

Accurate Modeling of Protein Structures by Homology

INAUGURALDISSERTATION

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

von
Marco Biasini
aus
Altdorf (UR) und Italien

Basel, 2013

genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von
Prof. Dr. Torsten Schwede und Prof. Dr. Andrew Torda

Basel, den 26. März 2013

Prof. Dr. Jörg Schibler
Dekan






Attribution-Noncommercial-No Derivative Works 2.5 Switzerland

You are free:

to share — to copy, distribute and transmit the work

Under the following conditions:

-  **ATTRIBUTION.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
-  **NONCOMMERCIAL.** You may not use this work for commercial purposes.
-  **NO DERIVATIVE WORKS.** You may not alter, transform, or build upon this work.

With the understanding that:

WAIVER. Any of the above conditions can be waived if you get permission from the copyright holder.

PUBLIC DOMAIN. Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.

OTHER RIGHTS. In no way are any of the following rights affected by the license:

- Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
- The author's moral rights;
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

This is a human-readable summary of the Legal Code (the full license) available at: <http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de>

SOURCE: <http://creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en>

The way to create something beautiful is often to make subtle
tweaks to something that already exists, or to combine existing
ideas in a slightly new way.

Paul Graham, *Hackers & Painters*

List of Abbreviations

EM	electron microscopy
C α	central carbon α atom of amino acids
CAMEO	c ontinuous a utomated m odel e valuation
CASP	c ritical a ssessment of techniques for protein s tructure p rediction
COM	contact overlap map
FM	free modeling
GDT-HA	global distance test (high accuracy)
GDT-TS	global distance test
HMM	hidden Markov model
LDDT	local distance difference test
NMR	nuclear magnetic resonance
PDB	protein databank
PDF	probability density function
PMF	potential of mean force
RMSD	root mean square deviation
SCOP	structural classification of proteins
SMTL	SWISS-MODEL template library
TBM	template-based modeling

Table Of Contents

List of Abbreviations	1
Table Of Contents	3
Abstract	9
Introduction	11
1.1 <i>Protein Structure</i>	11
1.1.1 Secondary Structure	12
1.1.2 Tertiary and Quaternary Structure	12
1.1.3 Experimental Methods	13
1.1.4 Resources for Experimental Structures	14
1.2 <i>Sequence and Structure</i>	15
1.2.1 Sequence and Profile Alignments	15
1.2.2 Structure-Sequence Relationship	16
1.2.3 Structural Genomics and the Sequence - Structure Gap	17
1.3 <i>Protein Structure Prediction</i>	17
1.3.1 Benchmarking Existing Structure Prediction Methods	18
1.3.2 Secondary Structure Prediction	18
1.3.3 Template-Based Protein Structure Prediction	19
1.3.4 De-Novo Structure Prediction	20
1.3.5 Model Refinement	21
1.4 <i>Model Quality Assessment</i>	21
1.4.1 Chemical Plausibility Checks	22
1.4.2 Physics-based	22
1.4.3 Potential of Mean Force	22
1.4.4 QMEAN	25
1.4.5 Consensus	25
1.5 <i>Objectives</i>	26
OpenStructure: A Flexible Software Framework For Computational Structural Biology	29
2.1 <i>Introduction</i>	29
2.2 <i>Implementation</i>	30
2.3 <i>Application Example</i>	31
2.4 <i>Acknowledgement</i>	32
OpenStructure — An Integrated Software Framework for Computational Structural Biology	33
3.1 <i>Introduction</i>	33
3.2 <i>Architecture</i>	34
3.3 <i>Molecular Structures</i>	35

3.3.1	Working with Subsets of Molecular Structures	35
3.3.2	The Query Language – Making Selections	36
3.3.3	Selection Example: Superposition	36
3.3.4	Mapping User-defined Properties on Molecular Structures	37
3.3.5	Connectivity and Topology	37
3.3.6	Loading and Saving Molecular Structures	38
3.4	<i>Sequences and Alignments</i>	39
3.4.1	Efficient Mapping of Structure and Sequence-based Information	39
3.4.2	Algorithms for Sequences and Alignments	39
3.4.3	Example: Ligand Binding Site Annotation	40
3.5	<i>Density Maps and Images</i>	41
3.5.1	Correlating Backbone Fragments with local Electron Density	42
3.5.2	Visualization	43
3.5.3	Visual Data Exploration Example: Proteomics Cross-Links	44
3.6	<i>Graphical User Interface</i>	46
3.7	<i>Conclusions</i>	48
Local Distance Difference Test - A Robust, Superposition-Free Protein Structure Similarity Measure		49
4.1	<i>Introduction</i>	49
4.2	<i>Methods</i>	51
4.2.1	The Local Distance Difference Test	51
4.2.2	Multi-reference Local Distance Difference Test	52
4.2.3	Structure Quality Checks	53
4.2.4	Determination of the Optimal Inclusion Radius	53
4.2.5	Validation of Baseline Scores for Different Folds	53
4.2.6	Implementation and Availability	55
4.3	<i>Results and Discussion</i>	55
4.3.1	Determination of the optimal inclusion radius.	55
4.3.2	Sensitivity analysis vs. relative domain movements.	56
4.3.3	Validation of IDDT Score Baselines for Different Protein Folds	56
4.3.4	Local Model Accuracy Assessment	59
4.3.5	Stereo-Chemical Realism Assessment	59
4.3.6	Multi-Reference Structure Comparison	61
4.4	<i>Conclusions</i>	61
Graph-based Constraint Selection for Multi-template Modeling		65
5.1	<i>Introduction</i>	65
5.2	<i>Materials & Methods</i>	66
5.2.1	The Core of Domain-Find	67
5.2.2	Global and Local Domain-Find	68
5.2.3	Fast Update of Edge Weights	69
5.2.4	Determining the Optimal Threshold Value	71
5.2.5	Consistent Constraints for Multi-Template Modeling (MTM)	72
5.3	<i>Results</i>	75

5.3.1	Domain-Find on Pairs of Experimental Structures	75
5.3.2	Identifying Correctly Predicted Residues in Models	78
5.3.3	Constraint Consistency for Multi-Template Modeling	79
5.4	<i>Conclusions</i>	82
Toward the Estimation of the Absolute Quality of Individual Protein		
Structure Models		
85		
6.1	<i>Introduction</i>	85
6.2	<i>Methods</i>	87
6.2.1	QMEAN	87
6.2.2	Datasets	87
6.2.3	QMEAN Z-score	88
6.2.4	Cross-validation	89
6.2.5	Comparison of predicted protein stability between thermophilic and mesophilic organisms	89
6.2.6	Implementation	89
6.3	<i>Results</i>	89
6.3.1	Normalization of the statistical potentials	89
6.3.2	PDB reference set and QMEAN Z-score concept	91
6.3.3	QMEAN Z-score analysis of experimental structures	92
6.3.4	Comparison of homologous proteins from thermophilic and mesophilic organisms	97
6.3.5	Analysis of theoretical models using normalized QMEAN scores and QMEAN Z-scores	97
6.4	<i>Conclusions</i>	99
6.5	<i>Acknowledgements</i>	100
QMEANdist - QMEAN Enhanced with Distance Restraints from		
Alignments		
101		
7.1	<i>Introduction</i>	101
7.2	<i>Materials & Methods</i>	103
7.2.1	Datasets	103
7.2.2	Model Quality Assessment Method	103
7.2.3	Template Identification	103
7.2.4	Clustering of Template Sequences	104
7.2.5	Distance Constraints	104
7.2.6	Scaling of Constraints from Structures with Similar Sequences	105
7.2.7	Scoring of Models	106
7.2.8	Combination of Distance Score with QMEAN	107
7.2.9	Implementation	108
7.3	<i>Results & Discussion</i>	108
7.3.1	Results on CASP8 data	108
7.3.2	Results on CASP9 data	111
7.3.3	The Importance of the Similarity Measure	115
7.4	<i>Conclusions</i>	116

Automated Modeling in SWISS-MODEL Next Generation	119
8.1 <i>Introduction</i>	119
8.2 <i>Materials & Methods</i>	120
8.2.1 Method Overview	120
8.2.2 Datasets	120
8.2.3 Template Identification	122
8.2.4 Template Properties	122
8.2.5 Choice of Structural Similarity Measure	124
8.2.6 Derivation of IDDT PDFs for Properties	124
8.2.7 IDDT Prediction	125
8.2.8 Model Building	125
8.2.9 Scoring of Models with QMEAN and QMEANdist	126
8.3 <i>Results & Discussion</i>	126
8.3.1 Coverage-Dependence	126
8.3.2 Discussion of the IDDT Predictors	127
8.3.3 Normalization of HH Scores	127
8.3.4 Choice of Bandwidth Parameter	129
8.3.5 Discussion of Template Selection Performance on Training Sets	131
8.3.6 Effect of QMEAN	133
8.3.7 Discussion of Template Selection Performance on CAMEO Test set A ..	133
8.3.8 Global IDDT Prediction	135
8.3.9 Comparison of MODELLER and PROMOD-II	135
8.3.10 Performance on CAMEO test set B	140
8.4 <i>Conclusions</i>	143
SMNG Web Interface	145
9.1 <i>Introduction</i>	145
9.2 <i>Implementation</i>	145
9.3 <i>Template Library</i>	146
9.3.1 Biological Units	146
9.3.2 Sequence and Profile Databases	146
9.3.3 Annotation of Cloning Artifacts	147
9.3.4 Web Access to the Template Library	148
9.4 <i>Modeling Interface Walkthrough</i>	149
9.4.1 Input	149
9.4.2 Template Search Results	150
9.4.3 Modeling Results	153
9.5 <i>Conclusions</i>	154
Knowledge-Based Extension of Fragmented Models at Low Resolution in ARP/wARP	157
10.1 <i>Introduction</i>	157
10.2 <i>Methods</i>	159
10.2.1 Identification of Gaps in Intermediate Models	160
10.2.2 FRAGRA	162

10.3	<i>Results</i>	166
10.3.1	Data	166
10.3.2	Application of the Method in Absence of Coordinate Error	166
10.3.3	Incorporation of FittOFF into ARP/wARP Protein Model Building	167
10.4	<i>Conclusions</i>	170
10.4.1	Software Availability	171
10.4.2	Acknowledgements	171
	FRAGRA - Knowledge-based Backbone Conformation Sampling	173
11.1	<i>Introduction</i>	173
11.2	<i>Fragment Database</i>	174
11.3	<i>Sampling Procedure</i>	175
11.3.1	Optimizing the Backbone Geometry	178
11.4	<i>Loop Ranking</i>	179
11.4.1	Density Correlation	179
11.4.2	Filtering Clashing Backbone Candidates	179
11.4.3	Ranking with Statistical Potentials	180
11.5	<i>Conclusions</i>	181
	Summary and Outlook	183
	Acknowledgement	185
	References	187
	OpenStructure — Technical Annex	205
A.1	<i>The Query Language</i>	205
A.1.1	Features of the Query Language	205
A.1.2	Implementation of the Queries	205
A.2	<i>Crystallographic Density Maps in OpenStructure</i>	207
A.2.1	Crystal lattice	207
A.2.2	Space Groups and Symmetry Operators	208
A.2.3	Implementation of Crystallographic Density Maps in OpenStructure	209

Abstract

Proteins are macromolecules which play a crucial role in virtually any process in the living cell. The determination of the 3-dimensional structure of a protein is a key component in understanding its function and mode of action. Preferably, the structure is solved by an experimental technique such as X-ray crystallography, nuclear magnetic resonance (NMR), or electron microscopy (EM). In many instances, experimental structures are unavailable or can not be readily determined. To the rescue come computational modeling techniques, e.g. comparative modeling, which are producing structures at a fast pace. State of the art methods are capable of generating accurate models down to the level of sidechains. These models are a useful tool in designing experiments, e.g. site-directed mutagenesis, virtual screening and identifying proteins of similar function. Despite the recent advancements, comparative modeling still has substantial room for improvement in many areas. In the course of this thesis, we aim at developing techniques which address some of the shortcomings of today's methods. As a solid foundation for this work, the OpenStructure software framework is developed, which allows to conveniently implement new methods and seamlessly integrate them with existing programs.

