly assess the abilities and weaknesses of current protein structure prediction methods (see

Introduction on page 19 for more details).

This section starts with the description of the overall performance of the pipeline at CASP7 followed by a detailed analysis of the results. Since an extensive evaluation of the performance of the first 3 steps in the modelling pipeline (*i.e.* template identification, target-template alignment and model building) would go beyond the scope of this work, the performance of the methods is discussed on the basis of some selected examples. The results are chosen in the attempt to highlight strengths and limitations of the methods used in the pipeline and to discuss possible future improvements.

3.1.2 Overview on the results

The CASP experiment was used as a testing ground for the pipeline established during the first two years of this project. Setting up a complete comparative modelling pipeline was a basic prerequisite for dealing with loop prediction and model quality assessment. Since we joined CASP for the first time, our primary intention was to investigate whether it is possible to build reasonable models with the pipeline and whether the scoring function is able to disciminate between good and bad models in the task of model quality assessment. The results exceeded all our expectations: several top ranking models have been built (rank 2, 4 and 6 of over 130 predictions) and the scoring function was among the top-ranking model quality assessment programs [113, 169]. The results are accessible from the official CASP website[a].

During the prediction season of approximately 3 months, the participating groups could submit up to 5 models for each of the 95 accepted targets. The predictors themselves rank the 5 models according to their belief which model is closest to the target structure (denoted as *model 1*). Our group (*i.e.* the author of this work) submitted a total number of 68 models to the tertiary structure prediction category and 65 predictions to the model quality assessment category. Due to the limitations in time and resources not more than 18 targets could be processed. Table 3.1 provides an overview on the ranking of all 18 models designated as model 1 (*i.e.* the model assumed to be closest to native).

[a] `http://predictioncenter.org/casp7/`

Table 3.1: Overview on the CASP7 results of the 18 models designated as *model 1*.

model quality	fraction	comment
top10 models	3 of 18	rank 2, 4 and 6 of over 130 participating groups
above average	11 of 18	above the community average at CASP7
below average	4 of 18	bad performance because too few residues modelled

If a target consists of more than one domain, the assessors additionally analysed the quality of each domain (denoted with subscript D1 and D2 in the first column of Table 3.2). The quality of most of the predicted models was above the community average and three of them were among the top 10 predictions for model 1. The best models were on rank 2, 4 and 8 of more than 130 participating groups. The bad results for the remaining 4 models can be attributed to the low target coverage of these models (*i.e.* not the full target has been modelled). In the CASP assessment, the models are ranked according to the GDT_TS score (see definition on page 62), which reflects the average percentage of residues alignable below different distance thresholds. As a consequence, models which do not cover the entire target sequence automatically get a lower score, since the missing residues are counted as "not alignable". The 4 bad models mentioned above all have some residues missing at the chain ends (target coverage 87.1% to 98.5%). A closer inspection of the models revealed that two of these models were actually very good in terms of all-atom RMSD (rank 14 and 21). A more detailed analysis follows in Chapter 3.1.5 with a ranking based on the all-atom RMSD for all 18 targets (see Table 3.4). At the beginning of the CASP7 prediction season, our pipeline was not yet able to model chain ends. At a later point of time (for models after target T0345), a modified version of the loop modelling protocol was used in order to model chain ends.

As it can be seen from Table 3.2, the 18 targets for which models have been submitted cover a wide range of modelling difficulty as expressed by the sequence identity between the target sequence and the template used to build the model. Two of the three easy modelling cases with sequence identity above 50% could be modelled with all-atom RMSD around 1.5 Å. The three outstanding predictions mentioned above (targets T0341 [domain 1], T0373 and T0379) are highlighted in bold and represent difficult modelling targets with a sequence identity around 20%. The results for these three targets are discussed in detail later. The last two columns in Table 3.2 show the ranking

of the best model (out of the maximum five models submitted per target) compared to all models of all predictors. As it can be seen, the best models are consistently better than average over all targets.

Table 3.2: Detailed analysis if the quality of the models submitted to CASP7.

Target	$\%_{id}$[a]	GDT_TS	RMSD[b]	$\%_{cov}$[c]	$rank_1$[d]	$\%_{rank}$	$rank_{all}$[e]	$\%_{rank}$
T0303	21.8	73.89	3.4	100	21/128	16.41	35/482	7.26
$T0303_{D1}$		83.84	2.45	100	18/128	14.06	41/482	8.51
$T0303_{D2}$		72.4	4.1	100	31/128	24.22	43/482	8.92
T0334	55	89.97	2.89	99.8	55/131	41.98	195/488	39.96
T0340	58.7	90.85	1.53	96	101/145	69.66	244/541	45.1
T0341	22.8	73.31	2.99	95	34/133	25.56	112/508	22.05
$T0341_{D1}$		78.38	2.25	92.6	66/133	49.62	237/508	46.65
$T0341_{D2}$		81.97	3.35	100	**6/133**	**4.51**	**28/508**	**5.51**
T0345	62.2	94.19	1.58	98.4	81/131	61.83	231/483	47.83
T0359	38.1	82.78	3.09	97.8	51/145	35.17	156/543	28.73
T0360	16.3	67.01	5.77	100	29/136	21.32	89/502	17.73
T0362	21.2	72.4	4.05	94.4	80/139	57.55	114/534	21.35
T0364	16.7	68.37	3.26	87.1	72/137	52.55	197/528	37.31
T0370	20.1	63.88	3.7	88.2	45/131	34.35	103/514	20.04
T0371	25.5	59.1	3.98	93.6	62/130	47.69	214/511	41.88
$T0371_{D1}$		72.69	2.99	88.9	67/130	51.54	236/511	46.18
$T0371_{D2}$		66.73	3.58	100	29/130	22.31	84/511	16.44
T0373	19.7	68.58	3.84	100	**2/138**	**1.45**	**13/525**	**2.48**
T0374	22.5	66.56	4.18	96.2	39/144	27.08	112/547	20.48
T0375	17.2	62.25	4.31	97	41/134	30.6	133/515	25.83
T0376	24.3	67.16	3.79	99	53/131	40.46	173/522	33.14
T0379	20.2	68.01	4.18	100	**4/135**	**2.96**	**18/516**	**3.49**
$T0379_{D1}$		78.22	3.35	100	4/135	2.96	13/516	2.52
$T0379_{D2}$		66.41	4.6	100	32/135	23.7	85/516	16.47
T0380	24.8	73.77	3.07	95.8	58/138	42.03	97/535	18.13
T0384	18.2	64.53	4.46	98.7	49/135	36.3	171/524	32.63

[a] Percent sequence identity between target and template.
[b] All-atom root mean square deviation.
[c] Fraction of target residues present in the model.
[d] Rank of model 1 among all other models designated as model 1.
[e] Rank of the best model (of maximum 5 submitted) among all models from all groups.

As mentioned in the beginning, the CASP experiment was used as a testing ground in order to identify bottlenecks in the prediction pipeline and to compare the performance with other methods. Even during the CASP prediction season the pipeline was constantly improved and new features were added (*e.g.* the ability to model chain ends where only one anchor group is present). This, in fact, complicates the evaluation

process but enormously pushed the whole project. The main purpose of the following sections is to highlight what went right in the different modelling steps and where is room for improvement. The lessons learnt during CASP and after CASP, when the evaluation of the assessors was available, will be addressed in detail.

3.1.3 Template identification

As described in Methods (see section 2.1.2), templates are identified using the PDB-BLAST protocol which uses a sequence profile (generated by PSI-BLAST) representing the protein family of the target protein in order to scan the PDB for possible templates.

In Figure 3.1, an extract of the PSI-BLAST output (first 10 hits) for the CASP7 target T0288 is shown as an example. The query sequence, a protein involved in signaling, consists of 93 amino acids and represents a target of the Structural Genomics Consortium.

```
                                                                    Score    E
Sequences producing significant alignments:                         (bits) Value

1Z87A 263 NMR NA NA NA Alpha-1-syntrophin <SWS SNA1_MOUSE> [MUS ...    95   1e-20
1UM7A 113 NMR NA NA NA synapse-associated protein 102 <GB BAA865...    92   7e-20
1QAVA 90 XRAY 1.90 0.208 0.259 ALPHA-1 SYNTROPHIN (RESIDUES 77-1...    88   1e-18
2FNEA 117 XRAY 1.83 0.194 0.243 Multiple PDZ domain protein <SWS...    88   1e-18
2FNEB 117 XRAY 1.83 0.194 0.243 Multiple PDZ domain protein <SWS...    88   1e-18
2FNEC 117 XRAY 1.83 0.194 0.243 Multiple PDZ domain protein <SWS...    88   1e-18
1TP3A 119 XRAY 1.99 0.233 0.296 Presynaptic density protein 95 <...    88   2e-18
1TP5A 119 XRAY 1.54 0.193 0.229 Presynaptic density protein 95 <...    88   2e-18
1TQ3A 119 XRAY 1.89 0.238 0.296 Presynaptic density protein 95 <...    88   2e-18
1BE9A 119 XRAY 1.82 NA NA PSD-95 <SWS DLG4_RAT> [RATTUS NORVEGICUS]    87   3e-18
```

Figure 3.1: Extract of the PSI-BLAST output for target T0288 of CASP7.

The output is structured as follows (from left to right): PDB idenitfier including chain identifier, number of amino acids, experimental method (NMR spectroscopy or X-RAY crystallography), resolution, R value, R free value, description of the protein and finally bit score and E-value.

In order to decide which template(s) to choose, the E-value, reflecting the reliability of the hit, is the most valuable criteria. Since BLAST [5] (Basic Local Alignment Search Tool), as the name suggests, only produces local alignments or matches, the coverage of the target by the selected template has to be checked. Templates with low E-value

but covering only a short fraction of the target are of little practical value (at least as a single template, but possibly in combination with others). In the presence of a variety of possible candidates, the quality of the template structure should be investigated by analysing resolution, R value and unresolved residues in the structure (see description of experimental methods in the Introduction on page 13). In our pipeline, 3-5 template structures are manually selected based on the criteria described above.

For many template-based modelling targets from CASP7, a simple BLAST search against the database of sequences from PDB structures is sufficient to detect suitable templates. But in some cases, BLAST is not sensitive enough to detect the homology as show exemplarily for target T0360 (141 amino acids). In Figure 3.2 the more or less random hits (E-value ≈ 1) identified by BLAST cannot be used as templates. An inspection of the corresponding alignment reveals that only approximately one-third of the query sequence are covered.

```
                                                                    Score    E
Sequences producing significant alignments:                         (bits) Value

2GLFD 450 XRAY 2.80 0.168 0.239 Probable M18-family aminopeptida...    29   0.99
2GLFC 450 XRAY 2.80 0.168 0.239 Probable M18-family aminopeptida...    29   0.99
2GLFB 450 XRAY 2.80 0.168 0.239 Probable M18-family aminopeptida...    29   0.99
2GLFA 450 XRAY 2.80 0.168 0.239 Probable M18-family aminopeptida...    29   0.99

Query: 12  KSAVQTMSKKKQTEMIA----DHIYGKYDVFKRFKPLALGIDQDLIAALPQYD 60
           K AV+T   K   EM      D + G+ +V   F P  +G+D+ LI A  Q D
Sbjct: 198 KEAVKTNVLKILNEMYGITEEDFVSGEIEVVPAFSPREVGMDRSLIGAYGQDD 250
```

Figure 3.2: Hits identified by a simple BLAST search for target T0360.

PDB-BLAST on the other hand identifies one temnplate with a reasonably good E-value for target T0360 which covers the whole target (Figure 3.3).

```
                                                                    Score    E
Sequences producing significant alignments:                         (bits) Value

1DVOA 152 XRAY 2.00 0.197 0.224 FERTILITY INHIBITION PROTEIN O <...    78   2e-15
1APYB 141 XRAY 2.00 0.169 0.224 ASPARTYLGLUCOSAMINIDASE <SWS ASP...    27   4.2
1APYD 141 XRAY 2.00 0.169 0.224 ASPARTYLGLUCOSAMINIDASE <SWS ASP...    27   4.2
1APZB 141 XRAY 2.30 0.212 0.291 ASPARTYLGLUCOSAMINIDASE <SWS ASP...    27   4.2
1APZD 141 XRAY 2.30 0.212 0.291 ASPARTYLGLUCOSAMINIDASE <SWS ASP...    27   4.2
1ZZMA 259 XRAY 1.80 0.171 0.231 putative deoxyribonuclease yjjV ...    26   7.8
```

Figure 3.3: Hits identified by a PSI-BLAST search for target T0360.

According to the evaluation of the CASP assessors (see Figure 3.4), the template identified by PDB-BLAST (1dvo) turned out to be the best available template (*i.e.* the template closest to the target as expressed by RMSD and LGA-score). LGA (Local/Global Alignment) is a standard tool in the CASP assessment and analyses the local and global structural similarity between two structures. Based on the structural superposition, the distance of the corresponding residues (according to the sequence) in target and model are analysed and defined as *correctly aligned* if they meet a certain distance threshold (here: Cα-distance below 5 Å). The LGA-score reflects the percentage of alignable residues among those of the whole target.

As it can be seen from Table 3.3, in at least 4 cases BLAST could not detect a suitable template for building a model. For the 3 targets marked with yes in brackets, the template could be identified but only with an E-value $> 10^{-3}$.

The PDB-BLAST protocol not only identifies more templates as compared to a simple BLAST but also identifies them with a clearly lower E-values. With PDB-BLAST, the “real" templates get considerably lower E-values than the apparently random hits whereas this is often not the case for BALST.

For hard template-based modelling targets (*i.e.* when only very remote homologous templates or only analogues are available), profile-to-sequence based homology detection methods such as PDB-BLAST reach their limitation. In this case, more sensitive profile-profile or HMM-HMM search methods have to be applied. Threading

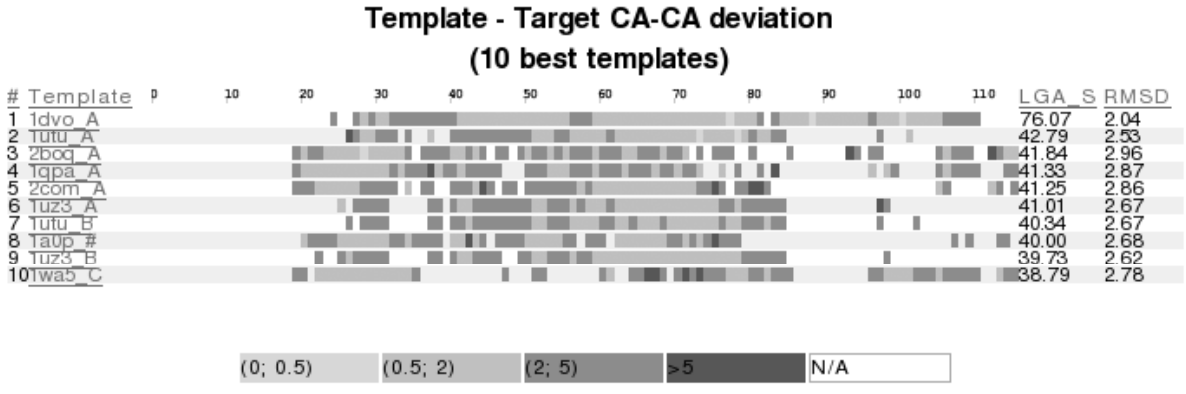


Figure 3.4: Coverage of the target T0360 by the top 10 templates [115].

Table 3.3: Template detectability by a simple BLAST run among the 18 processed targets.

target	% sequence identity	BLAST detectable[a]
T0303	21.8	yes
T0334	55.0	yes
T0340	58.7	yes
T0341	22.8	yes
T0345	62.2	yes
T0359	38.1	yes
T0360	16.3	no
T0362	21.2	(yes)
T0364	16.7	no
T0370	20.1	no
T0371	25.5	yes
T0373	19.7	(yes)
T0374	22.5	no
T0375	17.2	(yes)
T0376	24.3	yes
T0379	20.2	yes
T0380	24.8	yes
T0384	18.2	yes

[a] (yes): Only templates with E-value $> 10^{-3}$ are detected.

algorithms, asseessing the compatibility of the sequence to folds in a fold library, can be used in order to detect possible analogous folds in the absence of homology (see section "fold recognition" in the Introduction on page 20).

If no significant hits can be identified with PDB-BLAST, fold recognition servers such as HHPRED[b] [204] or 3D-PSSM[c] [110] can be consulted. The probably best starting point is the BioInfoBank meta server[d] which provides access to various fold recognition servers and translates the collected information (*i.e.* identified templates and corresponding alignments) into a uniform format.

As advanced template detection methods require a lot of time and resources, a hierarchical approach for template detection is advisable, especially for automatic

[b] `http://toolkit.tuebingen.mpg.de/hhpred`
[c] `http://www.sbg.bio.ic.ac.uk/~3dpssm/`
[d] `http://meta.bioinfo.pl`

servers:

- first try BLAST (sequence-to-sequence). If no suitable template has been identified, use
- PDB-BLAST (profile-to-sequence),
- otherwise, use advanced fold recognition methods (profile-profile and HMM-HMM, respectively)

It should be noted here that especially in the presence of very remote homologues, the coverage of the target sequence with respect to the template structure is usually very low which makes it difficult to build a reasonable model based on a single template. In this case, the combination of multiple templates potentially leads to better models.

Although being the second best model submitted to CASP7, our model 1 for target T0373 could have been further improved by combining two templates. The top scoring model has been built based on template 1s3j_A (*i.e.* PDB identifier 1s3j, chain A) and shows a very good overall quality except for the N-terminus as shown in Figure 3.5 a). The thick tube represents the native structure of the target and the thin tube the model. The regions colored in green mark corresponding residues in target and model which are below a certain distance threshold (here: Cα-distance below 5 Å). The other regions are either incorrect because of alignment errors or incorrect modelling. Alignment errors are discussed in the next section.

Figure 3.5 b) shows our second best model which has been built with template 1jgs_A. The model covers perfectly the N-terminal chain end which could not be accurately modelled with the first template. It becomes apparent, that a combination of both templates could lead to a considerably better model covering both chain ends perfectly.

In a future version of the modelling pipeline, the ability to use information from multiple templates for one model should be implemented. Due to the object-oriented implementation of the software, this can be done with minor effort. The difficulty which then arises is to decide which region to use from which template. Since the combination of multiple templates was not in the scope of this work, the models are currently built based on one template which represents a reasonable approach for many targets.

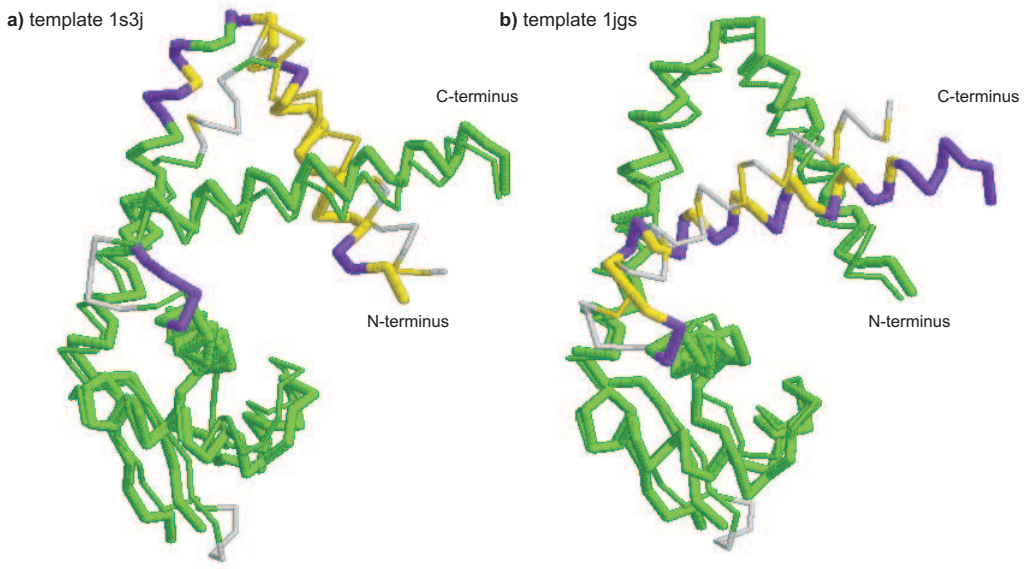


Figure 3.5: Two models for T0373 built based on different templates illustrating the potentials improvement possible by combining multiple templates [115].

3.1.4 Target-template alignment

As described in Methods (see section 2.1.3), the alignments between the target sequence and the template are generated with a profile-profile alignment protocol. The alignment algorithm has been optimised as part of the project thesis of Oscar Bortolami and showed a comparable performance in comparison to other state-of-the-art alignment programs *(data not shown)*. As mentioned in the Introduction, alignment errors are still, beside loop prediction, the major source of errors in comparative modelling. In this work, the performance of the alignment algorithm is evaluated qualitatively based on a detailed inspection of all our models submitted to CASP7 and the corresponding alignments. As an example, the alignment shift in target T0341 is described in more detail in order to point out the structural consequences of alignment errors.

Analysing the alignment quality of the models submitted to CASP7 is not a trival task. Since only the final models are submitted and not the corresponding alignments (which would be difficult to evaluate, if multiple templates are used), the assessors "calculated" the alignments quality indirectely by comparing the model with the corresponding experimental structure. The following procedure was used: Target and models were

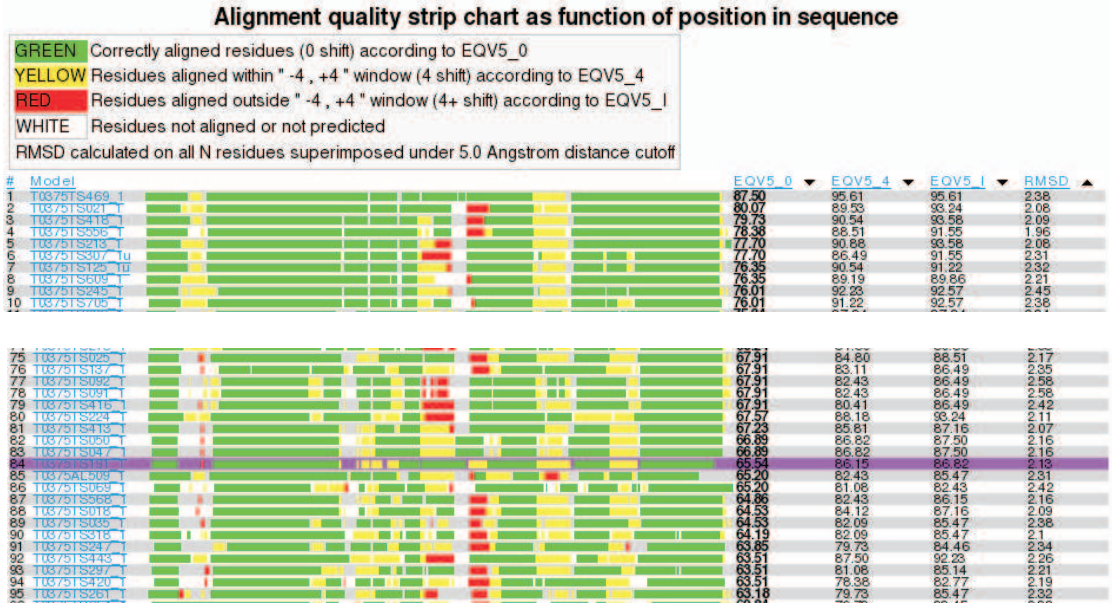


Figure 3.6: Alignment quality strip chart for target T0375 [115].

structurally superimposed in a sequence independent manner using the LGA algorithm [244]. The alignments based on the structural superposition are subsequently ranked according to the percentage of correctly aligned residues (Cα-distance below 5 Å) among those of the whole target. Residues not present in the model are defined as not aligned. This makes it difficult to decide based on a single quality number whether a certain alignment scored worse because of alignment errors or just because of some missing residues in the model. Beside alignment errors, local model errors can arise if structurally variable regions (mainly loops) of the template have not been re-modelled or have been modelled incorrectly, respectively. These errors cannot be distinguished from alignment errors without knowledge of the alignment and the corresponding templates used.

Nevertheless, the alignment quality strip charts (see Figure 3.6) as provided by the CASP assessors are useful means in order to compare models and identify regions of errors. Regions in the model with 'correctly aligned residues are marked in green. Regions colored in yellow and red highlight residues of the model which are, based on the superposition, shifted with respect to the position in the experimental structure, with yellow for shifts within 4 residues and red for shifts greater than 4 residues.

Alignment errors are a consequence of miss-placed gaps, *e.g.* if a long gap should be splitted into two shorter ones. As a consequence, in the region between the gaps, the residues of the model are shifted relative to the real position in the model (*i.e.* the target residues are mapped on the wrong region of the template backbone).

In the post-evaluation of the CASP models, the following procedure is used in order to identify alignment errors:

- The alignment quality strip charts are inspected in order to identify regions of structural divergence between target and model (see Figure 3.6).
- In these regions, the alignment used to build the model is compared to a structural alignment between target and template in order to identify possible differences in the gap placement.
- The structural superposition of target and model is used in order to determine those regions of the model, which are incorrect because of alignment shifts and not as a consequence of wrong loop modelling or structural divergence between target and template (*i.e.* structurally variable regions, which have not been remodelled).
- Alignment shifts appear as regions in the structural superposition where the backbone of model and native structure coincide (*i.e.* this part origines from a structurally conserved region of the template) but the corresponding sequence is shifted (*i.e.* the residues closest in space in the superposition are not identical).

A detailled inspection of all our models submitted to CASP7 revealed that the alignments are generally very accurate and worse alignment scores compared to the other groups, can be mostly attributed to either a low target coverage (*i.e.* chain ends have not been modelled) or inaccurate loop prediction (*e.g.* difficult long loops which could not be modelled accurately, non-conserved loops which should have been remodelled).

For the following targets, alignments error could be identified (in brackets the sequence identity between target and template): T0341 (22.8%), T0364 (16.7%), T0373 (19.7%), T0374 (22.5%), T0375 (17.2%), T0376 (24.3%). All these targets represent difficult modelling cases as reflected by the sequence identity between target and template being

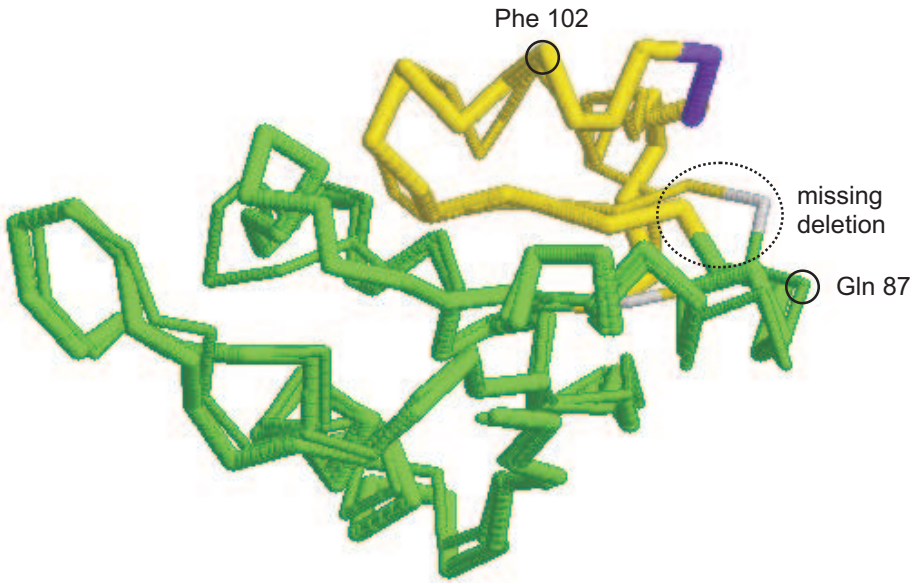


Figure 3.7: Superposition of model and target T0341: Structural consequences of alignment errors (yellow segment).

around 20%. In target T0375 for example, multiple alignment shift were observed in the models of nearly all groups. The available templates show a high structual similarity with the target such that a large fraction of the template could have been used for the model. But, as a consequnce of the low sequence conservation, most groups failed to accurately position the gaps resulting in multiple alignment shifts as reflected by the yellow regions in the alignment strip chart (see Figure 3.6).

Exemplarily, the alignment shift observed in our model for target T0341 domain 2 is described here in detail. Actually, this was one of our top scoring models, which suggests, that most of the groups as well had problems with the alignment for this target. The alignment error consists of a misplacement of the deletions after residue Glu-87 which caused an alignment shift of one residue for the following 16 residues until the next gap (marked in yellow in Figure 3.7).

A comparison of the alignment used to build the model (Figure 3.8) and a structure-based sequence alignment (Figure 3.9) between target and template generated by CE [192] reveals that two residues instead of one should have been deleted after *glutamine 87*. By looking at the superposition of target and model in Figure 3.7,

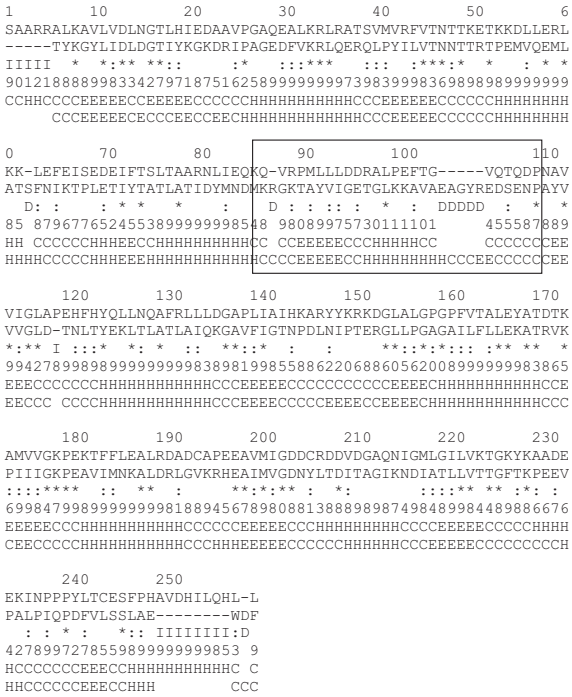


Figure 3.8: Original sequence alignment between target T0341 and template 1wvi_A. The region of the alignment error is marked with a box.

the missing deletion can be clearly identified and one can observe that the region between the two deletions (until approximately *phenylalanine 102*) is structurally highly conserved and the backbone therefore could have been copied from the template.

The structure-based sequence alignment is shorter since CE produces only local alignments based on the maximum common substructure. The location of the other two gaps (*i.e.* a deletion after residue 62 and an insertion at position 118) agree well between the two alignments.

The sequence identity between target and template (PDB code: 1wvi) is approximately 23% which represents a rather difficult modelling task. As it can be seen from Figure 3.8, the alignment error occured in a region of extremely low sequence conservation which makes it difficult for alignment algorithms to seperate the signal from the noise in this region. Here, purely sequence-based alignment algorithms reach their limit of accuracy and only algorithms integarting structural information (*e.g.* by the use

```
Chain 1: /biochem/mirror/pdb/all/pdb2ho4.ent:A (Size=259)
Chain 2: /biochem/mirror/pdb/all/pdb1wvi.ent:A (Size=257)

Alignment length = 241 Rmsd = 2.11A Z-Score = 7.0 Gaps = 8(3.3%) CPU = 1s Sequence identities
= 24.2%

Chain 1:    7 LKAVLVDLNGTLHIEDAAVPGAQEALKRLRATSVXVRFVTNTTKETKKDLLERLKK-LEFEISEDEIFTS
Chain 2:    3 YKGYLIDLDGTIYKGKDRIPAGEDFVKRLQERQLPYILVTNNTTRTPEMVQEMLATSFNIKTPLETIYTA

Chain 1:   76 LTAARNLIEQKQV--RPXLLLDDRALPEF-TGVQTQD---PNAVVIGLAPEHFHYQLLNQAFRLLLDGAP
Chain 2:   73 TLATIDYMNDMKRGKTAYVIGETGLKKAVAEAGYREDSENPAYVVVGLDTN-LTYEKLTLATLAIQKGAV

Chain 1:  140 LIAIHKARYYKRKDGLALGPGPFVTALEYATDTKAXVVGKPEKTFFLEALRDADCAPEEAVXIGDDCRDD
Chain 2:  142 FIGTNPDLNIPTERGLLPGAGAILFLLEKATRVKPIIIGKPEAVIMNKALDRLGVKRHEAIMVGDNYLTD

Chain 1:  210 VDGAQNIGXLGILVKTGKYKAADEEKINPPPYLTCESFPHAV
Chain 2:  212 ITAGIKNDIATLLVTTGFTKPEEVPALPIQPDFVLSSLAEWD
```

Figure 3.9: Structure-based sequence alignment between target T0341 and template 1wvi_A produced by CE.

of predicted secondary structure and solvent accessiblity of the target sequence or environment-specific gap penalties) can go beyond that.

A visual inspection of the alignment can help identifying potential alignment errors. Gaps within secondary structure elements are usually an evidence for alignment errors: In target T0379, for example, a gap has been moved out of the secondary structure element manually which is one of the reasons (beside the accurate extension of the N-terminal helix) of the high rank achieved by this model. The detection of alignments error can be automated. Several approaches have been described in literature which allow to detect reliable regions in alignmnets *e.g.* by analysing the variation among different sub-optimal alignments [229] or the sequence variation in the profiles [224].

3.1.5 Modelling

In this section, an in-depth analysis of the models submitted to CASP7 is performed. All model designated as model 1 have been evaluated and some general conclusions are drawn concerning the "lessons learnt" at CASP7 with the attempt to highlight possible aeras for future improvements. Some of the top-scoring models are discussed in more detail.

In Table 3.4, a summary of the performance (reflected by the rank of the model and the rank of the corresponding alignment) of all 16 models designated as *model 1* is given. In the last colomn, explanations for the good or bad performance are provided as keywords, since a detailed description of all models would go beyond the scope of this work.

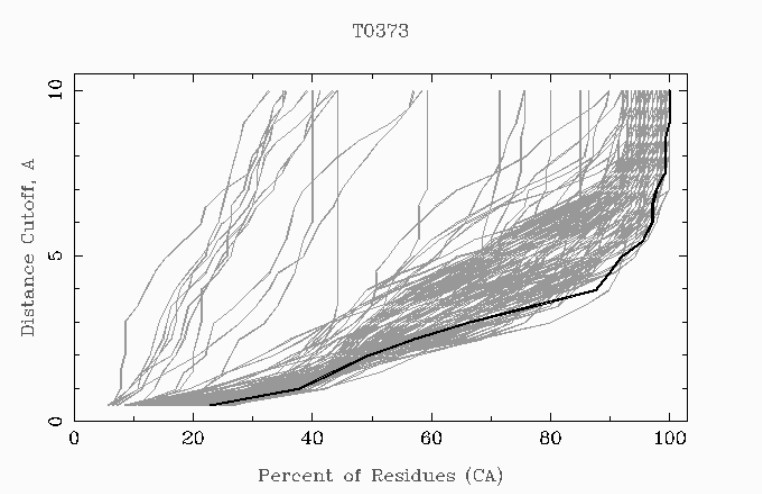


Figure 3.10: GDT plot for T0373: fraction of model residues superimposable with the experimental structure using variable distance thresholds [115].

In order to visualise and compare the quality of all models of a specific target, GDT plots (as shown in Figure 3.10) are provided on the CASP7 website which reflect the percentage of residues from the model which fall below a certain distance cutoff after a (sequence-dependent) superposition on the experimental structure of the target. The lower the run of the curve the better a model provided that enough target residues have been modelled. The GDT plots of all our models submitted to CASP7 is shown in the appendix.

Table 3.4: Detailed analysis of the quality of the models submitted to CASP7 with comments.

Target[a]	$\%_{cov}$	r_{GDT}[b]	r_{RMS}[c]	r_{aln}	Comment
T0303	100	21	33	20	Good alignment; 3 loops: 2 modelled very accurately
$T0303_{D1}$	100	18	13	3	Nice alignment; best available template identified
$T0303_{D2}$	100	31	58	52	Bad template selection for this domain (same template for both domains)
T0334	99.8	55	95	66	Inaccurate loop prediction for 8-residue insertion (difficult)
T0340	96	101	46	112	Alignment perfect, but bad coverage (chain ends missing)
T0341	95	34	15	33	Too few residues modelled (95% modelled); alignment error
$T0341_{D1}$	92.6	66	17	77	Good alignment, but bad coverage at C-term; non-conserved loop not modelled
$T0341_{D2}$	100	6	22	53	Alignment error: Wrong location of deletions in region of low sequence identity
T0345	98.4	81	21	88	Alignment good; 3 residues missing at N-terminal chain end
T0359	97.8	51	64	38	Alignment good, but only 97.8% modelled
T0360	100	29	44	28	Alignment good; bad modelling of chain ends
T0362	94.4	80	57	170 (47)	Bad model selection (model 3 much better); difficult 8-residue insertion
T0364	87.1	72	14	87	Too few residues modelled (<90%); alignment error at C-terminal end
T0370	88.2	45	7	32	24-residue insertion at C-terminal end not modelled
T0371	93.6	62	10	60	Too few residues modelled in domain 1
$T0371_{D1}$	88.9	67	11	83	N-terminus not modelled; difficult insertion around position 220
$T0371_{D2}$	100	29	28	22	Best available template used; nice alignment; 2 non-conserved loops not modelled
T0373	100	2	21	12	Good alignment; N-terminus perfect; C-terminus minor alignment shift
T0374	96.2	39	21	36	Suboptimal template selection; 2 difficult long loop regions; minor alignment error
T0375	97	41	15	83	Difficult alignment: multiple shifts; large movement of β-sheet in interface region
T0376	99	53	42	47	Minor alignment error; structurally var. helix and nonconserved loop not modelled
T0379	100	4	13	2	Alignment very good; accurate extension of N-terminal helix
$T0379_{D1}$	100	4	10	3	Alignment very good; accurate extension of N-terminal helix
$T0379_{D2}$	100	32	57	18	Alignment OK
T0380	95.8	58	20	55	Alignment good, but missing residues at C-terminal chain end
T0384	98.7	49	41	63	better templates available; huge insertion difficult to model; one alignment shift

[a]Subscript D1 and D2 specify domain 1 and 2 in multi-domain proteins.
[b]Rank based on GDT_TS (total number of models ∼130).
[c]Rank based on the all-atom RMSD between experimental structure and model.

3.1.5.1 Loop prediction at CASP7

Loop prediction at CASP7 has been performed using the fragment database described in Methods (Chapter 2.3). At the time of the prediction season, only a preliminary version of the scoring function used for loop ranking was implemented. The loops were ranked based on a combined scoring function consisting of a torsion energy term as well as a solvation and pairwise interaction energy term considering only the C_α atoms. Based on this ranking, loops have been manually selected by additionally taking into account sequence conservation between the target loop and the fragment extracted from the database. In the actual version of the scoring function, an all-atom implementation of the pairwise interaction potential and the solvation potential are used. The general performance of the current loop modelling routine is described in Chapter 3.3.

Nevertheless, in many cases, the simple scoring function was able to identify suitable loops from the fragments database. Due to the fact that human intervention has been used in loop modelling, a detailed evaluation of all loops in all CASP models is not given here, but instead, the loop modelling results of two selected targets are shown here exemplarily which clarify the strengths as well as the limitations of the loop prediction protocol.

In our first model submitted to CASP7 (target T0303), 3 insertions had to be modelled as it can be seen from the alignment between target and template 1ah5_A in Figure 3.11. Target and template have a sequence identity of about 23%. A comparison between the experimental structure of the target (PDB code: 2hsz) and the final model revealed that a nonconserved segment between Leu-195 and Pro-209 should have been remodelled as well (see superposition of target and model in Figure 3.12). In this nonconserved segment, two mutations involving glycine (position 199) and proline (position 203) can be observed which is most likely the reason for the observed local refolding.

Loop 1 (anchor group positions 55 and 74) represents a very difficult modelling case involving a huge insertion of 11 residue. Insertions of that size can usually not be modelled since most loop prediction programs are limited to loops of length 12 or 15 and, more importantly, the quality of loop predictions rapidly decreases with loops longer than approximately 7 or 8 residues. The limitations are discussed in more detail

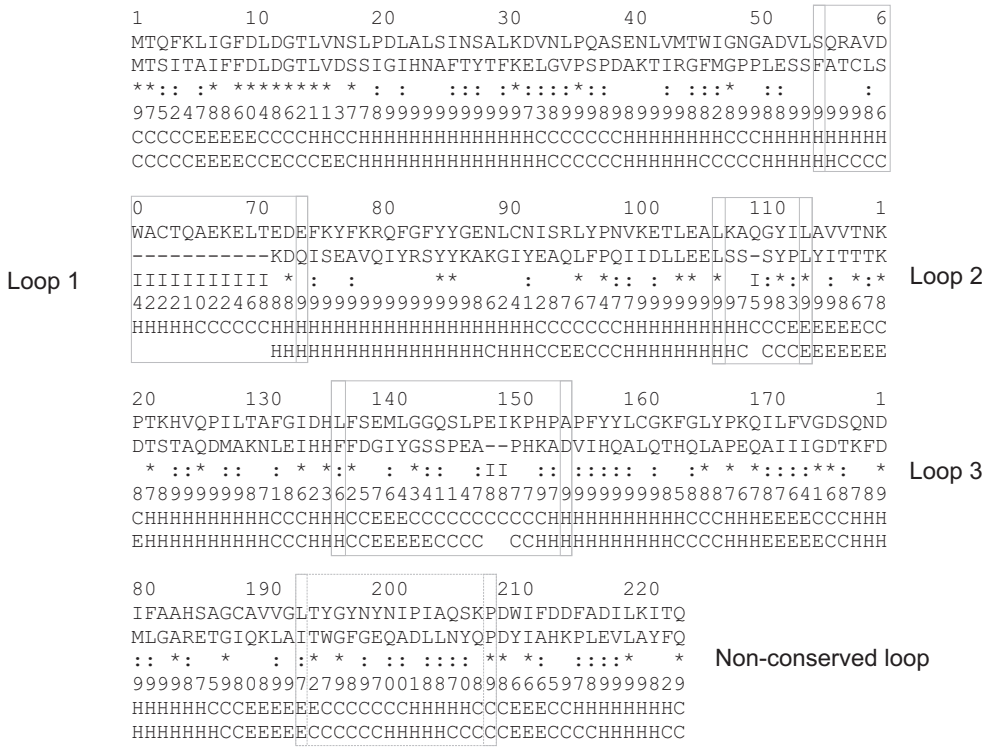


Figure 3.11: Alignment between target T0303 and template 2ah5_A: 3 insertion and 1 non-conserved loop.

in Chapter 3.3. Furthermore, long insertions and to some extent also deletions lead tendentially to distorsion of the anchor region, *i.e.* the region on both sides of the insertion is less structurally conserved between target and template, such that an even longer part need to be remodelled. It has been tried to model the insertion with a fragment of the length 20 (including anchor groups) but the prediction, as expected, failed completely: the backbone RMSD of the loop (without anchor residues) between experimental structure and model was 7.36 Å (see Figure 3.12). But, in some cases, long cases long fragments can be predicted rather accurately, as we will see for *loop 3*.

Loop 2 (anchor group positions 107 and 114) is a 6 residues loop and models a 1-residue insertion between a helix on the N-terminal side and a beta strand on the C-terminal side. This loop has been modelled rather accurately as reflected by the very similar backbone geometry between model and experimental structure (Figure 3.12). The backbone RMSD of 1.82 Å is acceptable for the modelling case, where the

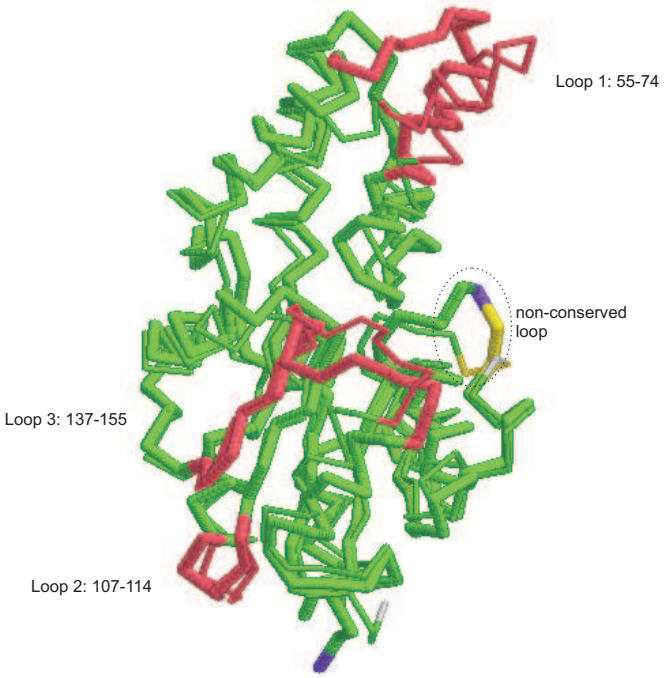


Figure 3.12: Superposition of model (thin tube) and target T0303 (thick tube): loop prediction [115].

anchor residues, on which the fragment is fitted, are inexact to a certain extent. In this case both anchor residues (leucine 107 and isoleucine 114) had an RMSD of around 1 Å (but they were the best anchors in this region).

Loop 3 involves modelling of an insertion of 2 amino acids as it can be seen from the alignment shown Figure 3.11. Since we were not sure if the N-terminal beta strand belongs to the structurally conserved region and can therefore be used from the template, it has been decided to put the anchor group before the beta strand at leucine 137. The anchor group on the C-terminal side of the insertion was set at alanine 155 since the two mutations involving proline just before were expected to have structural consequences. Finally, 17 residues have been remodelled with an exceptionally good RMSD (for this loop length) of 2.37 Å. This can be mainly attributed to the fact that a fragment from a homologue of the target could be used to build the loop (*i.e.* the fragment origines from a structure with 28.3% sequence identity to the target based on the BLAST local alignment). The beta strand mentioned before was indead partly

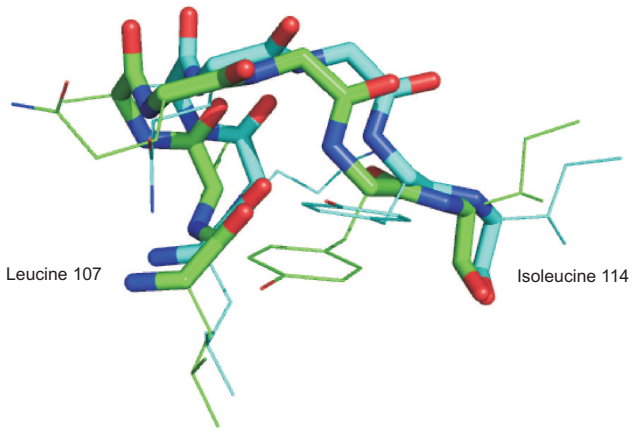


Figure 3.13: Superposition of model (light green) and target T0303 (light blue): loop 2 (residues 107-114).

structurally conserved, such that 5 residues less could have been remodelled, but still a loop of 12 residues needed to be modelled.

During CASP7, the anchor groups in the loop modelling process have been defined manually by placing them in regions on both sides of the gap which are expected to be structurally conserved between target and model. A rather conservative approach was used for the definition of the anchor groups leading to potentially longer fragments to be remodelled as necessary (for a more detailed description of the approach, see Chapter 2.2.2) in Methods. The trade-off between accuracy of the anchor groups and lenght of the fragment to be remodelled is adressed in the next section.

Figure 3.15 shows the very accurate prediction of a beta hairpin structure in target T0364. The alignment shows a 2-residue insertion between two beta strands as it can be seen in Figure 3.16.

The anchor groups were placed in the conserved region (in terms of sequence conservation) of the strands on both sides of the insertion (arginine 97 and leucine 104). The six residues have been modelled with an excellent backbone RMSD of 0.57 Å. Figure 3.15 shows, that the backbone superimposes almost perfectly between target and model and most of the sidechains point into the right direction.

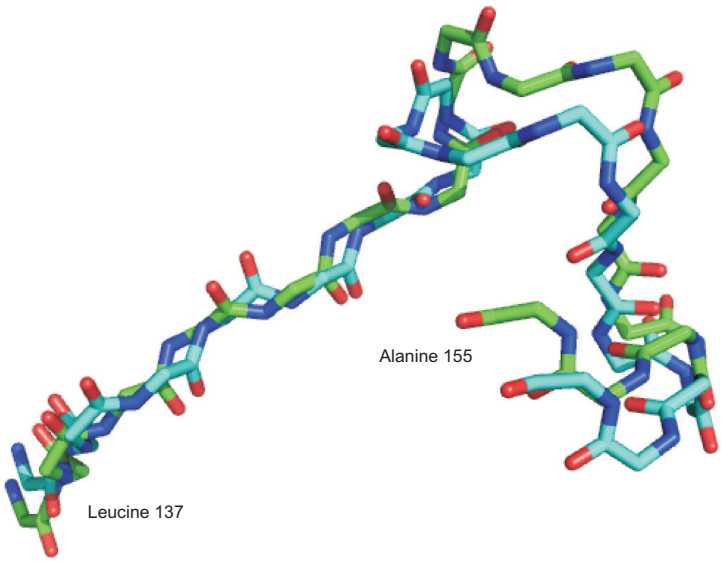


Figure 3.14: Superposition of model (light green) and target T0303 (light blue): loop 3 (residues 137-155).

The loop modelling cases described above, point out general problems in comparative modelling and loop prediction but also show some advantages of the method presented here compared to other loop prediction programs:

- **Remodelling of loop with no insertions and deletions**: Loop regions without insertions and deletions sometimes deviate substantially between target and template as a consequence of multiple amino acid substitutions (*i.e.* low sequence conservation) in this region or of considerable differences in the structural environments, *e.g.* the loop in the template is part of an interface region whereas in the target not, therefore the loop can indepedently adopt its conformation. The evaluation of the CASP7 models showed that non-conserved loops containing mutations involving glycine and proline have to be treated with caution and potentially need to be remodelled. The question, whether to remodel a certain non-conserved loop or not is difficult to answer and it has to be taken into account that loop prediction itself is only possible with a certain accuracy depending on the loop length. Investigating the local conformational energy

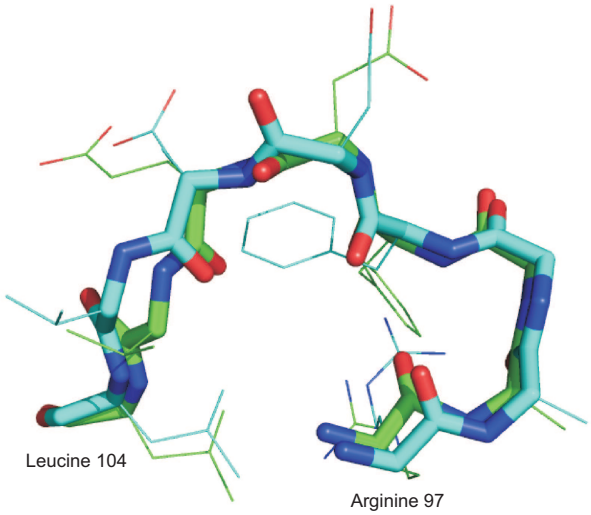


Figure 3.15: Very accurate prediction (light green) of a beta hairpin structure in target T0364.

```
           0        70        80        90          100       110       1
T0364      AHINYLHEVKLGTEVWVQTQILGFDRKRLHVYHSLHRAGFDEVLAASEQMLLHVDLAGPQ
1z54A      LGLTFRAPARFGEVVEVRTRLAELSSRALLFRYRVER--EGVLLAEGFTRHLCQV--GER
conserv    ::::    :::*  * * * :: :  : * : : : *II   :** : :  *:  II*
psipred    EEHHHHHHCCCCCEEEEEEEEEECCCEEEEEEEEEECCCCCEEEEEEEEEEEECCCCC
dssp       EEEEECCCCCCCCEEEEEEEEEECCCEEEEEEEEE  CCEEEEEEEEEECEE  CCC
```

Figure 3.16: Extract of the alignment between target T0364 and the corresponding template.

in this region can support the decision. There is still an urgent need for tools assessing the local model quality as recently underlined in the CASP7 assessment report of the quality assessment category [49]. Local model quality assessment, as described in Chapter 3.4, is a step in this direction.

- **Using fragments from homologues to the target**: Generally, if fragments from homologous structures to the target are present among the top scoring fragments, these should be prefered. Fragments from homologous structures have a higher probability to be correct since they tend to have a similar amino acid constitution compared to the target loop and origine from a similar structural environment. As shown in Chapter 3.3 describing the general performance of the loop prediction routine, fragments from homologues are almost always found on the top ranks.

- **Modelling of loop motifs**: With the given method, frequently occuring structural loop motifs are generally easier to predict than rare ones. In this case, a variety of suitable fragments are present in the database, which increases the chance of indentifying good candidates in the selection process. As a consequence of the statistical nature of the scoring function used for loop ranking, frequently occuring motifs potentially get assigned lower energies.

- **Features of the fragment database**: The fragment database presented in this work differs in many respects from other fragment databases described in literature (FREAD [53], LIP [139], methods by Fernadez-Fuentez *et al.* [70]) and several features of the database have shown to be advantageous in the modelling process during CASP7. The most important advantage is the fact that not only pure loop segments are stored in the database but *all* fragments from a representative set of high-resolution proteins structures. This allows the modelling of fragments containing secondary structure elements or parts of them. This is often necessary if, for example as a consequence of a long insertion, the surrounding secondary structure elements are extended or new secondary structures are formed in the loop region. This situation can typically not be processed with pure loop databases. Another situation in which parts of secondary structure elements need to be remodelled is the kink observed in helices as a consequence of proline [14, 130]. Helices with mutations between target and template involving proline can be remodelled using the fragment database. As described in Chapter 2.3.1 in Methods and later in Chapter 3.1.5.3 concerning the modelling of chain ends, the MySQL database allows to specifically search for fragments showing a certain secondary structure or sequence pattern. In the case of the proline induced helix kink described above, the database can be specifically scanned for fragments which contain an initial helix segment followed by some loop residues (since the subsequent loop probably is remodelled as well) and which have a proline residue at a given fixed position in the helix. The ability to model not only loops but any structural segment represents an overlap to fragment assembly methods successfully used in *ab initio* modelling and highlights the potential of the given methods to be applied in areas beyond pure comparative modelling.

3.1.5.2 Manual anchor group prediction at CASP7

The standard approach in comparative modelling is to place the anchor groups near the end points of the surrounding secondary structure elements of the template (typically 1-2 residues inside). At CASP7, we additionally took into account the agreement between the calculated positions of the secondary structure elements in the template with the potential location of the secondary structure elements in the target based on a consensus of 3 state-of-the-art secondary structure prediction programs (see Methods on page 39). This can provide evidence whether a secondary structure element is possibly extended or truncated with respect to the situation in the template. The sequence conservation between target and template in the anchor region is taken into account as well.

During CASP7, as mentioned in the previous section, the anchor groups have been often positioned further away from the gap as necessary resulting in longer fragments which are more difficult to model. As it can be seen in Chapter 3.3, the loop modelling accuracy rapidly drops for loops longer than 7 residues. At CASP7, often different anchor group combinations have been used for loop prediction if the situation was not clear. In most cases, this approach resulted in a set of alternatives models from which the best ones were selected based on the predicted model energy. But in a few cases, a selection of the anchor groups and the corresponding loop was made based on a comparison of the loop ranking output files: if for one anchor group combination only loops with similar scores are found on the top ranks but for the other combination a loop with a considerably better score than the rest was found on the first rank (*e.g.* because the fragment origins from a homologous structure), the later was chosen for all models. Loops with significantly higher scores than the rest of the fragments are potentially promising candidates. Thus, inspecting different alternative anchor groups seems to be indeed a reasonable approach especially for knowledge-based loop prediction protocols (see Chapter 3.4.2 for a more detailed discussion).

3.1.5.3 Modelling of chain ends

Chain ends are often highly flexible, particularly if they do not establish regular secondary structures. But if the chain ends are not flexible, a methods is needed

which can model these regions. Most of the existing loop prediction programs are not able to model chain ends since they are specialised on loops. Futhermore, the majority of knowledge-based loop prediction programs use the RMSD between the anchor group residues and the terminal fragment residues after fitting as the main scoring function term which cannot be used here. In this situation, only one anchor group is given and the RMSD of all fragments after fitting will be more or less the same. In the method described in this work, the ranking is performed based on a statistical potential scoring function investigating the interactions with the structural environment (see Methods).

At the beginning of the CASP7 prediction season, our pipeline was not able yet to model chain end (only loops, where two anchor groups are given). As a consequence, most of our models show a low target coverage which strongly influenced the ranking based on GDT_TS. As it can be seen from the overview table on page 93, missing chain ends were the main reason why some of the models did not score better. In a ranking based on all-atom RMSD, two third of the models designated as model1 ranked among the top 25 predictions (among approximately 130 groups).

Chain ends are modelled with an adapted version of the loop prediction routine: fragments from the database are fitted on *one* anchor group which results, as a consequence of the missing distance constraints (*i.e.* C*alpha* distance of the endpoints and RMSD of the anchor groups), in an enormous amount of possible candidates (actually all fragments of the given length present in the database). The following procedure was used in order to reduce the number of possible candidates:

- The clash filter which searches for overlaping van der Waals spheres between the fragment backbone atoms and the rest of the protein removes the majority of the candidate fragments.
- Only a certain fraction of the fragments is retained based on the "goodness of fit" on the anchor region (*i.e.* RMSD over the anchor group atoms). Three fitting strategies have been implemented: fitting on the backbone atoms of one residue or two residues and, alternatively, fitting on three consecutive Calpha atoms. Fitting of more than one residue turned out to be the best strategy for this task. If, for example, a terminal helix needs to be extended, fitting on more than one residues increases the chance that the helix fragment has the right orientation.

- In order to restrain the number of possible fragments in the initial selection, regular expression pattern on the sequence or on the secondary structure constitution of the fragment can be defined. For example for the extension of a helix element, only fragments consisting of an initial helix segment are needed. In analogy, for example in the presence of a conserved proline, only fragments with proline at the given position are retrieved from the database. This allows to reduce the number of candidates by several order of magnitude and therefore greatly improves the run time and the accuracy of the prediction.

- Ranking has been performed with the same scoring functions as for loop prediction.

- Furthermore, comparing the sequence conservation of the top scoring fragments (*i.e.* the agreement between the sequence of the segment in the target and the sequence of the original fragment) as well as a visual inspection of the top scoring solutions in a molecular graphics viewer such as Pymol provide additional evidence for the final selection.

The structure prediction of the N-terminal chain end in target T0373 is described here exemplarily. As it can be seen from the alignment extract in Figure 3.17, the target contains an insertion with respect to the template and all three secondary structure programs indicate that the terminal helix present in the template (last line) is most probably extended in the target. For a detailed describtion of the single data lines, visite Methods on page 42.

```
          1        10        20        30        40        50        6
T0373     MPTNQDLQLAAHLRSQVTTLTRRLRREAQADPVQFSQLVVLGAIDRLGGDVTPSELAAAE
1jgsA     L-FNEIIPLGRLIHMVNQKKDRLLNEYLSPLDITAAQFKVLCSIRCAAC-ITPVELKKVL
conserv   :I *  : *   ::   :   * *   :::  ::: *: **: *  : :I:** **  :
conf      98862056888899899999889886414268885788889998743788777789864
consensus CCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHH
psipred   CCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHH
SSpro     CCCHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHH
phd       CCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHH
dssp      C CCCCCCHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHCC ECHHHHHHHH
```

Figure 3.17: Extract of the alignment for target T0373 (N-terminal chain end).

A closer inspection of the target protein family revealed that the leucine at position

nine was rather conserved. Therefore, the following regular expressions have been used for the selection (the underscore stand for an arbitrary character):

```
regular expression for SSE:                CCC__HHHHH
regular expression for the sequence:       ________L_
```

The restaint selection resulted in an initial set of 8764 fragments from which all loops which a anchor group RMSD Z-score above one standard deviation are removed. The top 10 fragments with the lowest energy are shown in Figure 3.18. The 10 fragments show a high structural diversity although they have a comparable energy. This reflects the uncertainities associated with modelling of chain ends.

Figure 3.18: Structural diversity among the 10 top scoring fragments for the N-terminal chain end of T0373.

As a consequence of the correct assumption concerning the secondary structure constitution of the target structure (the experimental structure indeed contains 5 additional residues in helix conformation as compared to the template), the N-terminal chain end of target T0373 was modelled very accurately as it can be seen from the superposition of target and template in Figure 3.19 and this is probably the main reason why this model was the second best prediction at CASP7 (among the models designated as model1).

In the absence of secondary structure elements, modelling of chain ends can be a very difficult task because of the vast amount of possible conformations and the limited

ability of energy functions to identify the native conformation. Since chain ends are less constraint by the structural environment as compared to for example regions in the structural core. Their conformations are to a greater extent determined by the sequence itself and less by local (in sequence) and non-local structural constraints. Therefore, fragments from the database having a similar amino acid constitution and origin from similar environments (*i.e.* also chain ends) can be promising candidates. The fragment database described in this work contains an entry for each fragment specifing whether the fragment is part of a chain end. Additionally, information about the solvent exposure in the original environment is stored. This information could potentially be used in this context.

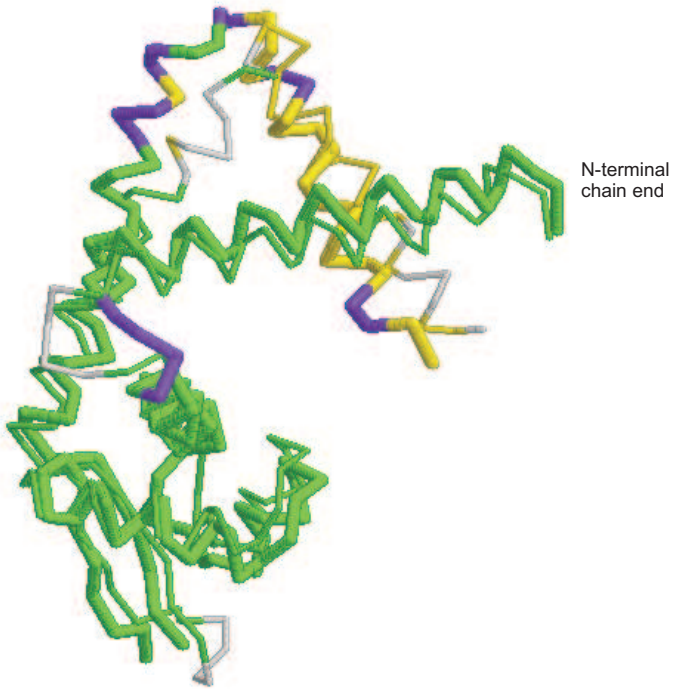


Figure 3.19: Superposition of model and experimental structure of target T0373: The N-terminal chain end has been modelled very accururately.

3.1.5.4 Modelling of sidechains

While establishing the modelling pipeline, it has been decided to use a conservative approach for sidechain modelling by leaving the sidechains conformation of conserved residues (*i.e.* identical residues between target and template in the alignment) untouched and to only re-model sidechains of residues differing between target and template and of course residues of regions which have been remodelled (*i.e.* loops and chain ends). The SCWRL software [31] was used in order to calculate the sidechain conformations. This turned out to be a good strategy: "Group 191 (Schomburg-group) has the best results for rotamer accuracy, but it should be noted that this group only submitted predictions for 6 of the 28 target domains" [169]. Figure 3.20 shows a comparison of the sidechain accuracy of the top performing groups in the category high-accuracy template-based modelling (HA-TBM). The fraction of sidechains (χ_1 angle in Figure a), χ_1 and χ_2 in Figure b)) modelled within 30 degree from the native conformation have been investigated and Z-scores over all groups are calculated in order to compare the performance.

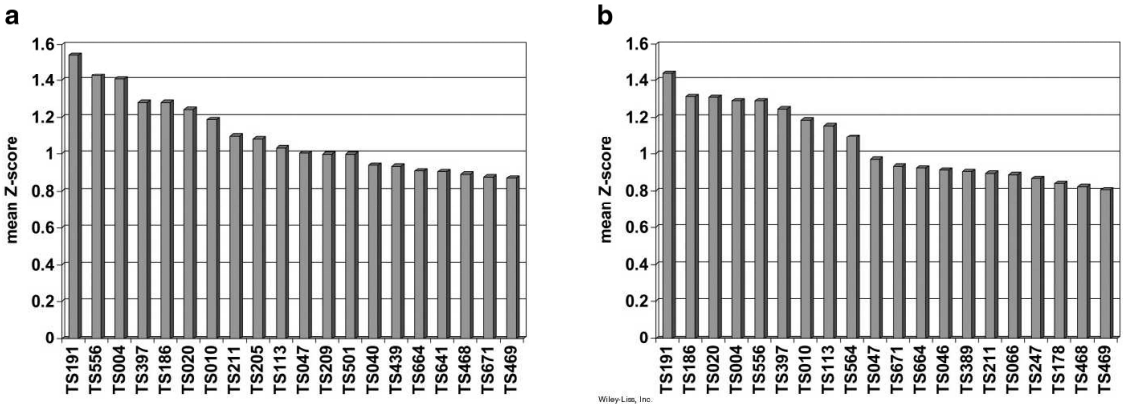


Figure 3.20: Accuracy of sidechain modelling (Z-scores) of sidechain torsion angle chi-1 (a) and over chi-1 and chi-2 (b) (Schomburg-group: TS191) [169].

The targets of the high-accuracy template-based modelling (HA-TBM) category are defined in the following manner:

- A suitable template was present in the PDB with LGA-S $>$ 80 (LGA-S is a sequence-idependent measure of structural similarity).

- At least one prediction with GDT_TD > 80 was submitted to CASP7.
- A total number of 24 HA-TBM targets were evaluated.

As mentioned above, only 6 of the total 28 HA-TBM domains have been processed which should be taken into account when comparing the performance with other groups. Nevertheless, since we did not just pick the easiest targets from the 24 possible one but could not solve all of them due to time constraints, the picture would be more or less the same. Beside the fact that SCWRL did a very good job, the decision to only remodel sidechains of non-conserved residues seems to be the crucial factor since the majority of the groups most probably used SCWRL as well. Using as much information of the templates as possible is indeed one of the lessons which has been learnt during the last CASP rounds. Currently, still no group is able to consistently produce models better than the best template although there are an increasing number of cases where improvement over the templates are shown [49].

3.2 Model quality assessment

Assessing the quality of model is a vital step in protein structure prediction as pointed out in the Introduction (see Chapter 1.2.4). Depending on the method and on the modelling difficulty, usually a certain amount of alternative models is generated ranging from a few alternative models (*e.g.* in comparative modelling) up to thousands or ten thousands of models (*e.g.* for *ab initio* methods based on fragment assembly in this context). A scoring function (typically called model quality assessment program) is needed which is able to discriminate between good and bad models and can potentially select the best model.

As a part of the modelling pipeline described above, a composite scoring function based on 3 statistical potential terms as well as two other terms has been developed [16]. The scoring function was named QMEAN which stands for Qualitative Model Energy ANanlysis. An early version of QMEAN was used at the CASP7 experiment in order to rank our own models and to identify the best models for submission. Additionally, we participated in the *quality assessment category* [49] which was newly introduced in CASP7 in order to test the performance model quality programs. The predictors were asked to estimate the quality of all models predicted by automatic servers. Motivated by the good results (we were among the top scoring methods solely relying on the coordinates of a single model), we decided to further extend and optimise the scoring function. The performance of the optimised scoring function (*i.e.* QMEAN) are described in the following.

The section is structured as follows: First, the results of the optimisation of the different statistical potentials terms is presented. Afterwards, it is descibed how the terms are combined in order to build the final composite scoring function QMEAN. In the subsequent section, QMEAN is compared to five well-established model quality assessment programs using several comprehensive test sets. The section ends with a concluding discussion of the results obtained on the different test sets and with a description of areas of possible future improvements.

3.2.1 Optimisation of the statistical potentials

All statistical potentials were extracted from a non-redundant protein data set of 1,471 high-resolution structures from the Protein Data Bank (PDB) [18]. The selection of the structures was performed with the PISCES server [236] and additional quality filters were applied as described in Methods (see page 58). The parametrisation of the different potentials as well as the optimisation of the weighting factors for the combined potential were both performed on the CASP6 decoy set by analysing the regression between the GDT_TS score of the models and the predicted score provided by the energy function. The CASP6 training set consists of all models submitted to CASP6 with a GDT_TS score above 20. Models with a score below 20 can be considered as more or less random and are therefore useless for training purposes.

For the purpose of providing an overview, Table 3.6 shows a short description of all scoring function terms mentioned in this section and the different versions of QMEAN which were built in order to assess the influence of the two agreement terms. In the following, QMEAN, unless specified with an index, always indicates the original scoring function consisting of 5 terms (*i.e.* QMEAN5).

For the three statistical potentials entering the QMEAN function a variety of alternative implementations have been investigated. The Pearson's correlation coefficients for the different implementations of the statistical potentials as well as for the agreement terms are given below (Table 3.7-3.11).

The correlation between the score from different implementations of the residue-level pairwise interaction potential and the GDT_TS score are shown in Table 3.7. The data underline the superior performance of the potentials based on Cβ atoms compared to the Cα implementation. Deriving the interaction potentials in a secondary structure specific manner further improves the correlation whereas taking into account solvent accessibility does not add any value (see Chapter 3.2.4.5 in the discussion section). In the secondary structure specific implementation, the contacts of helix, strand, and loop residues are counted separately, which seems to capture some characteristic features of the environment of residues belonging to the different secondary structure states. The final implementation of the residue-level distance-dependent pairwise potential is

Table 3.6: Short description of the terms and their combinations used in this in this work.

scoring function	description
torsion single	Ordinary torsion potential based on phi and psi propensities of single amino acids. Bin size: 10 degree
torsion 3-residue	Extended torsion potential over 3 consecutive residues. Bin sizes: 45 degree for the center residue, 90 degree for the 2 adjacent residues
pairwise Cα / pairwise Cβ	Residue-specific pairwise distance-dependent potential using Cα or Cβ atoms respectively as interaction centers . Range 3...25 Å, step size: 0.5 Å
pairwise Cβ/SSE	In analogy to pairwise Cβ, but a secondary structure specific implementation was used both for the derivation and application of the potential.
solvation Cβ	Potential reflecting the propensity of a certain amino acid for the a certain degree of solvent exposure based on number of Cβ atoms within a sphere of 9 Åaround the center Cβ.
SSE X	Agreement between the predicted secondary structure of the target sequence (using method X, or consensus of 3 methods) and the observed secondary structure of the model as calculated by DSSP. QMEAN uses X=PSIPRED
ACCpro	Agreement between the predicted relative solvent accessibility using ACCpro (2 states buried/exposed) and the relative solvent accessibility derived from DSSP (>25% accessibility => exposed)
QMEAN3	weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ
QMEAN4	weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED
QMEAN5	weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED, ACCpro

based on Cβ atoms as interaction centers and the radial distribution between 3 and 25 Å (bin size 0.5 Å) is taken into consideration (with secondary structure specificity).

An all-atom pairwise potential was established which investigates the interactions between all 167 atom types occuring in proteins (*i.e.* each non-hydrogen atom in the 20 amino acids belongs to a different atom type). As for the residue-level potentials, the secondary structure specific implementation results in a better correlation as compared to the normal one (see Table 3.8). All "interactions" in the interval from 3 to 20 Å (bin size 0.5) are taken into account. Interestingly, ignoring all contacts closer than 3 Å results in a considerably better correlation to GDT_TS. In this way, hydrogen bonds are completely ignored since the distance between the two atoms participating in a

Table 3.7: Correlation between GDT_TS and the residue-level pairwise potential on the CASP6 training set.

implementation	C_α	C_β	$C_{\beta,SSE}$	$C_{\beta,SSE,ACC}$
range: 0-20 Å, bin size: 0.5 Å	-0.272	-0.365	-0.454	-0.473
range: 0-25 Å, bin size: 0.5 Å	-0.365	-0.445	-0.514	-0.528
range: 0-30 Å, bin size: 0.5 Å	-0.430	-0.498	-0.531	-0.539
range: 3-20 Å, bin size: 1 Å	-0.452	-0.532	-0.598	-0.598
range: 3-25 Å, bin size: 1 Å	-0.520	**-0.562**	**-0.608**	**-0.608**
range: 3-20 Å, bin size: 0.5 Å	-0.457	-0.519	-0.582	-0.587
range: 3-25 Å, bin size: 0.5 Å	**-0.521**	-0.558	-0.601	-0.603
range: 3-20 Å, bin size: 0.2 Å	-0.444	-0.507	-0.546	-0.557

Table 3.8: Correlation between GDT_TS and all-atom pairwise potential on the CASP6 training set.

implementation	all-atom	all-atom$_{SSE}$
range: 0-15 Å, bin size: 0.5 Å	-0.247	-0.286
range: 0-20 Å, bin size: 0.5 Å	-0.302	-0.353
range: 3-15 Å, bin size: 0.5 Å	-0.471	-0.536
range: 3-18 Å, bin size: 0.5 Å	-0.519	-0.581
range: 3-20 Å, bin size: 0.5 Å	-0.540	**-0.600**
range: 3-15 Å, bin size: 0.2 Å	-0.462	-0.519
range: 3-20 Å, bin size: 0.2 Å	-0.557	-0.589

hydrogen bond is typically below 3 Å. Given the fact that hydrogen bonds are one of the main contributors to the overall protein stabilty, this may look strange at first sight. But it has to be taken into account that models, and not exact experimental structures are analysed. Especially for very coarse models (*e.g.* model from *ab initio* structure prediction), not the exact location of the single atoms shall be investigated but the overall correctness of the fold. Therefore, the high contribution of the hydrogen bonding term would potentially hide the signal of the other non-covalent energy contributions. Including hydrogen bonding in the scoring function would potentially favour models with more secondary structure elements (since these are stabilised by hydrogen bonds). The energy function would be very sensitive concerning small perturbations in the location of the atoms with the consequence, that a small shift of *e.g.* 0.5 Å away from the ideal hydrogen bonding distance would result in a dramatic increase in the interaction energy.

In the final version of QMEAN, the all-atom potential has not been integrated. Over the entire range of modelling difficulty, the residue-level potential performs better than the all-atom implementation. A comparison of the performance of the all-atom interaction potential on models from different CASP7 categories suggests that the strength of this potential is the assessment of template-based models and not of imprecise models from the free modelling category. An optimal integration of both potentials described above using machine learning algorithms (*i.e.* support vector machine or neural network) is currently under development.

For the solvation potential, which reflects the propensity of an amino acid to be found buried in folded proteins, the solvent accessiblity is approximated by counting the number of Cβ within 9 Å around the Cβ of a given amino acid. As it can be seen from Table 3.9, sphere radii of 9 and 12 Å result in equally good correlations and it has been decided to use the smaller radius since the same information content seems to be captured.

Table 3.9: Correlation between GDT_TS and residue-level solvation potential on the CASP6 training set.

implementation	**C$_\alpha$**	**C$_\beta$**
radius of sphere: 5 Å	-0.200	-0.153
radius of sphere: 6 Å	-0.431	-0.426
radius of sphere: 7 Å	-0.525	-0.551
radius of sphere: 8 Å	-0.542	-0.562
radius of sphere: 9 Å	**-0.559**	**-0.568**
radius of sphere: 10 Å	-0.541	-0.554
radius of sphere: 11 Å	-0.552	-0.559
radius of sphere: 12 Å	**-0.559**	**-0.569**
radius of sphere: 13 Å	-0.552	-0.562
radius of sphere: 14 Å	-0.547	-0.557

All chains present in the coordinate files have been taken into account in order to calculate the solvent accessibility. A potential improvement by considering the biological units is disscused later in Chapter 3.2.4.5.

A coarse-grained torsion angle potential using the phi/psi angles of three consecutive residues was developed. The bin sizes are 45 degrees for phi and psi of the center

residue and 90 degrees for the neighbouring torsion angles. Table 3.10 underlines the considerably better correlation of the 3-residue torsion angle potentials with the GDT_TS score as compared to the regular single residue torsion angle potential. For comparison purposes, the performance of the single residue torsion potential is shown.