Computational modeling often requires comparisons of models and/or template structures. Standard structure similarity measures, such as RMSD and GDT are based on global superposition of structures, and their results are not meaningful when applied to structures exhibiting domain movements. For unsupervised comparison of structures on a large scale, a similarity measure based on internal distances was developed, which, to a large extent, is insensitive to domain movements. In analogy to the global distance test, the similarity measure is referred to as local distance difference test (IDDT).

A critical step of template-based modeling is the selection of suitable template structure information. For well characterized protein families, often many alternative experimental template structures are available. While all templates may share a similar overall topology, the relative orientation of sub-domains often differs significantly. Such intrinsic movements limit the assignment of consistent structural constraints for the comparative modeling step. An efficient and robust procedure to identify stable structural building blocks in ensembles of structures using contact-overlap map consistency (COM) is proposed.

The ability of a structural model to answer a particular biological research question is strongly influenced by its accuracy. Since models may contain substantial errors, reliable quality estimates are fundamental to determine their usefulness. We develop techniques to assign quality estimates to models, which expand on the typical potential of mean force (PMF) formalism used in the field. By relating the protein's PMF energy to energy of experimental structures, we obtain a Z-score of the model's structure being of comparable quality to experimentally determined structures. In a second scoring function, the PMF scores are complemented with distance restraints from evolutionary related experimental structures. These restraints are helpful in discriminating between correct and incorrect folds and greatly improve the accuracy of the scoring function.

A novel modeling pipeline for the SWISS-MODEL expert system for comparative modeling is presented. For template and model selection, the pipeline builds on scoring functions developed in this thesis, and combines them with probability-based reliability estimates. The pipeline is embedded into a new web-interface, leveraging on capabilities of modern web browsers to perform the modeling in an interactive manner.

Finally, computational models are often improved by incorporating experimental restraints, e.g. from electron density maps, proteomics cross-links, mutation studies etc. Likewise, at resolutions below 2.5Å, X-ray density maps are often insufficiently defined to allow completely automated model building and can benefit from the incorporation of computational techniques. We explore the application of computational sampling techniques to the automated model building with ARP/wARP at low resolution with the aim to improve model completeness and to reduce fragmentation.

Introduction

1 Protein Structure

Polymers are a reoccurring theme in biological systems. They are built from a limited alphabet of residues and are much more complex than the parts they are made of. Proteins are one such class of polymers and are involved in virtually any process of living organisms. Proteins consist of one or more polypeptides, each of which is a linear chain of α amino acids. The atoms of the amino acids are grouped into backbone (N, $C\alpha$, C and O, H) and sidechain atoms. The α carbon atom of the i th amino acid in a polypeptide chain is connected to the nitrogen of the $i + 1$ th amino acid via a peptide bond. The peptide bond resonates between a charged and a neutral conformation, which gives it a partial double bond character¹. Free rotation around the C-N bond does not readily occur, since this would destroy the π -orbital overlap. This means that the ω dihedral angle [$C\alpha(1)$ - $C(1)$ - $N(1')$ - $C\alpha(1')$] assumes one of two values: $\omega = 0$ (*cis*) and $\omega = 180^\circ$ (*trans*). The *trans*-conformation is slightly lower in energy due to steric hindrance of the sidechains. Only around 0.3 % of peptide bonds occur in the *cis* conformation, 87% of which are peptide bonds preceding a proline residue²⁻³. The rigid planarity of the peptide bond is vital to the functioning of proteins, as it greatly reduces the degrees of freedom of the polypeptide chain. Torsional rotation of the protein backbone is limited to the two dihedral angles ϕ [$C(1)$ - $N(1')$ - $C\alpha(1')$ - $C(1')$] and ψ [$N(1)$ - $C\alpha(1)$ - $C(1)$ - $N(1')$]. The allowed combinations of ϕ/ψ -angles for a given residue have been theoretically calculated by Ramachandran based on steric hindrance of the sidechains. However, the ϕ/ψ pairs in real structures may deviate from the theoretical conformations, since the conformation is influenced by other interactions (van-der Waals, electrostatic, etc.) as well.

The wide range of sidechain chemical properties makes amino acids more versatile than nucleic acids for catalysing reactions. In addition, proteins spontaneously fold into stable 3-dimensional structures. This, among others, may have been a major driving force for the evolution of proteins as catalysts of living cells⁴. Despite their differences, the amino acids can be categorized according to the chemistry of their sidechains. The first class is formed by hydrophobic amino acids, and predominantly occur in the hydrophobic core of proteins. Hydrophobic amino acids are important during the folding as well as for the general stability of the protein as interaction of hydrophobic residues with water molecules are entropically not favourable⁴⁻⁵. Hydrophilic amino acids are predominantly found to be solvent-accessible (asparagine, glutamine). Charged amino acids are often in active sites, as their chemistry is amenable for interactions with other active biomolecules. In addition, they are able to form salt-bridges, which are important for the stability of the protein.

Secondary Structure

Corey and Pauling were the first to describe structure elements that are stabilized by a regular network of hydrogen bonds⁶. The first of these elements, the α -helices are rod-like, wound structures whose inner core is formed by the backbone of the polypeptide, with the sidechains pointing outwards. Hydrogen bonds are formed between the backbone CO group of i th and the NH of the $i + 4$ th amino acid in the sequence⁴. Helices are usually no longer than 45Å. However, in some cases, they entwine to form long, stable helical structures (coiled-coil). The second structural element described by Corey and Pauling are β -strands. Here, the hydrogen network is formed involving more distant residues. The backbone of the polypeptide is fully extended. In anti-parallel β -sheets, the pairing strands run in opposite directions, whereas in parallel β -sheets, the strands have the same direction. These regular secondary structure elements are connected by loops.

A standardized vocabulary of secondary structures has been introduced by Kabsch and Sander in their DSSP program⁷. The program assigns secondary structures states to each residue based on hydrogen bonding patterns. In addition to the above-mentioned α -helix (denoted 'H') and β -strand (denoted 'E'), DSSP introduced the π -helix ('I'), three-turn helix ('G'), turn ('T'), β -bridge('B'), bend ('S') and coil ('C'). Many programs use a simplified 3-state scheme, in which residues are grouped into helical, extended and coil states⁸⁻⁹. This is justifiable, since the other types of secondary structure are very rare¹⁰.

Tertiary and Quaternary Structure

The 3-dimensional arrangement of a polypeptide chain, including its secondary structure elements are referred to as the tertiary structure. Water soluble proteins fold into compact, globular, structures. Hydrophobic sidechains are buried in the core and thus shielded from the water⁵. Hydrophilic sidechains are predominantly found on the surface of the protein. Some proteins fold into several, independently stable regions, termed *domains*.

At the highest level of organisation, the quaternary structure, multiple polypeptide chains arrange into stable and semi-stable complexes. The complexes are stabilized by the hydrophobic effect or electrostatic interactions between residues of the polypeptides chains. Homo-oligomers consist of multiple peptides with the same sequence, hetero-oligomers have at least two different polypeptide sequences.

Oligomers are abundant in the living cell and serve a multitude of functions⁴. First, many structural proteins form oligomeric complexes. Some of these structures assemble into highly symmetric structures with a fixed number of copies, e.g. viral capsids, or the proteasome. For others, e.g. actin filaments, association of monomers into oligomeric filaments is a dynamic process. Here, oligomerisation occurs as response to external stimuli or progression in the cell cycle. Apart from structural reasons, oligomers are supposed to reduce errors in protein translation¹¹. Since the probability of a translation error scales linearly with the number of residues of a polypeptide, the number of units which can be translated without error is higher for smaller proteins. Additionally, the genetic information required to encode a single monomeric unit as opposed to encoding all copies of

a homo-oligomeric complex in the DNA is drastically reduced. Last, changes in the relative orientation of subunits can have regulatory effect on protein function. One example is the allosteric regulation of hemoglobin¹².

Experimental Methods

Over the years, there have been several methods developed to obtain structural information at atomic and near-atomic resolution. The 3 most important techniques are X-ray crystallography¹³, nuclear magnetic resonance (NMR)¹⁴, and electron microscopy (EM)¹⁵. They are all briefly introduced below.

X-RAY CRYSTALLOGRAPHY | X-ray crystallography exploits the properties of highly ordered crystals to obtain structural information of biological macromolecules at atomic resolution. Structure determination is a four-step process: After expressing and purifying the protein in sufficiently large quantities, protein crystals are grown. The crystalline sample is placed in front of a X-ray detector and irradiated with a X-ray beam. The X-ray wave interacts with the electrons of the sample and is diffracted by them. Due to the crystalline nature of the sample, the resulting diffraction pattern has non-zero intensity only at specific positions, the reflections. The reflections are related to the electron density of the sample via a Fourier relation. However, the diffraction pattern is a power spectrum, meaning that the observed intensity is proportional to the square of the amplitude of the waves. The phase information of the waves, which is also important for the determination of the structures is not available from the experiment. For very-high resolution structures, e.g. small molecules, properties of the inter-atomic distances are sufficient to determine the phases. For typical resolutions of data of biological macromolecules, the phases need to be obtained by other means. This is called the phase problem of crystallography. Commonly, molecular replacement, in which the phases are transferred from a protein of supposedly similar structure, or multiple anomalous dispersion are applied¹⁶.

The high degree of automation and availability of sophisticated refinement programs, currently make X-ray crystallography the method of choice to obtain protein structures at atomic resolution.

NUCLEAR MAGNETIC RESONANCE (NMR) | Nuclear Magnetic Resonance (NMR) is a spectroscopic technique to obtain information on the spatial arrangement of atoms in macromolecules in solution. It relies on the energy difference between spin states of nuclei with an uneven number of protons and neutrons in a magnetic field, e.g. the nuclei of hydrogen, ¹³C, or ¹⁵N atoms. By using a radio pulse, state transitions between the low and high energy spin state can be induced. Due to chemical shielding by electrons, the magnetic field perceived by nuclei differ. These differences can be detected in the spectrum. For proteins, typically higher-dimensional spectra are required, since chemical shifts of atoms can overlap and are indistinguishable from each other. By using sophisticated pulse-patterns, the signal is split across multiple dimensions. From all these spectra, distance constraints are extracted which are used to simulate possible conformations for

proteins. When enough of these distance constraints are known, the protein structure can be readily determined. In regions with enough distance constraints, the models typically agree well, for parts where not enough distance constraints are known, the models show large fluctuation in atomic positions. One advantage of NMR is to follow molecules in solution, e.g. to observe conformational changes. However, NMR is restricted to relatively small proteins only.

ELECTRON MICROSCOPY (EM) | Due to the smaller wave-length of electrons, electron microscopes can go beyond the resolution limit of conventional light microscopes. While electron microscopes can be used to obtain information at atomic resolution for metallic compounds (e.g. gold), the signal obtained by electron microscopy of biological samples is severely limited by the sensitivity of biological material to radiation damage. Much lower electron doses need to be used which leads to a smaller signal to noise ratio. Thus, the information from several copies of biological macromolecules need to be averaged in order to obtain high resolution density maps. There are two modes of operation for the electron microscope to obtain high-resolution information for proteins. The first, and more widely-used technique, is *single particle averaging*. To overcome the low signal-to-noise ratio, several images of particles (proteins) are collected. Each imaged particle is a projection of the particle's density. Using Radon back-projection, the two-dimensional images are assembled into a 3-dimensional density map¹⁵. Currently, the use of single particle averaging is limited to large particles. Work has been performed on the ribosome, which gave insight into the process of protein translation¹⁷. For highly symmetric assemblies, near-atomic resolutions have been obtained. But typical resolutions for non-symmetric particles are in the 10-15Å range¹⁵. As a second mode of operation, diffraction patterns of two-dimensional protein crystals can be collected. This technique is however limited to membrane proteins¹⁵.

Resources for Experimental Structures

Experimentalists deposit structures of polypeptides and poly-nucleotides in the Protein Data Bank (PDB). The PDB was established in the early seventies to make the small but growing number of solved protein structures available to the scientific community¹⁸. The atomic coordinates are deposited together with information associated with the crystallized polymer, e.g. the oligomeric state, references and experimental details such as unit cell size and refinement parameters. Each experimental structure is assigned a unique four-letter code (the PDB identifier) as well as a digital object identifier (DOI). The number of structures in the PDB has been growing exponentially. While, in the beginning, there were only a few structures deposited every year, today more than 80'000 entries are available. Part of it can be attributed to the high amount of automation in solving structures and to the efforts of the structural genomics projects¹⁹. The majority of structures are solved by X-ray crystallography, followed by NMR and electron microscopy.