Table 3.10: Correlation between GDT_TS and torsion potential over 3 residues on the CASP6 training set.

implementation	correlation
bin size central residue: 30°, bin size adjacent residues: 45°	-0.498
bin size central residue: 30°, bin size adjacent residues: 90°	-0.515
bin size central residue: 45°, bin size adjacent residues: 45°	-0.511
bin size central residue: 45°, bin size adjacent residues: 90°	**-0.517**
bin size central residue: 90°, bin size adjacent residues: 90°	-0.504
single residue torsion potential: 10°	-0.350

Table 3.11: Correlation between GDT_TS and agreement terms on the CASP6 training set.

description	correlation
agreement DSSP - PSIPRED	**-0.561**
agreement DSSP - ProfSec	-0.514
agreement DSSP - SSpro	-0.543
agreement DSSP - consensus (PSIPRED, ProfSec, SSpro)	-0.555
agreement DSSP - ACCpro	**-0.529**

Two terms reflecting the agreement between predicted features of the target sequence and calculated features from the model enter the final version of QMEAN. A term called "SSE PSIPRED" in the further course of this work describes the agreement between the predicted secondary structure of the sequence by PSIPRED [103] and the observed secondary structure from the model as calculated by DSSP [107]. Two further secondary structure prediction programs have been investigated (ProfSec [177] and SSpro [35]) as well as the use of a consensus of the three, but did not result in a better regression. The solvent accessiblility agreement term is based on the predicted solvent accessibilty of ACCpro [35] and the calculated of the model by DSSP. In the composite scoring function (QMEAN5), both terms lead to a significant improvement in the performance as compared to the version solely based on statistical potentials (see Table 3.12 in the next section).

3.2.2 QMEAN: Generation of the composite scoring function

Table 3.12 contains regression coefficients achieved in a regression of the models GDT_TS scores and the QMEAN scores. Two different regression schemes were investigated: A direct correlation of the scores (Pearson's correlation coefficient) and a rank correlation (Spearman's *rho*) in the hope of taking into account a possible non-linear relationship. As an alternative, the scores are transformed into Z-scores by comparing the given model to 1000 other models with the same structure but randomly shuffled sequences. Shuffling the order of the residues has been shown [137] to work almost as good as randomising the structure as originally proposed by Sippl [198]. Furthermore, two different strategies for the optimisation of the weighting factors have been investigated: First, an optimisation of the regression on a target-specific basis by maximising the average of the regression coefficients achieved on the individual targets and second, a global approach in which the regression is optimised by using all models from all the targets at once.

Table 3.12: Absolute values of the Pearson correlation coefficients obtained in a regression of the GDT_TS score against the predicted score.

	Pearson's correlation coefficient				**Spearman's c. c.**	
scoring function	**global**	**global/ Z-score**	**target averaged**	**target average/ Z-score**	**target averaged**	**target average/ Z-score**
torsion single	0.35	0.39	0.25	0.3	0.23	0.24
torsion 3-residue	0.52	0.5	0.35	0.39	0.32	0.31
pairwise Cα	0.54	0.57	0.42	0.54	0.37	0.42
pairwise Cβ	0.57	0.59	0.47	0.56	0.43	0.46
pairwise Cβ/SSE	0.61	0.6	0.49	0.58	0.45	0.48
solvation Cβ	0.58	0.55	0.5	0.52	0.46	0.43
SSE PSIPRED	0.57	0.57	0.52	0.54	0.48	0.48
SSE ProfSec	0.53	0.53	0.49	0.52	0.45	0.45
SSE SSpro	0.56	0.56	0.5	0.52	0.45	0.45
SSE consensus	0.57	0.57	0.51	0.53	0.46	0.46
ACCpro	0.53	0.53	0.47	0.51	0.47	0.47
QMEAN 3terms	0.66	0.64	0.56	0.58	0.52	0.52
QMEAN 4terms	0.71	0.69	0.62	0.64	0.57	0.58
QMEAN 5terms	0.72	0.69	0.64	0.65	0.59	0.6

The regression coefficients achieved for the different scoring function terms and their combinations do not differ much between the six optimisation strategies and all show the same tendency. QMEAN5, which is a linear combination of five terms (see Table 3.4), consistently achieves the highest regression coefficients for all optimisation strategies, directly followed by QMEAN4. QMEAN3, consisting only of statistical potential terms, shows a slightly worse correlation but is still better than any other single term. A Pearson's correlation coefficient of 0.72 was observed for QMEAN5 in the global approach in which the regression is optimised over all models of all targets at once. The scatter plot in Figure 3.21 shows a clear trend but also the presence of some outliers.

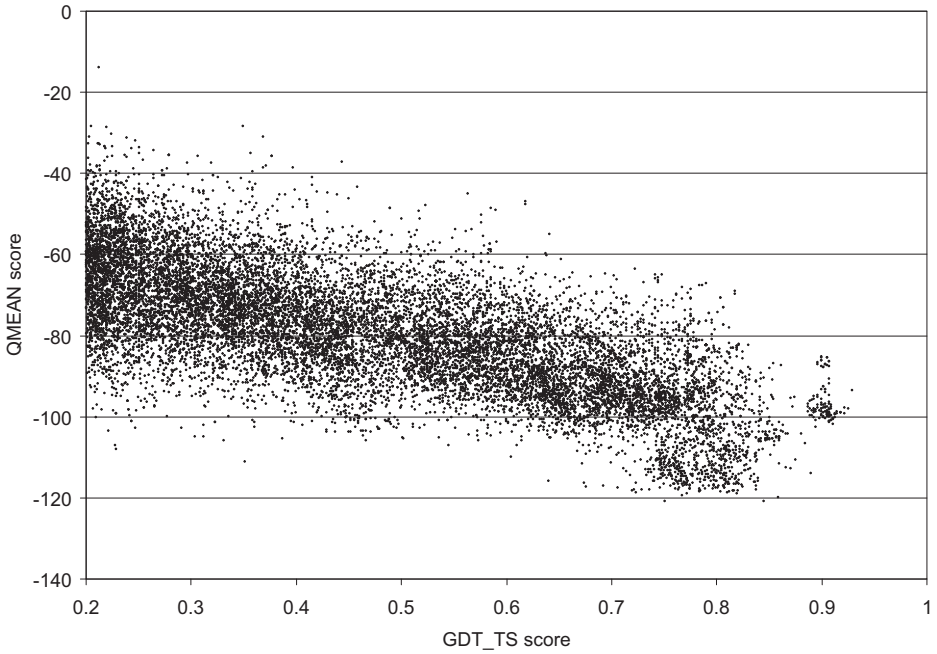


Figure 3.21: Correlation between GDT_TS and the composite score (QMEAN5) on the models in the CASP6 traing set. Models with GDT_TS < 0.2 are not considered.

The weighting factors achieved in the two target-specific approaches (Spearman and Pearson) are quite similar to each other. In comparison to those in the global strategy, lower weights were assigned for the torsion and pairwise term (*data not shown*). In any case, the performance differences when applying the weights of the six strategies to the decoy sets described in the next two sections are overall negligible.

For the sake of simplicity, the weights of the global optimisation strategy are used throughout:

$$QMEAN5 = 0.3 * Score_{torsion\ 3-residue} + 0.17 * Score_{pairwise\ C\beta,SSE} + 0.7 * Score_{solvation\ C\beta} + 80 * Score_{SSE\ PSIPRED} + 45 * Score_{ACCpro} \quad (3.1)$$

Table 3.13 shows the cross-correlation between QMEAN and its component terms as well as some additional terms for comparison purposes. It can be seen that the secondary structure specific implementation of the pairwise interaction potential does not have a significantly higher cross-correlation to any of the other terms than the regular one.

Table 3.13: Cross-correlation analysis of the terms entering the combined score (QMEAN) and some selected scores for comparison. The Pearson's correlation coefficients are based on the global optimisation strategy without Z-scores.

	torsion single	**torsion 3-residue**	**pairwise Cβ**	**pairwise Cβ/SSE**	**solvation**	**SSE PSIPRED**	**ACCpro**	**QMEAN3**	**QMEAN5**	**GDT_TS**
torsion single	1	0.81	0.41	0.43	0.34	0.35	0.31	0.59	0.54	-0.35
torsion 3-residue	0.81	1	0.58	0.6	0.5	0.48	0.41	0.78	0.73	-0.52
pairwise Cβ	0.41	0.58	1	0.97	0.71	0.43	0.58	0.89	0.83	-0.57
pairwise Cβ/SSE	0.43	0.6	0.97	1	0.72	0.44	0.62	0.92	0.85	-0.61
solvation	0.34	0.5	0.71	0.72	1	0.48	0.62	0.87	0.81	-0.58
SSE PSIPRED	0.35	0.48	0.43	0.44	0.48	1	0.42	0.54	0.81	-0.57
ACCpro	0.31	0.41	0.58	0.62	0.62	0.42	1	0.65	0.64	-0.53
QMEAN3	0.59	0.78	0.89	0.92	0.87	0.54	0.65	1	0.93	-0.66
QMEAN5	0.54	0.73	0.83	0.85	0.81	0.81	0.64	0.93	1	-0.72
GDT_TS	-0.35	-0.52	-0.57	-0.61	-0.58	-0.57	-0.53	-0.66	-0.72	1

The solvation potential shows a relatively high cross-correlation to the pairwise potentials which can be assigned to the similarity of their implementation. The correlation to the ACCpro term is lower than could be expected.

The integration of the SSE PSIPRED terms results in an increase of the regression coefficient of at least 0.05 in all the optimisation strategies (Table 3.12) while having no

noticeable cross-correlation to any of the other terms and QMEAN3 (Table 3.13). The ACCpro term, describing the agreement between the predicted and observed solvent accessibility, only leads to a minor increase of the regression coefficients of QMEAN5. ACCpro shows a cross-correlation around 0.6 to the distance-dependent potentials and the solvation potential and a comparison of the correlation to QMEAN3 and QMEAN5 would suggest that ACCpro does not add much value to the combined score. However, Table 3.16 proves that the opposite is true: ACCpro shows a very good performance according to the enrichment quality measures and is responsible for the constant improvement in all quality measures of QMEAN5 over QMEAN4.

According to Table 3.13, a major part of the discriminatory power of QMEAN3 can be assigned to the pairwise Cβ/SSE and to the solvation potential. The correlation of the 3-residue torsion angle potential is still rather high (regression coefficient 0.78). The secondary structure agreement term shows a significantly higher correlation to QMEAN5 than ACCpro.

3.2.3 QMEAN: Comparison with other methods

Three comprehensive test sets were used in order to assess the performance of QMEAN and compare it to other state-of-the-art methods. The first test set consists of three standard decoy sets from Decoys 'R' Us [182] which have been frequently used in literature in order to test scoring functions. Decoys are computer generated conformations of protein sequences that possess some characteristics of native protein structures, but are not biologically real. The second test set consists of conformations generated during a molecular dynamics (MD) simmulation and allow a comparison of QMEAN with a molecular mechanics (MM) force field. The third test set consists of all server models submitted to CASP7 and represents the same databasis which has been used for the quality assessment category of the last CASP [49].

3.2.3.1 Performance on three standard decoy sets

In order to compare the performance to several well-established statistical potentials, QMEAN was tested on three standard decoy sets from Decoys 'R' Us [182]. As

Table 3.14: Comparison of QMEAN with other methods in the performance of selecting the native structure in some standard decoy sets from Decoys 'R' us.

	4state_reduced		lattice_ssfit		LMDS	
	rank1[a]	**Znat**[b]	**rank1**	**Znat**	**rank1**	**Znat**
ProQ	5/7	4.1	7/8	12.1	4/10	3.7
Errat	1/7	2.5	3/8	5.1	5/10	3.1
ProsaII	5/7	2.7	8/8	5.6	6/10	2.5
Verify3D	4/7	2.6	7/8	4.5	2/10	1.4
SNAPP	3/7	2.6	5/8	3.5	2/10	1.1
AKBP	7/7	3.2	8/8	6.6	3/10	−0.5
DFIRE	6/7	3.5	**8/8**	**9.5**	7/10	0.9
RAPDF	7/7	3	8/8	7.2	3/10	0.5
FRST	**7/7**	**4.4**	8/8	6.7	6/10	3.5
torsion 3-residue	7/7	3.6	6/8	5	**7/10**	**3.7**
pairwise Cβ/SSE	3/7	2	7/8	5.1	1/10	0.4
solvation	0/7	1.6	3/8	3.1	0/10	1.1
SSE PSIPRED	0/7	1.6	7/8	5.4	2/10	1.3
ACCpro	1/7	2	5/8	3.7	3/10	1.9
QMEAN3	4/7	2.7	8/8	6.2	2/10	2.3
QMEAN4	3/7	2.4	8/8	7.5	4/10	2.3
QMEAN5	4/7	2.5	8/8	7.7	6/10	2.7

[a] rank1: Number of decoy set in which the native structure was found on the first rank.
[b] Znat: Z-score of the native structure compared to the ensemble of structure in the decoy set.

can be seen from Table 3.14, the 3-residue torsion angle potential shows the overall best performance in selecting the native structure and outperforms all other terms of QMEAN as well as all QMEAN versions. Except for the `lattice_ssfit` decoy set, the torsion angle potential also produces the highest Znat scores.

The pairwise potential performs comparably well on *lattice_ssfit*, shows a moderate performance on *4state_reduced* and fails on *LMDS*. The solvation potential only produces reasonable Z-scores on the *lattice_ssfit* but fails completely on the other two sets. Comparing the performance of QMEAN5 on the 3 decoy sets, it seems that QMEAN5 performs best on *lattice_ssfit*. In general the performance of QMEAN5 is comparable to the other methods taking into account the fact that QMEAN has been

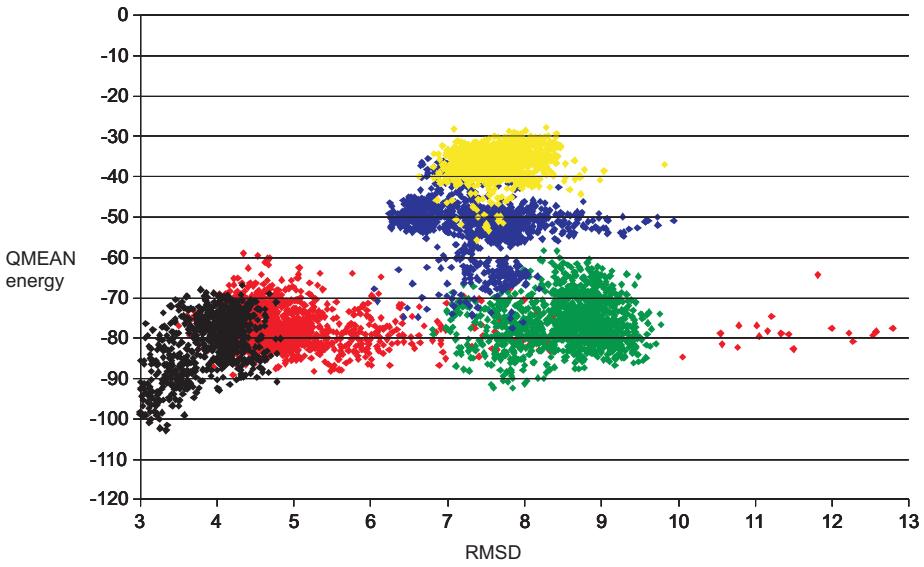


Figure 3.22: Correlation between GDT_TS and the composite score (QMEAN5) on the models of the molecular dynamics simulation decoy set of Fogolari *et al.* [81].

trained for model quality assessment and not specifically for the task of identifying native structures. The advantage of QMEAN5 as a combined scoring function over energy functions based on a single term is the decreased chance to fail on some decoy sets generated based on a specific method. Although the data basis is too sparse for well-founded conclusions, Table 3.14 suggests that the performance of a certain scoring function is dependent on the decoy set. More precisely, how a given decoy set has been built appears to allow some terms to perform better on one decoy set than on another.

3.2.3.2 Performance on a molecular dynamics decoy set

The decoy set generated by Fogolari and co-workers [81] consists of 6,255 snapshots from 5 different molecular dynamics simulations of the thermostable subdomain from the chicken villin headpiece. Since one simulation started from the native structure and the other 4 from alternative minimised conformation, this yields a wider range of RMSD values compared to the previously mentioned decoy sets which typically have only few conformtion close to native. The other advantage is that it allows a direct comparison with molecular mechanics force fields.

Table 3.15: Comparison of QMEAN and its terms with three molecular mechanics energy functions, a contact potential and FRST.

scoring function	$logP_{B1}$ [a]	$logP_{B10}$ [a]	**F.E.** [b]	r^{2} [c]	**RMSD** [d]
contact	-1.08	-1.08	13.8	0.62	3.03
FRST	-1.38	-1.94	23.2	0.48	2.61
MM[e]	-0.25	-1.39	10.6	0.21	7.45
MM/GBSA[e]	-1.71	-2.02	29.6	**0.66**	2.4
MM/PBSA[e]	-1.79	-2.02	23.2	0.58	2.35
QMEAN3	-1.5	**-3.5**	36.5	0.53	2.52
QMEAN4	-1.71	-2.8	90.2	0.56	2.4
QMEAN5	-1.51	**-3.5**	88	0.57	2.51
torsion 3-residue	-1.26	-2.8	58.4	0.57	2.71
pairwise Cβ/SSE	-1.02	-1.41	35.5	0.64	3.34
solvation	-0.32	-0.98	6.1	0.2	7.15
SSE PSIPRED	-1.32	-1.32	**91.2**	0.55	2.58
ACCpro	**-3.5**	**-3.5**	63	0.5	**1.84**

[a] $logP_{B1}$ and $logP_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the ten best-scoring models, respectively.

[b] F.E. stands for fraction enrichment.

[c] Person's correlation coefficient

[d] RMSD of the structure with the lowest score assigned by the energy function.

[e] Scoring by a molecular mechanics (MM) force field by using the Generalized Born surface area (GBSA) or the Poisson-Boltzmann surface area (PBSA) method for solvation effects.

As can be seen from Figure 3.22, QMEAN consistently assigns low energies to the near-native conformations of the simulation starting from the native structure (colored in black). Especially the decoys from the native simulation show a clear correlation between the RMSD and the score predicted by QMEAN5. Although the native structure was not predicted to have the lowest energy, several conformations around 2 Å RMSD get quite low energies. This is also reflected by the excellent $\log P_{B10}$ value of QMEAN5 as shown in Table 3.15. A description of the quality measures is given in the footer of Table 3.15 and more detailed in Methods on page 64ff.

The solvent accessibility agreement term seems to be quite good in identifying near-native structures and to a certain extent also the torsion angle potential over three residues, as reflected by the low logPB10 value and the high fraction enrichment score. The secondary structure agreement term produces a fraction enrichment of over 90%

which indicates that there were no major changes in secondary structures during the simulation starting from the native structure. The RMSD values of the conformation with the lowest score are more or less the same for all three QMEAN versions whereas ACCpro is able to pick the second best conformation. The solvation potential produces bad results across all quality measures. In comparison to the three versions of molecular mechanics (MM) energy functions, QMEAN shows comparable correlation coefficients and logPB1 values but performs significantly better in the enrichment of near-native solutions.

3.2.3.3 Performance on the CASP7 decoy set

A different, and perhaps more realistic, test case is presented by the decoys from the CASP7. In Table 3.16 QMEAN and its component scoring function terms are compared to five widely-used model quality assessment programs (MQAPs). The following executable programs could be downloaded from the CAFASP4 website [e]: Modcheck [162], RAPDF [184], FRST [215] and ProQ [233]. DFIRE [249] was requested from the author. ProQ was executed both with and without PSIPRED secondary structure prediction.

Table 3.16 shows the average performance of the methods over all targets using different quality measures. Most of the quality measures have been previously introduced and described [225, 237], but a detailed definition can be found in Methods on page 64. The last three columns describe the scoring functions ability in identifying the native structure out of the ensemble of models for a specific target whereas all other measures describe different aspects of model quality assessment. The opposite algebraic sign of Modcheck and ProQ observed for the Pearson's correlation coefficients and for the Znat scores can be ascribed to the fact that these two tools use an inverse scaling compared to the other scoring function by assigning the highest scores to the best models.

The statistical significance of the performance differences between the methods was validated using the the 2-sided t-test on paired samples (see Methods on page 66) in analogy to the method used in the assessment of CASP4 [131]. A 95% confidence level was used and the corresponding results are summarised in Figure 3.23. White squares

[e] `http://www.cs.bgu.ac.il/~dfischer/CAFASP4/`

Table 3.16: Performance of different scoring functions in predicting the quality of the server models submitted for all 95 targets of CASP7. Comparison of QMEAN with other well-known model quality assessment programs.

	regression[a]		**enrichment**[b]		**best predicted model**[c]			**best GDT_TS model**[d]			**native structure**[e]	
Method	r^2	*rho*	*F.E.*	$E_{15\%}$	r_{10}	$logP_{B1}$	$logP_{B10}$	ΔGDT_TS	r_1	r_{10}	Z_{nat}	r_1	r_{10}
Modcheck	0.64	0.59	0.33	2.7	17	-0.7	-1.67	-0.18	**6**	27	1.99	47	69
RAPDF	-0.5	0.5	0.31	2.44	17	-0.91	-1.67	**-0.08**	4	17	-2.09	55	77
DFIRE	-0.39	0.53	0.32	2.59	19	-0.93	-1.68	**-0.08**	5	18	-1.25	**59**	72
ProQ	0.36	0.26	0.13	1.22	5	-0.32	-0.99	-0.22	0	**6**	1.51	9	32
$ProQ_{SSE}$	0.54	0.43	0.19	1.71	8	-0.51	-1.21	-0.16	2	11	1.76	14	42
FRST	-0.57	0.53	0.3	2.36	21	-0.91	-1.74	-0.09	6	22	-2.41	56	72
QMEAN3	-0.65	0.58	0.33	2.57	16	-0.8	-1.83	-0.12	1	35	-2.27	**59**	75
QMEAN4	-0.71	0.63	0.38	2.76	28	-1.02	-1.9	**-0.08**	5	39	-1.86	55	69
QMEAN5	**-0.72**	**0.65**	**0.4**	**2.9**	**30**	**-1.05**	**-1.94**	**-0.08**	**6**	**40**	-1.89	56	71
torsion single	-0.44	0.39	0.22	1.76	6	-0.6	-1.5	-0.13	0	13	-2.09	51	67
torsion3-residue	-0.53	0.44	0.22	1.86	13	-0.76	-1.51	-0.11	1	10	**-2.64**	**59**	**79**
pairwise$C\beta$	-0.58	0.51	0.3	2.51	17	-0.7	-1.7	-0.18	4	27	-1.96	39	69
pairwise$C\beta$/SSE	-0.59	0.52	0.34	2.58	22	-0.84	-1.8	-0.13	5	36	-2.16	45	71
solvation	-0.55	0.49	0.29	2.31	10	-0.55	-1.65	-0.24	2	27	-1.3	18	45
SSEPSIPRED	-0.65	0.52	0.24	2.03	9	-0.63	-1.43	-0.13	3	17	-0.89	7	25
ACCpro	-0.59	0.56	0.35	2.75	21	-0.85	-1.66	-0.11	6	33	-1.38	20	44

[a] Pearson's correlation coefficient r^2 and Spearmans's rank correlation coefficient *rho*

[b] *F.E.* stands for fraction enrichment and $E_{15\%}$ is the enrichment among the top 15% best predicted models as compared to a random selection.

[c] r_{10} are the number of targets for with the top-scoring models is among the top10 best models (based on GDT_TS). $logP_{B1}$ and $logP_{B10}$ are the log probability of selecting the highest GDT_TS model as the best model or among the ten best-scoring models, respectively.

[d] GDT_TSloss is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set. $r1$ and $r10$ are the number of targets in which the best model based on GDT_TS, excluding the native structure was found on the first rank or among the top 10 predictions.

[e] Z_{nat} is the Z-score of the native structure as compared to the ensemble of models. $r1$ and $r10$ are the number of targets in which the native structure was found on the first rank or among the top 10 predictions.

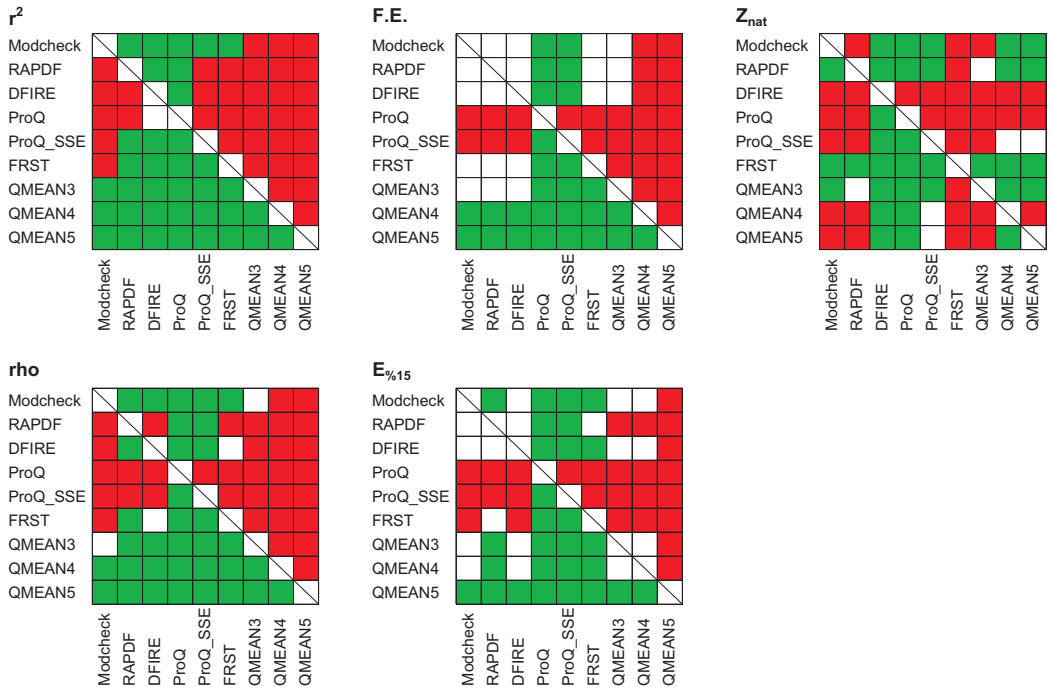


Figure 3.23: Statistical analysis of the performance differences between the methods at the confidence level of 95%. Green (red) stands for a better (worse) performance.

indicate that the performance difference between two methods is not statistically significant on a 95% confidence level whereas coloured squares mark statisticaly significant differences. In case of a green square, the corresponding method denoted in the on the left side of the plot performs better than the one on the bottom.

In general, QMEAN5 consistently outperforms the other five MQAPs with respect to almost all tested quality measures on both categories (free modelling (FM) and template-based modelling (TBM), see Table 5.2 and 5.4 in the Appendix) and over all targets (see Table 3.16). The specific evaluation of the free modelling (FM) and the template-based modelling (TBM) targets shows a similar trend as for all target: QMEAN outperforms the other methods over nearly all quality measures and the difference is potentially more pronounced in the template-based modelling category.

On the two regression and enrichment quality measures, QMEAN5 performs significantly better than all other methods tested (see Figure 3.23). DFIRE, together with QMEAN3 and the 3-residue torsion angle potential, identify to highest number of native

structures whereas DFIRE has significantly worse Znat scores compared to all other methods (see Figure 3.23). FRST produces better Znat scores than QMEAN3 but never better than the torsion angle potential over 3 residues which shows an extraordinary good performance in recognising the native structure.

For the model quality assessment task described by the other quality measures, the 3-residue torsion angle potential does mostly better than the ordinary single residue potential. Modcheck generates statistical significantly better regression coefficients than the other methods except the 3 QMEAN functions. Consistently over all quality measures (except for the Pearson's correlation coefficient), ProQ performs significantly worse than the other methods tested even after the integration of PSIPRED secondary structure prediction. The only exception is the good average Znat scores achieved on the free modelling targets which reflects the fact that ProQ has been trained specifically on fold recognition models (see Table 5.4 in the Appendix).

The secondary structure agreement term shows on average the highest Pearson correlation coefficient of all single terms and a reasonable performance on all the other model quality assessment measures. The solvent accessibility agreement term on the other hand reaches the highest enrichment values and rank correlation coefficients and is very valuable for the selection of good models. Over all quality measures and in both categories the secondary structure specific pairwise potential reaches significantly better scores than the regular one for the model quality assessment task as well as in the identification of the native structure. The analysis of the statistical significance of the QMEAN component terms can be found in Figure 5.2 in Appendix.

The differences in the results achieved for the free modelling and template-based modelling targets are frequently easy to explain but sometimes appear to be contra-intuitive. For the task of identifying the native structure, the solvent accessibility agreement term (and to a certain extent also SSE PSIPRED) performs considerably better on the FM targets than on the TBM category. In contrast to the secondary structure agreement term, the ACCpro score can help to identify the native structure in the case of free modelling targets where it recognises 7 out of 18 native structures with an average Z-score of the native structure of more than 2 standard deviations. Over all targets (Table 3.16), QMEAN3 is slightly better than QMEAN4 and QMEAN5 as a consequence of the inability of the secondary structure agreement term in recognising

the native structure which is reflected by the low Z-scores of the native structure and the rank measures (rank1 and rank10). An explanation for this observation is given in a separate discussion section on page 130.

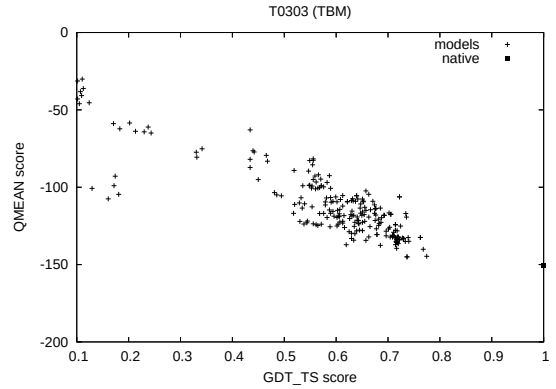
As expected, the regression coefficients for TBM targets are on average higher than for FM targets. A slightly better enrichment is possible with FM targets, since the models in this category tend to be less similar to each other than for example in the high-accuracy template-based modelling category in which a large fraction of the models can be more or less identical as it can be seen in Figure 3.24 b. Of the free modelling targets, the pairwise and solvation potentials as well as ACCpro all produce high enrichment values whereas on the template-based modelling targets the performance of the solvation potential is significantly worse compared to the others over most quality measures. For the FM targets, the native structures are recognised with better Z-scores on average but, surprisingly, the relative number of native structures ranked as number one is lower (9 out of 18) as compared to the TBM targets (51 out of 77) (see Supplementary Material).

Figure 3.24 shows the correlation between GDT_TS and QMEAN score for the models of four selected targets belonging to the TBM and FM target category. The scatter plots on the left-hand side (Figure 3.24 a and c) represent two examples in which both the regression and the identification of the native structure went fine. The scatter plots for all of the 95 targets are shown in the Appendix.

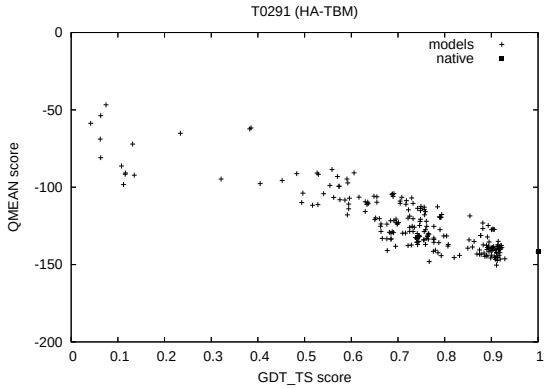
Sometimes the native structure can be easily identified (target T0321, Figure 3.24 c) but sometimes the native structure is hidden among the bulk of the models (target T0300, Figure 3.24 d) even though the regression can be reasonably good. This is quite astonishing, since for most of the FM targets, no submitted model had a GDT_TS score of more than 50 and one should expect the native structure to be easy to identify. On the other hand, the enrichment for FM targets works rather well with enrichment values (E15%) on the order of factor 3 achieved on average.

3.2.3.4 Estimating overall performance

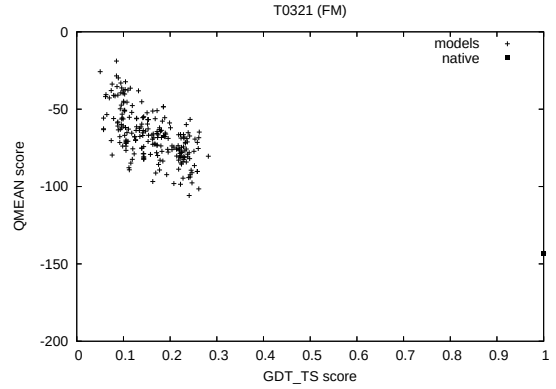
Fraction enrichment curves [217] are useful to compare and visualise the performance of different MQAPs in analogy to receiver operator characteristic (ROC) curves frequently



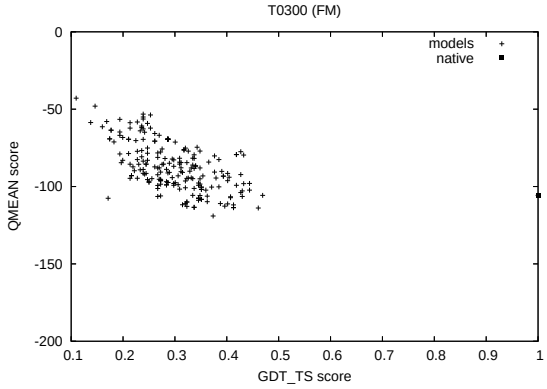
(a) Target T0303 (template-based modelling category)



(b) Target T02901 (high-accuracy TBM category)



(c) Target T0321 (free modelling category)



(d) Target T0300 (free modelling category)

Figure 3.24: Scatter plots showing the correlations between GDT_TS and QMEAN5 for four selected examples.

used in benchmarks of fold recognition and alignment programs. They implicitly cover several quality measures used in Table 3.16, *e.g.* enrichment and regression. Where ROC curves require the somewhat arbitrary definition of a threshold to distinguish good from bad models, fraction enrichment curves measure the added value of MQAPs in ranking different models.

Figure 3.25 and 3.26 show the fraction of best models (based on GDT_TS) found among a certain fraction of the top scoring models as predicted by the scoring function (fraction enrichment). The calculations are performed on the server models of CASP7 after removing the native structures. The curves in the upper part of Figure 3.25 reflect

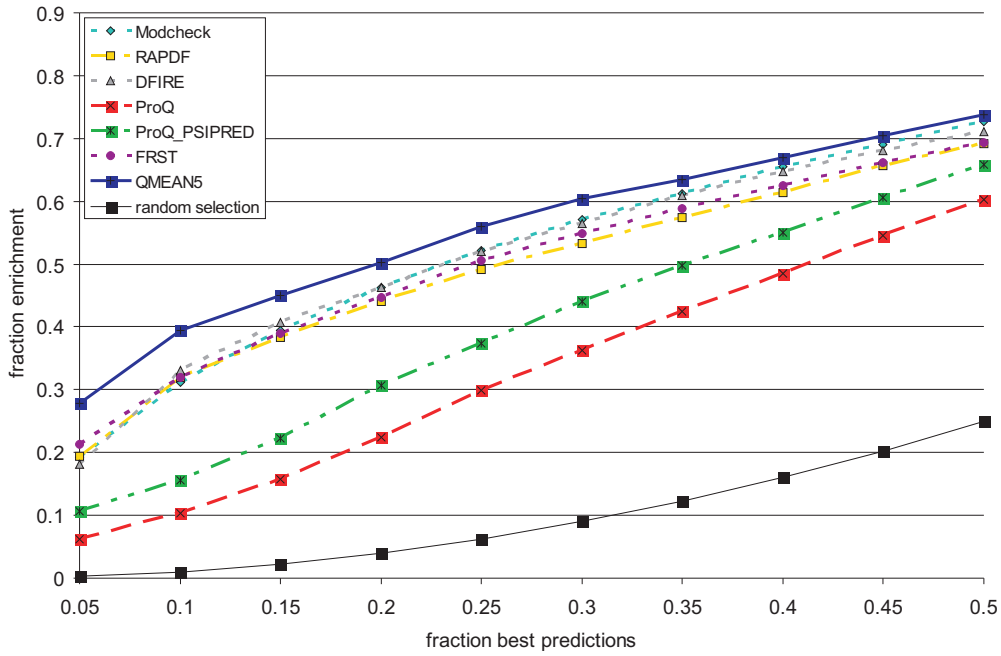


Figure 3.25: Target-specific fraction enrichment curves showing the percentage of top x% highest GDT_TS models among the top x% best-scoring structures (averaged over all CASP7 targets).

the ability of the scoring function to identify the best models among all models for a given target (averaged of all targets) and are a measure for the scoring functions ability to predict the relative model quality. The steeper the progression of the curve, and the larger the area under the curve, the better a scoring function agrees with the measured model quality. The average fraction enrichment over the individual targets for cutoffs ranging from 5% to 50% is shown. QMEAN consistently shows the best performance over the whole range but especially between 5% and 15%, underlining its strength in recognising the best models. Modcheck, RAPDF, DFIRE and FRST show a quite similar behavior over the first 3 thresholds. Above 20 percent, the curve obtained for Modcheck and DFIRE are slightly higher which agrees with its good rank correlation coefficients and enrichment values in Table 3.16. ProQ performs significantly worse than the others.

The global fraction enrichment curves shown in Figure 3.26 are obtained by pooling together the models of all targets and calculating the fraction enrichment over the whole

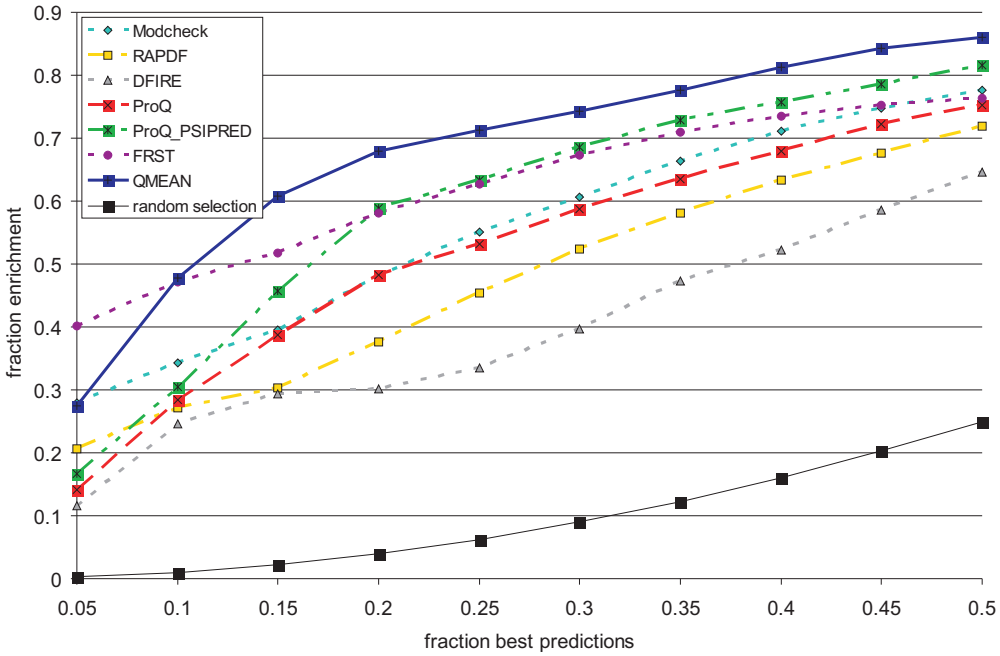


Figure 3.26: Global fraction enrichment curves over all model from all CASP7 targets.

set. In this way, the scoring function's ability to predict the absolute model quality (*i.e.* to estimate the degree of "nativeness" of a model) is investigated. In contrast to the results in Figure 3.25, the performance of RAPDF and especially DFIRE are strikingly low compared to Modcheck and FRST. FRST shows the best fraction enrichment within the first 5 percent and appears to be good in recognising native and native-like structures. This is also reflected by the low average Z-scores of the native structure (Znat) shown in Table 3.16. In the global enrichment, ProQ shows a reasonable performance which can be mainly attributed to the secondary structure information included as the difference between ProQ and ProQ PSIPRED suggests. Above a fraction of 0.1, QMEAN consistently generates the highest fraction enrichments of all MQAPs tested. For example, among the 15% best QMEAN predictions more than 60% of the 15% best models are identified. The high enrichments are an evidence of a good global correlation between the QMEAN score and the effective model quality.

Slope and intercept from the regression between GDT_TS and QMEAN score obtained on the training set can be used in order to derive a predicted GDT_TS. Figure 3.27

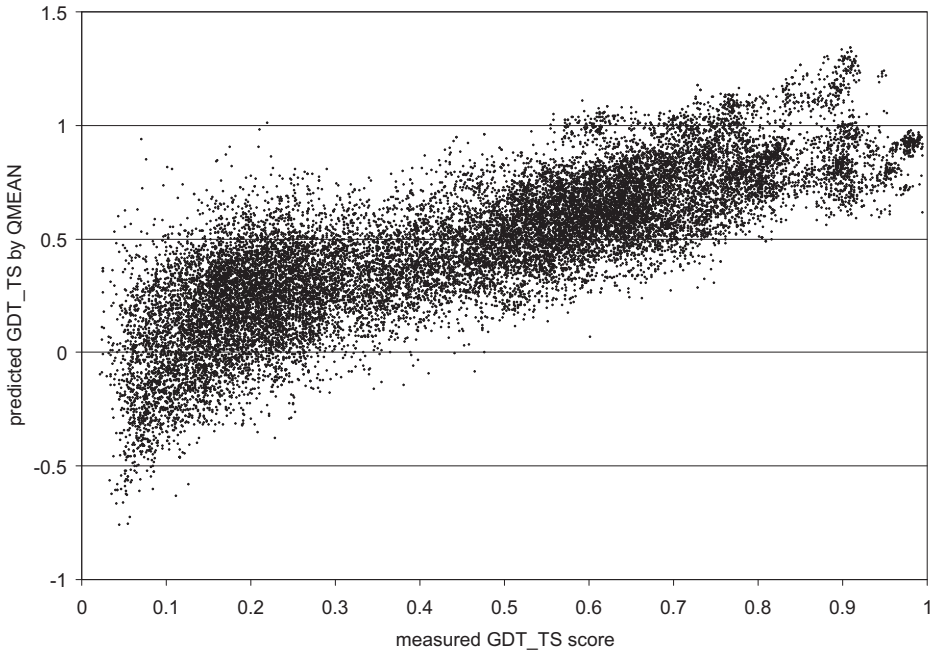


Figure 3.27: Regression between GDT_TS and the composite score (QMEAN5) of the models in the CASP7 test set.

shows the correlation between measured GDT_TS and predicted GDT_TS based on QMEAN on the CASP7 test set. Although the correlation is quite good, the data show that a prediction of the absolute GDT_TS of a given model is only possible with a certain accuracy. An improved global correlation will be definitively achieved by using machine learning approaches in order to combine the terms (as first results with a neural network suggest).

3.2.4 QMEAN: Discussion and outlook

3.2.4.1 General performance

The QMEAN scoring function has been shown to be a valuable tool for model quality assessment by distinguishing good from bad models and for the identification of the native structure among decoy sets generated by a variety of methods. On the comprehensive set of 22,420 server models of CASP7, QMEAN consistently outperforms the five model quality assessment programs over nearly all quality measures and model difficulty ranges.

3.2.4.2 Agreement between predicted and measusred features

Only in two decoy sets from Decoys 'R' us, `lattice_ssfit` and `LMDS`, did the integration of the secondary structure agreement term result in an improved ability of the combined scoring function in identifying the native structure compared to the statistical potential terms only (QMEAN3). This can be possibly attributed to the greater overall variability of the decoy structures in these sets and the absence of native-like structures: `lattice_ssfit` conteins structures with RMSD ranging from 5.68 to 13.23 Å and `LMSD` from 4.05 to 11.5Å. On the other hand, the `4state_reduced` set on which the two agreement terms failed in recognising the native structure covers structures between 1.15 and 8.80 Å. The CASP7 test set shows a similar trend: for free modelling targets slightly better Znat scores are obtained than for template-based modelling targets using the secondary structure agreement term and solvent accessibility terms performs considerably on targets of the FM category.

In contrast to this observation, the secondary structure agreement term turned out to be a valuable contributor to the good performance of QMEAN in the model quality assessment task. The different performance on these two tasks can, especially in the case of the CASP7 set, tentatively be ascribed to the fact that the secondary structure composition of the native structure can only be predicted with a certain accuracy, typically around 76-80%. A theoretical limit of prediction accuracy of 88% percent was proposed by Rost [176] arguing that minor variations in structures even between homologous proteins can result in different secondary structure assignments

made by tools such as DSSP. It is therefore rather unlikely that the secondary structure agreement between PSIPRED and DSSP achieves 100 percent for the native structure and more likely that there is a tendency for models generated by methods taking implicitly advantage of predicted secondary structure information to receive better scores than the native structure.

The same argument given above holds for the solvent accessibility agreement term, although the effect seems to be less pronounced as reflected by the higher Z-scores of the native structure (Znat) achieved in the CASP7 decoy set. This might be explained by the significantly reduced sensitivity of this term toward minor differences in the structures, since it is based on a binary classification of solvent accessibility (buried/exposed) as provided by ACCpro. Thus, near-native structures would tend to have solvent accessibility agreement values (*e.g.* packing) similar to the native structure but bad models do not, which would explain the moderate Znat scores to some extent.

In contrast to the observation described above, both agreement terms turned out to be valuable contributors to the good performance of QMEAN in the model quality assessment task as reflected by the consistently better performance of QMEAN5 compared to the version using statistical potential terms only (QMEAN3).

3.2.4.3 Torsion angle potential over 3 residues

The torsion angle potential over three residues turned out to be a very powerful term for the identification of the native structures out of a variety of decoy sets, suggesting that the 3-residue torsion angle potential describes the propensity of a certain amino acid for a certain local geometry considerably better than the single residue torsion angle potential. The final bin sizes of 45 degree for the phi and psi angles of the center residue and 90 degree for the neighbouring torsion angles are surprisingly coarse-grained, but can possibly be explained by reasonable binning of the Ramachandran plot [167] in 90 and 45 degrees and how these values represent a trade-off between resolution and number of states, reducing the danger of over-fitting. The resulting number of 327,680 ($= 20 * (360/45)^2 * (360/90)^2 * (360/90)^2$) possible states is in the same order of magnitude as observed in some all-atom potentials. Betancourt and Skolnick [19] have shown that the dihedral angles of a residue are influenced by the identity and

conformation of the adjacent residues. This effect is especially pronounced in loop regions and near the end of β-sheets. The 3-residue torsion angle potential seems to capture this effect to a certain extent. In contrast to the potential introduced by Betancourt and Skolnick, the 3-residue potential described in this work does not take into account the identity of the adjacent residues and is attractive in its simplicity. It basically reflects the propensity of a certain amino acid type for a given local geometry (as described by six torsion angles) as compared to other 19 amino acids.

3.2.4.4 Secondary structure specific pairwise potential

The secondary structure specific implementation has shown to lead to a statistically significant improvement of the performance over all quality measures compared to the regular residue-level pairwise potential. Loops are primarily located at the protein surface and are to a greater extent influenced by non-local interactions in contrast to helices and sheets which are mainly determined by the local potential [19]. As loops have fewer contacts to the rest of the protein than helices and sheets, which are at least partially surrounded by more residues, it can be speculated that pairwise statistical potentials tend to be biased towards interaction patterns observed in the protein core. As a consequence, some motifs observed only in loop regions receive a slightly too high energy. A specialised potential compiled and applied in a secondary-specific manner may counteract this.

3.2.4.5 Solvation potential

The calculation of the solvent accessibility solely based on the atoms present in the coordinate file is problematic. As described in Methods, the solvent accessiblity is approximated by counting the number of β atoms with 9 Å arrount the β of the given residue. Although all chains are taken into account in the calculation, the structure in the PDB file often does not represent the biologically active molecule. For example in the case of homo-multimers (*i.e.* proteins consisting of several identical subunits in the quaternary structure), typically only one subunit is present in the coordinate file. As a consequence, some residues which are buried in the native complex are considered as exposed leading to inaccuracies in the resulting potentials. This is a possible

explanation for the bad performance of the solvation term and also for the observation, that a solvent accessiblity specific implementation of the pairwise interaction potential did not improve the results.

To the best of this author's knowledge, non of the statistical solvation potentials described in literatur does take into account the biological unit of the protein in the derivation of the potentials. To some extent, statistical potentials are tolerant concerning minor error in the derivation of the observed frequencies as a consequence of their statistical nature. But, in the case of the solvation potential, the errors introduced by not considering the biological unit can most probably not be neglected.

In a future implementation of the solvation potential, the information of the biological unit of the proteins will be taken into account *e.g.* by using either structures from the Protein Quaternary Structure (PQS) server[f] or by only using monomeric structures. Both approaches are associated with inaccuracies as well (*e.g.* because the biological unit is often assigned wrong [242]), but including information about the quaternary structure is probably the better alternative than ignoring it.

3.2.4.6 Training and evaluation process

In order to reduce a possible over-fitting of any of the potentials, all structures with detectable homology (based on a BLAST search) to any of the structures of the two CASP decoy sets were removed from the protein data set used to build the potentials. In this way, several 100 percent sequence identity hits have been removed. Remarkably, comparing the results before and after adjusting the potentials, no considerable change has been observed even for the task of detecting the native fold *(data not shown)*. This can be explained by the rather large number of structures used to compile the potentials, where the influence of one specific (even identical) structure is diminished by the others. In model quality assessment in particular, models with significant errors, not the actual structures, are evaluated, further reducing a possible bias from the presence of homologous structures in the data set.

Parameterising and optimising the single term as well as their combination on CASP decoys represents a reasonable approach since a variety of methods and the entire range

[f] `http://pqs.ebi.ac.uk/`

of modelling difficulty is covered. The good performance of QMEAN on all decoy sets and the fact that the targets of two CASP rounds are completely different indicates that QMEAN has not been specifically trained to assess models produced by CASP participants but instead is applicable to the variety of methods.

Although the strategy to derive the weighting factors for the composite score based on the regression coefficient represents a reasonable starting point (assuming a correlation between energy and degree of "nativeness"), this approach also has some disadvantages. Some terms showing a medium correlation to GDT_TS can still perform better on other quality measures and their discrimination power tends to be underestimated. A good example is the solvent accessibility agreement term which shows lower correlation to GDT_TS than the secondary structure agreement term (Table 3.12) but performed consistently better in the CASP7 decoy set over a wide range of conditions (Table 3.16). A possible underestimation is also reflected by the low correlation to the QMEAN5 score as shown in Table 3.13. The fact that some of the other terms show varying discrimination power depending on the modelling difficulty may warrant specialised versions of the scoring function *e.g.* for free modelling or template-based modelling targets. In particular, it remains to be seen why decoys for certain free modelling targets have lower energy than the native structure.

3.2.4.7 Global and target-specific prediction of model quality

QMEAN shows a consistently better enrichment performance based on the fraction enrichment curves shown in Figure 3.25 and 3.26 compared to other MQAPs for both the relative prediction of model quality for models of the same target as well as for the global quality prediction over all targets. Since MQAPs are routinely used to assess ensemble of models for the same target, the target-averaged fraction enrichment curves are probably of greater practical interest since they reflect the ability of the scoring function in discriminating good from bad models. On the other hand, the need for scoring functions predicting the absolute quality of a model has only recently been highlighted by the CASP7 assessors [49]. QMEAN represents a further step towards the prediction of the absolute quality of protein models.

3.3 The loop prediction routine

3.3.1 General performance

The knowledge-based loop modelling protocol described in this work basically consists of 3 steps (see schematic representation on page 45 in Methods): selection of fragments from the fragment database which approximately fit to the geometry imposed by the anchor groups, filtering of the initial selection in order to remove unfavourable candidates and, finally, ranking of the remaining loops according to a scoring function. The optimisation of the parameters and thresholds used in the selection process as well as for the different filters (anchor geometry filter, clash filter, torsion energy filter and backbone energy filter) is described in detail in Methods on page 45ff. In this section, the results of the loop ranking process are described and compared to other loop prediction methods (section 3.3.2).

The loop modelling accuracy of knowledge-based approaches is determined by two distinct factors: first, the availability of suitable conformations in the fragment database based on experimentally solved protein structures and, second, the ability of the scoring function to identify fragments which are close to the native conformation. In contrast to *ab initio* methods, in which the loop conformation is incrementally built up in the given protein framework, in knowledge-based approaches the candidate fragments are fitted on the anchor groups located on the N-terminal and C-terminal side of the loop. Therefore, not only the local conformation of a fragment is important (as expressed by the local RMSD between the fragment and the native loop after superposition), but also its orientation in the protein framework (as expressed by the global RMSD between native loop and candidate loop after fitting on the anchor groups).

As described in Methods (Chapter 2.3), a maximum number of 3000 fragments are retained after the application of all filters. In a subsequent step the sidechains are added to the loop backbone and the loops are ranked based on an all-atom distance-dependent interaction potential which investigates the compatibility of the loop with the given structural environment. The evaluation of different scoring functions is described later.

In Figure 3.28 the average (a) and the median (b) global RMSD of the top-ranking

loops together with the lower and upper bounds of the prediction accuracy are shown for loops of length of 4 to 12 residues of the test set of Rossi *et al.* [174]. All RMSD values shown in this section are calculated based on the four backbone atoms of the loop without the anchor group residues. The lower bound is determined by the loop with the lowest global RMSD present among the 3000 candidate fragments, averaged over the different test cases. This represents the maximum possible prediction accuracy which could be achieved by a "perfect" scoring function, *i.e.* if the scoring function would consistently choose the fragment closest to the native conformation. The upper bound is defined by randomly selecting a conformation out of the 3000 candidates. Detailed results for loop of length 4, 6, 8 residues are shown later in Table 3.19-3.21.

In the majority of the test cases for loops of length 4-7 residues, a fragment with a global RMSD below 1 Å is present in the final selection of 3000 conformations. For loops below 8 residues, the scoring function shows a good performance in the selection of near native conformations and works considerably better than the random selection. For loops of 8 residues and longer the median RMSD of the best fragment in the final selection increases which reflects the decrease in coverage of the conformational space.

In Figure 3.28, only fragments originating from protein structures showing no measurable sequence identity to the protein in which the loop is modelled have been used. This allows to avoid trivial predictions and guarantees a fair comparison to other methods. However, in a realistic application case, depending of the modelling difficulty (*i.e.* the sequence identity of the query protein to its templates), fragments of remote homologous proteins are present and can be used. Figure 3.29 underlines the influence of the presence of fragments from homologous proteins on the prediction quality. The median RMSD of the top ranking loops is shown using different sequence identity cutoffs in order to filter out fragments from homologues of the query protein (*i.e.* the protein in which the loop is modelled). The homology is detected by a BLAST [5] search of the query protein sequence against the set of proteins used to build the fragment database. Since BLAST provides local alignments, the percentage sequence identity over the entire structure can be considerably lower and therefore the prediction accuracy for a given cutoff even better.

Figure 3.29 shows that the median RMSD is consistently lower if fragments from homologous proteins are accepted, suggesting that they are often found on the first

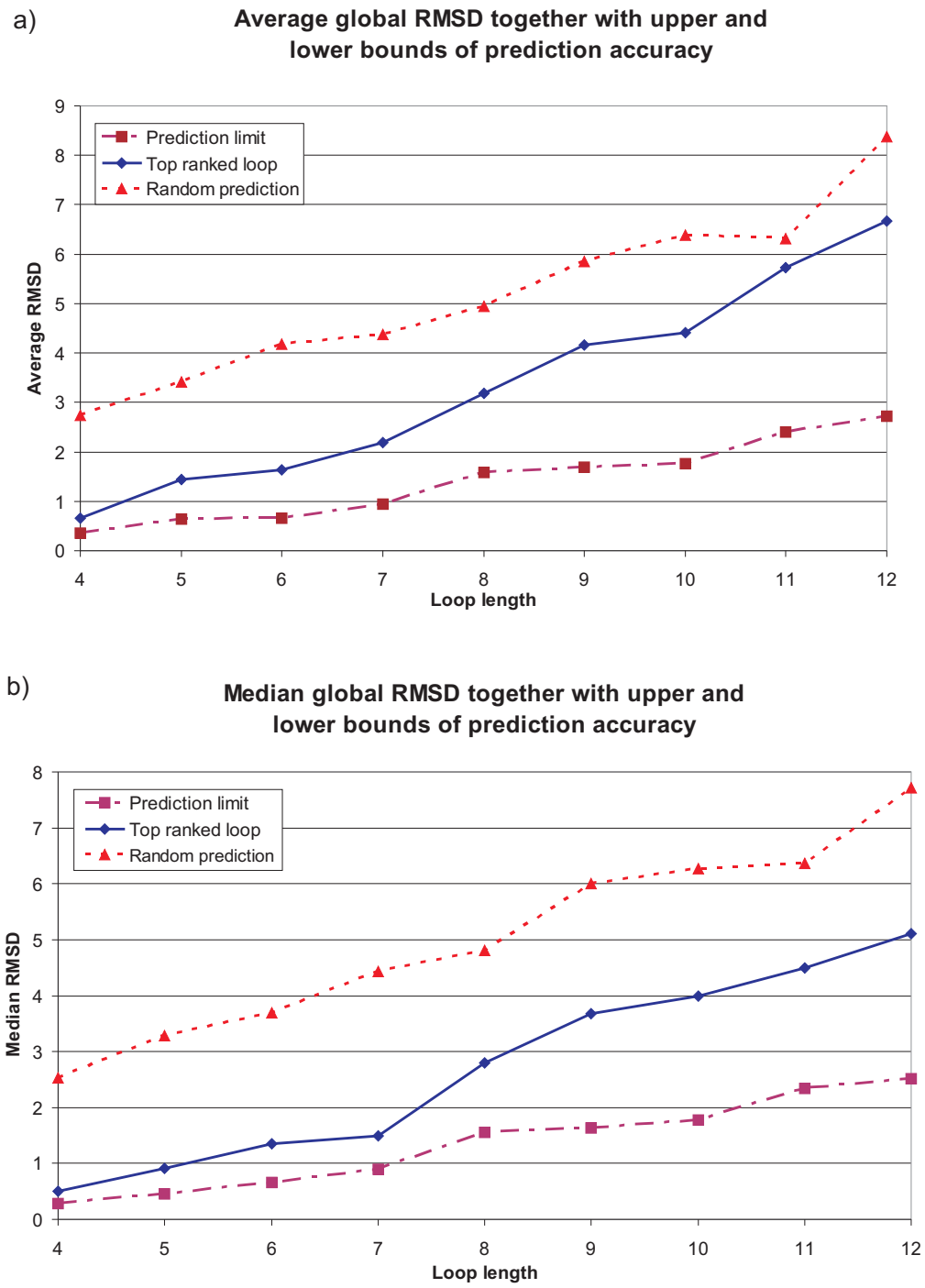


Figure 3.28: Average (a) and median (b) RMSD of the top-ranking loops per loop length as well as upper and lower bound of loop prediction accuracy on the test set of Rossi *et al.*

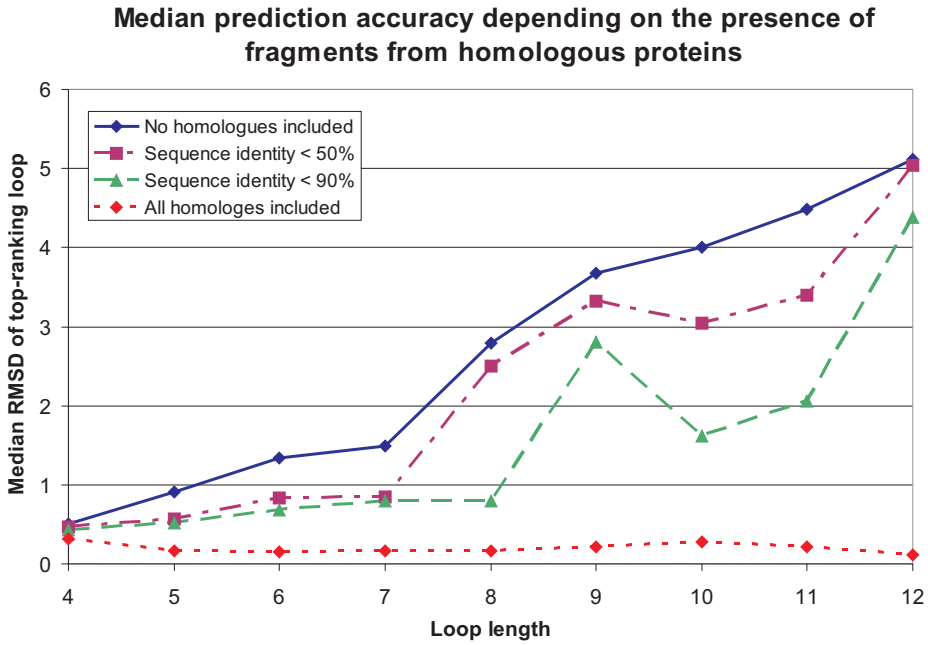


Figure 3.29: Median RMSD of the top-ranking loops per loop length in presence of fragments originating from homologues of the loops in the test set of Rossi *et al.* [174].

rank. Fragments from homologous proteins were exposed to a similar structural environment and potentially have anchor geometries comparable to those observed in the protein in which the loop is modelled. This increases to probability that an analogous local fold is adopt and that the orientation of the fragment with respect to the protein framework is approximately correct. If no homology filter is applied, the median of the RMSD drops significantly (lowest curve in Figure 3.29). In this case, fragments of the native loop conformation itself or of a very close homologue are ranked first. Since a non-redundant set of protein structures clustered at 95% sequence identity has been used to generate the fragment database, the loop of the native structure itself is often not present in the database. These results prove that the all-atom interaction potential used for loop ranking is able to consistently identify loops having a very similar or identical conformation compared to the one observed in the native structure and that these loops are in most cases ranked first.

However, in a realistic modelling situation the local loop conformation is only approx-

imately correct (for an evaluation of the anchor region see Chapter 3.4.2) and the orientation of the fragment in the protein framwork after fitting on the anchor groups is rarely ideal. Since the geometry of the anchor groups and the terminal residues of the fragments are different even for fragments with a local conformation close to the native one, the fragments are slighly misoriented in the protein framework. The resulting rotation has a much stronger effect on longer loops as a consequence of the longer radius. This problem will be addressed in detail at the end of this section and a possible solution is discussed.

Furthermore, even minor distorsions of the protein backbone with respect to the native conformation can lead to considerable differences in the orientation of the sidechains resulting in unfavourable interactions of the loop with its environment (see *e.g.* [31] for the description of the backbone-dependent rotamer libraries used in sidechain modelling). On the other hand, ranking the loops without considering sidechain interactions is too imprecise since especially the conformation of longer loops is mainly determined by interactions with the structural environment rather than by the local geometry (*i.e.* by torsion angle preferences of the amino acids of the loop) [19]. The torsion angle potential, for example, but also as the residue-level interaction potential based on Cα atoms (definitions in Methods page 58) are both able to roughly discriminate between good and bad fragments but fail in recognising near native solutions. This is the reason why they are used as filters and not in the scoring process.

For the final scoring step, a variety of implementations for the all-atom interaction potential and combinations with other statistical potential terms (torsion angle potential, all-atom solvation potential) have been investigated. A combination with the anchor group RMSD (describing the "goodness of fit" of the fragment to the geometry imposed by the anchor residues) has been tried as well. Table 3.18 shows some of the best performing scoring functions tested in the evaluation process. The average global RMSDs on the parametrisation test set are shown for different loop length.

Overall, the all-atom interaction potential shows the best performance in scoring loop conformations, approximately as good as the combination of the of three statistical potentials (torsion angle potential, all-atom solvation potential, all-atom interaction potential) together with the anchor group RMSD (RMSa). This can be partly

attributed to the fact that some of the terms of the combined scoring function have been previously used in the filtering step. The information captured by the all-atom solvation potential is to some extent covered by the all-atom interaction potential: the propensity of a loops to form contacts with the protein framework insteed of being solvent exposed described by the solvation potential (*e.g.* the burial of hydrophobic residues) is also reflected by the interaction potential. Loops lying against the protein body tend to have also more favourable interactions and, as a consequence, potentially lower energies.

Table 3.18: Comparison of different scoring functions on the parametrisation set for loops of length 4, 6, 8 and 12. A desciption of the terms can be found in Methods on page 67.

	Loop length			
scoring function	**4**	**6**	**8**	**12**
RMSa[a]	0.95	2.1	3.16	5.98
RMSa + sequence conservation	1.01	2.19	3.06	5.83
all-atom 2-10 Å (default)	**0.94**	**1.95**	**3.03**	**5.96**
all-atom 2-10 Å (environment sidechains rebuilt)[b]	0.91	1.85	3.28	5.62
all-atom 3-10 Å	0.91	2.02	3.2	5.68
all-atom 0-10 Å (environment sidechains rebuilt)[a]	0.85	1.91	3.13	6.22
RMSa + all-atom	1.76	3.08	3.54	5.65
all-atom + solvation	1.72	2.84	3.26	5.7
all-atom + torsion	1.37	2.31	3.75	6.5
all-atom + solvation + torsion	1.32	2.37	3.47	5.88
all-atom + solvation + torsion + RMSa	0.98	1.9	3.08	5.53
Cα-pairwise + Cα-solvation + RMSa[c]	1.95	2.81	3.48	5.63

[a] RMSD between the terminal fragment residues and the anchor group residues after fitting.
[b] In a second round, the sidechains of surrounding residues within 5 Å are rebuilt simultaneously with the loop sidechains.
[c] Scoring function only relying on the loop backbone (used in the backbone energy filter).)

The average RMSD values for four alternative implementations of the all-atom interaction potential are shown in Figure 3.18. A lower distance cutoffs of 2 Å performs slightly better than 3Å for medium loop lengths. In the former implementation, hydrogen bonding is taken into account typically occurring at distances between approximately 2.5 Å - 3Å [231]. In two implementations, the structural environment is allowed to

relax in that the sidechains of all residues having an atom within 5 Å around the loop after the intitial sidechain modelling process are rebuilt in a subsequent step together with the loop sidechains. Slighly better RMSDs are obtained in this approach for small loops up the length 6. If no lower distance cutoff is used, the repulsive term at close distances improves loop ranking for smaller loops but not for longer ones. This can be attributed to the higher probability of clashes at longer loop lengths. Overall, the performance differences of the four alternative implementations are only marginal. Since rebuilding the structural environment results in an increase of the run-time, the version investigating contacts between 2 Å and 10 Å (highlighted in bold) is used in the following. At the end of this section, the application of a subsequent energy minimisation step based on a molecular mechanics force field is suggested. This would allow to relax the loop, and, a sidechain rebuilding process would not be necessary.

Using solely the all-atom potential for scoring without considering the RMSa has the advantage that the scoring function is more generally applicable. Loop prediction methods are typically tested in self-prediction experiments, which means that a loop is cut out from a experimental protein structure and rebuilt in the given exact environment. In the modelling case, the situation is quite different: the environment is only approximately correct and especially the anchor geometry is usually slightly distorted (see section 3.4.2) leading to a different orientation of the fragment after fitting. Whereas in the self-prediction case the RMSa can to some extent indicate whether a fragment has the correct orientation with respect to the framework, this is hardly the case in the modelling situation. Therefore this term should not be used for scoring as done in many knowledge-based approaches described in literature [53, 90, 139].

In the following, the performance of the loop prediction routine on the test set by Rossi *et al.* [174] is described in detail. A comparison to other methods is described in the next section. In Table 3.19-3.21, the results for loops of length 4, 6 and 8 are shown. The results for the other loop length can be found in the Appendix Table 5.6-5.11.

For loops of length 4 the average (median) prediction accuracy is 0.66 Å (0.51 Å) if all fragments from homologous structures are excluded. More than 90% of the loops are predicted with a global backbone RMSD below 1 Å. In column 6 the rank of the

Table 3.19: Results for loops of length 4 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop[a]				
PDB ID	residues	best loop[b]	random 20000[c]	random 3000[d]	rank Top10[e]	no homolgues	all homologues	<90%	<50%	<30%
1aaj	82-85	0.28	2.21	1.61	6	0.61	0.17	0.37	0.44	0.61
1ads	99-102	0.22	3.67	1.83	15	0.24	0.33	0.33	0.33	0.24
1cbs	21-24	0.26	4.7	0.91	3	0.34	0.34	0.34	0.34	0.34
1frd	59-62	0.29	3.58	2.87	6	0.43	0.06	0.39	0.43	0.43
1gpr	123-126	0.34	3.63	1.03	7	2.12	0.07	2.12	2.12	2.12
1nfp	37-40	0.95	5.31	2.54	1	0.95	0.95	0.95	0.95	0.95
1pbe	117-120	0.38	2.63	1.38	2	0.42	0.29	0.42	0.42	0.42
1pda	139-142	0.26	1.91	0.9	17	0.32	0.32	0.32	0.32	0.32
1plc	74-77	0.53	1.94	2.24	16	0.81	0.06	0.21	0.58	0.81
1ppn	42-45	0.28	3.48	0.41	79	0.55	0.55	0.55	0.55	0.55
1rcf	111-114	0.11	0.6	0.25	4	0.46	0.46	0.46	0.46	0.46
1thw	194-197	0.36	0.69	3.57	1	0.43	0.43	0.43	0.43	0.43
1tib	46-49	0.32	2.55	4.05	1	0.53	0.53	0.53	0.53	0.53
1tml	42-45	0.87	2.09	2.16	110	2.11	2.11	2.11	2.11	2.11
1xif	82-85	0.32	1.77	1.29	26	0.6	0.1	0.6	0.6	0.6
2exo	116-164	0.29	4.83	2.47	7	0.51	0.51	0.51	0.51	0.51
2sil	220-223	0.4	1.92	1.74	6	0.51	0.18	0.51	0.51	0.51
2tgi	72-75	0.24	2.11	1.57	4	0.71	0.06	0.5	0.71	0.71
4enl	335-338	0.15	2.53	2.85	2	0.24	0.31	0.31	0.24	0.24
4gcr	116-119	0.34	3.64	3.25	3	0.4	0.11	0.4	0.4	0.4
7rsa	47-50	0.28	1.7	2.08	12	0.47	0.35	0.35	0.47	0.47
average	-	0.36	2.74	1.95	-	0.66	0.39	0.61	0.64	0.66
median	-	0.29	2.53	1.83	-	0.51	0.32	0.43	0.47	0.51

[a] RMSD of the top ranking loop after removing fragments from homologues above a given cutoff.
[b] Best nonhomologues loop present among the 3,000 candidate fragments after all filtering steps.
[c] Random selection of a fragment from the maximum 20,000 loops present after application of the torsion energy filter.
[d] Random selection of a fragment from the maximum 3,000 loops present after application of the backbone energy filter.
[e] Rank of the first Top10 fragment according to RMSD.

first Top10 solution (accoriding to the RMSD) is shown. In majority of the test cases a Top10 fragments is found among the first 10 ranks. But even if this is not the case the prediction can be still accurate which confirms that a variety of near native fragments are present and that the fragment database shows good coverage of the conformational space at this loop length. Two test cases were predicted with an RMSD above 2 Å: in the first case (PDB identifier 1gpr, residues 123-126), two good loops can be found on rank 3 (0.55 Å) and rank 7 (0.35 Å). For the second loop only 2 loops with an RMSD below 1 Å are present in the selection. On rank 7, a loop with an RMSD of 1.31 Å is found.

Table 3.20: Results for loops of length 6 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
1ads	149-154	0.15	8.39	2.23	1	0.15	0.15	0.15	0.15	0.15
1ads	150-155	0.27	4.38	3.04	5	0.3	0.18	0.42	0.42	0.3
1brt	174-179	0.73	4.41	3.53	21	1.63	0.05	0.39	1.63	1.63
1brt	253-258	0.76	2.06	4.44	77	1.24	0.06	0.33	0.33	1.24
1cbs	66-71	0.66	6.82	5.6	2	0.66	0.41	0.41	0.66	0.66
1dim	318-323	0.28	2.33	1.57	5	0.67	0.3	0.67	0.67	0.67
1dts	146-151	0.51	4.05	2.43	2	0.81	1.67	0.81	0.81	0.81
1ede	180-185	1.14	3.47	4.4	87	2	0.21	2	2	2
1gca	100-105	0.57	3.63	0.86	5	1.63	0.06	1.63	1.63	1.63
1mrp	233-238	0.34	3.91	3.55	4	1.76	1.76	1.76	1.76	1.76
1nif	211-216	0.76	3.33	2.03	115	3.8	0.18	0.25	3.8	3.8
1noa	25-30	0.61	3.31	2.71	7	3.55	0.05	0.62	0.62	3.55
1onc	12-17	0.94	5.28	4.44	51	2.18	2.18	2.18	2.18	2.18
1rge_A	73-78	0.96	3.28	2.84	359	3.58	3.58	3.58	3.58	3.58
1rhs	50-55	0.68	2.22	3.36	7	1.45	0.07	1.45	1.45	1.45
1tca	38-43	0.65	1.35	1.51	2	0.65	0.08	0.65	0.65	0.65
1tca	94-99	0.66	4.04	4.04	7	1.72	0.06	1.72	1.72	1.72
1tys	66-71	0.87	4.94	5.73	17	3.17	0.15	0.35	0.84	3.17
1xyz_A	633-638	0.86	2.97	3.67	5	0.91	0.06	0.43	0.43	0.91
1xyz_A	711-716	0.49	2.6	2.18	10	0.64	0.07	0.26	0.26	0.64
2ayh	81-86	0.86	3.77	3.12	4	0.95	0.06	0.22	0.95	0.95
2mnr	308-313	0.53	6.51	1.41	15	2.1	0.13	2.1	2.1	2.1
2ran	40-45	0.33	3.25	1.79	10	0.57	0.26	0.57	0.57	0.57
2sil	176-181	1.07	2.89	2	4	1.07	0.18	0.74	0.74	1.07
3pte	131-136	0.53	6.73	4.05	2	0.7	0.14	0.7	0.7	0.7
3pte	256-261	0.98	7.32	6.26	3	1.03	0.18	0.82	0.82	1.03
5p21	104-109	0.82	6.65	3.84	7	3.61	3.61	3.61	3.61	3.61
8abp	65-70	0.56	3.28	3.02	16	3.14	0.06	3.14	3.14	3.14
average	-	0.66	4.18	3.2	-	1.63	0.57	1.14	1.37	1.63
median	-	0.66	3.7	3.08	-	1.35	0.15	0.69	0.83	1.35

If only non-homologous fragments are accepted, an average (median) RMSD of 1.63 Å (1.35 Å) is obtained for loops of length 6. 39% of the loops in the test set are modelled with an RMSD below 1 Å and 54% below 1.5 Å. If homologues with a sequence identity of less than 50% are included, the percentage of loops modelled below 1 Å increases to over 57% and the median RMSD drops to 0.83 Å. For the vast majority of loop test cases, a Top10 loop can be found on the first ranks.