In the last few years, the efforts of the PDB are managed by a world-wide consortium of scientists²⁰. They are responsible for identifying the requirements of the research

community as well as defining data exchange dictionaries. The structures themselves are made available through mirror sites, e.g. RCSB²¹, PDBe²² and PDBj²³.

Since the quality of the structures depends on both the experimental data and the refinement protocols, structures solved decades ago can often be improved by using modern refinement protocols. PDBredo is making the efforts of re-refining the structures available to the research community²⁴. Likewise, Paul Adams has reported that already deposited structures are often improved by using newer version of the PHENIX package²⁵.

Many other databases are derived from the PDB such as CATH²⁶ and SCOP (structural classification of proteins)²⁷, which classify protein structures in families based on their folds

2 Sequence and Structure

Sequence and Profile Alignments

The importance of evolutionary relationship between protein sequences for many bioinformatics and computational biology methods has led to the development of increasingly sophisticated descriptions of sequence similarity^{28–33}.

Sequence identity is the crudest and least sensitive of similarity measures. It is calculated as the fraction of conserved amino acids divided by the number of aligned residues. Because of its simplicity, sequence identity is often used to categorize the evolutionary distance of two proteins. Not all mutations have the same impact since some amino acids are chemically more related than others. Mutating an alanine into a valine is on average a smaller change than mutating a tyrosine into a glycine. Substitution scores take into account how favourable a certain amino acid substitution is. A substitution score penalizes and rewards mutations according to a scoring matrix $S(a, b)$. $S(a, b)$ is positive if the mutation of a to b is observed more often than would be expected by a chance. Vice-versa, negative elements of $S(a, b)$ denote mutations from a to b that are less often observed and thus less favourable than a random null-model. All substitution scores can be understood as log-odd scores of co-occurrence³⁴, e.g. $S(a, b) = \log f(a, b) / f(a)f(b) = \log f(a|b) / f(a)$. Many substitution matrices have been generated, some better suited to measure evolution of closely related protein sequences, some targeted at more distant pairs of protein sequences^{28,35–36}. The scores for the most widely-used substitution scoring matrix BLOSUM62 have been estimated based on co-occurrence probabilities of amino acids in columns from a large multiple sequence alignment with sequences sharing less than 62 percent identity²⁸. Other scoring functions have been derived from pairwise contact potentials³⁶.

The most sensitive of the currently available alignment programs represent the query sequence as a Hidden Markov Model (HMM). The amino acid emission probabilities of each column are estimated from a multiple sequence alignment generated for the

query sequence. The transition probabilities for match, insertion and deletion states are estimated from the multiple sequence alignment as well. The query HMM is then either aligned against sequences (HMM-sequence alignment^{31,37}), or against a database of HMMs (HMM-HMM alignment)^{32–33}. HMMs have greatly improved the detection of remote homologs. In some rare cases, HMM-HMM programs are able to detect homologs with less than 15% sequence identity.

Structure-Sequence Relationship

The work of Anfinsen in 1973³⁸ revealed that a major determinant of the 3-dimensional structure of a protein is its primary sequence. Folding is driven by thermodynamic stability. Today, this concept is still the basis of our understanding, though it is known that *in-vivo* folding is much more complex³⁹. Ensuring proper folding of proteins in the crowded intra-cellular milieu requires the interplay of hundreds of genes expressing chaperones, degradation pathways and post-translational modifications. They protect the unfolded and semi-folded intermediates from interactions that could lead to misfolding.

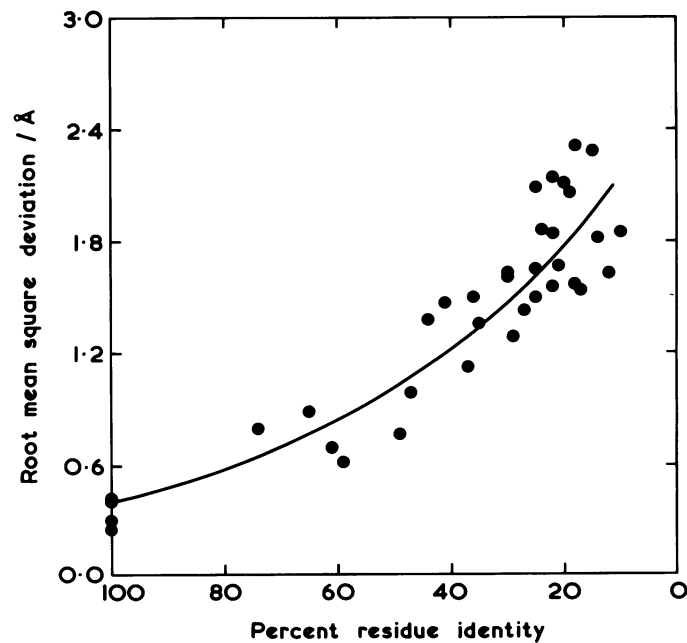


Figure 1.1 The sequence-structure relationship states that proteins with similar sequence adopt a similar structure⁴⁰.

The importance of the amino acid sequence on folding and the tertiary structure led Chothia and Lesk⁴⁰ to compare experimental X-ray structures of evolutionary related proteins. They calculated the RMSD of the conserved core of 20 proteins and plotted it against the sequence identity of the protein pair (**figure 1.1**). A clear relation between the root-mean-square deviation and the sequence identity could be seen. With increasing sequence identity between the two proteins, the RMSD decreases. The relation is clearly

non-linear and the RMSD increases more rapidly with increasing dissimilarity between the two protein sequences. In their work, they were limiting the structural comparison to what they called the *structural core* of the protein, made necessary by the use of RMSD as the similarity measure. However, similar trends can be seen when comparing the complete structures using more robust structural similarity measures such as GDT⁴¹.

Structural Genomics and the Sequence - Structure Gap

Even though the number of experimentally determined protein structures has increased exponentially, deep sequencing technology has led to an enormous increase of available genome sequences. As a result, the difference in numbers between experimentally available protein structures and sequences — the sequence - structure gap — is becoming larger⁴². To counteract these trends, the structural genomics initiative has made efforts to further automate the structure determination processes. Proteins were not selected on biological relevance, but on homology (or lack thereof) to already solved protein structures. A sequence has been selected if it shares less than 30% sequence identity to existing protein structures⁴³. Despite these efforts, it is unlikely that the number of structures will keep up with the ever-increasing number of available sequences.

The question remains as to how well the currently solved structures cover the fold-space, e.g. how likely is it that a newly solved structure shares considerable structural similarity to an already known protein structure. There has been a gradual decline in newly discovered folds, e.g. as seen in the SCOP database. Can we assume that the fold-space is completely covered by the PDB? To address this question, *in-silico* reduced polypeptide models have been built using a pair-wise attractive potential, hydrogen bonding terms and excluded volume to guide the sampling process⁴⁴. The researchers then compared the built models to the existing protein structures. They found, that for any built model, there is an experimental structure that shares significant similarity on the fold level. Also, the inverse was true: for any experimental structure from a set of 150 PDB structures, a structure was found in the library of sampled models with significant structural similarity. This suggests a strong upper limit of folds that single-domain proteins can adapt. They argue that the limited number of folds in the PDB is a direct effect of geometric constraints imposed by regular secondary structure elements. For proteins with less secondary structure content, the degrees of freedom increase and, as a result, many more folds become possible.

Based on the sequence-structure relationship and the increasing coverage of fold space, the sequence-structure gap can be bridged by computational modeling methods.

3 Protein Structure Prediction

Protein structure prediction uses two fundamentally different concepts: template-based structure prediction techniques obtain a 3-dimensional, atomistic model by exploiting

the structure–sequence relationship described above. *De-novo* structure prediction generates vast sets of alternative structures and selects the best scoring model using sophisticated scoring functions. Before turning our attention to computational structure prediction methods, efforts to objectively benchmark competing approaches are introduced in the next section.

Benchmarking Existing Structure Prediction Methods

Critical assessment of structure prediction (CASP) is a community-wide double-blind benchmark for protein structure prediction and related methods, taking place every second year^{45–53}. Participants in the protein structure prediction category are sent amino acid sequences of proteins whose structures have been experimentally determined but are not publicly available yet, and are asked to return 3-dimensional models. After the prediction season, which typically lasts 3 months, the submitted models are compared to the experimental structures using a variety of structural similarity measures. Methods which perform particularly well are then highlighted at the CASP meeting.

As an alternative to CASP, since 2011, the Continuous Automated Modeling Evaluation (CAMEO) web server benchmarks computational approaches for protein structure and ligand binding site prediction⁵⁴. Each Friday, sequences of the PDB pre-release are sent to the predictors. The following week, the predictions are compared to the experimental structures. The number of targets per week is usually between 15 and 40, meaning that it takes between 3 to 8 weeks to have a target number comparable to CASP. In contrast to the model evaluation of CASP, evaluation of CAMEO targets is completely automated and does not involve human intervention.

Secondary Structure Prediction

Since regular secondary structure elements are an essential part of protein structures, a major contributor to protein stability, and arguably even to the folding-pathway, there has been a strong interest in predicting the secondary structure elements from the primary sequence. Early attempts at secondary structure prediction were based on the observation that some amino acids are more commonly found in secondary structure elements than in others⁵⁵. However, these prediction methods never reached an accuracy higher than 70%. Modern secondary structure prediction programs rely on sequence profiles to improve the accuracy of the prediction. For example, today's most widely-used secondary structure prediction program PSIPRED⁸ generates a multiple sequence alignment for the protein of interest with PSI-BLAST²⁹. Each column in the alignment is converted to a vector of amino acid probabilities, derived from the frequencies of occurrence. The resulting set of probability vectors is then used as input to a neural network. PSIPRED has repeatedly been shown to be one of the top-scoring secondary structure prediction programs in the EVA live benchmark⁵⁶. The reliability measures of PSIPRED agrees well with actual errors. The residues marked most accurate reach overlap of >90% to the DSSP states. In general, the accuracy of secondary structure prediction programs

to predict α -helices is higher than for β -strands. The hydrogen bonding structure of α -helices and β -strands can partially explain the differences in performance. For α -helices, the hydrogen donor and acceptors pairs are at fixed offset in the primary amino acid sequence. Identifying the hydrogen donors and acceptors in β -sheets on the other hand is a much more complex problem, since the offset between the donor and acceptor is not fixed. Additionally, the stabilizing interactions are less local than for α -helices.

Template-Based Protein Structure Prediction

Template-based modeling techniques are based on the sequence-structure relationship first outlined by Chothia and Lesk⁴⁰, and exploit experimentally available structures to obtain a protein structure model. In contrast to *de-novo* methods, they are primarily based on evolutionary information and only secondly on energy functions. For these methods, the most important step is to identify related experimental protein structures.

Most comparative modeling procedures consist of four consecutive steps: (a) identification of protein structures related to the target sequence with a target/template alignment, (b) modeling of the target structure based on the information of the template, (c) refinement of the model, (d) evaluation of the model quality and ranking of generated models. These steps might be repeated iteratively until a satisfactory model is obtained⁵⁷.

In traditional comparative modeling, local alignment algorithms such as BLAST⁵⁸ are used to obtain an alignment between the target sequence and experimentally determined structures. For sequence alignments above 40%-50% sequence identity, the alignments are very accurate and the fold between the target and the template is conserved. When no close homologs are detected, more sophisticated homology detection algorithms are required. Successful approaches are based on sequence-profile⁵⁹, sequence-HMM⁶⁰, or HMM-HMM alignments³²⁻³³. Several research groups have reported improved model accuracy when combining multiple homology detection programs⁶¹⁻⁶³.

Other programs for remote homology detection thread the protein sequence through template structures⁶⁴⁻⁶⁶. These programs have traditionally been into classified as fold-recognition methods. However, with the advent of more sensitive sequence-based homology detection programs, the distinction has started to blur. Especially in the twilight zone for protein sequence alignment⁶⁷, improvements in alignment quality are possible by combining sequence and structural information. For example, RaptorX adjusts the importance of sequence and structural information based on a non-linear regression tree⁶⁸⁻⁶⁹. For high-sequence identity alignments, alignment scoring is mainly driven by sequence features, whereas the importance of structural *threading* features increases for remote targets. Similar adjustments are performed when the profile generated for the target sequence has a low number of effective sequences, that is a low entropy. For such profiles, the amount of information is insufficient to approximate the evolutionary events from the sequence alignment. This can be compensated by increasing the relative importance of structural information⁶⁸.