As could be seen from Figure 3.28, the prediction accuracy drops considerably between loops of length 7 and 8. The data suggest that this can be mainly attributed to the incompleteness of the fragment database concerning fragments with a similar local geometry and orientation after fitting. Whereas for loops of length 7 in 50% of the test

Table 3.21: Results for loops of length 8 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
1a62	71-78	2.41	4.17	3.89	9	3.99	0.13	3.99	3.99	3.99
1ads	274-281	1.17	4.53	2.08	71	3.56	0.29	0.29	0.47	3.56
1alc	34-41	3.1	6.59	5.56	203	4.24	0.11	0.89	0.66	4.24
1arb	136-143	1.53	3.27	3.18	110	2.66	0.07	2.66	2.66	2.66
1cvl	148-155	1.86	5.23	7.28	842	4.33	0.06	4.33	4.33	4.33
1gof	606-613	0.79	6.37	4.11	1	0.79	0.79	0.79	0.79	0.79
1hbq	31-38	1.55	6.8	4.9	394	3.57	1.22	1.22	3.57	3.57
1hfc	119-126	1.42	7.75	5.84	44	2.5	0.07	0.38	2.5	2.5
1hfc	142-149	0.59	4.81	3.42	9	0.59	0.51	0.51	0.51	0.52
1nar	192-199	1.3	6.02	3.67	106	2.13	0.05	2.13	2.13	2.13
1nif	221-228	2.73	6.77	5.53	62	3.04	0.31	0.26	4.85	4.85
1nif	279-286	0.67	3.73	4.71	5	0.82	0.46	0.46	0.89	1.17
1nls	97-104	0.58	6.22	2.28	5	0.58	0.07	0.41	0.58	0.58
1nwp_A	84-91	1	2.99	4.89	704	1.91	0.18	0.31	7.6	7.6
1oyc	80-87	1.56	2.57	1.91	2	1.91	0.07	1.91	1.91	1.91
1prn	150-157	2.56	3.41	7.1	71	5.14	0.26	5.14	5.14	5.14
1thw	18-25	1.87	6.2	6.3	26	7.79	0.17	7.79	7.79	7.79
1tml	187-194	1.59	2.92	4.69	3	2.79	0.49	0.49	0.49	2.79
2ayh	194-201	1.7	3.56	4.27	15	2.52	0.1	0.25	2.52	2.52
average	-	1.58	4.94	4.51	141.16	2.89	0.28	1.8	2.81	3.3
median	-	1.55	4.81	4.69	44	2.66	0.17	0.79	2.5	2.79

cases a fragment with RMSD below 1.5 Å is present in the final selection, the percentage drops 21% for loops of length 8. Only 4 loops are predicted with an RMSD below 1 Å (21%). If homologues are excluded, a median RMSD of 2.66 Å is achieved which drops to 0.79 Å if a homology cutoff of 90% is used. By applying no homology filter (column 8 in Table 3.21, the scoring function consistently ranks near native fragments on the top which underlines that sampling of the conformational space is the main limitation in modelling of longer loops not scoring.

The scoring function is unable to discriminate between solutions which are approximately correct and fragmets which have a few favourable interactions but point into the wrong direction. This holds for both the all-atom interaction potential but also for scoring functions consisting of multiple terms. For example, a loop establishing only one or two hydrogen bonds to the environment but having a completely wrong orientation can still have a considerable lower energy than a loop which has an approximately correct conformation but several unfavourable interactions (*e.g.* overlaps of Van der Waals spheres or atom-atom distances slightly too long for hydrogen bonding). A correlation between interaction energy of the loop with its environment and RMSD

can only be expected for conformations close to the native solution.

The conformational space for short loops is restricted by the geometrical constraints imposed by the anchor region. For longer loops, as the ratio between loop length and distance between the end points increases, the number of available conformations increases exponentially [251]. The rapid growth in the available alternative conformations is challenging both for *ab initio* methods (extensive sampling needed) and knowledge-based approaches (coverage by the fragment database decreases). Furthermore, the chance for false positive conformations increases by interactions with other regions of the protein framework. For knowledge-based approaches, the fitting process represents another source of errors as a consequence of the difference in the geometry of the anchor groups and the terminal fragment residues. Several fitting strategies have been investigated (*e.g.* fitting of two residues on both sides or fitting on three consecutive Cα atoms) but did not result in a better performance.

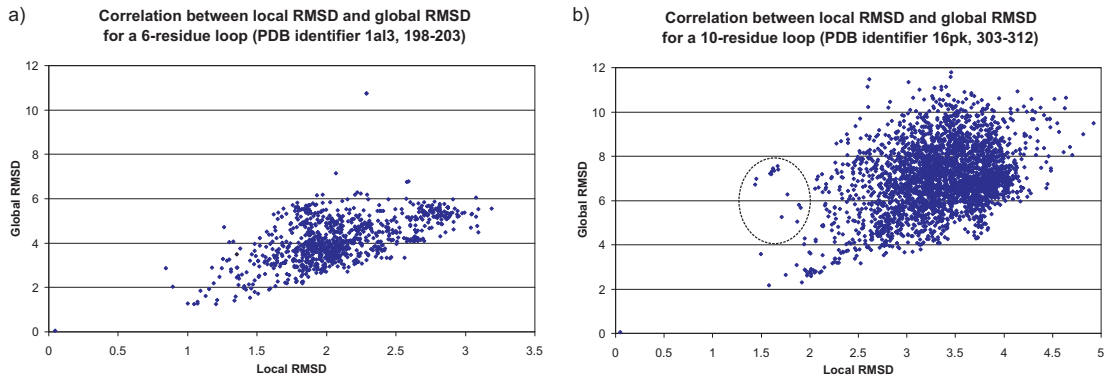


Figure 3.30: Regression between local RMSD and global RMSD for two loop prediction test cases of length 6 (a) and 10 (b), respectively.

Figure 3.30 shows the correlation between local RMSD (based on the fitting the fragment on the native loop conformation) and global RMSD (based on the orientation of the fragment after fitting on the anchor groups) for two loop prediction test cases: On the left hand side, the correlation for the first loop prediction test case of length 6 of the parametrisation set is shown (PDB identifier 1al3, residues 198-203) and in analogy, on the right hand side, the first test case of length 10 (PDB identifier 16pk, residues 303-312). For the longer loop prediction, the correlation is considerably worse compared to the one obtained for the loop of length 6. Several fragments with low

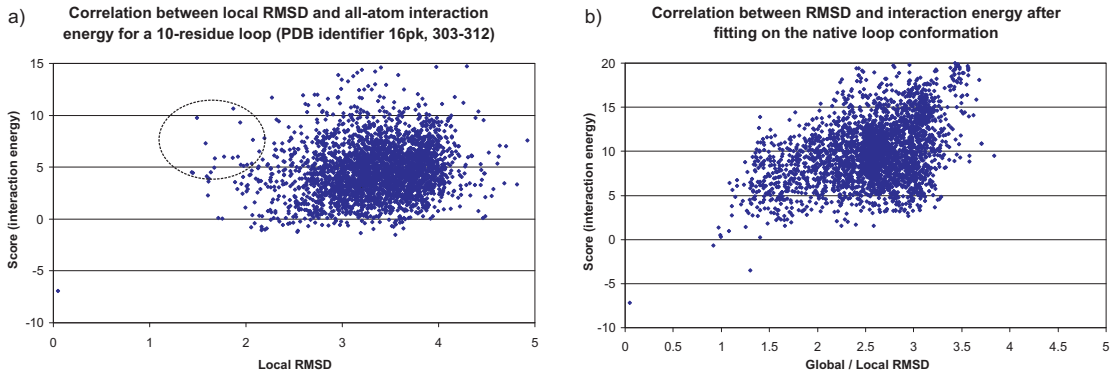


Figure 3.31: Correlation between local RMSD and loop energy calculated after fitting the fragments on the anchor groups (a) and on the native loop (b), respectively.

local RMSD have a wrong orientation with respect to the native loop as reflected by the high global RMSD (see highlighted area).

Figure 3.31 exemplifies that the poor loop prediction accuracy for longer loops is mainly a consequence of the misorientation of the fragments in the protein framework (beside the decreasing database coverage) and not a problem of loop ranking. Two alternative regressions between the local RMSD of the fragments and their energy are shown for a loop prediction test case of length 10 (PDB identifier 16pk, residues 303-312). In Figure 3.31 a) a regression between the local RMSD of the fragment with respect to the native conformation and the score of the fragment (after fitting of the anchor groups) is shown. Virtually no correlation exists and several fragments with low local RMSD have energies higher than the average of the ensemble. In Figure 3.31 b) each fragment has been fitted on the native loop conformation in order to enforce an approximately correct orientation (at least for fragments having a similar local geometry compared to the native loop). This respresents only a hypothetical example, since the native loop is, of course, not known in the application case. As it can be seen, a correlation exists for loops close to the native one and most of the low RMSD loops get assiged scores considerably lower than the rest of the fragments. Furthermore, several near-native loops around 1 Å RMSD are not oberved on the plot on the right hand side since they have been filtered out by the clash filter as a consequence of the wrong orienation with respect to the structural environment.

A reasonable extension of the current loop prediction protocol represents the appli-

cation of a molecular mechanics force field for a subsequent energy minimisation step (not in the scope of this work). Energy minimisation of the loop and possibly the sidechains of the surrounding structural environment could counteract several inherent problems of knowledge-based approaches. The fitting of a rigid fragment in a fixed protein framework results in very unfavourable bond lengths and angles between the anchor residues and the first loop residues which should be relaxed. Annealing the loop with the anchor residues and simultaneously relaxing the loop in the given structural environment can adjust the orientation of the fragment with respect to the protein framework. Thereby atomic clashes are removed and favourable interactions can be established such as hydrogen bonds and salt bridges.

The following strategy could be used in a future implementation:

- Application of the loop prediction protocol described here for the selection of candidate fragments and for an initial ranking.
- Energy minimisation of the top ranking fragments (*e.g.* to top 20 predictions).
- Optionally, re-scoring according the force field energy (with implicit treatment of solvation effects for example by the Generalized Born solvation model [82]).

Such a strategy most probably improves the prediction quality for longer loops and extends the applicability of the knowledge-based approach described in this work which seems to be limited to loops of up to length 7 according to the results shown above. For loops of up to length 10, a fragment below 2 Å is present in the final selection in at least 70% of the test cases but this percentage drops to 23% and 11% for loops of length 11 and 12. Although the data basis is too sparse for well-founded conclusions, this observation suggests that for loops up to a length of approximately 10 residues, fragments from the database could be used as reasonable staring points for a subsequent energy minimisation. Vlijmen and Karplus [226] conlcude in 1997 that candidate segments can be used as suitable starting points for loops of length up to nine. In contrast to the strategy described above, Vlijmen and Karplus selected the candidate fragments for energy minimisation (using the CHARMM [25] non-bonded energy function) from the 50 loops closed to native (which are not known in the application case). Therefore, using the current method to preselect suitable

fragments represents a very promising strategy. Recently, Soto *et al.* used the statistical potential DFIRE [249] in order to reduce the number of conformation generated in an *ab initio* search based on the Direct Tweak algorithm [241] and subsequently scored the candidates with the OPLS force field [106].

3.3.2 Comparison with other methods

In the following, the loop prediction routine presented in this work is compared to other methods based on two different test set:

- A comprehesive test set of approximately 200 loops of length 4-12 used recently by Rossi *et al.* [174] in order to benchmark 4 commercial loop prediction programs.
- A set of 14 test cases covering loops of length 4-9 which has been frequently literature used for the evaluation of different loop modelling algorithms (*e.g.* in [53, 139]). The complete test set in available online[g].

For the test set of Rossi *et al.* the prediction results of the 4 commercial programs were requested from the author directely (Karen A. Rossi). Two *ab initio* methods (Prime, Modeler) and two knowledge-based loop modelling protocols are compared in this study [174]. The 4 methods are briefly described here:

- The Loop Refinement module in *Prime* 2.5 (Schrödinger, LLC) extensively samples the conformational space by a dihedral-angle-based buildup procedure and uses the OPLS-2001 force field [106] together with the Generalized Born solvation model [82] in order to minimise and rank the loop candidates.
- The Refine Loop functionality implemented in *Modeler* (Accelrys Software Inc.) relies on conjugate gradients and molecular dynamics with simulated annealing [77] and uses the CHARMM-22 force field [25] combined with statistical potential terms.

[g] `http://www.drug-redesign.de/LIP/LIP_WebseiteErgebnisse.html`

- The Loop Sampling option in *ICM* 3.4-8 (Molsoft LLC) uses fragments extracted from a nonredundant subset of the PDB and ranks the fragments based on geometrical fit of the loop ends and sequence similarity.
- The Protein Loop Search module in *Sybyl* 7.1 (Tripos) uses a fragment database constructed from the PDB and selects the candidates based on the geometrical fit to the anchor groups. If no suitable fragments are identified an *ab initio* protocol is used.

For the two knowledge-based approaches, all fragments from proteins sharing more than 90% sequence identity to the protein of the loop test set are excluded in the study of Rossi *et al.*. Despite this rather permissive cutoff, the results (average global backbone RMSD) for both knowledge-based approaches but also for Modeler are astonishingly bad (Figure 3.32). The loop prediction method presented in this work performs consistently better than these 3 methods but slightly worse than Prime which can be attributed to the extensive sampling strategy and especially the advanced scoring function for energy minimisation and ranking used in this method.

For Prime and Sybyl as well as for the present method, the prediction accuracy drops rapidly for loops longer than 7 residues. The median of the global RMSD for all methods is greater than 2 Å for loops of length 8. If fragments originating from proteins sharing less than 50% sequence identity to the proteins of the test set are included, the performance of the present methods becomes comparable to Prime. If a cutoff of 90% is used as in the other to knowledge-based approaches, this method outperforms Prime for some loop length (length 7,8,10 and 11).

The second test set consists of 14 short and medium loops of length 4-9 and has been previously used in literature in order to test loop prediction methods [53, 90, 139]. The first two methods (column 4 and 5 in Table 3.22) are knowledge-based approches, the next three are *ab initio methods* and, finally, the method by Deane and Blundell ist a combination of both. The different methods are not described in detail here. The results of the two knowledge-based approaches need to be treated with caution and the approaches are therefore briefly described here: In LIP [139], loops are extracted from a fragment database and ranked according to the geometrical fit to the anchor residues but a very permissive filter in order to remove loops from homolues has been

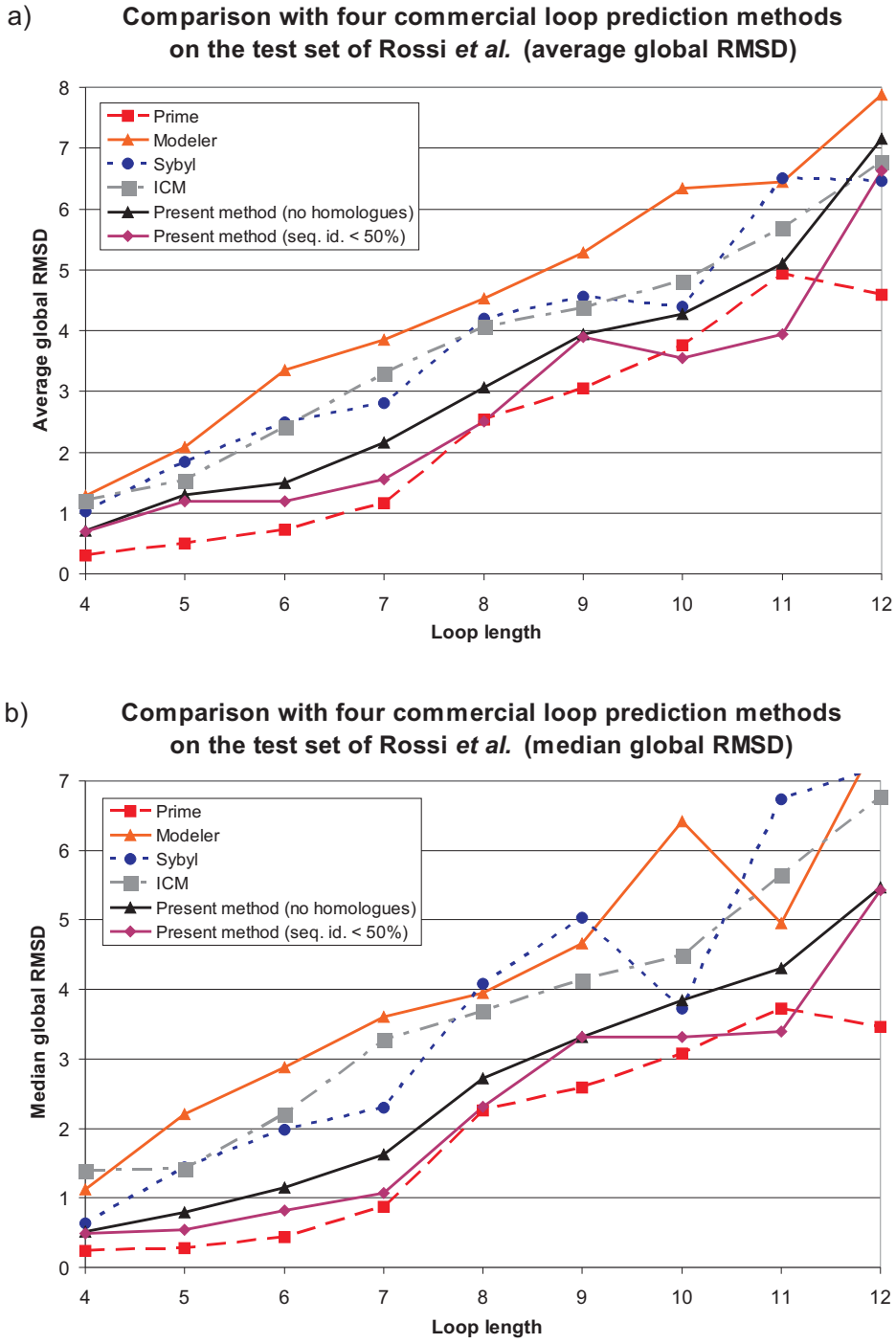


Figure 3.32: Comparison to four commercial loop prediction programs: Average (a) and median (b) RMSD on loops of lenth 4-12 of the test set of Rossi *et al.* [174].

Table 3.22: Comparison with other methods on 14 loops of length 4-9 [53, 139].

Length	PDB ID	Residues	Vlijmen *et al.* [226] *[a]	LIP [139]	Fiser *et al.* [77] *	ModLoop Server [78]	RAPPER Server [50, 56]	Deane *et al.* [53] *	CODA Server [53]	Present method
4	3dfr	20-23	2.6	1.3	1.2	1.8	1	0.4	-	1.3
5	3dfr	89-93	1.6	3.3	1	1	1.1	0.6	1.3	0.9
5	3dfr	120-124	0.5	2.1	0.3	0.4	0.6	0.7	0.7	0.7
5	3blm	131-135	0.8	0.2	0.2	0.2	0.1	0.2	0.4	0.4
6	8abp	203-208	0.3	0.8	0.4	0.4	0.5	0.8	0.8	0.7
7	8tln_E	32-38	3.7	0.3	2	3.5	3.3	1.9	2.2	2.8
7	3grs	83-89	4.6	2.4	0.4	0.6	0.4	1.4	5.3	5.9
7	5cpa	231-237	2.1	0.3	1	5.8	0.7	0.2	2.8	2.5
7	2fb4_H	26-32	1.6	0.2	4.2	4.4	0.6	0.4	0.4	0.3
7	2fbj_H	100-106	0.5	9.2	0.8	3.1	1	1.4	1.7	2.7
8	2apr	76-83	5.2	0.5	1.3	2.7	0.6	2.2	5.3	1.7
8	2act	198-205	1.6	0.1	2	2.8	3.5	3.1	6.2	5.9
8	8tln_E	248-255	1.8	0.6	0.9	3.3	0.8	1.8	3.7	2.0
9	3sgb_E	199-211	1.8	0.2	0.3	0.7	0.3	-	-	0.9

[a] Methods marked with an asterisk use an RMSD based on only 3 backbone atoms (without oxygen).

applied such that the results probably do not reflect the performace of the method in a modelling application. As mentioned in the last section, in the approach of Vlijmen *et al.* [226], the 50 loops from a database search being closest to the native loop (being unknown in the application case) are subjected to a subsequent energy minimisation using a molecular mechanics force field.

In general, the present method shows comparable results to the other methods especially for shorter loops. For some loops of length 7 and 8 (for which most of the other methods had problems as well) bad results are obtained. It sould be mentioned here, that the methods marked with an astesisk in Table 3.22 use an RMSD based only on three backbone atoms (without the oxygen) with is typically slightly lower than the RMSD over all backbone atoms. For the first loop which was predicted with a RMSD above 5 Å (3grs, 83-89) a fragment with 1.36 Å was found on rank 4. The second outlier (2act, 198-205) represents a difficult test case for the given method since it involves the formation of a disulfide bridge of the first N-terminal residue (the cystein) with the environment. As a consequence, many fragments clashed with the environment, since the protein framework was extremely close to the N-terminal anchor in this example. Given that the presence of a disulfide bridge is known before, the present method would

have benefitted from a subsequent energy minimisation step allowing the fragment to relax in the environment, adjust its orientation for the disulfide bridge.

Recently, remarkably accurate predictions have been reported also for long loops with RMSD values below 1.5 Å for loops of length 11-13 residues [251]. These results were possible if extensive sampling is used and if crystal contacts are taken into account in the scoring which reflects that conformations of longer loops observed in protein structures determined by X-ray crystallography are sometimes not native conformations observed in solution. The CPU time (AMD processor with 1.4 GHz or 900 MHz) needed for the calculation of a loop of length 11 (12, 13) took on average 12 days (19 days, 31 days) in this study! The loop prediction routine presented in this work needs on average less than 2 hours per loop prediction test case independent of the loop length (Intel Xeon 2.80 GHz). In knowledge-based loop prediction, the CPU time scales only marginally with the loop length in constrast to *ab initio* methods which often show an exponential relationship. The vast majority of the computation time in the present method is spent on the calculation of the sidechain orientations for the 3000 loops in the final selection. The speed of sidechain prediction step highly depends on the presence of close atoms (potential clashes) in the structural environment. The selection of the fragments from the MySQL database as well as the application of all filters takes typically only a few minutes depending on the network connection since a considerable amount of data (mainly of the loop coordinates) have to be transfered. The computation time can be accelerated if stricter cutoffs are used in the filtering step and therefore fewer sidechain orientations have to be predicted.

3.4 Local model quality assessment and anchor group prediction

In this section the applicability of statistical potentials for the assessment of the local model accuracy is discussed briefly, since an extensive evaluation was not the scope of this work. The aim is to show that a local model quality analysis is possible. Furthermore it is analysed whether local model energy profiles can used in order to predict the location of anchor groups serving as starting points for the loop prediction process.

3.4.1 Local model quality assessment

As an example, the energy profile of our first model submitted to the CASP7 target T0373 is shown in Figure 3.33 together with the residue-specific backbone RMSD between the model and the corresponding experimental structure (lower curve).

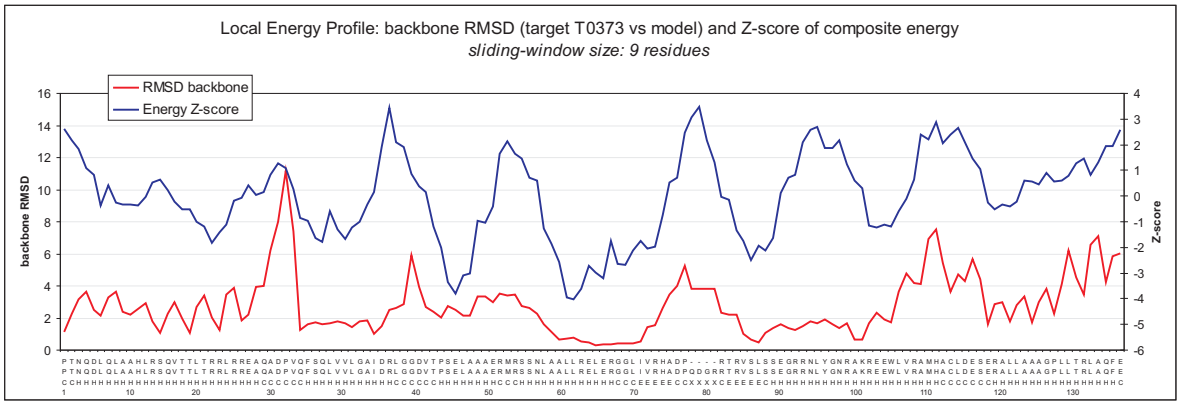


Figure 3.33: Example of a model energy profile for model 1 submitted for target T0373. The per-residue RMSD is given in the lower curve.

The energy profile was obtained, as described in Methods (Chapter 2.4.5), by adding up the per-residue energies in a sliding window of size 9 and by combining the three statistical potenial terms (torsion angle potential over three residues, all-atom solvation potential and short-range all-atom interaction potential) based on Z-scores over the entire model. The x-axis shows the sequences of the experimental structure and of the

model, respectively (the gap indicates that four residues have not been resolved in the experimental structure), together with the secondary structure of the target.

A clear relation between energy and model accuracy can be observed: the peaks in the upper curve, representing regions of high energy, coincide with the local model accuracy expressed by the structural deviation between target and model. Similar results have been obtained for other models. The correlation between peak height and extend of structural deviation is less pronounced which can be partly attributed to the simple strategy used to combine the different statistical potential terms based on Z-scores. Especially the predicted model accuracy based on the interaction potential (and also the solvation potential) should be treated with caution: Since interaction potentials are two-body potentials (in contrast to single-body potentials such as the torsion angle potential), the high energy resulting from a unfavourable interaction is assigned to both partners. For example, a solvent exposed loop lying against the wrong region of the protein surface gets assigned high energies as a consequence of the unfavourable interactions and the loop regions is therefore predicted to be of low accuracy. On the other hand, the same holds for the residues in contact with the loop although the high energies can to some extent be compensated by other, more favourable interactions with the structural environment (*e.g.* with residues of the protein core). In this given situation, the location of only one interaction partner is wrong and therefore the high energy (*i.e.* the predicted low model accuracy) should be assigned to one of the interaction partners, in this case to the loop.

The second last peak in the energy profile given in Figure 3.33 represents such an example: The helix in this region (residues 95-111) is approximatielly correct, despite a small shift with respect to the experimental structure. The residues have a backbone RMSD below 2 Å, but since the helix is in contact with a loops showing serious deviations from the native conformation (residues 27-35), this region gets assigned a high energy. An extract of the structural superposition of the model and the corresponding experimental structure is shown in Figure 3.34 (Cα atoms only). The wrong loop as well as the part of the nearby helix which both got assigned high energies in the profile shown above are marked in bold.

Single-body potentials, such as the torsion angle potential, do not have this problem. A possible strategy could be to use the torsion angle energy of the interaction partners

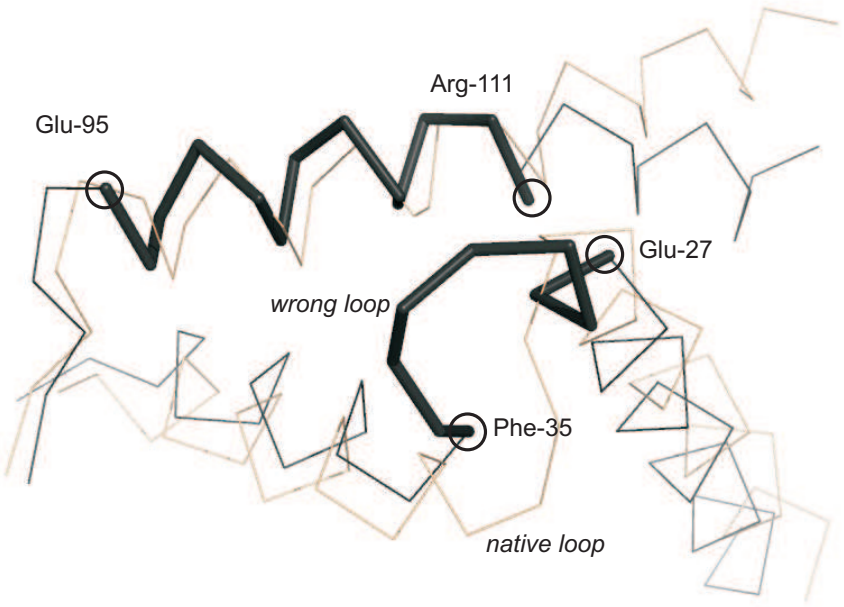


Figure 3.34: Extract of the superposition between the experimental structure of target T0373 (light grey) and the model with an incorrect loop in contact with a nearby helix (dark grey).

in order to assign the high interaction energy to one of the participating residues. The secondary structure constitution of both regions can also be taken into account, since loop regions are more likely incorrect than helix and sheets which are usually part of the structural core. Anyway, the preliminary but promising results indicate that the statistical potentials developed in this work can be used in the analysis of the local model accuracy. In future developments the combination of the terms should be optimised on a comprehensive test set. Two recent publications concerning local model quality assessment use support vector machines [68] and artificial neural networks [234], respectively, in order to combine multiple terms. The use of machine learning algorithms in order to combine different terms in a composite scoring function is surely a resonable approach. The authors do not address the problem of two-body potentials for local model quality assessment although machine learning algorithms can possibly cope with this situation if implemented correctly. A future implemtation of the local energy function should take this into account.

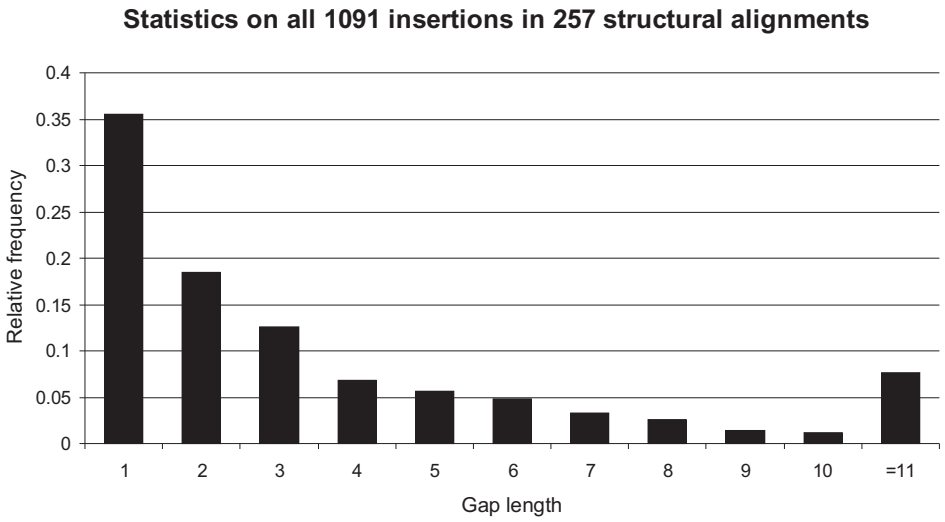


Figure 3.35: Statistical analysis of insertions in a set of 257 structural alignments between pairs of homologous proteins.

3.4.2 Analysis of the anchor region around gaps

In this section, a statistical analysis of the length of insertions and deletions occurring in typical modelling situations is performed based on a comprehensive set of structural alignments obtained from the HOMSTRAD database [142] (see Methods on page 70). Furthermore, the structural consequences of isolated insertions and deletions in loops is investigated and the region around the gaps is analysed for the location of suitable anchor groups. Several strategies for the prediction of anchor groups are discussed.

Figure 3.35 and 3.36 show the distribution of gap lengths for 1091 insertions and 945 deletions extracted from a non-redundant set of 257 structural alignments between pair of homologous proteins sharing less than 40% sequence identity representing realistic modelling situations. More than 35% of all gaps are of length 1. 73% of all insertions and 77% of all deletions are smaller than 5 residues. The distribution of the gap lengths for insertion and deletions is quite similar.

In Table 3.23, the results of the analysis of the local structural environment around the gaps is shown. The analysis of the 257 structure-based sequence alignments reveals that approximately 10% of the insertions and 15% of the deletions are located in

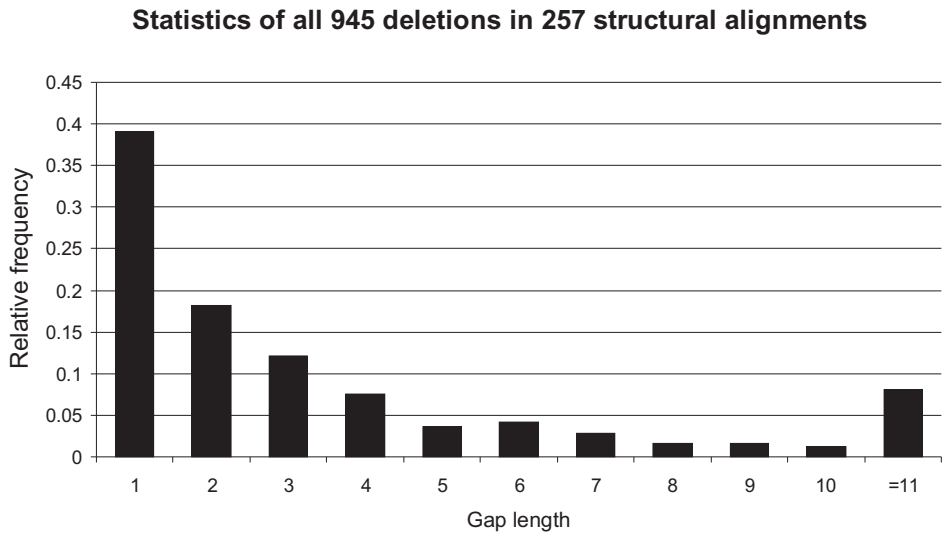


Figure 3.36: Statistical analysis of deletions in a set of 257 structural alignments between pairs of homologous proteins.

within secondary structure elements. Among those, 58% of the insertions and 55% of the deletions are close of the end of the secondary structure elements (*i.e.* not more then 2 residues apart from the next loop region). These results underline the advantage of being able to remodel parts of secondary structure elements (*e.g.* by extending or truncating secondary structure elements as part of the loop prediction process). In contrast to most knowledge-based loop modelling procedures described in literature which are specialisied on the prediction of "pure" loop regions, the method described in this work is able to model any structural segment.

The majority of the gaps are located within loop regions. From those, 642 (632) of the insertions (deletions) have secondary structure elements within 10 residues on both sides. The remaining gaps are located in longer loops (of at least 10 residues), 119 (157) of them are longer than 20 residues in the insertion (deletion) test set.

The region around the gaps has been inspected for possible anchor groups. In analogy to Lessel and Schomburg [121], at least 3 consecutive residues with an RMSD below 1.8 Å with respect to the corresponding residues in the alignment have to be present on both sides of the gap. In the given test set, only 16% of the insertion 23% of

Table 3.23: Analysis of the structural environment of 1091 insertions and 945 deletions in 257 structural alignments.

description	# insertions	# deletions
gaps in secondary structure elements (SSE)	108	145
gaps within SSE but with 2 residues of SSE-end	63	80
gaps within loops with SSE begin within 10 residues	642	632
gaps with 3 alignable residues on both sides[a]	177	214
gaps with 2 alignable residues on both sides	266	295
gaps with no residue $<$ 1.8 Å RMSD within 10 residues	216	179
gaps with neighbouring gap within 10 residues	504	442
gaps with neighbouring gap within 8 residues	258	259
gaps with neighbouring separated by $<$ 4 residues	50	51
final number of gaps in "anchor group test set"	112	124
total number of gaps	1091	945

[a] At least 3 consecutive residues with an RMSD below 1.8 Å are found on both sides of the gap.

the deletion fulfill this condition. The percentages raise to 24% and 31% if only 2 residues on both sides are required. The different percentages observed for insertions and deletions confirm the expected stronger influence of insertions on the structural environment compared to deletions. For approximately 20% of the gaps, non of 10 residue on both sides has an RMSD below 1.8 Å. These results show that there are often considerable local deviations between pairs of homologous proteins in the potential anchor regions. This can be partly attributed to the presence of remote homologues in the test set representing difficult modelling test cases (one quarter of the pairs have a sequence identity below 20%). Furthermore, as the sequence identity decreases, the secondary structure elements of the structural core are often slightly displaced between the homologues. If multiple homologues (templates) are present in the modelling process, using different parts of different templates can potentially improve the coverage and bring the model closer to the experimental structure of the target. The identification of regions where the model can benefit from fragments of other templates is not a trivial task. A local scoring function, as described in the last section, can potentially support the decision.

46% of the insertions and 47% of the deletions have a neighbouring gap within 10 residues. If the neighbouring gap is close (*e.g.* separated by less than 4 residues as

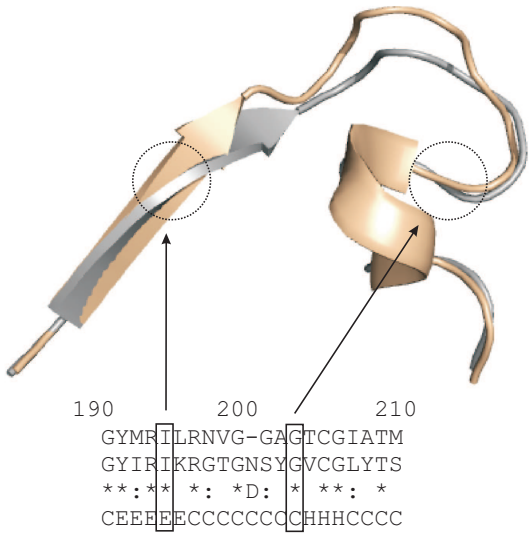


Figure 3.37: Schematic representation of the anchor group prediction problem.

observed for a total of 101 gaps in the test set) they would be definitively merged and modelled in one step. Otherwise, it has to be decided in the modelling process whether these gaps are merged and modelled by a longer loop or whether they are treated separately. In the later case, structurally conserved residues have to be present between the gaps serving as anchor groups. As can be seen from Table 3.23 this situation occurs quite often. The analysis of the local energy profile can possibly help indentifying structurally conserved residues.

A subset of 112 insertions and 124 deletions has been extracted from all gaps from the test set by applying the criteria described in Methods (Chapter 2.4.6.2). The regions around the gaps are analysed and different strategies for the positioning of anchor groups are compared in the following. A schematic representation of the anchor group prediction problem is given in Figure 3.37. An extract of the structural alignment and the corresponding sequence alignment of a pair of distantly homologous proteins is shown. The target structure is coloured in grey and refers to the first sequence in the alignment. The superposition points out the structural consequences of the 1-residue deletion observed in the loop region.

Anchor group prediction refers to the attempt to identify those regions on both sides

of the gap (or any structurally non-conserved loop to be remodelled) where the target structure begins to deviate from the template and therefore the backbone coordinates cannot be simply copied. In the given example the anchor groups are positioned close to the end points of the surrounding helix and sheet, respectively, and sequence conservation has been taken into account. On the C-terminal side of the deletion, glycine 204 has been used as anchor group, which, by looking at the superposition, turned out to be a good decision. On the N-terminal side, the anchor group has been placed within the strand, resulting in 8 residues to be remodelled. The conserved arginine immediatelly after the strand represents another possible anchor and would reduce the number of residues to model by two and a shorter loop can potentially be predicted more accurately (see Chapter 3.3).

This highlights the problemtic situation in anchor group prediction: a reasonable compromise between accuracy of the anchor groups and length of the fragment to be remodelled has to be found which is not a trivial task and difficult to automate. As shown exemplarily in Figure 3.33, regions of low energy in the energy profile of a model often correspond to structurally conserved segments representing promising anchor groups for the loop modelling process. The energy profiles are based on a sliding window of size of 5 using the central residue together with the 4 neighbouring residues in direction away from the gap. A variety of other implementations have been tested but resulted in a worse performance. Figure 3.38 shows that there is indeed a correlation between the local structural deviation as expressed by the S-score (see definition in Methods on page 72) between target and template and the local energy, although not very pronounced.

Table 3.24 and Table 3.25 show the average loop lengths and RMSDs of the anchor group residues between target and template for different anchor group prediction strategies on the test sets of 112 insertions and 124 deletions. Approaches with and without the use of information obtained from the energy profiles are compared and related to an "optimal" anchor group positions (*i.e.* if the RMSD between taget and template is assumed to be known). For the insertion test set, an average backbone RMSD of 0.87 Å is achieved if the anchors with minimal RMSD within 10 residues on both sides of the gap are taken. This results in an average loop length of 14.59 residues which is too long for accurate loop modelling. If the first anchor groups (starting from

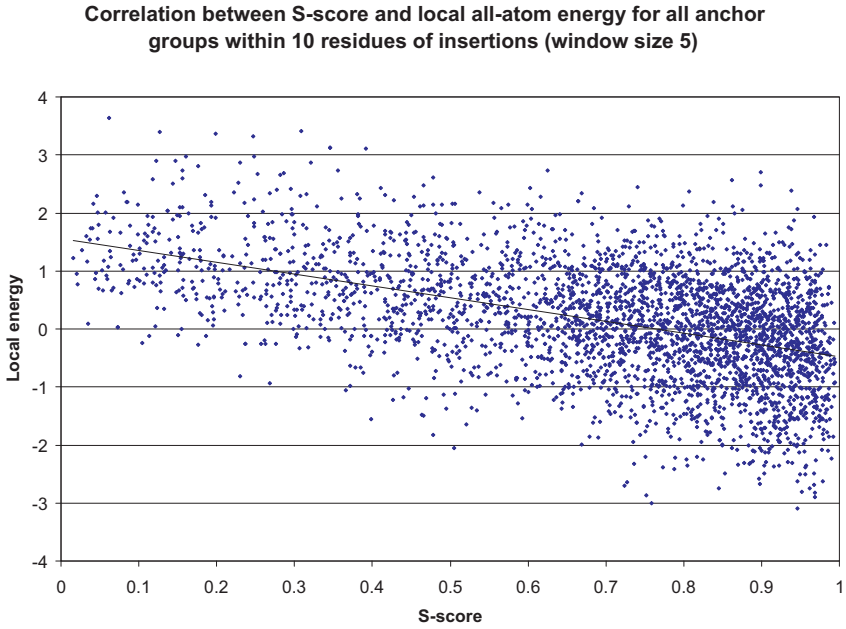


Figure 3.38: Regression between S-score (as a measure for the local deviation between target and template, definition in Chapter 2.4.6.2) and local energy.

the gap) with an RMSD below 1.5 Å (2Å) are used, an average loop length of 9.26 (7.24) residues is achieved which are reasonable loop length for modelling. For the deletion test set, the average loop lengths (and also the RMSDs) are lower as expected since the "gap residues" are not modelled in this case. Even these "optimal" anchor groups show on average considerable deviations from the native structure. This has to be taken into account in the loop ranking process of knowledge-based approaches: Loop ranking methods with only rely on the geometrical fit of the fragments on the anchor groups are potentially not applicable in realistic modelling situations. In the loop prediction method described in this work, this criteria has not been used (in constrast to most existing algorithms) and the ranking has been performed based on the interaction potentials as described in Chapter 3.3.

The results for the deletion test set are not discussed in detail here. Deletions are typically much easier to model than insertion since the structural consequences of deletions on the surrounding residues are less pronounced. A simple strategy of using

Table 3.24: Comparison of different anchor group prediction strategies on a test set of 112 insertions.

strategy used for anchor group positioning	ØRMSD	Øloop length
fixed distance from gap: 1 residue	3.11	3.42
fixed distance from gap: 2 residues	2.40	5.42
fixed distance from gap: 3 residues	2.00	7.42
fixed distance from gap: 4 residues	1.67	9.42
energy minimum within 3 residues (all-atom)[a]	2.22	6.17
energy minimum within 3 residues (3 terms)[b]	2.29	6.11
energy minimum within 4 residues (all-atom)	1.95	7.42
energy minimum within 4 residues (3 terms)	2.05	7.21
fixed depth in SSE: 0 residues (SSE begin)	1.92	8.49
fixed depth in SSE: 1 residues	1.66	10.49
energy minimum around SSE end (all-atom)	1.92	8.22
energy minimum around SSE end (3 terms)	2.09	7.75
global energy minimum within 10 residues (3 terms)	1.39	13.36
anchors with lowest RMSD	0.87	14.59
first anchors with RMSD < 1.5 Å	1.46	9.26
first anchors with RMSD < 2 Å	1.69	7.24

[a] The minimum in the energy profile based on the all-atom interaction potential is taken.
[b] A combination of the all-atom interaction potential, the torsion potential and the solvation potential is used.

Table 3.25: Comparison of different anchor group prediction strategies on a test set of 124 deletions.

strategy used for anchor group positioning	ØRMSD	Øloop length
fixed distance from gap: 1 residue	3.24	2
fixed distance from gap: 2 residues	2.27	4
fixed distance from gap: 3 residues	1.76	6
fixed distance from gap: 4 residues	1.46	8
energy minimum within 3 residues (all-atom)	1.91	5.13
energy minimum within 3 residues (3 terms)	2.03	4.99
energy minimum within 4 residues (all-atom)	1.72	6.68
energy minimum within 4 residues (3 terms)	1.75	6.39
fixed depth in SSE: 0 residues (SSE begin)	1.98	6.27
fixed depth in SSE: 1 residues	1.52	8.27
energy minimum around SSE end (all-atom)	1.76	6.68
energy minimum around SSE end (3 terms)	2.06	5.88
global energy minimum within 10 residues (3 terms)	1.36	12.48
anchors with lowest RMSD	0.76	13.15
first anchors with RMSD < 1.5 Å	1.35	7.21
first anchors with RMSD < 2 Å	1.63	5.04

anchor groups approximately 3-4 residues away from the gap results in better anchor groups than any other, more sophisticated approach. The average length of the loop to be remodelled in this approach is between 6 and 8 residues.

The first four lines in Table 3.24 show the average loop lengths and RMSDs if fixed anchor group positions relative to the gap are used for distance of 1 to 4 residues. For insertions, the probably best compromise between loop length and RMSD of the anchor groups is approximately 3 residues away from the gap (average RMSD 2Å loop length 7.42). If the energy profile is taken into account the RMSD or the loop length can be slighly lowered. If the anchor groups are positioned within the surrounding secondary structure elements, lower RMSDs can be achieved but only at the cost of longer loops. This can be attributed to the fact that (for longer loops) the secondary elements can be far away. Depending on the structural conservation, anchor groups closer to the gap can possibly be used. If the energy profile is taken into account (using a combination of three statistical potential terms), the average loop length can be reduced from 8.49 to 7.75 at the cost of a slightly higher RMSD. Additional characteristics of the potential anchor residues, such as hydrophobicity, solvent accessibility and sequence conservation have been also taken into account (as suggested by Wohlfahrt *et al.* [238]) but did not improve the prediction over the statistical potentials. This can be attributed to the fact that these factors are to some extent covered by the statistical potential terms. The approach of simple adding Z-scores of the terms is also not optimal.

Generally, the use of information about the local energy of the candidate anchor groups, did not result in a considerably better predictions. Local energy functions are possibly to imprecise for the prediction of exact locations (on the level of single residues) and are more appropriate for the identification of segments of structural deviation which can be subjected to refinement in order to bring the model close to the experimental structure or for loop prediction.

Another factor complicating the automation of the anchor group prediction task is closely connected with the knowledge-based approach to loop prediction used in this work: the spacial orientation of the database fragment after fitting on the anchor group atoms, is highly sensitive to distortions of the anchor geometry. Thus, it is not only important to position the anchor groups near the end of the structurally conserved region of the template, but also to take into account that a suitable fragment with

a similar overall geometry and showing a correct orientation after fitting has to be present in the database. A worse anchor group in terms of backbone deviation from the target structure can still result in better loop modelling results if a loop with a better orientation after fitting is present in the database or if the gap can be bridged by a shorter fragment which can potentially be predicted more accurately. For the knowledge-based loop prediction routine presented in this work (Chapter 3.3), the prediction quality decreases considerably between loops of length 7 and 8 residues.

The best strategy to cope with the uncertainties concerning anchor groups selection and loop modelling is to use multiple alternative anchor groups and a set of top-scoring loops for each combination in the modelling process and to subsequently select the best prediction based on the quality of the final model. The QMEAN scoring function [16] presented in this work (Chapter 3.2) can be used for this task since it is both fast and reliable in discriminating good from bad models.

4 Conclusions and Outlook

The prediction of the 3-dimensional structure of a protein from its sequence is greatly facilitated by the presence of proteins with experimental structure sharing an evolutionary relationship to the target protein (homology modelling). The aim of this work was to establish a loop prediction methods which optimally takes advantage of the growing number of proteins present in the database of known protein structures. Furthermore, scoring functions need to be implemented which can be used for the ranking of candidate fragments in loop modelling and for the assessment of the quality of the generated models. Both tasks are of crucial importance for the final applicability of the models. As a framework in order to deal with loop prediction and model quality assessment, a complete homology modelling pipeline has been established.

The homology modelling pipeline has been tested at the seventh round of the community-wide CASP experiment in summer 2006. The results on the 18 investigated targets confirmed that the modelling pipeline is able to produce very accurate homology models: 3 extraordinarily good predictions have been submitted (rank 2, 4 and 6 of over 130 participating groups) and the vast majority of remaining targets have been modelled above the community average. Several factors are responsible for these results: beside a good strategy for template identification and alignment building, the ability to not only remodel loop regions but any structural segment (*e.g.* chain ends or segments containing secondary structure elements) is an important ingredient together with the scoring function used to assess to quality of the produced models and to select of the most reliable candidate.

A composite scoring function (called QMEAN) has been presented consisting of three statistical potential terms covering the major aspects of protein stability and two additional terms describing the agreement of predicted and calculated secondary structure and solvent accessibility, respectively. QMEAN has been shown to be a valuable tool for the discrimination of good from bad models and performs significantly better than five well-established methods on a comprehensive test set of 22,420 models from CASP7. Some of the scoring function terms turned out to be more specialised for

a specific task (e.g. the torsion angle potential over 3 consecutive residues developed in this work turned out to be very effective in recognising the native fold) whereas other factors are more widely applicable. The results confirm that a combination of multiple terms increases the performance of the scoring function by taking advantage of the strengths of certain terms for a specific task while reducing a possibly negative contribution of other terms. The statistically significant improvement in performance of QMEAN over five methods gets even more pronounced when taking into account that a simple linear combination was used in order to combine the different terms to the final scoring function. The performance of the QMEAN scoring function can potentially be improved by the application of machine learning algorithms for the combination of the terms and by using specialised versions of the scoring function depending on the resolution of the models (*e.g.* by using a fine-grained all-atom implementation for the assessment of models generated by comparative modelling and residue-level potentials for the analysis of rough models predicted by *ab initio* methods).

The loop modelling routine presented in this work combines a knowledge-based approach for conformational sampling based on a comprehensive fragment database with a knowledge-based approach for scoring of the selected fragments based on an specialised all-atom interaction potential. In contrast to other database loop prediction approaches described in the literature, loop ranking is performed based on the complete loop including sidechains. The presented method is able to accurately model loops of length up to 7 residues and outperforms 3 of 4 commercial loop prediction programs on a comprehensive test set of over 200 loops of length 4-12 residues. An average (median) global backbone RMSD of 0.66 Å (0.51 Å) and 1.63 Å(1.35 Å) is obtained for loops of length 4 and 6, respectively. If fragments from proteins sharing less than 50% sequence identity to the proteins in the loops test set are included, the median prediction accuracy drop below 1 Å per loop length for loops up to 7 residues. For loops longer than 8 residues the prediction accuracy drops as a consequence of the database incompleteness and the fact that the orientation of the fragments after fitting in the protein framework is only approximately correct resulting in an atomic displacement increasing with the loop length. A subsequent energy minimisation step using a molecular mechanics force field can counteract the inherent problems of database loop prediction approaches. In this way, the loop can be annealed with the anchor

groups and at the same time the loop conformation can be relaxed in the structural environment. Energy minimisation and re-ranking of the top scoring loops generated with the given method represents a very promising strategy to extend the applicability of knowledge-based loop prediction approaches toward longer loop lengths.

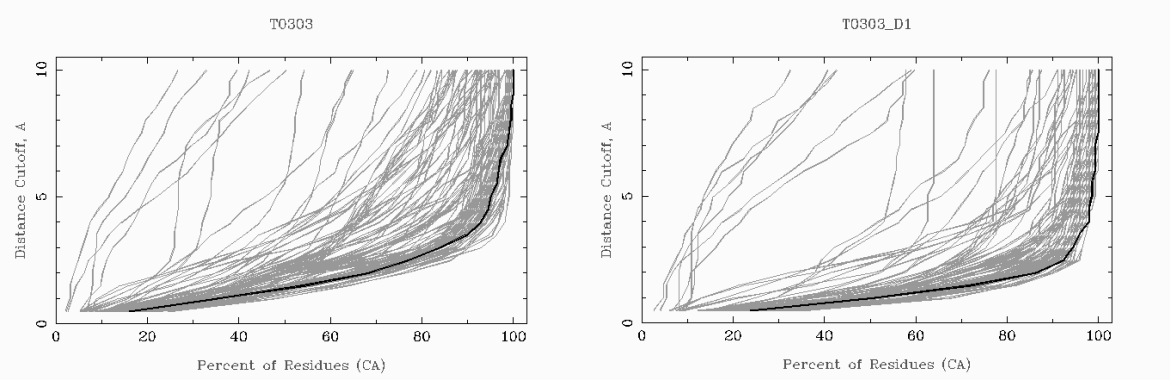
A prediction of suitable anchor groups serving as starting points for loop prediction based on the analysis of the local model energy around insertions and deletions turned out to perform only marginally better than placing the anchor groups at a fix distance from the gap and near the end of the surrounding secondary structure elements. Anchor groups should be placed at the end of the structurally conserved region of the template structure (*i.e.* in the region where target and template begin to deviate) and at the same time, the length of the loop to be remodelled should be kept as short as possible. In the context of knowledge-based loop prediction, another factor influences the location of the optimal anchor groups: A fragment with a locally correct geometry needs to be present in the database which, after fitting on the anchor groups, approximately shows a correct orientation with respect to the protein framework. Due to the interplay of all these factors, the best approach is to use several alternative anchor groups in the modelling process.

A reasonable future extensions of this work represents the automation of the whole modelling process. The best strategy in order to cope with the multitude of factors influencing the accuracy of protein structure models is to generate a vast amount of alternative models (e.g. by using multiple templates, alternative alignments, different anchor groups and several loop conformations) and to subsequently select the final model based on the scoring function described in this work.

5 Appendix

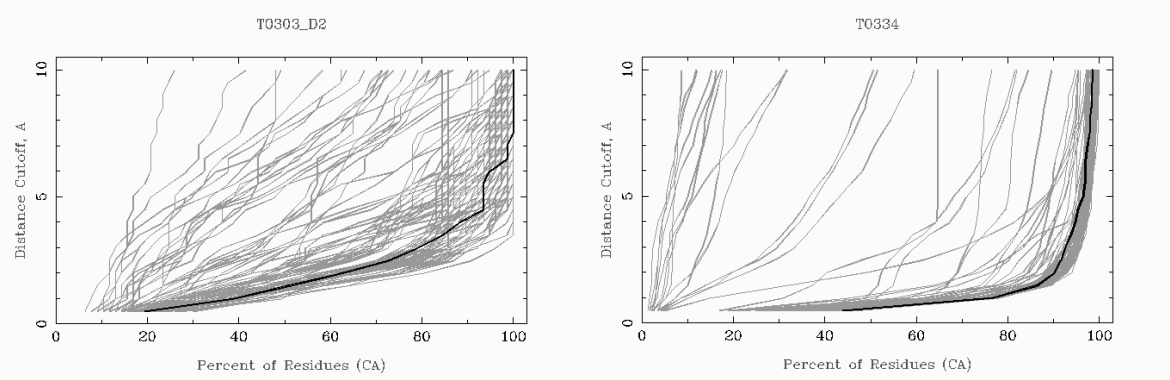
Table 5.1: Classification of the 95 target of CASP7 according to their difficulty in free modelling (FM), template-based modelling (TBM) and high-accuracy template-based modelling (HA-TMB) targets. HA-TBM are a subsection of TBM targets.

category	**targets**
FM	T0287, T0296, T0300, T0304, T0307, T0309, T0314, T0316, T0319, T0321, T0347, T0348, T0350, T0353, T0356, T0361, T0382, T0386
TBM	T0283, T0284, T0285, T0286, T0288, T0289, T0290, T0291, T0292, T0293, T0295, T0297, T0298, T0299, T0301, T0302, T0303, T0305, T0306, T0308, T0311, T0312, T0313, T0315, T0317, T0318, T0320, T0322, T0323, T0324, T0325, T0326, T0327, T0328, T0329, T0330, T0331, T0332, T0333, T0334, T0335, T0338, T0339, T0340, T0341, T0342, T0345, T0346, T0349, T0351, T0354, T0357, T0358, T0359, T0360, T0362, T0363, T0364, T0365, T0366, T0367, T0368, T0369, T0370, T0371, T0372, T0373, T0374, T0375, T0376, T0378, T0379, T0380, T0381, T0383, T0384, T0385
HA-TBM	T0288, T0290, T0291, T0292, T0295, T0302, T0305, T0308, T0311, T0313, T0315, T0317, T0324, T0326, T0328, T0332, T0334, T0340, T0345, T0346, T0359, T0366, T0367



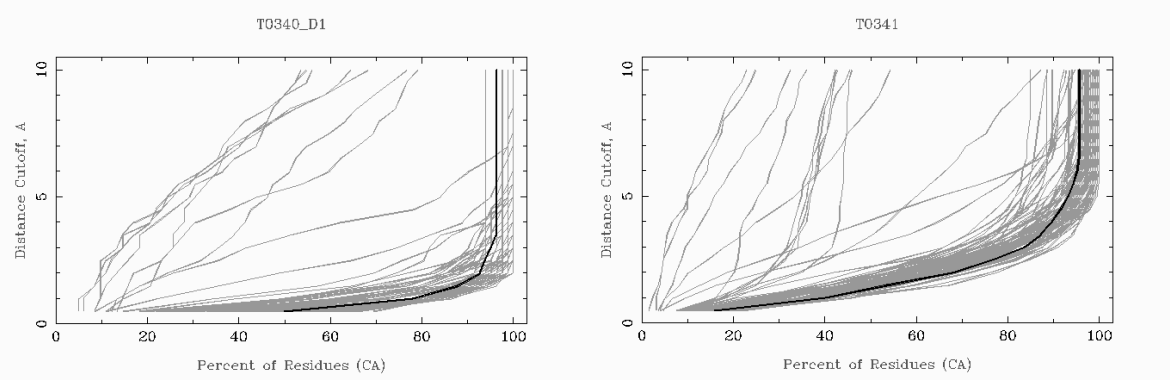
(a) T0303

(b) T0303, domain 1



(c) T0303, domain 2

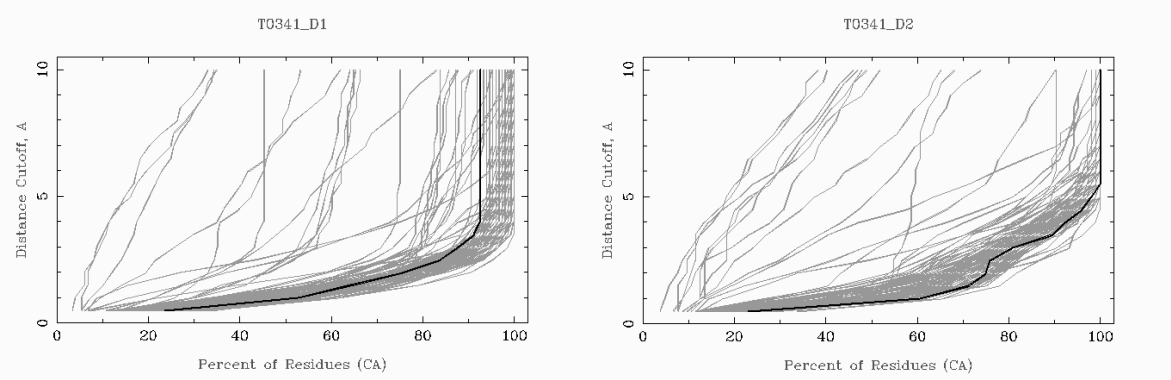
(d) T0334



(e) T0340

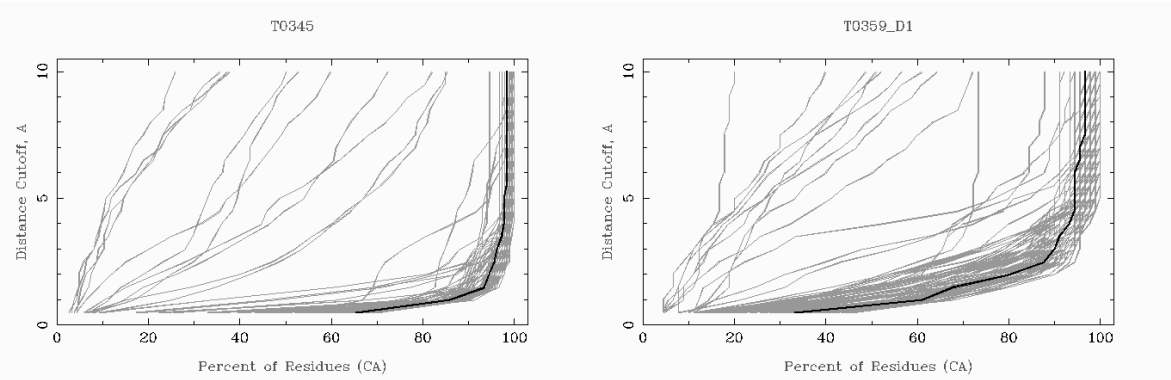
(f) T0341

Figure 5.1: GDT plot of all targets processed by our group (1/5).



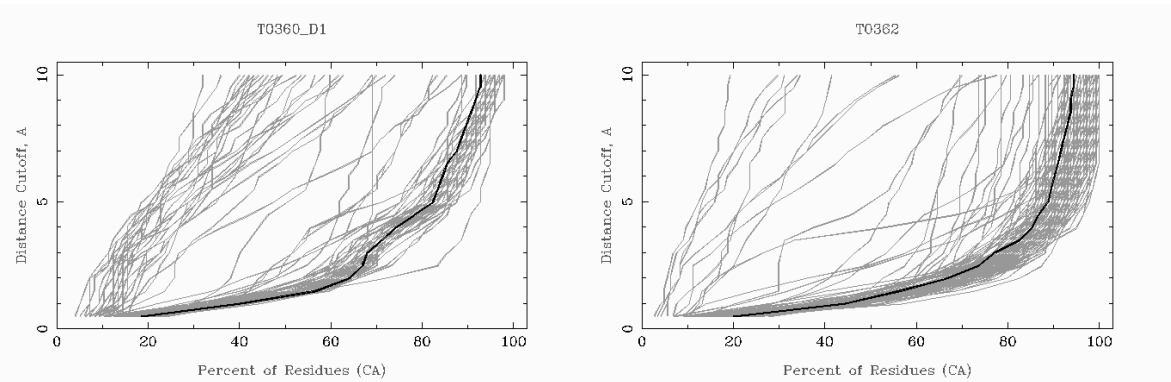
(g) T0341, domain 1

(h) T0341, domain 2



(i) T0345

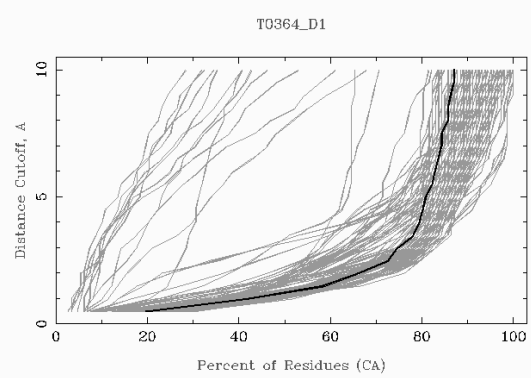
(j) T0359



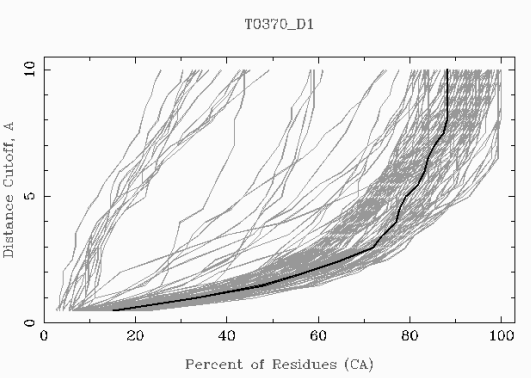
(k) T0360

(l) T0362

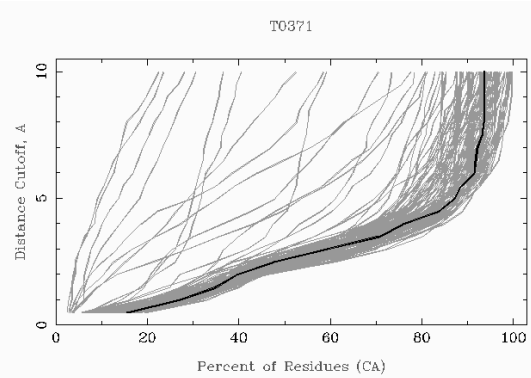
Figure 5.1: GDT plot of all targets processed by our group (2/5).



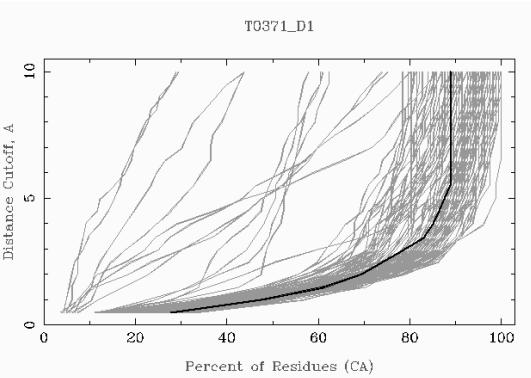
(m) T0364



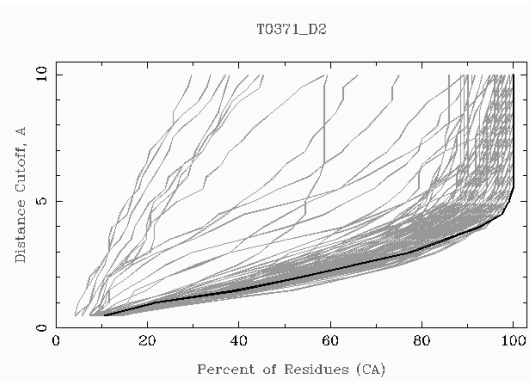
(n) T0370



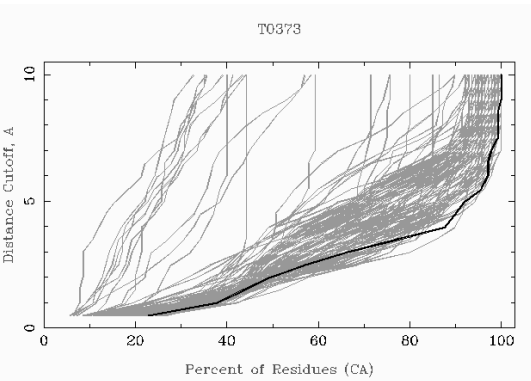
(o) T0371



(p) T0371, domain 1

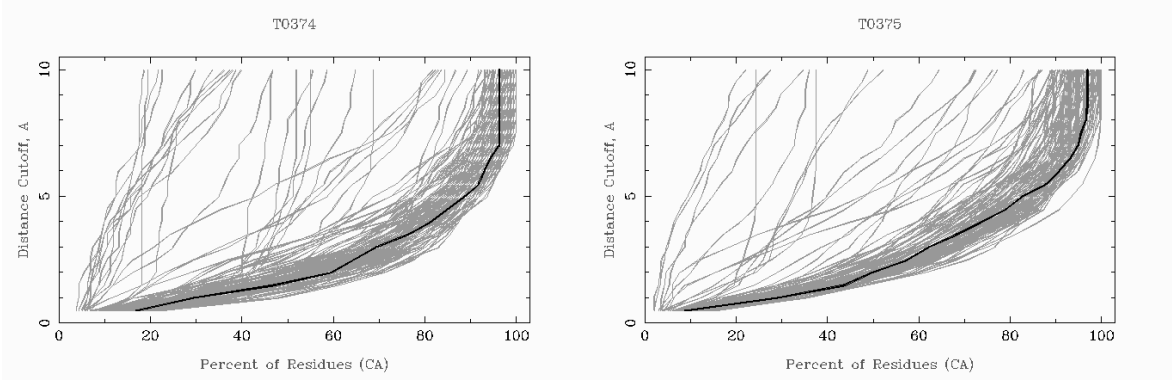


(q) T0371, domain 2



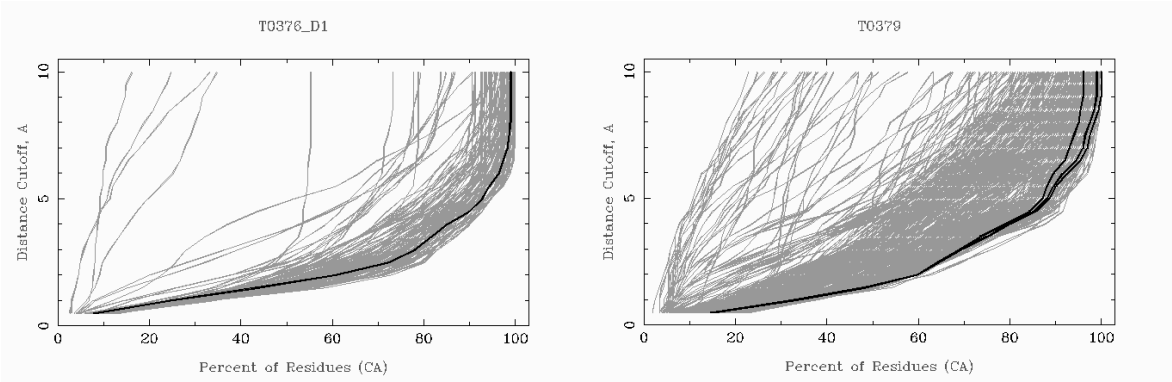
(r) T0373

Figure 5.1: GDT plot of all targets processed by our group (3/5).



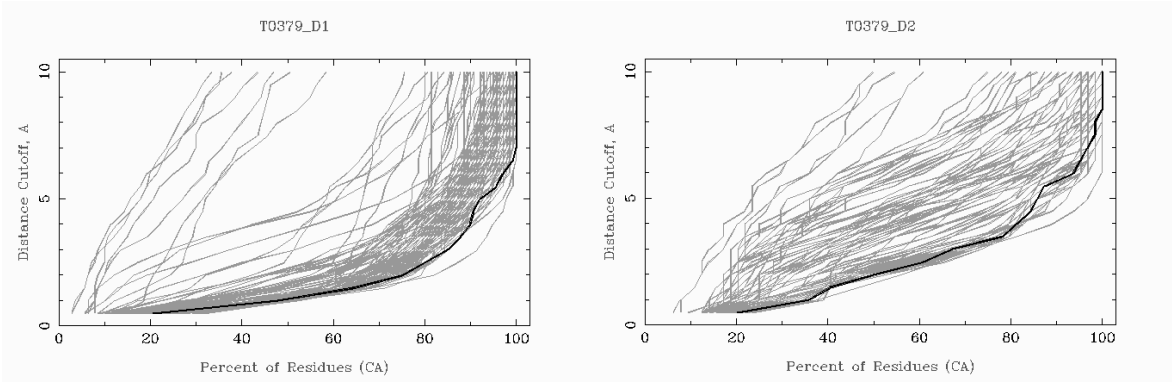
(s) T0374

(t) T0375



(u) T0376

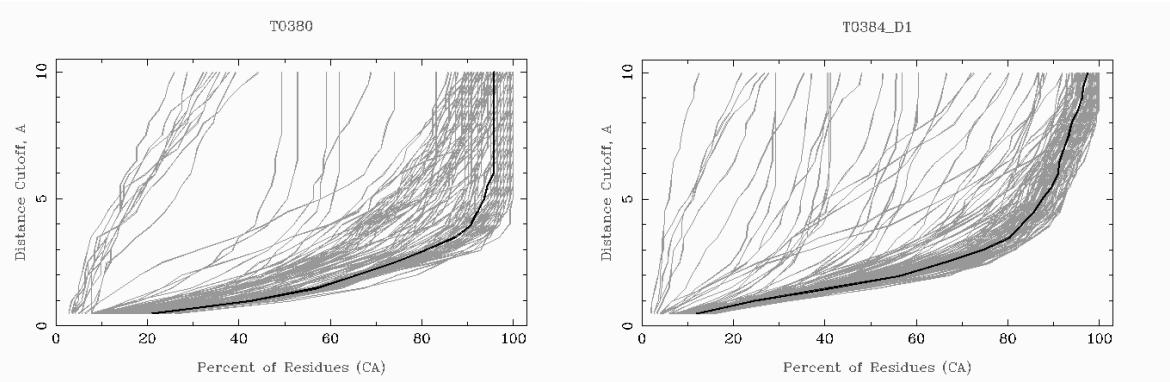
(v) T0379



(w) T0379, domain 1

(x) T0379, domain 2

Figure 5.1: GDT plot of all targets processed by our group (4/5).



(y) T0380 (z) T0384

Figure 5.1: GDT plot of all targets processed by our group (5/5).

Table 5.2: Performance of different scoring functions in predicting the quality of the server models submitted for the 77 CASP7 targets of the category template-based modelling.

	regression[a]		**enrichment**[b]		**best predicted model**[c]			**best GDT_TS model**[d]		**native structure**[e]		
Method	r^2	*rho*	*F.E.*	$E_{15\%}$	r_{10}	$logP_{B1}$	$logP_{B10}$	ΔGDT_TS	r_1	r_{10}	Z_{nat}	r_1	r_{10}
Modcheck	0.68	0.61	0.32	2.63	12	-0.66	-1.63	-0.2	5	22	1.87	39	58
RAPDF	-0.53	0.52	0.31	2.48	13	-0.86	-1.64	-0.08	3	13	-1.97	46	63
DFIRE	-0.41	0.56	0.31	2.66	16	-0.96	-1.67	**-0.07**	4	14	-1.18	47	58
ProQ	0.39	0.28	0.12	1.1	3	-0.3	-0.96	-0.23	0	6	1.39	9	24
$ProQ_{SSE}$	0.57	0.44	0.17	1.59	7	-0.49	-1.12	-0.17	2	8	1.55	10	32
FRST	-0.6	0.55	0.29	2.27	18	-0.91	-1.72	-0.08	**6**	18	-2.37	49	60
QMEAN3	-0.69	0.62	0.32	2.48	15	-0.8	-1.8	-0.13	1	28	-2.16	50	61
QMEAN4	-0.76	0.66	0.37	2.73	22	-0.97	-1.91	-0.08	4	32	-1.76	47	56
QMEAN5	**-0.77**	**0.67**	**0.39**	**2.87**	**24**	**-1.01**	**-1.93**	-0.08	5	**33**	-1.76	47	58
torsion single	-0.48	0.42	0.22	1.76	6	-0.62	-1.47	-0.12	0	11	-2.17	47	60
torsion3-residue	-0.57	0.47	0.21	1.8	9	-0.72	-1.49	-0.12	1	8	**-2.64**	**51**	**65**
pairwise$C\beta$	-0.62	0.54	0.28	2.42	15	-0.66	-1.68	-0.19	4	21	-1.84	32	56
pairwise$C\beta$/SSE	-0.63	0.56	0.32	2.52	17	-0.78	-1.8	-0.14	5	29	-2.04	38	56
solvation	-0.59	0.52	0.26	2.22	6	-0.47	-1.6	-0.27	0	20	-1.2	14	36
SSEPSIPRED	-0.71	0.54	0.23	2.03	7	-0.63	-1.44	-0.13	2	15	-0.83	6	20
ACCpro	-0.62	0.58	0.34	2.71	17	-0.85	-1.62	-0.11	5	25	-1.19	13	32

[a] Pearson's correlation coefficient r^2 and Spearmans's rank correlation coefficient *rho*

[b] *F.E.* stands for fraction enrichment and $E_{15\%}$ is the enrichment among the top 15% best predicted models as compared to a random selection.

[c] r_{10} are the number of targets for with the top-scoring models is among the top10 best models (based on GDT_TS). $logP_{B1}$ and $logP_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the ten best-scoring models, respectively.

[d] GDT_TS loss is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set.r1 and r10 are the number of targets in which the best model based on GDT_TS, excluding the native structure was found on the first rank or among the top 10 predictions.

[e] Z_{nat} is the Z-score of the native structure as compared to the ensemble of models. r1 and r10 are the number of targets in which the native structure was found on the first rank or among the top 10 predictions.