Once a target-template alignment is available, two conceptually different approaches exist to build a 3-dimensional protein model: modeling by assembly of rigid bodies⁷⁰⁻⁷¹, and modeling by satisfaction of spatial restraints⁷².

MODELING BY ASSEMBLY OF RIGID BODIES | In this approach, structurally conserved regions are directly copied from the template to the model. Variable regions such as insertions and deletions are then remodeled using a loop modeling protocol^{73–76}. The last step is modeling of sidechain conformations. For residues which are conserved between the target and the template, the sidechain coordinates can be copied from the template to the model. However, non-conserved residues need to be remodeled in any case. Typically, backbone-dependent rotamer libraries are used to sample possible sidechain conformations e.g. as done by the SCWRL software^{77–78}. Modeling by assembly of rigid bodies has the advantage of being very fast and accurate in the high sequence identity range.

MODELING BY SATISFACTION OF SPATIAL RESTRAINTS | A different approach to structure prediction has been introduced by Sali in 1993 in his MODELLER program⁷². The structure determination is described as an optimization of spatial restraints between atoms in the structure. Distances between atoms, angles and dihedrals are modelled as probability density functions (PDFs). PDFs can assume any form, provided that they integrate to one and are always positive. Probability density functions of MODELLER have been derived for many features, and a variety of sources: from known protein structures, force fields, or stereo-chemical considerations. The formulation of modeling as satisfaction of spatial restraints allows for flexible combination of structural information from multiple sources. For example, the use of restraints from multiple templates is a natural extension of modeling with a single template structure. Instead of expressing the distance between two atoms by a single PDF, two or more PDFs may be combined to obtain a *feature* PDF. Constraints from different template structures are additive as they represent a different conformation for the feature: Thus, the feature PDF is a weighted sum of the individual basic PDFs.

The molecular PDF is then the probability density function for the whole protein structure to be modelled. Additivity is assumed and thus the molecular PDF is given by the product of the individual feature PDFs. This is clearly incorrect as features, especially local ones, are highly correlated. To some extent, this can be corrected by introducing higher-order PDFs. However, the derivation of higher-order PDFs is limited by the available experimental information.

Structures that optimally satisfy the restraints are generated by a series of conjugate gradient optimizations of the molecular PDF. The molecular PDF is first approximated with only local terms enabled. This allows for local packing and folding of secondary structure elements. In each iteration, more terms are added until the target function is identical to the molecular PDF. An ensemble of models is obtained by choosing different initial conditions.

De-Novo Structure Prediction

Comparative, or template-based modeling techniques rely on an alignment of the target sequence to determined protein structures. When no template information is available for the whole, or parts of the sequence, *de-novo* methods may be used to predict the

3-dimensional arrangement of atoms. These methods are independent of aligned template structures and are thus able to predict the structure of proteins for any sequence. Even though *de-novo* methods are not using structures as a whole, the most successful techniques do incorporate information from experimentally available structures, e.g. in the form of fragments for backbone conformation sampling, or empirical energy functions derived from databases⁵⁷. Successful *ab initio* structure prediction methods include ROSETTA, which uses a fragment-guided sampling technique together with a sophisticated energy function⁷⁹, and threading of the target sequence through structures from the PDB combined with lattice-based simulations, e.g. as implemented in I-TASSER⁶⁶.

Overall, the quality of *de-novo* predictions is still rather poor and for many targets, the predictions do not reach fold-level accuracy⁸⁰. Thus, template-based methods are preferable under almost all circumstances, as they deliver more accurate results at a fraction of the time. Still, *de-novo* techniques play an important role in modeling of large insertions or deletions of structures which have otherwise been predicted by comparative modeling. In addition, energy functions developed for *de-novo* predictions may be used to refined comparative models.

Model Refinement

Template-based modeling protocols often incorporate a model refinement step. It serves two main purposes: first, the refinement step regularizes the structure, e.g. by removing clashes, and adjusting bond lengths and angles to chemically possible values. Second, in some cases, conformations closer to the native state can be identified by using conformational sampling. Many of the energy functions in use for *de-novo* structure prediction are also applied for model refinement. They are often able to distinguish between native and non-native conformations^{81–83}. However, the scoring functions are not able to distinguish between near-native and non-native conformations. The result is a *blind* search until the native state is visited as part of the conformational sampling. To avoid conformational drift, refinement protocols include information from template structures to re-restraint the possible conformations of the model⁶⁶. Still, even the best-performing servers at CASP are unable to improve upon the best available structural template in more than 30% of the cases, in only 20% by more than 2 GDT_HA points⁵³.

4 Model Quality Assessment

Any structural model, irrespective if it has been determined by X-ray crystallography or is purely computational, is just an approximative representation of the protein's true structure, for otherwise it would be the protein itself. The question is not *if* models have errors, but how large the errors are. Model quality assessment has set itself the task to analyse theoretical models and assign error estimates. The advancement in recent years to detect more remote homologs have made model quality assessment even more important, as

models may contain substantial errors. Applications of model quality assessment ranges from selecting the best model among a set of alternative model or the prediction of per-residue quality estimates on a global scale.

Model quality assessment routines can be broadly categorized into (a) chemical plausibility checks, (b) physics-based (c) knowledge-based and (d) consensus-based quality checks.

Even though some of the scoring functions described here have applications outside of model quality assessment (fold recognition, model ranking), we would like to focus the attention on their application to predicting errors for models.

Chemical Plausibility Checks

Chemical plausibility checks assess the chemical compliance of a protein structure. Bond-lengths and angle parameters are compared to values obtained from high-resolution structures, e.g. the set defined by Engh and Huber⁸⁴. Additionally, conformance of backbone torsions with the Ramachandran plot, planarity of rings and sterical clashes are checked. For experimental structures, such checks are routinely employed^{85–86} as part of the structure deposition process, and the results are deposited together with the atomic coordinates. For theoretical models, these checks have only recently been added to model evaluation protocols^{52–53}. In drawing an analogy to writing, chemical plausibility check the spelling of individual words, but are oblivious to the structure of sentences. In a sense, the plausibility checks are assessing an orthogonal quality of models, and a model with strong violations of these parameters can still be close to the target structure. For further use of models, e.g. molecular dynamics simulations or in-depth analysis of atomic interactions, adherence to chemical and physical laws is highly important.

Physics-based

Physics-based quality estimation methods rely on the thermodynamic hypothesis, that the native conformation of a protein lies in the free energy minimum⁸⁷. The energy of a protein structure is described using physics-based energy functions describing interactions between atoms and entropic contributions. The functions are parametrized by fitting experimental data or performing quantum chemical calculations. Others have performed molecular dynamics (MD) simulations to assess a model's quality^{88–91} by calculating the stability of a particular conformation. They claim that structures close to the native state are stable, that is the RMSD does not change much with respect to the initial conformation. Non-native conformations on the other hand, tend to drift away from the initial conformation⁹¹.

Potential of Mean Force

Statistics on how often a certain type of residue is buried inside the core of a protein, or the expected distance of a pair of atoms may be turned into knowledge-based scores or

potentials of mean force. They assess how well a given model agrees with our current knowledge of protein structure. Initially, potentials of mean force have been motivated by the inverse Boltzmann law, where state frequencies are turned into energies. More recently, a more versatile and intuitive description has arisen that is based on information theory^{92–94}. Others have motivated potentials of mean force from particle correlation functions⁹⁵.

The Boltzmann principle connects the energy state c_i of a conformation at equilibrium to the probability $p(c_i)$ of that conformation:

$$p(c_i) = \exp(-E(c_i)) / \sum_j \exp(-E(c_j)/kT)$$

where k is the Boltzmann constant, T is the absolute temperature, i refers to the conformational state of interest, and the sum j runs over all states of the system. The denominator, $Z(C) = \sum_j \exp(-E(c_j)/kT)$ is called the Boltzmann sum or partition function of the system. The inverse Boltzmann law allows to derive energies from the occurrences of a conformation:

$$E(c_i) = -kT \ln p(c_i) + kT \ln Z(C)$$

Rather than assigning absolute energies to a conformation, it is more practical to consider energy differences with respect to a reference conformation. Typical examples include the energy difference of an interaction between two particular types of sidechains and interactions of any kind. The specific interaction is denoted as $p(c_i|s_k)$, which translates to, the energy of conformation c_i , under the condition that we only consider components of the system in state s_k . The particular meaning of the state s_k and the conformation c_i are on purpose left open. Specific parametrisations for both c_i and s_k will be given below. The energy difference is given as

$$\Delta E(c_i|s_k) = E(c_i|s_k) - E(c_i) = -kT \ln[p(c_i|s_k)/p(c_i)] + kT \ln[Z(C)/Z(c|s)]$$

Under the assumption, that $Z(c)$ is equal to $Z(c|s)$, the energy difference simplifies to

$$\Delta E(c_i|s_k) = -kT \ln[p(c_i|s_k)/p(c_i)]$$

The probabilities to derive the energies can be estimated from experimental structures, e.g. high resolution structures from the PDB. To give us a better understanding of the net energy difference, it is useful to make a link to informatic quantities. The average energy difference over all states and conformations is

$$\langle \Delta E(C|S) \rangle = \sum_k \sum_i -kT p(c_i, s_k) \ln[p(c_i|s_k)/p(c_i)]$$

Upon expansion, the net energy difference can be written as

$$\begin{aligned} \langle \Delta E(C|S) \rangle &= -kT \sum_k p(s_k) \sum_i p(c_i|s_k) \ln p(c_i|s_k) + kT \sum_i \ln p(c_i) \sum_k p(c_i, s_k) \\ &= kT \sum_k p(s_k) H(C|s_k) + kT \sum_i p(c_i) \ln p(c_i) \\ &= kT [H(C|S) - H(C)] = -kT \cdot I_g(S, C) \end{aligned}$$

with

$$\begin{aligned}H(C) &= - \sum_i p(c_i) \ln p(c_i) \\H(C|S) &= - \sum_i \sum_k p(c_i|s_k) \ln p(c_i|s_k) \\I_g(C, S) &= H(C) - H(C|S)\end{aligned}$$

Here $H(\cdot)$ denotes the Shannon entropy⁹⁶. As can be seen, the net energy difference is up to a constant factor of kT identical to the information gain of the system^{92–94}.

The link to informatic quantities gives us some insight into the nature of potentials of mean force. Using this notation, it is clear that the choice of reference state will have an influence on the information gain and thus the discriminative qualities of the potential. In the literature, many reference states have been proposed. Some were derived from statistics extracted from experimental protein structures, others from theoretical considerations⁹⁵. No definitive answer to what the best reference state is has been found yet⁹⁷.

Apart from the reference state, the information gain also depends on the choice of c_i and s_k . According to Solis, these two parameters should be chosen in order to maximise the information gain of the system⁹³. This requires to identify components with similar distribution of c_i , as this will lead to the sharpest distributions. Typically, one would like to have as many s_k as possible. However, the choice of s_k is affected by data sparsity as well. When only limited data is available, grouping of components in the system might be required to improve performance.

Even though additivity does in general not hold for biological polymers⁹⁸, due to data sparseness and computational tractability for potentials of mean force, additivity is usually assumed. In the literature, there have been some correction factors suggested to reduce the magnitude of errors introduced by the additivity assumption.

In the following, let us turn the attention to particular potentials of mean force that have been proven useful for the assessment of model quality.

INTERACTION POTENTIAL | Interactions are parametrised on distances between atoms, e.g. $\Delta E_{int}(d, a_i, a_j) = - \ln p(d|a_i, a_j)/p(d)$, where d is the distance between two atoms of type a_i and a_j ^{99–107}. Several definitions of atom types have been used, but the most common one is use one atom type for every chemically distinguishable atom of the standard amino acids. Some potentials additionally introduce a sequence separation parameter s which has the effect that only interactions between atoms further than s apart in sequence are considered. This counter-balances the usually observed over-weighting of interactions of neighboring residues. Most widely used and successful potentials include DFIRE¹⁰⁶, RAPDF¹⁰³, DOPE¹⁰⁷ and QMEAN¹⁰⁸. In addition to the full-atom models, potentials have been introduced that operate on a subset of atoms, e.g. backbone atoms and an additional virtual center of sidechain, or $C\beta$ potentials using only one atom per residue¹⁰⁸. Recently, potentials have been developed that, in addition to the distance, include angular parameters between the interacting partners^{95,109}.