Table 5.4: Performance of different scoring functions in predicting the quality of the server models submitted for the 18 free modelling targets of CASP7.

	regression[a]		**enrichment**[b]		**best predicted model**[c]			**best GDT_TS model**[d]			**native structure**[e]	
Method	r^2	*rho*	*F.E.*	$E_{15\%}$	r_{10}	$logP_{B1}$	$logP_{B10}$	ΔGDT_TS	r_1	r_{10}	Z_{nat}	r_1	r_{10}
Modcheck	0.46	0.51	0.39	3.02	5	-0.88	-1.87	-0.13	1	5	2.5	8	11
RAPDF	-0.38	0.41	0.34	2.26	4	-1.1	-1.8	-0.07	1	4	-2.63	9	14
DFIRE	-0.32	0.43	0.34	2.27	3	0.8	-1.71	-0.11	1	4	-1.58	**12**	14
ProQ	0.2	0.18	0.17	1.73	2	-0.37	-1.09	-0.17	0	0	1.95	0	8
$ProQ_{SSE}$	0.38	0.42	0.25	2.21	1	-0.58	-1.59	-0.13	0	3	2.6	4	10
FRST	-0.42	0.44	0.33	2.71	3	-0.92	-1.81	-0.11	0	4	-2.56	7	12
QMEAN3	-0.46	0.45	0.4	2.99	1	-0.82	-1.95	-0.12	0	7	**-2.76**	9	14
QMEAN4	-0.48	0.53	0.42	2.87	6	-1.25	-1.87	-0.07	1	7	-2.29	8	13
QMEAN5	**-0.51**	**0.56**	**0.44**	**3.06**	**6**	**-1.22**	**-2**	**-0.07**	1	7	-2.43	9	13
torsion single	-0.27	0.29	0.2	1.73	0	-0.52	-1.65	-0.14	0	2	-1.74	4	7
torsion3-residue	-0.35	0.32	0.26	2.12	4	-0.91	-1.6	-0.1	0	2	-2.65	8	14
pairwise$C\beta$	-0.4	0.38	0.39	2.88	2	-0.88	-1.77	-0.12	0	6	-2.45	7	13
pairwise$C\beta$/SSE	-0.41	0.36	0.43	2.84	5	-1.03	-1.79	-0.09	0	7	-2.67	7	**15**
solvation	-0.36	0.38	0.39	2.71	4	-0.86	-1.87	-0.13	**2**	7	-1.69	4	9
SSEPSIPRED	-0.37	0.48	0.27	2.05	2	-0.62	-1.38	-0.15	1	2	-1.16	1	5
ACCpro	-0.44	0.51	0.39	2.93	4	-0.84	-1.83	-0.1	1	**8**	-2.21	7	12

[a] Pearson's correlation coefficient r^2 and Spearmans's rank correlation coefficient *rho*

[b] *F.E.* stands for fraction enrichment and $E_{15\%}$ is the enrichment among the top 15% best predicted models as compared to a random selection.

[c] r_{10} are the number of targets for with the top-scoring models is among the top10 best models (based on GDT_TS). $logP_{B1}$ and $logP_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the ten best-scoring models, respectively.

[d] GDT_TS loss is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set.r1 and r10 are the number of targets in which the best model based on GDT_TS, excluding the native structure was found on the first rank or among the top 10 predictions.

[e] Z_{nat} is the Z-score of the native structure as compared to the ensemble of models. r1 and r10 are the number of targets in which the native structure was found on the first rank or among the top 10 predictions.

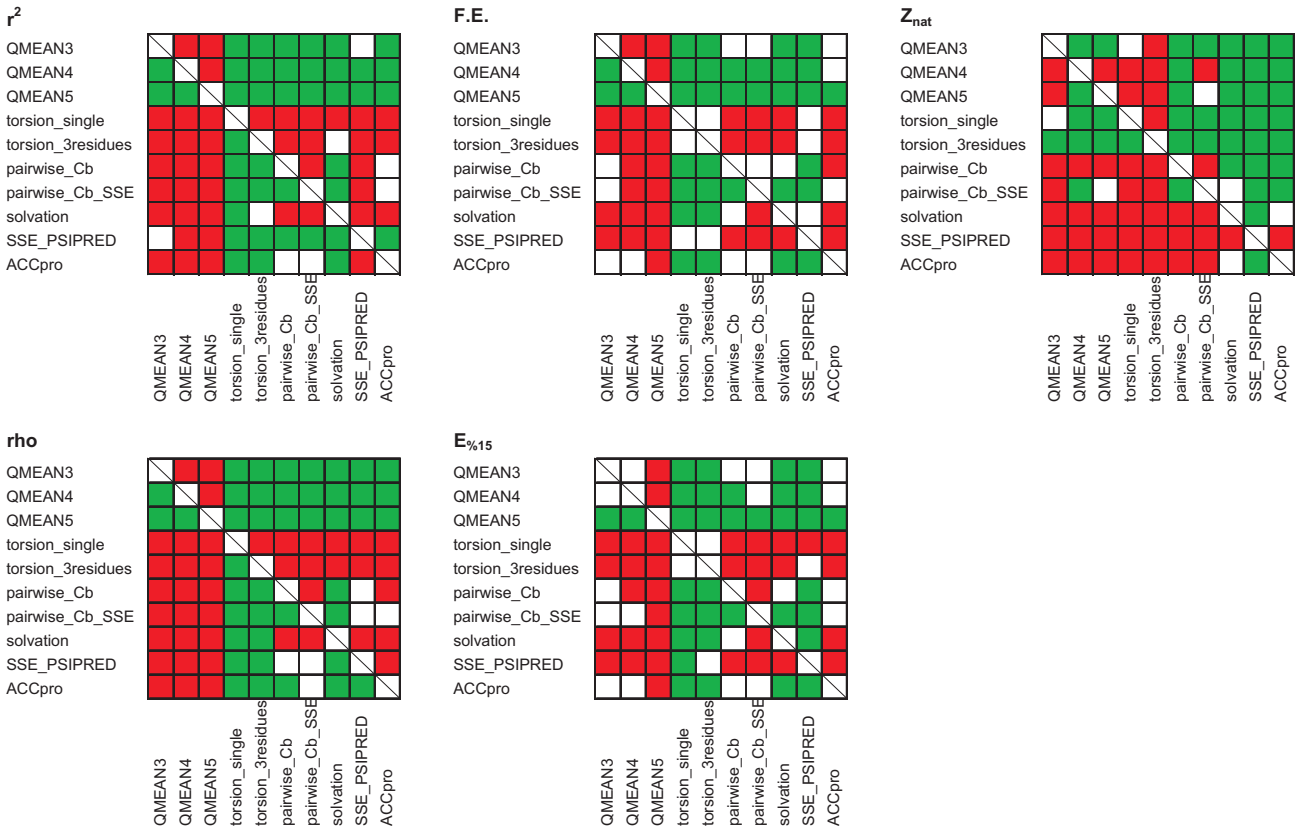
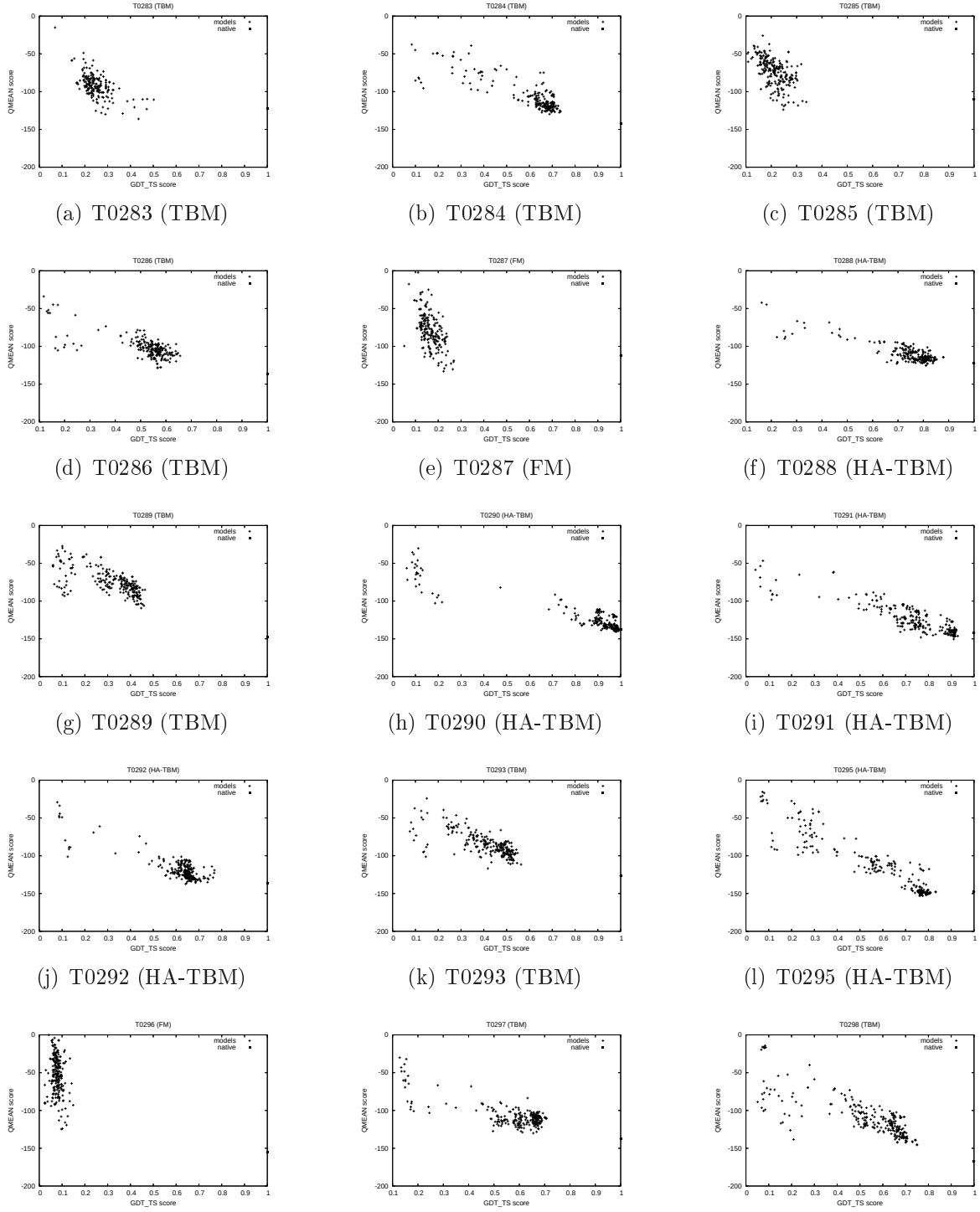


Figure 5.2: Statistical analysis of the performance differences between the different QMEAN terms at the confidence level of 95%.



(a) T0283 (TBM) (b) T0284 (TBM) (c) T0285 (TBM)

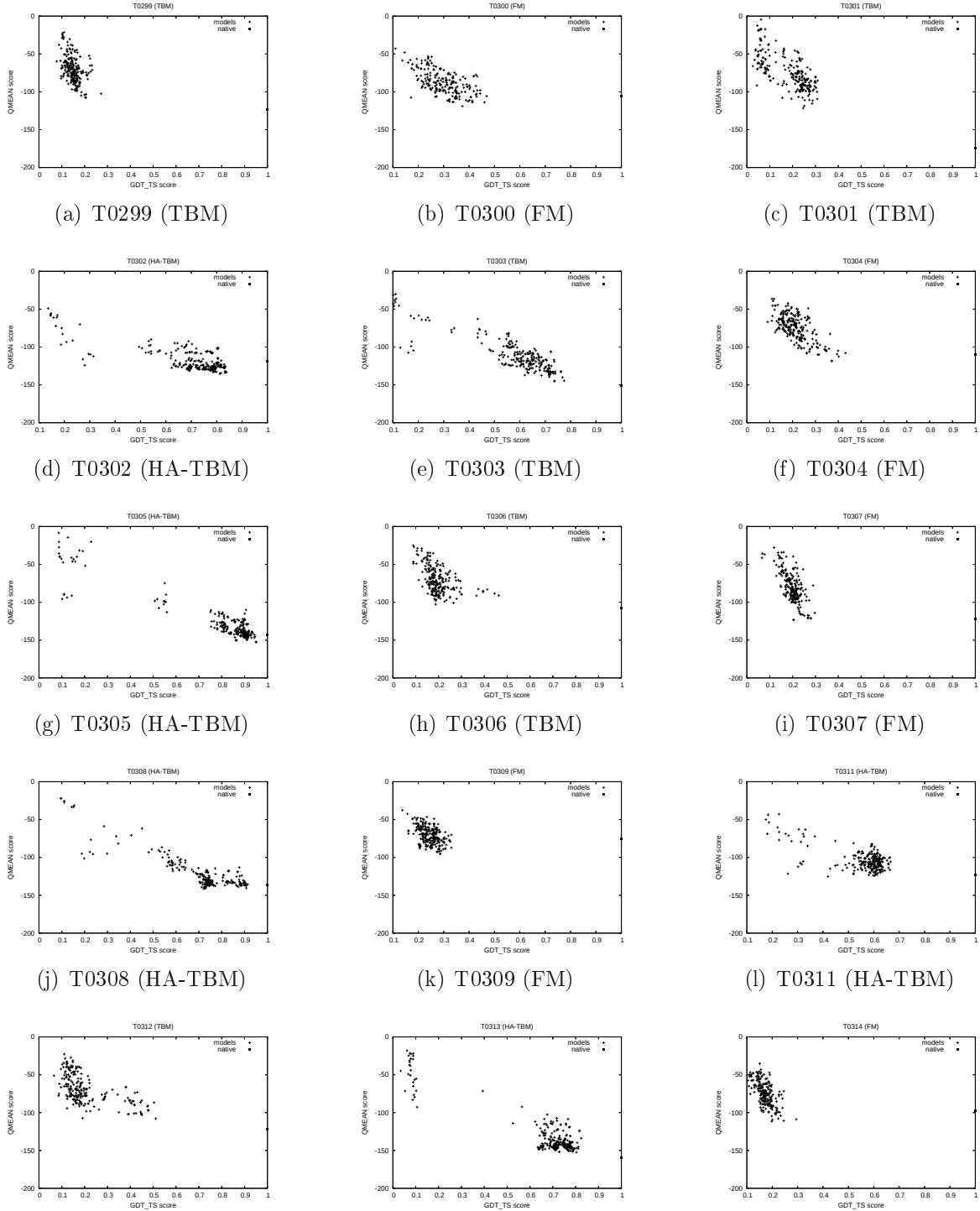
(d) T0286 (TBM) (e) T0287 (FM) (f) T0288 (HA-TBM)

(g) T0289 (TBM) (h) T0290 (HA-TBM) (i) T0291 (HA-TBM)

(j) T0292 (HA-TBM) (k) T0293 (TBM) (l) T0295 (HA-TBM)

(m) T0296 (FM) (n) T0297 (TBM) (o) T0298 (TBM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (1/7).



(a) T0299 (TBM) (b) T0300 (FM) (c) T0301 (TBM)

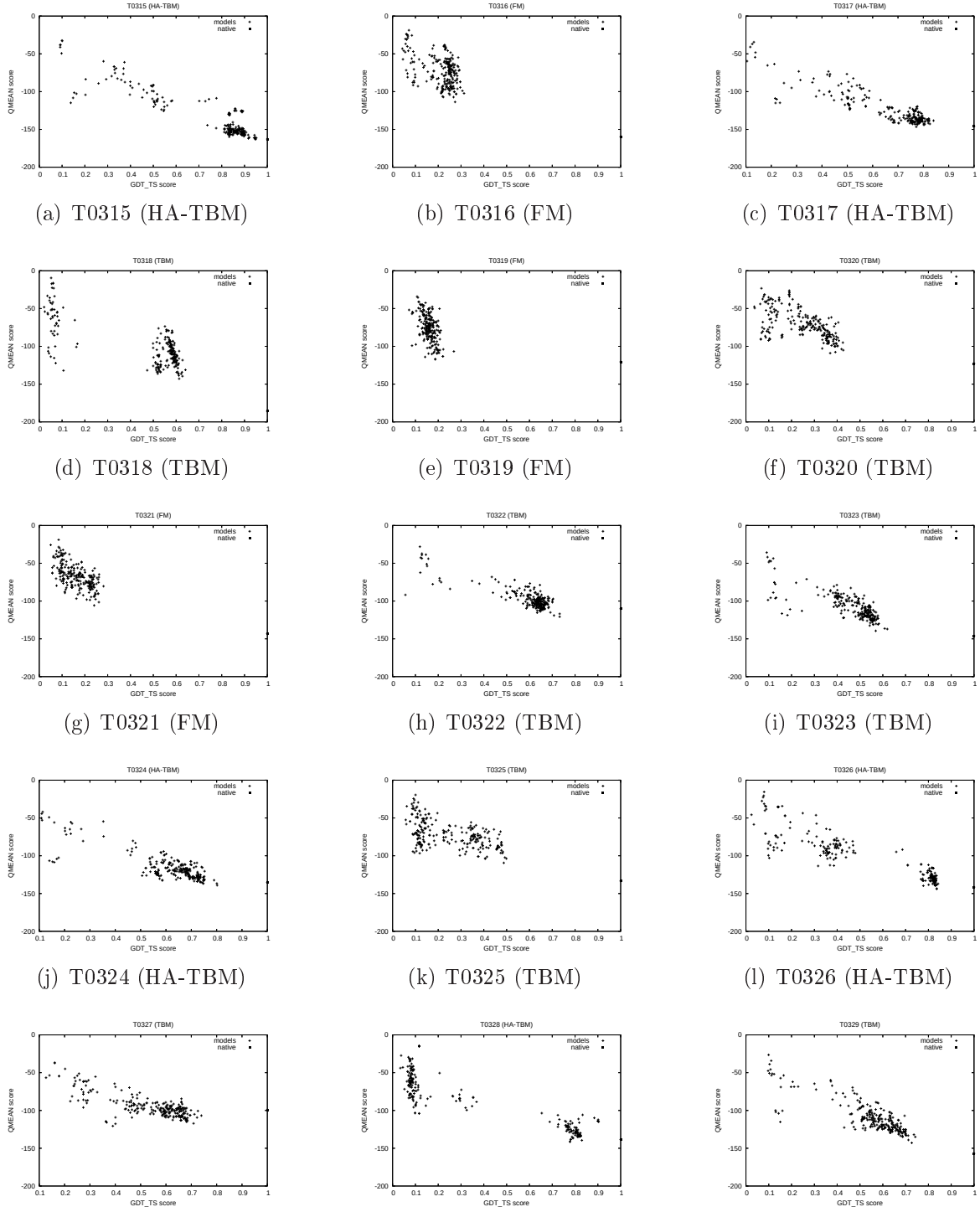
(d) T0302 (HA-TBM) (e) T0303 (TBM) (f) T0304 (FM)

(g) T0305 (HA-TBM) (h) T0306 (TBM) (i) T0307 (FM)

(j) T0308 (HA-TBM) (k) T0309 (FM) (l) T0311 (HA-TBM)

(m) T0312 (TBM) (n) T0313 (HA-TBM) (o) T0314 (FM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (2/7).



(a) T0315 (HA-TBM) (b) T0316 (FM) (c) T0317 (HA-TBM)

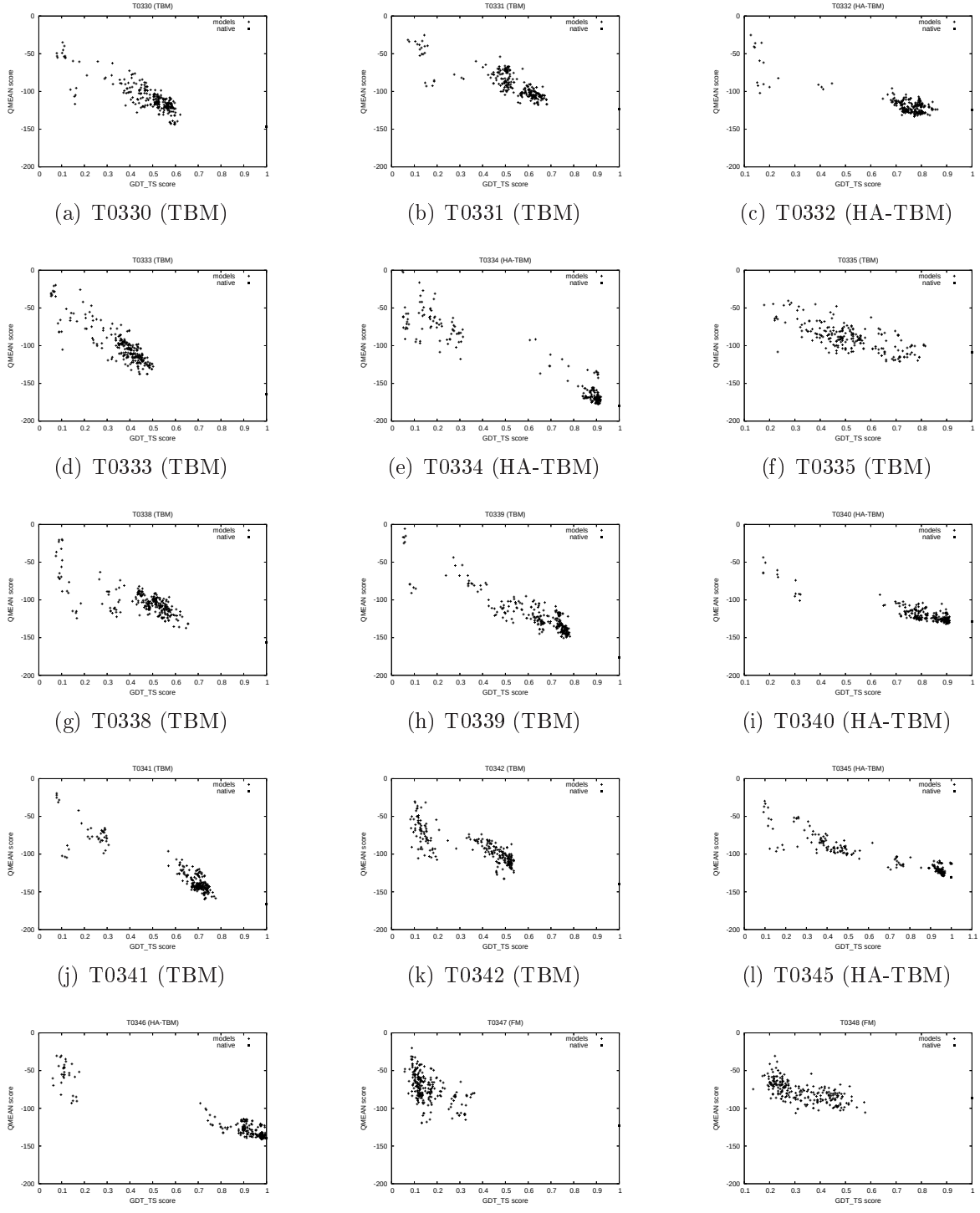
(d) T0318 (TBM) (e) T0319 (FM) (f) T0320 (TBM)

(g) T0321 (FM) (h) T0322 (TBM) (i) T0323 (TBM)

(j) T0324 (HA-TBM) (k) T0325 (TBM) (l) T0326 (HA-TBM)

(m) T0327 (TBM) (n) T0328 (HA-TBM) (o) T0329 (TBM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (3/7).



(a) T0330 (TBM) (b) T0331 (TBM) (c) T0332 (HA-TBM)

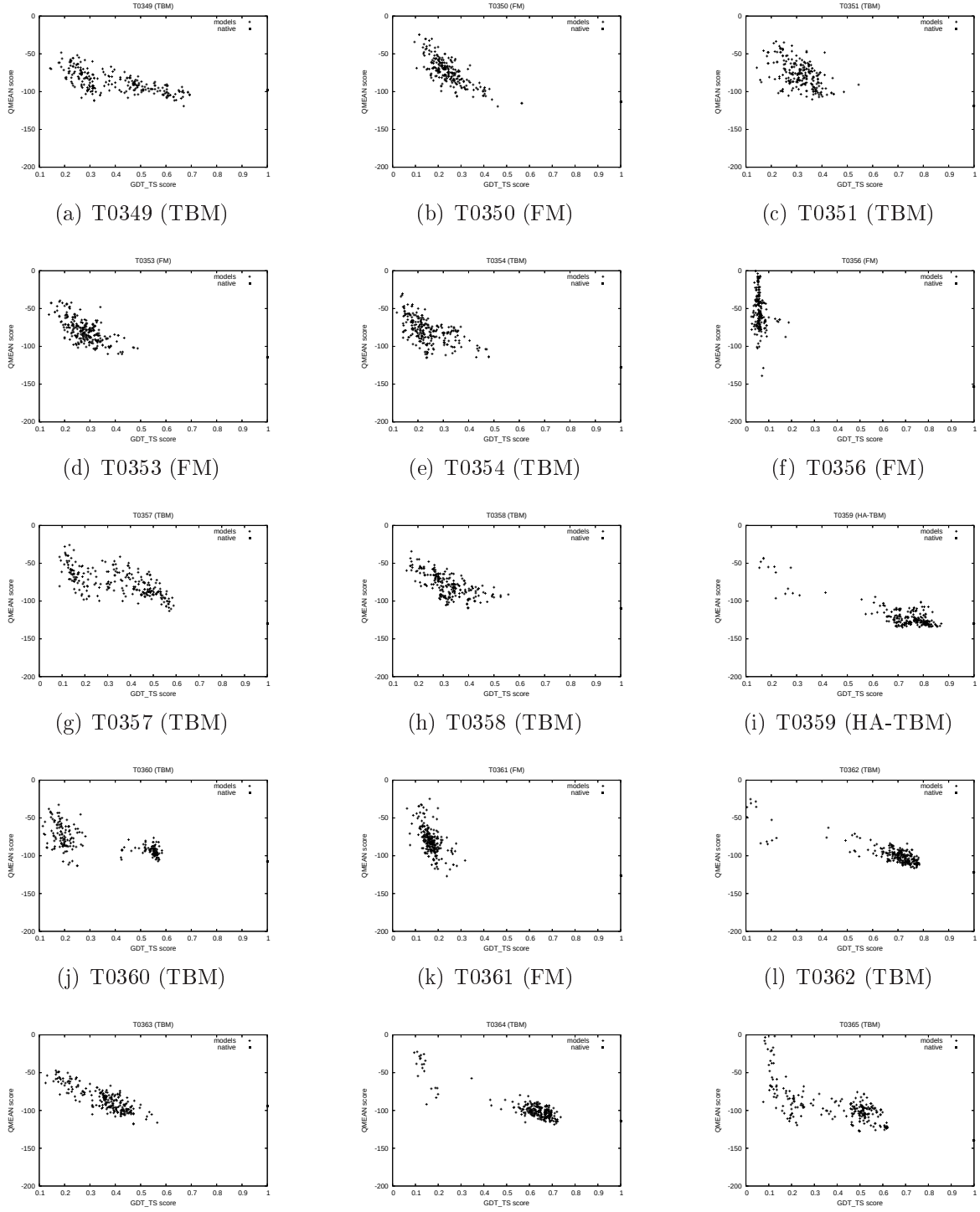
(d) T0333 (TBM) (e) T0334 (HA-TBM) (f) T0335 (TBM)

(g) T0338 (TBM) (h) T0339 (TBM) (i) T0340 (HA-TBM)

(j) T0341 (TBM) (k) T0342 (TBM) (l) T0345 (HA-TBM)

(m) T0346 (HA-TBM) (n) T0347 (FM) (o) T0348 (FM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (4/7).



(a) T0349 (TBM) (b) T0350 (FM) (c) T0351 (TBM)

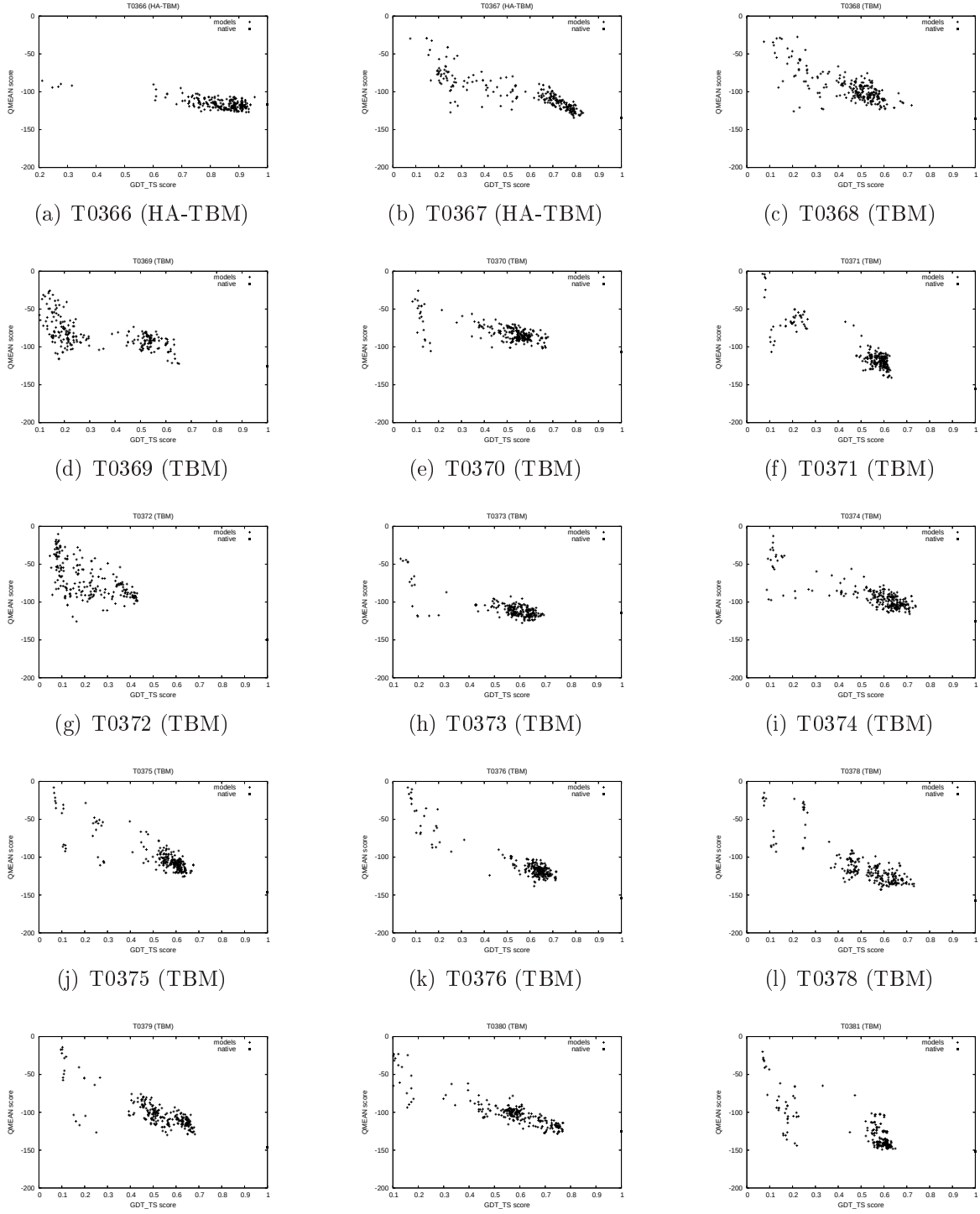
(d) T0353 (FM) (e) T0354 (TBM) (f) T0356 (FM)

(g) T0357 (TBM) (h) T0358 (TBM) (i) T0359 (HA-TBM)

(j) T0360 (TBM) (k) T0361 (FM) (l) T0362 (TBM)

(m) T0363 (TBM) (n) T0364 (TBM) (o) T0365 (TBM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (5/7).



(a) T0366 (HA-TBM) (b) T0367 (HA-TBM) (c) T0368 (TBM)

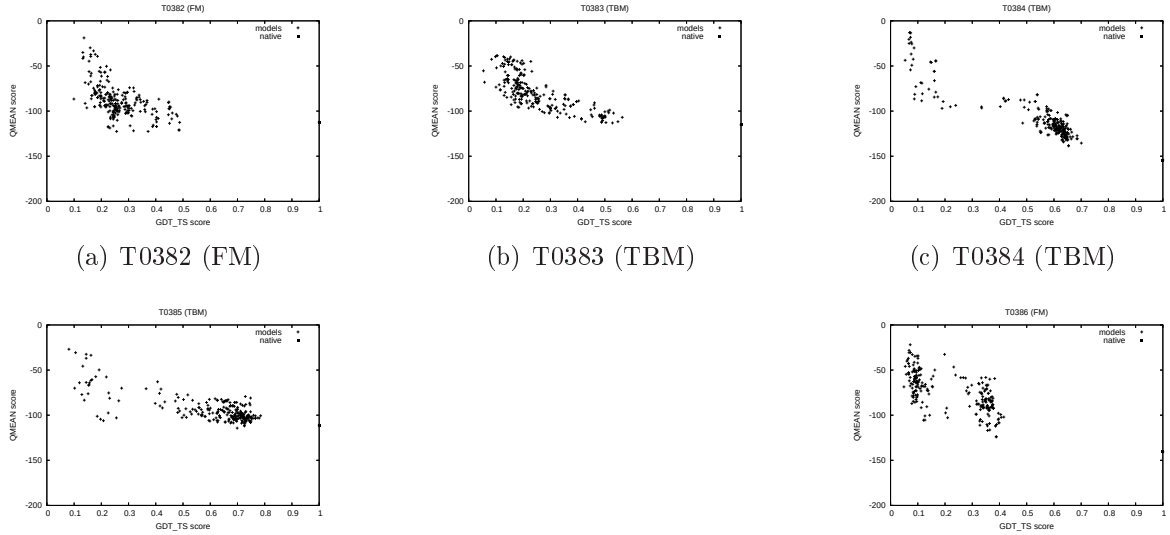
(d) T0369 (TBM) (e) T0370 (TBM) (f) T0371 (TBM)

(g) T0372 (TBM) (h) T0373 (TBM) (i) T0374 (TBM)

(j) T0375 (TBM) (k) T0376 (TBM) (l) T0378 (TBM)

(m) T0379 (TBM) (n) T0380 (TBM) (o) T0381 (TBM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (6/7).



(a) T0382 (FM) (b) T0383 (TBM) (c) T0384 (TBM)

(d) T0385 (TBM) (e) T0386 (FM)

Figure 5.2: Correlation between GDT_TS and QMEAN score for all server models of the 95 targets of CASP7 (7/7).

Table 5.6: Results for loops of length 5 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop[a]				
PDB ID	residues	best loop[b]	random 20000[c]	random 3000[d]	rank Top10[e]	no homolgues	all homologues	<90%	<50%	<30%
153l	131-135	0.27	4.54	2.87	7	0.91	0.28	0.91	0.91	0.91
1a2y_A	14-18	0.26	1.69	1.16	74	0.91	0.29	0.29	0.91	0.91
1a8e	197-201	0.3	1.97	3.59	10	0.48	0.21	0.21	0.48	0.48
1frd	83-87	0.38	3.3	2.55	5	0.5	0.09	0.18	0.2	0.5
1gpr	54-58	0.25	3.78	2.5	3	0.25	0.05	0.52	0.52	0.25
1hbg	19-23	1.16	4.15	5.44	14	1.99	0.09	1.99	1.99	1.99
1hbq	158-162	0.25	5.55	1.54	1	0.25	0.25	0.25	0.25	0.25
1kuh	37-41	0.64	3.9	2.63	2	0.78	0.16	0.78	0.78	0.78
1lit	131-135	0.68	2.92	4.55	1	0.81	0.81	0.81	0.81	0.81
1lit	51-55	0.37	2.99	2.6	6	0.4	0.1	0.3	0.3	0.4
1lkk_A	186-190	1.22	1.47	4.1	59	4.04	4.04	4.04	4.04	4.04
1mla	102-106	0.24	4.71	4.86	3	0.24	0.06	0.26	0.26	0.24
1mla	275-279	1.08	2.84	6.64	26	1.68	0.05	0.29	0.29	1.68
1nar	56-60	0.49	5.62	3.32	2	0.49	0.06	0.49	0.49	0.49
1nfp	95-99	0.42	1.93	1.18	6	1.37	0.08	0.38	0.38	1.37
1noa	88-92	1	2.28	3.51	50	1.91	1.91	1.91	1.91	1.91
1prn	187-191	0.46	2.86	3.23	8	5.01	0.33	5.01	5.01	5.01
1rie	149-153	1.49	5.49	5.52	11	3.8	0.06	3.8	3.8	3.8
1sbp	181-185	0.38	2.39	2.65	2	0.57	0.09	0.57	0.57	0.57
1tca	157-161	0.39	2.32	2.93	5	0.92	0.05	0.92	0.92	0.92
1tml	147-151	0.52	4.4	3.6	2	0.91	0.91	0.91	0.91	0.91
1vcc	63-67	0.28	0.61	1.8	290	1.96	1.96	1.96	1.96	1.96
1xyz_A	559-563	0.71	3.28	2.16	14	3.05	0.05	0.25	0.25	3.05
2cba	168-172	0.53	3.73	4.02	8	0.53	0.39	0.39	0.39	0.53
2cmd	188-192	0.31	5.15	3.71	2	0.31	0.08	0.31	0.31	0.31
2hbg	37-41	0.21	4	2.2	3	0.21	0.05	0.21	0.21	0.21
5p21	104-109	2.42	7.51	3.96	2	3.69	3.69	3.69	3.69	3.69
7rsa	75-79	0.6	1.23	2.06	2	0.6	0.39	0.39	0.39	0.6
8abp	65-70	1.12	2.63	3.04	102	3.17	3.17	3.17	3.17	3.17
average	-	0.64	3.42	3.24	-	1.44	0.68	1.21	1.24	1.44
median	-	0.46	3.28	3.04	-	0.91	0.16	0.52	0.57	0.91

[a] RMSD of the top ranking loop after removing homologues below a given cutoff.

[b] Best nonhomologues loop present among the 3000 candidate fragments.

[c] Random selection of a fragment from the maximum 20,000 loops present after application of the torsion energy filter.

[d] Random selection of a fragment from the maximum 3,000 loops present after application of the backbone energy filter.

[e] Rank of the first Top10 fragment according to RMSD.

Table 5.7: Results for loops of length 7 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
1a62	89-95	0.05	4.37	3.09	1	0.05	0.05	0.05	0.05	0.05
1bkf	64-70	0.37	1.2	0.69	5	0.4	0.06	0.29	0.4	0.4
1ads	186-192	1.33	6.7	5.32	17	4.91	0.29	0.29	0.35	4.91
1brt	226-232	0.34	4.44	4.82	4	0.49	0.11	0.37	0.37	0.49
1cvl	111-117	0.26	4.21	5.36	3	0.26	0.16	0.28	0.26	0.26
1dad	116-122	1.17	5	4.09	2	1.17	0.87	1.17	1.17	1.17
1dim	198-204	1.17	5.94	5.08	2	1.21	0.2	1.21	1.21	1.21
1edg	309-315	1.35	2.47	3.19	29	1.76	0.06	1.76	1.76	1.76
1gca	196-202	0.56	6.47	4.73	15	0.81	0.06	0.81	0.81	0.81
1hbg	46-52	1.31	7.96	4.64	8	3.25	0.1	3.25	3.25	3.25
1hfc	152-158	1.78	2.29	5.21	12	1.78	0.05	0.59	1.78	1.78
1iab	142-148	0.86	2.58	4.08	3	5.59	0.11	5.59	5.59	5.59
1lif	64-70	0.92	5.36	4.89	148	6.26	0.16	0.45	0.48	6.26
1mbd	17-23	0.47	5.03	2.67	1	0.79	0.79	0.79	0.79	0.79
1mla	80-86	1.36	6.09	3.86	2	1.99	1.99	1.99	1.99	1.99
1nif	65-71	1.35	5.91	5.54	6	1.35	0.31	0.31	0.42	1.35
1php	135-141	0.55	2.21	3.02	6	1.2	0.16	0.33	0.42	1.2
1rhs	21-27	1.52	3.21	3.91	88	4.04	0.07	4.04	4.04	4.04
1sgp_E	128-134	0.61	4.98	4.61	3	0.71	0.06	0.51	0.71	0.71
1tca	132-138	0.52	2.04	3.01	2	0.66	0.17	0.66	0.66	0.66
1tml	20-26	0.65	3.89	4.65	2	1.07	0.32	1.07	1.07	1.07
1xyz_A	689-695	2.02	2.41	6.87	177	5.28	0.89	0.89	0.89	5.28
2mnr	270-276	1.18	3.91	4.05	12	2.01	0.14	1.15	1.15	0.9
2pth	95-101	0.71	4.88	6.86	3	6.09	0.1	6.09	6.09	6.09
3tgl	159-165	1.25	5.92	3.74	1	2.07	2.07	2.07	2.07	2.07
5p21	83-89	0.71	4.44	5.89	59	1.63	0.22	0.22	0.17	0.37
average	-	0.94	4.38	4.38	23.5	2.19	0.37	1.39	1.46	2.09
median	-	0.89	4.44	4.63	4.5	1.49	0.16	0.8	0.85	1.21

Table 5.8: Results for loops of length 9 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
1arb	168-176	4.53	6	7.52	20	8.83	0.07	8.83	8.83	8.83
1arp	127-135	1.14	9.47	6.04	5	1.93	0.37	1.93	1.93	1.93
1aru	36-44	2.18	7.26	4.3	142	6.8	6.8	6.8	6.8	6.8
1cse_E	95-103	2.42	6.27	4.69	351	8.83	0.52	0.52	0.42	8.83
1csh	252-260	0.83	8.08	6.67	9	0.83	0.06	0.7	0.7	0.56
1ede	257-265	1.48	4.03	5.84	17	4.37	0.25	4.37	4.37	4.37
1fus	91-99	1.82	5.7	5.12	167	3.99	3.99	3.99	3.99	3.99
1lkk_A	142-150	1.7	6.22	8.68	8	3.67	0.1	1.64	3.67	3.67
1mla	194-202	2.11	5.06	6.16	100	3.32	0.2	3.32	3.32	3.32
1nls	131-139	0.76	4.67	4.11	405	5.88	0.06	5.88	5.88	5.88
1onc	70-78	0.96	6.66	5.78	8	2.94	2.94	2.94	2.94	2.94
1pda	108-116	1.41	8.8	8	7	6.41	0.21	6.41	6.41	6.41
1pgs	117-125	1.73	2.06	6.24	2	1.8	0.1	1.8	1.8	1.8
1php	91-99	1.55	6.47	7.67	638	6.08	0.15	0.71	6.08	6.08
1sgp_E	109-117	1.76	4.67	6.35	43	3.64	0.11	3.64	3.64	3.64
1xnb	116-124	1.53	7.35	5.16	1	1.88	1.88	1.88	1.88	1.88
1xnb	133-141	1.95	5.19	8.65	21	4.19	0.35	4.19	4.19	4.19
1xyz_A	795-803	1.64	9.83	5.69	1143	5.32	0.24	0.89	0.89	5.32
2ayh	169-177	1.24	2.02	2.46	12	3.08	0.1	0.34	3.08	3.08
2cpl	24-32	0.82	4.36	4.73	17	0.82	0.4	0.4	0.33	0.82
3pte	107-115	1.84	2.55	2.73	2	2.8	0.2	2.8	2.8	2.8
average	-	1.69	5.84	5.84	-	4.16	0.91	3.05	3.52	4.15
median	-	1.64	6	5.84	-	3.67	0.21	2.8	3.32	3.67

Table 5.9: Results for loops of length 10 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
135l	18-27	1.9	5.41	7.03	252	4	0.2	0.4	0.39	4
1ads	170-179	1.66	4.67	7.02	414	3.6	0.31	0.31	0.44	3.6
1ads	171-180	1.68	4.45	7.6	44	2.74	0.42	0.42	0.47	2.74
1amp	181-190	2.97	3.86	3.91	8	4.05	0.23	4.05	4.05	4.05
1arb	41-50	1.88	4.78	4.28	75	5.53	0.06	5.53	5.53	5.53
1arp	37-46	3.32	7.72	4.56	586	8.45	8.45	8.45	8.45	8.45
1aru	128-137	1.6	8.88	2.25	9	2.88	0.36	0.66	0.66	2.88
1btl	170-179	2.17	4.7	4.44	448	3.38	0.76	0.76	0.76	3.38
1dim	87-96	1.83	3.04	13.26	601	7.85	0.28	7.85	7.85	7.85
1fkf	63-72	0.54	6.57	6.57	7	0.54	0.35	0.43	0.47	0.54
1gpr	133-142	1.36	6.68	4.25	3	3.04	0.15	3.04	3.04	3.04
1gvp	49-58	1.2	8.66	8.16	9	3.68	0.06	3.68	3.68	3.68
1ixh	84-93	1.77	4.85	4.41	530	4.49	0.13	4.49	4.49	4.49
1knt	35-44	1.67	5.86	6.06	7	1.75	0.24	1.62	1.62	1.75
1mrj	173-182	1.94	4.98	5.33	373	6.34	0.06	6.34	6.34	6.34
1plc	42-51	1.58	6.82	7.55	58	6.46	0.57	0.57	1.41	6.46
1ppn	190-199	2.22	7.28	9.16	25	4.9	1.56	1.56	1.56	4.9
1scs	65-74	0.71	5.97	3.33	79	3.58	0.53	3.58	3.58	3.58
1tca	23-32	2.56	9.93	7.28	8	11.31	0.05	11.31	11.31	11.31
1whi	47-56	1.97	5.62	8.26	40	6.2	0.06	1.09	6.2	6.2
2cmd	57-66	1.44	8.23	9.21	3	2.99	0.11	2.99	2.99	2.99
2mnr	91-100	2.2	9.35	7.05	18	5.09	5.09	5.09	5.09	5.09
2sil	197-206	1.05	6.27	6.19	2	1.05	0.22	1.05	1.05	1.05
3hsc	28-37	1.98	7.8	5.98	8	4.05	0.27	0.64	4.05	4.05
7rsa	110-119	1.13	1.88	2.45	3	1.13	0.41	0.41	1.13	1.13
7rsa	33-42	2.02	7.19	3.22	197	7.68	0.37	0.91	7.68	7.68
7rsa	87-96	1.34	10.79	9.56	1	2.39	2.39	2.39	2.39	2.39
average	-	1.77	6.38	6.24	141.04	4.41	0.88	2.95	3.58	4.41
median	-	1.77	6.27	6.19	25	4	0.28	1.62	3.04	4

Table 5.10: Results for loops of length 11 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
153l	154-164	2.15	7.79	4.83	154	8.46	0.14	8.46	8.46	8.46
1a2p_A	76-86	2.42	5.7	8.09	164	5.48	5.48	5.48	5.48	5.48
1a2y_A	91-101	100	4.92	5.02	3000	2.23	0.26	0.96	1.12	2.23
1akz	211-221	2.73	7.15	6.22	317	4.31	0.19	0.24	0.84	4.31
1awq_A	1101-1111	2.63	6.39	5.14	26	9.51	0.87	0.87	0.58	9.51
1cvl	257-267	6.15	10.27	12.44	972	11.71	0.07	11.71	11.71	11.71
1dad	42-52	1.75	9.29	10.69	18	3.54	0.66	3.54	3.54	3.54
1fus	28-38	3	6.36	9.7	254	11.26	2.06	2.06	11.26	11.26
1ixh	120-130	2.25	3.19	3.41	12	3.4	0.06	3.4	3.4	3.4
1mla	9-19	1.11	3.67	4.36	3	1.11	0.21	0.98	0.98	1.11
1rcf	122-132	2.33	9.58	4.14	73	4.49	0.42	0.81	0.81	4.49
2pth	8-18	2.34	4.05	3.5	92	3.77	0.21	0.68	0.68	3.77
3pte	91-101	2.2	3.8	4.54	4	5.1	0.12	5.1	5.1	5.1
average	-	10.08	6.32	6.31	-	5.72	0.83	3.41	4.15	5.72
median	-	2.34	6.36	5.02	-	4.49	0.21	2.06	3.4	4.49

Table 5.11: Results for loops of length 12 residues from the test set of Rossi *et al.* [174].

						Global RMSD of the top ranking loop				
PDB ID	residues	best loop	random 20000	random 3000	rank Top10	no homolgues	all homologues	<90%	<50%	<30%
153l	98-109	3.53	7.72	8.89	363	8.95	0.17	8.95	8.95	8.95
1akz	181-192	2.07	5.25	6.32	154	5.11	0.71	0.71	0.91	5.11
1arb	74-85	2.37	7.52	3.92	357	5.82	0.06	5.82	5.82	5.82
1bkf	9-20	2.6	6.73	4.95	191	5.04	0.05	0.68	5.04	5.04
1cex	40-51	2.47	8.13	11.84	196	11.75	0.11	11.75	11.75	11.75
1dim	213-224	1.83	8.15	4.89	11	4.38	0.24	4.38	4.38	4.38
1ixh	161-171	4.31	14.32	9.18	128	11.97	0.08	11.97	11.97	11.97
1luc_A	158-169	2.86	5.38	5.39	2	2.86	0.07	2.86	2.86	2.86
2ayh	21-32	2.51	12.19	11.53	339	4.18	0.13	4.18	4.18	4.18
average	-	2.73	8.38	7.43	-	6.67	0.18	5.7	6.21	6.67
median	-	2.51	7.72	6.32	-	5.11	0.11	4.38	5.04	5.11

References

[1] Al-Lazikani, B., Sheinerman, F. B., and Honig, B. (2001). Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc Natl Acad Sci U S A*, 98(26):14796–14801.

[2] Alberts, B., Johnson, A., Lewis, J., Raff, M., R.s, K., and Walter, P. (2002). *Molecular Biology of the Cell - Fourth Edition*. Garland Science, New York.

[3] Albiero, A. and Tosatto, S. C. E. (2006 Mar). Fine-grained statistical torsion angle potentials are effective in discriminating native protein structures. *Curr Drug Discov Technol*, 3(1):75–81.

[4] Albrecht, M., Tosatto, S. C. E., Lengauer, T., and Valle, G. (2003). Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng*, 16(7):459–462.

[5] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[6] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

[7] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96):223–230.

[8] Aszodi, A. and Taylor, W. R. (1996). Homology modelling by distance geometry. *Fold Des*, 1(5):325–334.

[9] Bahar, I. and Jernigan, R. L. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol*, 266(1):195–214.

[10] Bairoch, A. and Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*, 19 Suppl:2247–2249.

[11] Baker, D. and Sali, A. (2001 Oct). Protein structure prediction and structural genomics. *Science*, 294(5540):93–6.

[12] Baldwin, R. L. (2007). Energetics of protein folding. *J Mol Biol*, 371(2):283–301.

[13] Barker, W. C., Garavelli, J. S., Huang, H., McGarvey, P. B., Orcutt, B. C., Srinivasarao, G. Y., Xiao, C., Yeh, L. S., Ledley, R. S., Janda, J. F., Pfeiffer, F., Mewes, H. W., Tsugita,

A., and Wu, C. (2000). The protein information resource (PIR). *Nucleic Acids Res*, 28(1):41–44.

[14] Barlow, D. J. and Thornton, J. M. (1988). Helix geometry in proteins. *J Mol Biol*, 201(3):601–619.

[15] Bassolino-Klimas, D. and Bruccoleri, R. E. (1992). Application of a directed conformational search for generating 3-D coordinates for protein structures from alpha-carbon coordinates. *Proteins*, 14(4):465–474.

[16] Benkert, P., Tosatto, S. C. E., and Schomburg, D. (2007). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, in press.

[17] Benson, D., Lipman, D. J., and Ostell, J. (1993). GenBank. *Nucleic Acids Res*, 21(13):2963–2965.

[18] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000 Jan). The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42.

[19] Betancourt, M. R. and Skolnick, J. (2004). Local propensities and statistical potentials of backbone dihedral angles in proteins. *J Mol Biol*, 342(2):635–649.

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

[21] Bonneau, R., Strauss, C. E. M., Rohl, C. A., Chivian, D., Bradley, P., Malmström, L., R.son, T., and Baker, D. (2002). De novo prediction of three-dimensional structures for major protein families. *J Mol Biol*, 322(1):65–78.

[22] Bork, P. and Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol*, 266:162–184.

[23] Bower, M. J., Cohen, F. E., and Dunbrack, R. L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol*, 267(5):1268–1282.

[24] Bowie, J. U., Lüthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170.

[25] Brooks, B., RE, B., Olafson, B., States, D., Swaminathan, S., and Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.*, 4:187–217.

[26] Bruccoleri, R. E. and Karplus, M. (1990). Conformational sampling using high-temperature molecular dynamics. *Biopolymers*, 29(14):1847–1862.

[27] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21(3):167–195.

[28] Burke, D. F. and Deane, C. M. (2001). Improved protein loop prediction from sequence alone. *Protein Eng*, 14(7):473–478.

[29] Burke, D. F., Deane, C. M., and Blundell, T. L. (2000). Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, 16(6):513–519.

[30] Canutescu, A. A. and Dunbrack, R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci*, 12(5):963–972.

[31] Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014.

[32] Carlacci, L. and Englander, S. W. (1993). The loop problem in proteins: a Monte Carlo simulated annealing approach. *Biopolymers*, 33(8):1271–1286.

[33] Chance, M. R., Fiser, A., Sali, A., Pieper, U., Eswar, N., Xu, G., Fajardo, J. E., Radhakannan, T., and Marinkovic, N. (2004). High-throughput computational and experimental techniques in structural genomics. *Genome Res*, 14(10B):2145–2154.

[34] Chandonia, J.-M. and Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science*, 311(5759):347–351.

[35] Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P. (2005 Jul). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res*, 33(Web Server issue):W72–6.

[36] Chinea, G., Padron, G., Hooft, R. W., Sander, C., and Vriend, G. (1995). The use of position-specific rotamers in model building by homology. *Proteins*, 23(3):415–421.

[37] Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, 357(6379):543–544.

[38] Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A. (2003). Evolution of the protein repertoire. *Science*, 300(5626):1701–1703.

[39] Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826.

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

[41] Chou, P. Y. and Fasman, G. D. (1974b). Prediction of protein conformation. *Biochemistry*, 13(2):222–245.

[42] Chung, S. Y. and Subbiah, S. (1995). The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. *Protein Sci*, 4(11):2300–2309.

[43] Claessens, M., Cutsem, E. V., Lasters, I., and Wodak, S. (1989). Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng*, 2(5):335–345.

[44] Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, 20(3):426–427.

[45] Cline, M., Hughey, R., and Karplus, K. (2002). Predicting reliable regions in protein sequence alignments. *Bioinformatics*, 18(2):306–314.

[46] Cohen, B. I., Presnell, S. R., and Cohen, F. E. (1993). Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci*, 2(12):2134–2145.

[47] Collura, V., Higo, J., and Garnier, J. (1993). Modeling of protein loops by simulated annealing. *Protein Sci*, 2(9):1502–1510.

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

[49] Cozzetto, D., Kryshtafovych, A., Ceriani, M., and Tramontano, A. (2007). Assessment of predictions in the model quality assessment category. *Proteins*.

[50] de Bakker, P. I. W., DePristo, M. A., Burke, D. F., and Blundell, T. L. (2003). Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins*, 51(1):21–40.

[51] de Hoon, M. J. L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

[52] Deane, C. M. and Blundell, T. L. (2000). A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins*, 40(1):135–144.

[53] Deane, C. M. and Blundell, T. L. (2001). CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci*, 10(3):599–612.

[54] Deane, C. M. and Lummis, S. C. (2001). The role and predicted propensity of conserved proline residues in the 5-HT3 receptor. *J Biol Chem*, 276(41):37962–37966.

[55] DePristo, M. A., Bakker, P. I. W. D., Shetty, R. P., and Blundell, T. L. (2003a). Discrete restraint-based protein modeling and the Calpha-trace problem. *Protein Sci*, 12(9):2032–2046.

[56] DePristo, M. A., de Bakker, P. I. W., Lovell, S. C., and Blundell, T. L. (2003b). Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins*, 51(1):41–55.

[57] Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155.

[58] Dill, K. A. (1997). Additivity principles in biochemistry. *THE JOURNAL OF BIOLOGICAL CHEMISTRY*, 272:701âĂŞ704.

[59] Dill, K. A. and Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol*, 4(1):10–19.

[60] Domingues, F. S., Lackner, P., Andreeva, A., and Sippl, M. J. (2000). Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol*, 297(4):1003–1013.

[61] Donate, L. E., Rufino, S. D., Canard, L. H., and Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci*, 5(12):2600–2616.

[62] Du, P., Andrec, M., and Levy, R. M. (2003). Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng*, 16(6):407–414.

[63] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755–763.

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

[65] Elofsson, A. (2002). A study on protein sequence alignment quality. *Proteins*, 46(3):330–339.

[66] Eramian, D., Shen, M.-y., Devos, D., Melo, F., Sali, A., and Marti-Renom, M. A. (2006 Jul). A composite score for predicting errors in protein structure models. *Protein Sci*, 15(7):1653–66.

[67] Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, 17(12):1242–1243.

[68] Fasnacht, M., Zhu, J., and Honig, B. (2007). Local quality assessment in homology models using statistical potentials and support vector machines. *Protein Sci*, 16(8):1557–1568.

[69] Fernandez-Fuentes, N. and Fiser, A. (2006). Saturating representation of loop conformational fragments in structure databanks. *BMC Struct Biol*, 6:15.

[70] Fernandez-Fuentes, N., Oliva, B., and Fiser, A. (2006a). A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res*, 34(7):2085–2097.

[71] Fernandez-Fuentes, N., Querol, E., Aviles, F. X., Sternberg, M. J. E., and Oliva, B. (2005). Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. *Proteins*, 60(4):746–757.

[72] Fernandez-Fuentes, N., Zhai, J., and Fiser, A. (2006b). ArchPRED: a template based loop structure prediction server. *Nucleic Acids Res*, 34(Web Server issue):W173–W176.

[73] Ferrada, E. and Melo, F. (2007). Nonbonded terms extrapolated from nonlocal knowledge-based energy functions improve error detection in near-native protein structure models. *Protein Sci*, 16(7):1410–1421.

[74] Fidelis, K., Stern, P. S., Bacon, D., and Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng*, 7(8):953–960.

[75] Finkelstein, A. V., Badretdinov, A. Y., and Gutin, A. M. (1995 Oct). Why do protein architectures have Boltzmann-like statistics? *Proteins*, 23(2):142–50.

[76] 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*, Suppl 3:209–217.

[77] Fiser, A., Do, R. K., and Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci*, 9(9):1753–1773.

[78] Fiser, A. and Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, 19(18):2500–2501.

[79] Fogolari, F., Brigo, A., and Molinari, H. (2003 Jul). Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys J*, 85(1):159–66.

[80] Fogolari, F. and Tosatto, S. C. E. (2005 Apr). Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. *Protein Sci*, 14(4):889–901.

[81] Fogolari, F., Tosatto, S. C. E., and Colombo, G. (2005). A decoy set for the thermostable subdomain from chicken villin headpiece, comparison of different free energy estimators. *BMC Bioinformatics*, 6:301.

[82] Ghosh, A., Rapp, C., and Friesner, R. (1998). Generalized born model based on a surface integral formulation. *J. Phys. Chem.*, 102:10983–10990.

[83] Gilis, D. and Rooman, M. (1996 Apr). Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. *J Mol Biol*, 257(5):1112–26.

[84] Gilis, D. and Rooman, M. (1997 Sep). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J Mol Biol*, 272(2):276–90.

[85] Ginalski, K., Grishin, N. V., Godzik, A., and Rychlewski, L. (2005). Practical lessons from protein structure prediction. *Nucleic Acids Res*, 33(6):1874–1891.

[86] Go, N. N. and Scheraga, H. (1970). Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3:178âĹŠ187.

[87] Godzik, A. (1996). Knowledge-based potentials for protein folding: what can we learn from known protein structures? *Structure*, 4(4):363–366.

[88] Greer, J. (1980). Model for haptoglobin heavy chain based upon structural homology. *Proc Natl Acad Sci U S A*, 77(6):3393–3397.

[89] Henikoff, S. and Henikoff, J. G. (1994). Position-based sequence weights. *J Mol Biol*, 243(4):574–578.

[90] Heuser, P., Wohlfahrt, G., and Schomburg, D. (2004). Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins. *Proteins*, 54(3):583–595.

[91] Hillisch, A., Pineda, L. F., and Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discov Today*, 9(15):659–669.

[92] Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci*, 1(12):1691–1698.

[93] Holm, L. and Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol*, 218(1):183–194.

[94] Holm, L. and Sander, C. (1992 May). Evaluation of protein models by atomic solvation preference. *J Mol Biol*, 225(1):93–105.

[95] Honig, B. (1999). Protein folding: from the levinthal paradox to structure prediction. *J Mol Biol*, 293(2):283–293.

[96] Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996). Errors in protein structures. *Nature*, 381(6580):272.

[97] Hoppe, C. and Schomburg, D. (2005 Oct). Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. *Protein Sci*, 14(10):2682–92.

[98] Jacobson, M. and Sali, A. (2004). *Comparative Protein Structure Modeling and its Applications to Drug Discovery*. ANNUAL REPORTS IN MEDICINAL CHEMISTRY, VOLUME 39.

[99] Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., and Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins*, 55(2):351–367.

[100] Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Sci*, 9(8):1487–1496.

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

[102] Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, 287(4):797–815.

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

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

[105] Jones, T. A. and Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J*, 5(4):819–822.

[106] Jorgensen, W., Maxwell, D., and Tirado-Rives, J. (1996). Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236.

[107] Kabsch, W. and Sander, C. (1983 Dec). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

[108] Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., and Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins*, Suppl 3:121–125.

[109] Keasar, C. and Levitt, M. (2003 May). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J Mol Biol*, 329(1):159–74.

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

[111] Kocher, J. P., Rooman, M. J., and Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol*, 235(5):1598–1613.

[112] Kolodny, R., Guibas, L., Levitt, M., and Koehl, P. (2005). Inverse kinematics in biology: The protein loop closure problem. *The International Journal of Robotics Research*, 24:151–163.

[113] Kopp, J., Bordoli, L., Battey, J., Kiefer, F., and Schwede, T. (2007). Assessment of casp7 predictions for template-based modeling targets. *Proteins*, in press.

[114] Krissinel, E. B., Winn, M. D., Ballard, C. C., Ashton, A. W., Patel, P., Potterton, E. A., McNicholas, S. J., Cowtan, K. D., and Emsley, P. (2004 Dec). The new CCP4 Coordinate Library as a toolkit for the design of coordinate-related applications in protein crystallography. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1):2250–5.

[115] Kryshtafovych, A., Prlic, A., Dmytriv, Z., Daniluk, P., Milostan, M., Eyrich, V., Hubbard, T., and Fidelis, K. (2007). New tools and expanded data analysis capabilities at the protein structure prediction center. *Proteins*.

[116] Kryshtafovych, A., Venclovas, C., Fidelis, K., and Moult, J. (2005). Progress over the first decade of CASP experiments. *Proteins*, 61 Suppl 7:225–236.

[117] Laskowski, R., MacArthur, M., D.S., M., and Thornton, J. (1993). Procheck: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291.

[118] Lazaridis, T. and Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol*, 288(3):477–487.

[119] Lazaridis, T. and Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*, 10(2):139–145.

[120] Lessel, U. and Schomburg, D. (1997). Creation and characterization of a new, non-redundant fragment data bank. *Protein Eng*, 10(6):659–664.

[121] Lessel, U. and Schomburg, D. (1999). Importance of anchor group positioning in protein loop prediction. *Proteins*, 37(1):56–64.

[122] Levinthal, C. (1968). Are there pathways for protein folding? *J. Chem. Phys.*, 65:44–45.

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

[124] Li, A. J. and Nussinov, R. (1998). A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins*, 32(1):111–127.

[125] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

[126] Li, W., Liu, Z., and Lai, L. (1999). Protein loops on structurally similar scaffolds: database and conformational analysis. *Biopolymers*, 49(6):481–495.

[127] Livingstone, C. D. and Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 9(6):745–756.

[128] Lu, H. and Skolnick, J. (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–232.

[129] Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364):83–85.

[130] MacArthur, M. W. and Thornton, J. M. (1991). Influence of proline residues on protein conformation. *J Mol Biol*, 218(2):397–412.

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

[132] Marti-Renom, M. A., Madhusudhan, M. S., and Sali, A. (2004). Alignment of protein sequences by their profiles. *Protein Sci*, 13(4):1071–1087.

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

[134] Martin, A. C. and Thornton, J. M. (1996). Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol*, 263(5):800–815.

[135] Melo, F. and Feytmans, E. (1997). Novel knowledge-based mean force potential at atomic level. *J Mol Biol*, 267(1):207–222.

[136] Melo, F. and Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol*, 277(5):1141–1152.

[137] Melo, F., Sanchez, R., and Sali, A. (2002). Statistical potentials for fold assessment. *Protein Sci*, 11(2):430–448.

[138] Mevissen, H. T. and Vingron, M. (1996). Quantifying the local reliability of a sequence alignment. *Protein Eng*, 9(2):127–132.

[139] Michalsky, E., Goede, A., and Preissner, R. (2003). Loops In Proteins (LIP)–a comprehensive loop database for homology modelling. *Protein Eng*, 16(12):979–985.

[140] Mittelman, D., Sadreyev, R., and Grishin, N. (2003). Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, 19(12):1531–1539.

[141] Miyazawa, S. and Jernigan, R. L. (2000 Jul). Identifying sequence-structure pairs undetected by sequence alignments. *Protein Eng*, 13(7):459–75.

[142] Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*, 7(11):2469–2471.

[143] Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, 15(3):285–289.

[144] Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, Suppl 5:2–7.

[145] Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, Suppl 3:2–6.

[146] Moult, J. and James, M. N. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins*, 1(2):146–163.

[147] Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23(3):ii–iv.

[148] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540.

[149] Naor, D. and Brutlag, D. L. (1994). On near-optimal alignments of biological sequences. *J Comput Biol*, 1(4):349–366.

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

[151] Oliva, B., Bates, P. A., Querol, E., AvilÃ©s, F. X., and Sternberg, M. J. (1997). An automated classification of the structure of protein loops. *J Mol Biol*, 266(4):814–830.

[152] Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–634.

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

[154] Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, Suppl 3:177–85.

[155] Panchenko, A. R. (2003). Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res*, 31(2):683–689.

[156] Panchenko, A. R. and Bryant, S. H. (2002). A comparison of position-specific score matrices based on sequence and structure alignments. *Protein Sci*, 11(2):361–370.

[157] Panchenko, A. R., Marchler-Bauer, A., and Bryant, S. H. (2000). Combination of threading potentials and sequence profiles improves fold recognition. *J Mol Biol*, 296(5):1319–1331.

[158] Park, B. and Levitt, M. (1996 May). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2):367–92.

[159] Parthiban, V., Gromiha, M. M., Hoppe, C., and Schomburg, D. (2007 Jan). Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins*, 66(1):41–52.

[160] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448.

[161] Petrey, D. and Honig, B. (2005). Protein structure prediction: inroads to biology. *Mol Cell*, 20(6):811–819.

[162] Pettitt, C. S., McGuffin, L. J., and Jones, D. T. (2005 Sep). Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics*, 21(17):3509–15.

[163] Ponder, J. W. and Richards, F. M. (1987). Internal packing and protein structural classes. *Cold Spring Harb Symp Quant Biol*, 52:421–428.

[164] Pruitt, K. D. and Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, 29(1):137–140.

[165] Przybylski, D. and Rost, B. (2004). Improving fold recognition without folds. *J Mol Biol*, 341(1):255–269.

[166] Ramachandran, G. N. and Mitra, A. K. (1976). An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. *J Mol Biol*, 107(1):85–92.

[167] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963 Jul). Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:95–9.

[168] Rapp, C. S. and Friesner, R. A. (1999). Prediction of loop geometries using a generalized born model of solvation effects. *Proteins*, 35(2):173–183.

[169] Read, R. and Chavali, G. (2007). Assessment of casp7 predictions in the high accuracy template-based modeling category. *Proteins*, in press.

[170] Reva, B. A., Finkelstein, A. V., Sanner, M. F., and Olson, A. J. (1997). Residue-residue mean-force potentials for protein structure recognition. *Protein Eng*, 10(8):865–876.

[171] Rohl, C. A., Strauss, C. E. M., Chivian, D., and Baker, D. (2004a). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*, 55(3):656–677.

[172] Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004b). Protein structure prediction using Rosetta. *Methods Enzymol*, 383:66–93.

[173] Rooman, M. J. and Wodak, S. J. (1995). Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng*, 8(9):849–858.

[174] Rossi, K. A., Weigelt, C. A., Nayeem, A., and Krystek, S. R. (2007). Loopholes and missing links in protein modeling. *Protein Sci*.

[175] Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94.

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

[177] Rost, B. (2003). Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem Anal*, 44:559–87.

[178] Rost, B., Schneider, R., and Sander, C. (1997). Protein fold recognition by prediction-based threading. *J Mol Biol*, 270(3):471–480.

[179] Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci*, 9(2):232–241.

[180] Sadreyev, R. I., Baker, D., and Grishin, N. V. (2003). Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Sci*, 12(10):2262–2272.

[181] Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815.

[182] Samudrala, R. and Levitt, M. (2000 Jul). Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci*, 9(7):1399–401.

[183] Samudrala, R. and Moult, J. (1998). Determinants of side chain conformational preferences in protein structures. *Protein Eng*, 11(11):991–997.

[184] Samudrala, R. and Moult, J. (1998 Feb). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275(5):895–916.

[185] Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68.

[186] Saqi, M. A., Bates, P. A., and Sternberg, M. J. (1992). Towards an automatic method of predicting protein structure by homology: an evaluation of suboptimal sequence alignments. *Protein Eng*, 5(4):305–311.

[187] Sauder, J. M., Arthur, J. W., and Dunbrack, R. L. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, 40(1):6–22.

[188] Shan, Y., Wang, G., and Zhou, H. X. (2001). Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. *Proteins*, 42(1):23–37.

[189] Shen, M.-Y. and Sali, A. (2006 Nov). Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11):2507–24.

[190] Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H. J., and Levinthal, C. (1987). Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers*, 26(12):2053–2085.

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

[192] Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747.

[193] Shortle, D. (2002 Jan). Composites of local structure propensities: evidence for local encoding of long-range structure. *Protein Sci*, 11(1):18–26.

[194] Shortle, D. (2003). Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci*, 12(6):1298–1302.

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

[196] Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268(1):209–225.

[197] Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–883.

[198] Sippl, M. J. (1993 Aug). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, 7(4):473–501.

[199] Sippl, M. J. (1993 Dec). Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17(4):355–62.

[200] Sippl, M. J. and Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, 13(3):258–271.

[201] Skolnick, J. and Kihara, D. (2001). Defrosting the frozen approximation: PROSPECTOR–a new approach to threading. *Proteins*, 42(3):319–331.

[202] Skolnick, J., Kolinski, A., and Ortiz, A. (2000). Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins*, 38(1):3–16.

[203] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

[204] Soeding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, 33(Web Server issue):W244–W248.

[205] Solis, A. D. and Rackovsky, S. (2006). Improvement of statistical potentials and threading score functions using information maximization. *Proteins*, 62(4):892–908.

[206] Sommer, I., Toppo, S., Sander, O., Lengauer, T., and Tosatto, S. C. E. (2006). Improving the quality of protein structure models by selecting from alignment alternatives. *BMC Bioinformatics*, 7:364.

[207] Soto, C., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. (2007). Loop modeling: Sampling, filtering and scoring. *Proteins: Struct. Func. Bioinform.*, in press.

[208] Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, 1(5):377–384.

[209] Tanaka, S. and Scheraga, H. A. (1976). Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins. *Macromolecules*, 9(1):142–159.

[210] Tang, C. L., Xie, L., Koh, I. Y. Y., Posy, S., Alexov, E., and Honig, B. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J Mol Biol*, 334(5):1043–1062.

[211] Thomas, P. D. and Dill, K. A. (1996 Mar). Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol*, 257(2):457–69.

[212] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680.

[213] Tobi, D. and Elber, R. (2000 Oct). Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins*, 41(1):40–6.

[214] Topham, C. M., McLeod, A., Eisenmenger, F., Overington, J. P., Johnson, M. S., and Blundell, T. L. (1993). Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J Mol Biol*, 229(1):194–220.

[215] Tosatto, S. C. E. (2005). The victor/FRST function for model quality estimation. *J Comput Biol*, 12(10):1316–1327.

[216] Tosatto, S. C. E., Albiero, A., Mantovan, A., Ferrari, C., Bindewald, E., and Toppo, S. (2006). Align: a C++ class library and web server for rapid sequence alignment prototyping. *Curr Drug Discov Technol*, 3(3):167–173.

[217] Tosatto, S. C. E. and Battistutta, R. (2007). TAP score: torsion angle propensity normalization applied to local protein structure evaluation. *BMC Bioinformatics*, 8:155.

[218] Tosatto, S. C. E., Bindewald, E., Hesser, J., and Männer, R. (2002). A divide and conquer approach to fast loop modeling. *Protein Eng*, 15(4):279–286.

[219] Tramontano, A. (2006). *Protein Structure Prediction*. WILEY-VCH Verlag GmbH & Co.

[220] 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*, 13(3):231–245.

[221] Tramontano, A. and Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins*, 53 Suppl 6:352–368.

[222] Tress, M., Ezkurdia, I., Grana, O., Lopez, G., and Valencia, A. (2005). Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins*, 61 Suppl 7:27–45.

[223] Tress, M. L., GraÃ±a, O., and Valencia, A. (2004). SQUARE–determining reliable regions in sequence alignments. *Bioinformatics*, 20(6):974–975.

[224] Tress, M. L., Jones, D., and Valencia, A. (2003). Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol*, 330(4):705–718.

[225] Tsai, J., Bonneau, R., Morozov, A. V., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 53(1):76–87.

[226] van Vlijmen, H. W. and Karplus, M. (1997). PDB-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol*, 267(4):975–1001.

[227] Venclovas, C. and Margelevicius, M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins*, 61 Suppl 7:99–105.

[228] Vingron, M. (1996). Near-optimal sequence alignment. *Curr Opin Struct Biol*, 6(3):346–352.

[229] Vingron, M. and Argos, P. (1990). Determination of reliable regions in protein sequence alignments. *Protein Eng*, 3(7):565–569.

[230] Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001). Completeness in structural genomics. *Nat Struct Biol*, 8(6):559–566.

[231] Voet, D. and Voet, J. (2004). *Biochemistry*. John Wiley & Son, Inc.

[232] von Ohsen, N., Sommer, I., and Zimmer, R. (2003). Profile-profile alignment: a powerful tool for protein structure prediction. *Pac Symp Biocomput*, pages 252–263.

[233] Wallner, B. and Elofsson, A. (2003). Can correct protein models be identified? *Protein Sci*, 12(5):1073–1086.

[234] Wallner, B. and Elofsson, A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci*, 15(4):900–913.

[235] Wang, G. and Dunbrack, R. L. (2004). Scoring profile-to-profile sequence alignments. *Protein Sci*, 13(6):1612–1626.

[236] Wang, G. and Dunbrack, R. L. J. (2003 Aug). PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–91.

[237] Wang, K., Fain, B., Levitt, M., and Samudrala, R. (2004). Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct Biol*, 4:8.

[238] Wohlfahrt, G., Hangoc, V., and Schomburg, D. (2002). Positioning of anchor groups in protein loop prediction: the importance of solvent accessibility and secondary structure elements. *Proteins*, 47(3):370–378.

[239] Wojcik, J., Mornon, J. P., and Chomilier, J. (1999). New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289(5):1469–1490.

[240] Xia, Y., Huang, E. S., Levitt, M., and Samudrala, R. (2000 Jun). Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol*, 300(1):171–85.

[241] Xiang, Z., Soto, C. S., and Honig, B. (2002). Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci U S A*, 99(11):7432–7437.

[242] Xu, Q., Canutescu, A., Obradovic, Z., and Dunbrack, R. L. (2006). ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics*, 22(23):2876–2882.

[243] Yona, G. and Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315(5):1257–1275.

[244] Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*, 31(13):3370–3374.

[245] Zhang, Y., Arakaki, A. K., and Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*, 61 Suppl 7:91–98.

[246] Zhang, Y., Kolinski, A., and Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J*, 85(2):1145–1164.

[247] Zhang, Y. and Skolnick, J. (2004 Dec). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–10.

[248] Zhang, Y. and Skolnick, J. (2004 May). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A*, 101(20):7594–9.

[249] Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*, 11(11):2714–2726.

[250] Zhou, H. and Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, 55(4):1005–1013.

[251] Zhu, K., Pincus, D. L., Zhao, S., and Friesner, R. A. (2006). Long loop prediction using the protein local optimization program. *Proteins*, 65(2):438–452.

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Dietmar Schomburg betreut worden.

Pascal Benkert

Teilpublikationen

Benkert P, Tosatto S C E, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins, *in press*.