SOLVATION POTENTIAL | Solvation potentials account for the preference of amino acids to be exposed to solvent^{108,110–111}. The burial status of residues is approximated as the number of interacting residues (e.g. C β atoms) in a sphere of radius r , e.g. $r = 9$ in the original QMEAN¹⁰⁸. The energy is then calculated as $\Delta E_{solv}(n, a_i) = -\ln p(n|a_i)/p(n)$, where n is the number of interacting C β atoms, and a_i is the amino acid type.

TORSION POTENTIAL | Torsion potentials are parametrised on ϕ/ψ backbone torsion angles and measure the propensity of a certain residue for a given torsion angle pairs compared to a background distribution. A single-residue torsion potential, can be thought of as the log-odd score derived from the residue-specific Ramachandran plot in comparison to the pooled Ramachandran plot of all residues. Since the torsional preference also depends on the neighbours of the residue, there have been various torsion potentials proposed that also include information from the neighboring residues^{108,112–115}. Typical parametrisation are based on the amino acid type before and after the residue of interest as well as torsion values. Torsion potentials are most affected by data sparseness, and torsion angles need to be binned heavily to achieve saturation.

QMEAN

QMEAN (Qualitative Model Energy ANalysis) is a composite scoring function that combines statistical potentials terms with evolutionary information to assess the quality of protein structures **figure 1.2**. QMEAN has been developed by Pascal Benkert during his PhD in the Schomburg group¹⁰⁸.

On the statistical potential side, QMEAN implements an all-atom interaction potential, an interaction potential based on β carbons only, a solvation potential and a torsion potential parametrised on 6 consecutive ϕ/ψ torsion angles. These four terms are complemented by two evolutionary agreement terms: The predicted secondary structure by PSIPRED⁸ is compared to the observed secondary structure by DSSP⁷. Similarly, the solvent accessibility predicted by ACCpro¹¹⁶ is compared to the solvent accessibility assigned by DSSP. The terms are linearly combined to obtain the QMEAN score for a certain model.

Even though the performance of QMEAN is well behind that of consensus-based quality assessment methods, QMEAN has repeatedly been shown to perform among the top non-consensus scoring functions in the quality assessment category of CASP^{117–118}.

Consensus

The idea of scoring structural features by their abundance can be extended to complete models as well. Consensus, or clustering methods, score models of the same target sequence by structural comparison, e.g. superposition¹¹⁷, or internal instance agreement⁶³. Regions with high similarity (e.g. small C α -atom deviations) are thought to be more accurate than regions with large deviations. Many variants of consensus methods have been

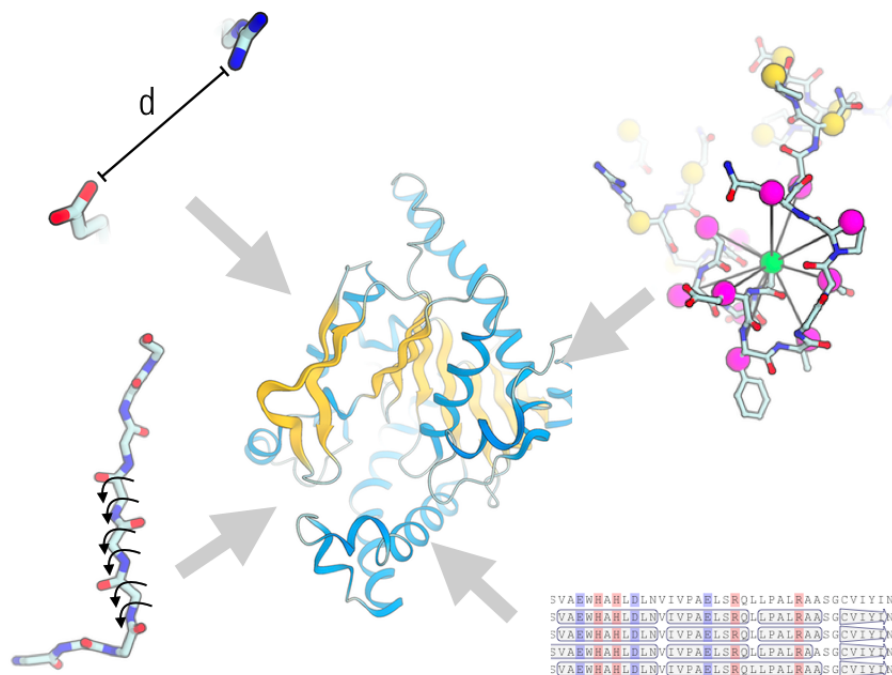


Figure 1.2 Graphical depiction of the statistical and agreement terms of the QMEAN scoring function: The all-atom and $C\beta$ interaction potentials (top-left) are combined with the solvation (top-right) and torsion potential (bottom-left). Additionally, agreement between predicted and observed secondary structure and solvent accessibility are taken into account.

proposed in the literature, e.g. ModFOLDclust^{63,119}, QMEANclust¹¹⁷, MQAPmulti (unpublished), Pcons and ProQ¹²⁰. In addition to structural consensus, they employ potentials of mean force, predicted features from the target sequence to rank models. However, their performance is clearly dominated by the structural comparisons of models. In the context of the quality assessment category of CASP, consensus methods are very successful. The high number of models built by independent modeling pipelines are ideal for consensus methods. Modeling errors tend to cancel out and the most populated states are usually closest to the target structure. As soon as the number of models becomes smaller, e.g. less than 50, the performance of consensus methods starts to degrade¹¹⁸. Since such large number of models are difficult to come by, especially generated by independent modeling pipelines, the application of consensus methods outside of CASP is limited.

5 Objectives

The main objectives of this thesis are to advance methods for protein structure prediction. Contributions in two major areas are made. First, as we have seen, assessment of model quality is of high relevance. We will expand on existing approaches to assign error estimates to models and improve their accuracy. Second, we develop methods to increase

the model accuracy itself by improving template selection and developing approaches to select consistent constraints from multi-template modeling.

This thesis is organised as follows: First, we will describe a computational software framework upon which the remainder of the work is built. Second, a similarity measure for superposition-free comparison of pairs of proteins is described. Third, an algorithm for the identification of common residues and its application to multi-template modeling is shown. Fourth, the QMEAN scoring function is extended: we transform the QMEAN scores into absolute quality scores by relating the pseudo-energy of models to that of experimental structures of similar size. In a second score, to improve the fold-level assessment of QMEAN, we combine the scoring function with distance constraints from alignments. Fifth, the work on the QMEAN scoring function is then used as the basis for a newly developed modeling pipeline for SWISS-MODEL. Sixth, the pipeline is made available to the research community through the SWISS-MODEL web server, which focuses on interactive modeling. Seventh, in a collaboration with Tim Wiegels from the EMBL in Hamburg, we combine methods from computational modeling with X-ray density map interpretation to improve model completeness at resolutions below 2.5Å.

OpenStructure: A Flexible Software Framework For Computational Structural Biology

This chapter has been published as:

Biasini M.^{1,2}, Mariani V.^{1,2}, Haas J.^{1,2}, Scheuber S.^{1,2}, Schenk A.D.³, Schwede T.^{1,2} and Philippsen A.¹ (2010) *OpenStructure: A flexible software framework for computational structural biology*, *Bioinformatics*, 26, 2626–2628.

¹ Biozentrum, University of Basel, Klingelbergstrasse 50 / 70, 4056 Basel, Switzerland

² SIB Swiss Institute of Bioinformatics, Basel, Switzerland

³ Department of Cell Biology, Harvard Medical School, 240 Longwood Ave, Boston, MA 02115, USA

MOTIVATION: Developers of new methods in computational structural biology are often hampered in their research by incompatible software tools and non-standardized data formats. To address this problem, we have developed OpenStructure as a modular open source platform to provide a powerful, yet flexible general working environment for structural bioinformatics. OpenStructure consists primarily of a set of libraries written in C++ with a cleanly designed application programmer interface. All functionality can be accessed directly in C++ or in a Python layer, meeting both the requirements for high efficiency and ease of use. Powerful selection queries and the notion of entity views to represent these selections greatly facilitate the development and implementation of algorithms on structural data. The modular integration of computational core methods with powerful visualization tools makes OpenStructure an ideal working and development environment. Several applications, such as the latest versions of IPLT and QMean, have been implemented based on OpenStructure—demonstrating its value for the development of next-generation structural biology algorithms.

AVAILABILITY: Source code licensed under the GNU lesser general public license and binaries for MacOS X, Linux and Windows are available for download at <http://www.openstructure.org>.

CONTACT: torsten.schwede@unibas.ch

SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

1 Introduction

We introduce OpenStructure, a flexible software framework for computational structural biology, a solid, yet flexible and versatile toolkit for rapid prototyping of new methods as well as their productive implementation. Typically, method development in structural bioinformatics involves combining different independent software tools, and significant effort is devoted to writing code for input/output operations and format conversions between different packages. This culminates when data and algorithms from different domains are to be combined, e.g. protein structures, protein sequence annotation and chemical ligands. Several software tools and frameworks are available today for molecular modeling, e.g. MMTK¹²¹, Coot¹²², MolIDE¹²³, Modeller¹²⁴, bioinformatics algorithms libraries, e.g. BALL¹²⁵, workflow automation tools, e.g. Biskit¹²⁶ or KN-

IME (www.knime.org) and visualization e.g. VMD¹²⁷, PyMol (www.pymol.org), DINO (www.dino3d.org), or SwissPdbViewer⁴².

OpenStructure is a flexible software framework tailored for computational structural biology, which combines a C++ based library of commonly used functionality with a Python layer and powerful visualization tools. While PyMol and VMD also combine a scripting environment with sophisticated visualization tools, they are primarily geared toward visualization and less on providing a clean application programmer interface (API) that is easy to use and allows for rapid development of new algorithms. OpenStructure is also designed to easily accommodate interfaces to already existing software. This allows for rapid visually enhanced prototyping of new functionality, making OpenStructure an ideal environment for the development of next-generation structural biology algorithms. For example, new versions of the QMean tools for model quality assessment^{117,128} are based on OpenStructure, as well as the structural analysis tools in ProteinModelPortal¹²⁹. Further, work is on the way to implement the next generation of the SWISS-MODEL pipeline using the OpenStructure framework^{130–131}.

2 Implementation

In OpenStructure, molecular or chemical entities, such as macromolecules, sequences, alignments or electron density maps, are represented as objects, offering a comprehensive set of functions for data manipulation and information querying. Typically, users interact with a high-level Python interface, while ‘power users’ with high computational requirements access the API at the level of C++.

Functionality in OpenStructure is grouped into modules. Each of these modules consists of a computational core as a shared library of C++ code and a set of Python modules built on top of the exported API. Parts of the computational core and the graphical user interface of the Image Processing Library and Toolkit IPLT¹³² have been incorporated into OpenStructure to offer versatile handling of image data with support for various algorithms in one, two and three dimensions. A graphics module for real-time rendering of molecules, density maps and molecular surfaces provides functionalities for data visualization.

Processing and visualization of molecular entities often requires filtering by certain selection criteria. These selections are implemented as so-called EntityViews, containing subsets of atoms, residues, chains and bonds of the respective EntityHandle chosen using selection statements (queries). The EntityView class shares a common interface with the EntityHandle class it points to, and hence they can be used interchangeably. This handle/view concept pertains to the full structural hierarchy, i.e. residue views will only contain the atoms that were part of the selection, etc. The query language supports sophisticated selection criteria (for example, distance-based selection, Boolean operators, selections based on user-defined properties, and so on).

In order to infer connectivity and topology when reading molecular coordinate files, we make use of the chemical components dictionary which is part of the official PDB

distribution²⁰. Thus, detailed information is available on any of the chemical components, allowing the framework to ensure correct connectivity and topology during the load process and issue appropriate warnings. The connectivity step is extensible and its behavior can be adapted by overloading functions. Additionally, a heuristic method is available as a fallback for loading unknown residues or to handle non-standard residue and atom names.

3 Application Example

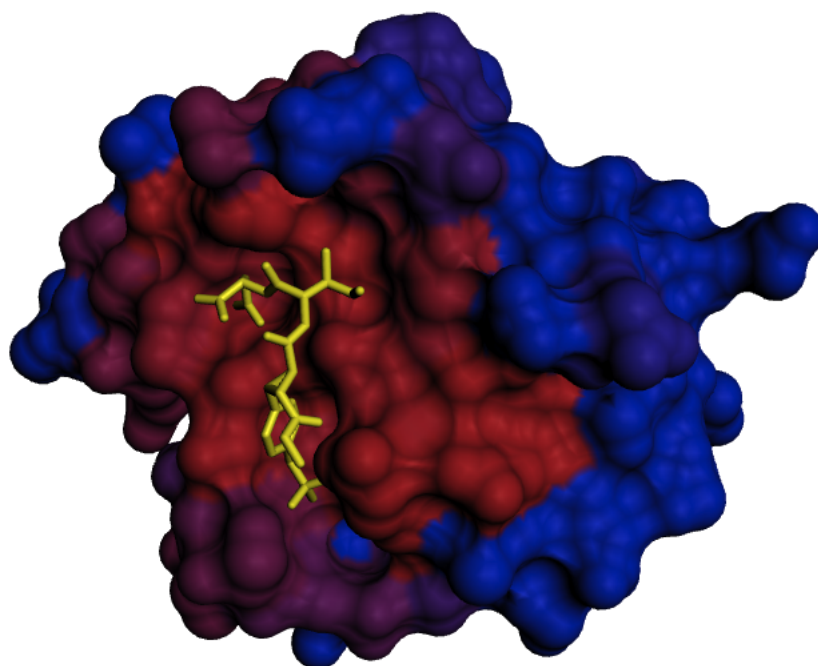


Figure 2.1 Molecular surface representation of a SH2 domain (PDB:3IMJ) colored by conservation of the positions in a multiple sequence alignment. The color scale ranges from red for conserved residues to blue for residues with high variability. The ligand peptide is shown as yellow stick representation. The image was rendered in OpenStructure, the molecular surface was calculated using MSMS¹³³. See Supplementary Table S1 for details on calculation of sequence conservation scores.

Most users will interact with OpenStructure using Python. The code fragment in Supplementary Table S1 illustrates the expressiveness of the OpenStructure API in combining data from different domains. In this example, we compare the sequence conservation of residues in contact with a ligand with the rest of the protein, quantifying the visually derived hypothesis that the binding-site residues of the SH2 domain are more conserved than the rest. This is achieved by mapping of a conservation score derived from a multiple sequence alignment of various SH2 domains ('sh2.aln') onto a representative structure (PDB: 3IMJ)¹³⁴ and identifying residues in direct contact with the ligand. Figure 1 shows

the results displayed in the DNG ('DINO/DeepView Next Generation') graphical user interface, using the conservation score to color a molecular surface representation.

The OpenStructure distribution contains several scripting examples to introduce new users to the functionalities and usage style of the tool kit, such as scripts to animate molecular dynamics trajectories, calculate electron density maps from atomistic structures or rank short peptide fragments according to their correlation with electron density. Exhaustive documentation and tutorials are provided on the web site. Mailing lists for OpenStructure users and developers provide a forum to ask questions, report problems or suggest new developments.

4 Acknowledgement

We would like to thank Andras Aszodi for inspiring discussion during the conception phase of the project, and Pascal Benkert and Tobias Schmidt for critical feedback. OpenStructure uses Eigen (eigen.tuxfamily.org), Boost (www.boost.org), FFTW (www.fftw.org), Qt4 (qt.nokia.com) and PyQt4 (riverbankcomputing.co.uk).

FUNDING: Development of OpenStructure was funded by the SIB Swiss Institute of Bioinformatics and the Biozentrum University of Basel. Implementation of the structure comparison for the Nature PSI SBKB Protein Model Portal based on OpenStructure was supported by the National Institutes of Health as a sub-grant with Rutgers University, under Prime Agreement Award Number: (3U54GM074958-05S2).

CONFLICT OF INTEREST: none declared.

OpenStructure – An Integrated Software Framework for Computational Structural Biology

This chapter has been published as:

Biasini M.^{1,2}, Schmidt T.^{1,2}, Bienert S.^{1,2}, Mariani V.^{1,2}, Studer G.^{1,2}, Haas J.^{1,2}, Johner N.³, Schenk A.D.⁴, Philippsen A.¹ and Schwede T.^{1,2} (2013) *OpenStructure - An Integrated Software Framework for Computational Structural Biology*, *Acta. Cryst.* 69(Pt 5):701-709

¹ Biozentrum, University of Basel, Klingelbergstrasse 50 / 70, 4056 Basel, Switzerland

² SIB Swiss Institute of Bioinformatics, Basel, Switzerland

³ Department of Physiology and biophysics, Weill Medical College of Cornell University, New York, NY 10065, USA.

⁴ Department of Cell Biology, Harvard Medical School, 240 Longwood Ave, Boston, MA 02115, USA

Research projects in structural biology increasingly rely on combinations of heterogeneous sources of information, e.g. evolutionary information from multiple sequence alignments, experimental evidence in the form of density maps, or proximity constraints from proteomics experiments. Previously, we have introduced the OpenStructure software framework, which allows the seamless integration of information of different origin. The software consists of C++ libraries which are fully accessible from the Python programming language. Additionally, the framework provides a sophisticated graphics module to interactively display molecular structures and density maps in three dimensions. In this work, we outline the latest developments in the OpenStructure framework. The extensive capabilities of the framework will be illustrated by using short code examples that show how information from molecular structure coordinates can be combined with sequence data and/or density maps. The framework has been released under the LGPL version 3 license and is available for download from www.openstructure.org.

1 Introduction

In computational structural biology, there is a growing demand for tools operating at the interface of theoretical modeling, X-ray crystallography, electron microscopy, nuclear magnetic resonance, and other sources of information for the spatial arrangement of macromolecular systems^{125,132,135}. Synergy between these fields has led to methods which e.g. combine electron density information with evolutionary information from sequence alignments and structural information from computational models^{79,136–139}. The need to combine heterogeneous data in incompatible formats is often found to be the reason why new methods in computational structural biology rely on custom-made ad hoc combinations of command-line tools built to perform specific tasks. Hence, to facilitate these inconvenient data conversions and to make the development of new methods more efficient, we have developed OpenStructure as a powerful and flexible platform for method development in computational structural biology¹³⁵. This open-source framework provides an expressive API and seamlessly integrates with external tools, e.g. for

structural superposition and comparison^{9,41,140–141}, secondary structure assignment⁷ or homology detection^{29,32,58}. OpenStructure has been consistently extended and its API matured to allow building complex software stacks on top such as homology modeling software, structure comparison methods⁵³ and model quality estimation packages¹⁴².

Since the previous paper on OpenStructure substantial improvements to graphics, performance and the user interface were made, and, support for molecular dynamics trajectories, integration of external software tools and data handling was significantly extended. Here, we first briefly describe the architecture of the OpenStructure framework at the code level. Then, we present the main components of the 1.3 release and individual modules to interact with molecular structures, density maps and sequence data. Code examples will be used to demonstrate the smooth integration of the OpenStructure components.

2 Architecture

OpenStructure was conceived as a scientific programming environment for computational protein structure bioinformatics with reuse of components in mind. The functionality of OpenStructure is divided into modules, dealing with a specific type of data: `mol` and `mol.alg` are concerned with molecular structures and the manipulation thereof. `conop` is mainly concerned with connectivity and topology of molecules. `seq` and `seq.alg` handle sequence data (alignments, single sequences). `img` and `img.alg` implement classes and algorithms for density maps and images. File input and output operations for all data types are collected in the `io` module. `gfx` provides functionality to visualize protein structures, density maps and 3-dimensional primitives. `gui` implements the graphical user interface.

The framework offers three tiers of access, where at the lowest level, the functionality of the framework is implemented as a set of C++ classes and functions, meeting both the requirement for computational efficiency and low memory consumption. The framework makes heavy use of open source libraries, including FFTW for fast Fourier transforms¹⁴³, Eigen for linear algebra¹⁴⁴, and Qt for the graphical user interface.

The middle layer is formed by Python modules, which are amenable to interactive work and scripting. This hybrid compiled/interpreted environment combines the best of both worlds: high performance for compute-intensive algorithms and flexibility when prototyping or developing applications. In fact this approach to multi-language computing has found favor with many in the scientific computing community^{125,145–146} and Python has established itself as the de-facto standard scripting language for scientific frameworks. In addition to general-purpose libraries, e.g. `numpy`¹⁴⁷, `SciPy`¹⁴⁸ and the plotting framework `matplotlib`¹⁴⁹, there are many bioinformatics and structural biology toolkits that are either completely implemented in Python, or provide a well-maintained Python wrapper to their functionality^{121,125,146,150–151}. The combination of general-purpose frameworks with specialized libraries allows developing new algorithms with very little effort.

At the highest level, we offer a graphical user interface with 3D rendering capabilities and controls to manipulate structures or change rendering parameters. The 3 layer architecture is one of the main strengths of OpenStructure and sets it apart from other commonly used tools in computational structure bioinformatics. Rapid prototyping can be done in Python, and if successful, the code can later on be translated to C++ for better performance. Since most of the functions in Python have a C++ counterpart, the Python/C++ adaption is straight-forward and can be completed in a very short time.

3 Molecular Structures

The software module mol implements data structures to work with molecular datasets. At its heart lie the EntityHandle and EntityView classes which represent molecular structures such as proteins, DNA, RNA and small molecules. Other classes deal with molecular surfaces as generated by MSMS¹³³ or other external tools. The EntityHandle class represents a molecular structure. The interface of entities is tailored to biological macromolecules, but this does not prevent it to be used for any kind of molecules: for example, an entity may also represent a ligand or a collection of water molecules - hence the rather generic name. An entity is in general formed by one or more chains of residues. These residues in a chain may be ordered, e.g. in a polypeptide, or unordered, e.g. a collection of ligands. A residue consists of one or more atom. The atoms store atomic position, chemical element type, anisotropic B-factor, occupancy, charge, atom bond list etc. The hierarchy of chains, residues and atoms is arranged in a tree-like structure rooted at the entity (figure 3.1). Atoms of an entity may be connected by bonds, which group the atoms of the entity into one or more connected components.

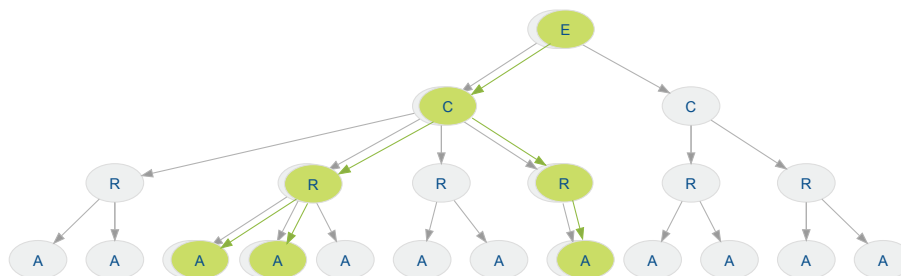


Figure 3.1 Schematic diagram of the components of entity handles and views. The molecular structure is represented in a tree-like structures rooted at the entity (E). The levels of the tree are formed by chain (C), residue (R), and atom (A). In green, an example entity view containing only a selected subset of elements is shown. The hierarchy of the entity view is separate from the handle, however, at every level, the view maps back to its handle giving access to its properties

Working with Subsets of Molecular Structures

Processing and visualization of molecular entities often requires filtering by certain crite-

ria. The result of these filtering operations are modeled as EntityViews (**figure 3.1**), which contain subsets of atoms, residues, chains and bonds of the respective EntityHandle. The entity view references the original data, e.g. modifications to atom positions in the original entity handle are also reflected in the entity view. This handle/view concept pertains to the full structural hierarchy, i.e. residue views will only contain the atoms that were part of the filtering, etc. The EntityView class shares a common interface with the EntityHandle class it points to, and hence they can be used interchangeably in Python. In C++, where type requirements are strict, we employ the visitor pattern¹⁵² to walk through the chain, residue, and atom hierarchy without having to reside to compile-time polymorphism through templates.

The use of entity views throughout the framework makes the implemented algorithms more versatile. For example, the same code to superpose two structures based on C α atoms can be used to superpose the sidechains of a binding site. The only difference is the view and thus the set of atoms that gets passed to the superposition function. These sets of atoms do not need to be consecutive and thus can be used to superpose arbitrary sets of atoms.

The Query Language – Making Selections

Entity views are conveniently created by using a dedicated mini-language. Filtering a structure and returning subsets of atoms, residues, chains and bonds is achieved by predicates which are combined with Boolean logic, often referred to as “selection”. Typical examples include selecting all backbone atoms of arginines, binding site residues, ligands or solvent molecules. Conceptually, the language is similar to selection capabilities of other software packages, e.g. VMD¹²⁷, Coot¹²² or PyMOL¹⁵³.

The predicates may use any of the available built-in properties defined for the atoms, residues, and chains. Examples include the atom name, residue number, chain name, or atom element. A complete list of built-in properties is given in the OpenStructure documentation. In addition, the predicates may refer to user-defined properties declared using generic properties (see below). The within-operator of the query language allows selecting atoms in proximity of another atom or another previously performed selection.

Since selection statements can be applied both to EntityHandles and EntityViews, complex selections can be carried out by chaining selection statements. For rare cases of highly complex selections, the user may assemble the view manually, e.g. by looping over the atoms and including atoms meeting some conditions

Selection Example: Superposition

As an example of how entity views make OpenStructure functions more versatile, we will now consider the binding sites of two heme-containing proteins. We will use the Superpose function of the mol.alg module to calculate rotation and translation operators that superpose the atoms of two structures, first based on the coordinates of the heme ligands and second on the residues binding the heme.

```

# Load two HEM-containing proteins from the PDB website
# into EntityHandles.
prot_one=io.LoadPDB('1mbo', remote=True)
prot_two=io.LoadPDB('1mbn', remote=True)

# Superpose the proteins based on the HEM ligands
mol.alg.Superpose(prot_one.Select('rname=HEM'),
                  prot_two.Select('rname=HEM'))

# Superpose based on HEM-binding residues using the within (<>)
# operator
mol.alg.Superpose(prot_one.Select('rname!=HEM and 3.0 <> [rname=HEM]',
                                  mol.MATCH_RESIDUES),
                  prot_two.Select('rname!=HEM and 3.0 <> [rname=HEM]',
                                  mol.MATCH_RESIDUES))

```

As can be seen, the superposition based on heme atoms or heme-binding residues use the same Superpose function. The only difference is the selection statement to prepare the subset of atoms used to superpose.

Mapping User-defined Properties on Molecular Structures

Many algorithms calculate properties for atoms, residues, chains, bonds or entities. Examples of such properties include sequence conservation of a residue or local structural similarity scores. OpenStructure includes a system to store these properties as key-value pairs in the respective handle classes: the generic properties. Classes deriving from `GenericPropertyContainer` inherit the ability to store properties of string, float, int, and bool type, identified by a key. For each of these data types, methods to retrieve and store values are available both in Python and C++. As with all other built-in properties, the view counterpart will reflect the generic properties of the handle. Since generic properties are implemented at a low-level of the API, they are accessible by the query language, and may e.g. be used for substructure selection or in coloring operations.

Connectivity and Topology

The `conop` module interprets the topology and connectivity of proteins, poly-nucleotides and small molecules. For example, after importing a structure from a PDB entry, bonds between atoms have to be inferred as well as missing information be completed. In addition, the `conop` module provides an infrastructure for consistency checks. OpenStructure supports two conceptually related, yet different approaches for deriving the connectivity information: A rule-based approach that connects atoms based on rules outlined in a database and a heuristic approach which uses a distance-based heuristic.

The rule-based approach to connectivity derivation is based on a set of rules that define the bonding partners for each atom based on its name. The rules are extracted from

the chemical component dictionary provided by the wwPDB²⁰ and stored in a compound library. Since this library has knowledge of all residues deposited in the PDB, deviations from the rules are easily detected and may be reported back to the user. For automatic processing pipelines that operate on large set of structures, strict settings when loading a structure are advised to limit surprises.

For structures from other sources, including molecular dynamic simulations and virtual screening studies with loosely defined naming conventions, the heuristic approach might be more appropriate. The heuristic builder uses lookup tables for the connectivity of standard nucleotides and standard amino acids, but falls back to a distance-based connection routine for unknown residues or additional atoms present in the structure. In contrast to the rule-based approach outlined above, the heuristic builder is meant to be used as a quick and dirty connectivity algorithm when working with structures interactively.

Loading and Saving Molecular Structures

OpenStructure contains the io module for importing and exporting structures from and to various file formats such as PDB, CRD, PQR. In the following, reading of molecular structures and molecular dynamics trajectory files is described in more detail.

File input is concerned with data from external sources. As such, importers are exposed to files of varying quality. For automated processing scripts, it is crucial to detect non-conforming files during the import, as every non-conforming file is a potential source for errors. For visualization purposes and interactive work on the other hand, one would like files to load, even if they are not completely conforming to standards. To account for these two different scenarios, OpenStructure introduced IO profiles in version 1.1. A profile aggregates flags that fine-tune the behavior of both the io and conop modules during import of molecular structures. The currently active IO profile controls the behavior of the importer upon encountering an issue. By default, the import aborts upon encountering a non-conforming file. This strict profile has been shown to work well for files from the wwPDB archive. Many files that could not be loaded using the strict settings exposed actual problems in the deposited files. These issues have been reported and resolved in the meantime by the wwPDB.

Molecular dynamics simulations generate a series of coordinate snapshot of the simulated molecule. These snapshots are often stored in binary files. OpenStructure supports reading of CHARMM formatted DCD files in two different ways: First, the whole trajectory may be loaded into memory. This is the recommended behavior for small, pre-processed trajectories. However, since trajectories may well be larger than the available RAM, loading the complete trajectory is not always an option. The second alternative is loading only a set of frames into memory. The remaining frames are transparently fetched from disk, when required. This allows to efficiently processing very large trajectories without consuming huge amounts of memory.

4 Sequences and Alignments

Since sequence and structure of a protein are intrinsically linked, scientific questions in computational structural biology often require the combination of structure and sequence data. In fact, for many applications, methods based on evolutionary information considerably outperform physics-based approaches^{53,118}. Thus, efficient and convenient mapping between sequence information and structural features of a protein is crucial.

In OpenStructure, the functionality for working with sequences, and the integration with structure data, is implemented in the seq module. The principal classes, SequenceHandle, AlignmentHandle and SequenceList represent the three most common types of sequence data. Instances of SequenceHandle hold a single, possibly gapped, nucleotide or protein sequence. These instances serve as a container for the raw one-letter-code sequence with additional methods geared towards common sequence manipulation tasks. The SequenceList is suited for lists of sequences, e.g. sequences resulting from a database search using BLAST⁵⁸. An AlignmentHandle holds a list of sequences, which are related by a multiple sequence alignment. The interface for alignments is focused on column-wise manipulation, e.g. insertion or removal of blocks or single columns. Importing alignments and sequences is supported for the FASTA, ClustalW or PIR formats, while exporting of sequence related data is implemented for FASTA and PIR formats.

Efficient Mapping of Structure and Sequence-based Information

The combination of structure and sequence information is embedded into the core of the sequence handle class. A structure may be linked to its matching amino acid sequence by simply attaching it, defining a relation between information associated with residues in the structure and information related to residues in the sequence. To determine the index of the residue in the protein sequence at the *n*th position in the alignment, the number of gaps prior to *n* needs to be subtracted. A naive mapping implementation counting the number of gaps prior to position *n* would scale linearly with *n*, which is suboptimal for long sequences. For efficiency, the sequence handle maintains a list of all gaps present in the sequence. Instead of traversing the complete sequence, traversal of the gap list yields the number of gaps before a certain position. Since the number of gaps is usually much smaller than the sequence length, a more efficient run time is thus observed when mapping between residue index and position in the alignment.

Algorithms for Sequences and Alignments

The seq.alg module contains several general-purpose sequence algorithms. To align two sequences using a local or global scoring scheme, the Smith-Waterman¹⁵⁴ and Needleman-Wunsch¹⁵⁵ dynamic programming algorithms have been implemented. Conservation of

columns in an alignment may be calculated with a variation of the algorithm from ConSurf¹⁵⁶, which considers the pairwise physico-chemical similarity of residues in each alignment column (for an example, see Biasini *et al.*¹³⁵). More sophisticated sequence and alignment algorithms are available through one of the available interfaces to external sequence search tools such as BLAST^{29,58}, ClustalW¹⁵⁷, kClust (A. Hauser, unpublished), or HHsearch³².

Example: Ligand Binding Site Annotation

The following example illustrates how annotation on ligand-interacting residues for a protein may be automatically inferred from a related protein structure.

Dengue fever is a neglected tropical disease, caused by a positive-sense RNA virus which contains a type-1 cap structure at its 5' end. The dengue virus methyltransferase is responsible for cap formation and is essential for viral replication¹⁵⁸. Thus, it is an attractive drug target. Four closely related dengue virus serotypes (DENV1-4) have been isolated, where each serotype is sufficiently different, that no cross-protection occurs¹⁵⁹. The structure of DENV2 methyltransferase (PDB-ID:1r6a) binds S-adenosyl-L-homocysteine (SAH) and ribavirin monophosphate (RVP) in two distinct binding sites. RVP is a weak inhibitor of the enzyme's activity (Benarroch *et al.*, 2004). In the structure of DENV3 methyltransferase (PDB-ID: 3p97) only the SAH binding site is occupied. We would now like to identify which residues in the second structure potentially interact with RVP. Since the two structures share sequence identity of 77% to each other, the two sequences can be aligned with high confidence using a pairwise sequence alignment algorithm. Using the mapping defined by the sequence alignment, we then transfer the ligand binding site information from the first to the second structure.


```

# load the two structures
structure_wi_lig = io.LoadPDB('data/structure-with-ligand.pdb')
structure_no_lig = io.LoadPDB('data/structure-without-ligand.pdb')

# set a generic property for all residues of the first structure in
# contact with RVP
for res in structure_wi_lig.Select('4.0 <> [rname=RVP]').residues:
    res.SetBoolProp('close_to_ligand', True)

# get the sequence for both structures
ligand_seq = seq.SequenceFromChain('ligand', structure_wi_lig.chains[0])
no_lig_seq = seq.SequenceFromChain('no lig', structure_no_lig.chains[0])

# align the two sequences using a global alignment algorithm with the
# BLOSUM62 substitution matrix and default gap extension and opening
# penalties. Global align returns a list of global alignments, but we
# only use the first one...
aln = seq.alg.GlobalAlign(ligand_seq, no_lig_seq, seq.alg.BLOSUM62)[0]

# print the alignment, to check that the alignment is reasonable
print aln

# the alignment essentially defines a mapping of residues in the first
# and the second structure. We will use GetMatchingBackboneViews to
# obtain two entity views which contain the corresponding residues
aln_wi_lig, aln_no_lig = aln.GetMatchingBackboneViews()

# iterate over the residue pairs and print residue names of residues
# which are part of the "predicted" binding site
print 'predicted binding site'
for lig_res, no_lig_res in zip(aln_wi_lig.residues, aln_no_lig.residues):
    if lig_res.HasProp('close_to_ligand'):
        print no_lig_res

```

This example illustrates how little effort it takes to map between information contained in two distinct structures. The results are visualized in **figure 3.3B**. Often, useful scripts can be built with only a few lines of descriptive OpenStructure Python code.

5 Density Maps and Images

The majority of available experimental protein structures have been determined using X-ray crystallography. This technique produces density maps into which an atomistic or semi-atomistic model is built. For high resolution structures, model building into density maps is completely automated^{146,160}. However, for low resolution, automated approaches usually fail and manual intervention is required. As shown repeatedly, the integration of theoretical modeling techniques is often able to improve the built models^{136,138}. The theoretical modeling field on the other hand can profit from the availability of density maps, even at low resolution to refine homology models.

To provide efficient and convenient access to density data, OpenStructure includes the `img` and `img.alg` modules. The core functionality of these two modules has initially been developed as part of the Image Processing Library and Toolkit (IPLT)^{132,161–162}. The IPLT package implements a complete processing pipeline to obtain density maps from recorded electron micrographs. As part of a joint effort to lower the maintenance burden for the two packages, the core data structures and algorithms of IPLT have been moved to OpenStructure. The two modules offer extensive processing capabilities for 1-, 2- and 3-dimensional image data. In this module, electron density maps are considered as 3D images and hence, the terms image and density map are used interchangeably.

The principal class of the image processing capabilities is the `ImageHandle`. It provides an abstraction on top of the raw pixel buffers and keeps track of pixel sampling, dimension and data domain. An `ImageHandle` can store an image either in real or reciprocal space. The image is aware of the currently active domain. This means, for example that one can apply a Fourier transformation (FT) to an `ImageHandle` containing a spatial image and the image will correctly identify the new active domain as frequency. The `ImageHandle` also supports applying a FT to an image with conjugate symmetry, resulting in a real spatial image, while applying a FT to a non-centrosymmetric one results in a complex spatial image.

Image and density data may be imported and exported from and to PNG, TIFF, JPK, CCP4, MRC, DM3 and DX files. Standard processing capabilities for images are provided in the `img.alg` module. This module contains filters, e.g. low- and high-pass filters, masking algorithms and algorithms to apply a Gaussian blur to an image. Additionally, the module contains algorithms to calculate density maps from molecular structures, either in real-space or Fourier space¹³⁶, which we will use in the following example.

Correlating Backbone Fragments with local Electron Density

We would like to illustrate the combined use of density maps and structure data in OpenStructure in the following paragraph. As an example, consider a protein structure where a segment of six residues has not been resolved. However, close inspection of the density map reveals that there is substantial experimental evidence to connect the two parts of the protein chain. We would now like to rebuild the missing part of the backbone. Possible conformations are sampled from a database of structurally non-redundant fragments compiled from the PDB. For scoring, the density for the fragment is calculated by placing a Gaussian sphere on the position of every atom. The resulting density map is then compared to the experimental density with real-spatial cross-correlation. **Figure 3.2** shows a few selected backbone conformations colored by correlation to the density map.

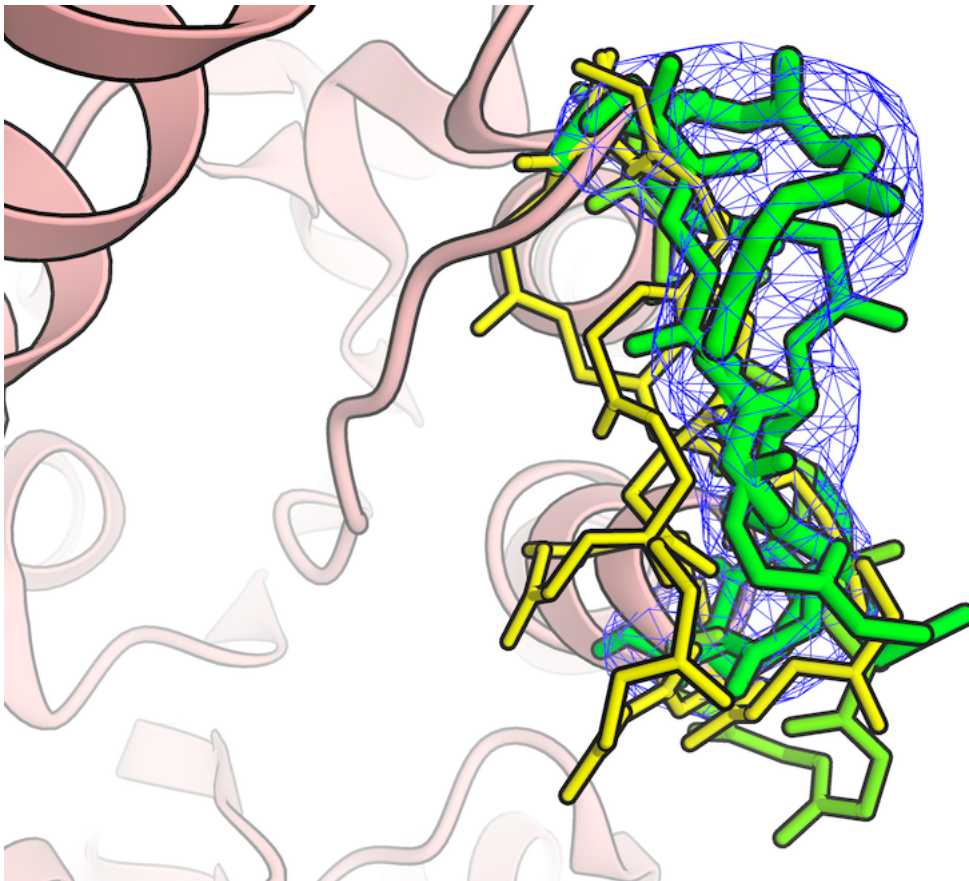


Figure 3.2 A selection of possible backbone conformations to bridge a fragmented chain. The fragments are colored by correlation with the density from yellow to green. The tube thickness to render the backbone fragment is scaled according to the density correlation.

```

# convert the candidate loop into a density and calculate the real-
# spatial
# cross correlation with the actual density. The correlation
# coefficient is stored as the generic property "correl".
# which could be later used to colour the loops

for index, candidate in enumerate(candidates):
    EntityToDensityRosetta(candidate.CreateFullView(), cmap,
                           HIGH_RESOLUTION, 5.0, True)

    correl=img_alg.RealSpatialCrossCorrelation(dmap, cmap, dmap.GetExtent())
    candidate.SetFloatProp('correl', correl)

```

Visualization

Solutions to challenging scientific and algorithmic problems often become obvious after an appropriate form to display the information has been found. Readily available visual-

ization tools are an enabling factor, both for science and algorithm development. OpenStructure features sophisticated visualization capabilities as part of the gfx module. The rendering engine is capable of producing publication-ready graphics. It has been used for the visualization of very long molecular dynamic simulations^{163–164}.

Each principal class of the mol and img modules has a renderer (graphical object) in the graphics module, responsible for turning the abstract data into a 3-dimensional rendering, supporting displaying of molecular structures, surfaces and density data. The separation of graphical objects from their corresponding counterparts keeps the orthogonal concepts of display and general manipulation/querying of structural data separate and saves memory when no visualization is required. Graphical objects are organized by the scene, a scene-graph like object. The scene manages the currently active graphical objects and is responsible for rendering them. In addition, the scene manages rendering parameters, such as light, fog, clipping planes, and camera position.

The rendering engine has been implemented with OpenGL. Typically, each of the graphical objects calculates the geometry, i.e. the vertices and triangle indices, once, and stores it in vertex buffers. Since the geometry of most objects does not change with every frame, storing the geometry allows for more efficient rendering of large structures. If possible, the vertex buffers are transferred to the video memory of the graphics cards to save round-trip time of sending the geometry over the system bus. For advanced effects, the gfx module uses the OpenGL shading language (GLSL). The fixed-pipeline shaders of OpenGL are replaced by custom shaders, which implement special lighting models, e.g. cartoon or hemi-light shading, shadows or ambient occlusion effects.

Figure 3.3 contains two images generated with the graphics module of OpenStructure. The scripts to generate these images are contained in supplementary materials. **Figure 3.3A** is inspired by a recent analysis of modeling performance within the Continuous Automated Model Evaluation assessment framework (<http://www.cameo3d.org>; CAMEO). The target structure is shown in tube representation (white color, larger tube radius) together with 3 theoretical models (thin tubes). The models are colored in a traffic-light gradient from red to yellow to green using a superposition-free all-atom structural similarity measure called the local distance difference test (LDDT)⁵³. The combination of outline render mode with hemi-light shading gives this image a very clear style. **Figure 3.3B** shows the structures of the methyltransferase of two different dengue virus serotypes as described in the example “Ligand Binding Site Annotation”. On top, the enzyme is in complex with the inhibitor ribavarin monophosphate whereas at the bottom, no ligand is present in this binding pocket. The enzyme is represented by its molecular surface as calculated by MSMS¹³³ and the inhibitor is shown in sticks representation. The surface of observed (top) or predicted residues (bottom) interacting with the ligand are highlighted in blue or red, respectively.

Visual Data Exploration Example: Proteomics Cross-Links

The following example illustrates how visualization of structure-based predictions can help to rationalize the planning of proteomics crosslinking experiments. Large macro-

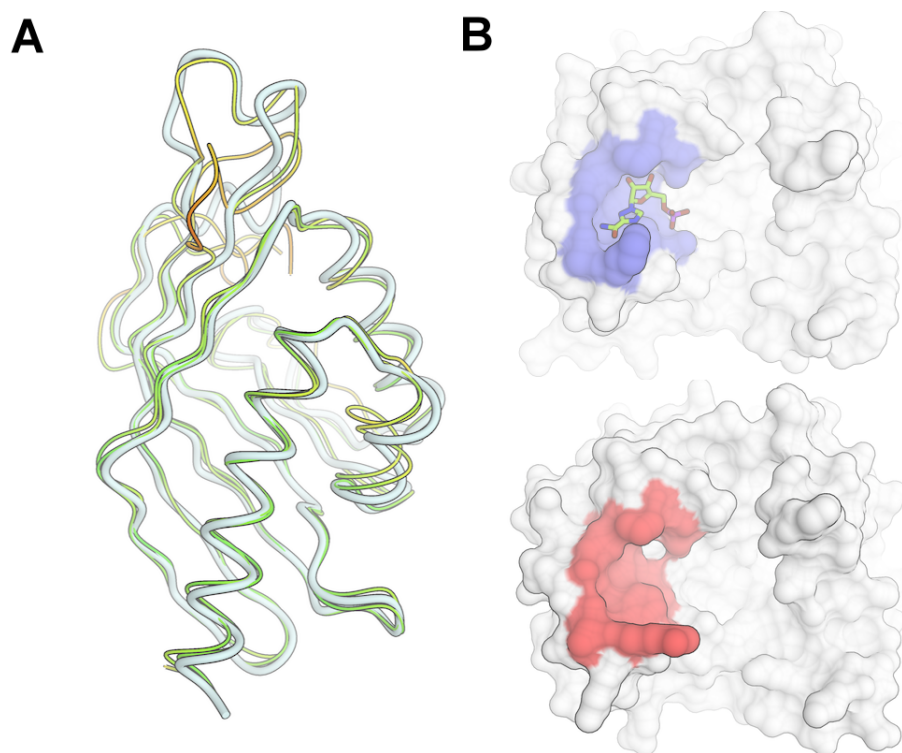


Figure 3.3 Two distinct visualization styles illustrating the graphical capabilities (see text for more detailed description); Panel A: hemi-light shading with outline mode; Panel B: simplified enzyme representation by its molecular surface together with an inhibitor.

molecular structures are difficult to crystallize and often only diffract to limited resolution where it is unfeasible to determine the structure at atomic detail. It is thus common practice to solve the structure of individual components separately and use other experimental techniques to identify the relative orientations of the components. Proteomics cross-links are one such experimental technique¹⁶⁵ where isotope-labeled cross-linkers such as Disuccinimidyl Glutarate (DSG) or Disuccinimidyl suberate (DSS) are added to the sample. The cross-linking reaction chemically connects primary amines, i.e. terminal amines of lysine side chains, which are in close proximity. After protein digestion with Trypsin cross-linked fragments are then identified using mass spectrometry.

Urease of *Y. enterocolitica* is a large oligomeric complex, vital to the pathogenicity of the bacteria. The enzyme catalyzes the cleavage of urease to ammonia at the expense of protons to reduce the acidity during its passage through the stomach. To investigate which cross-links are theoretically possible for this protein, we have built a homology model based on the X-ray structure of the urease from *H. pylori* (PDB ID: 1e9y)¹⁶⁶, sharing 50% sequence identity. Possible cross-linking sites have then been identified using Xwalk¹⁶⁷. Visualizing the cross-links by connecting the lysine atoms with a straight line does not lead to conclusive results as the straight lines pass through the protein. To overcome this visualization problem, we have used OpenStructure to simulate the cross-links as strings of beads. By introducing a force that drives the beads away from the center of the protein, their positions are optimized. The cross-links appear as red loops sticking

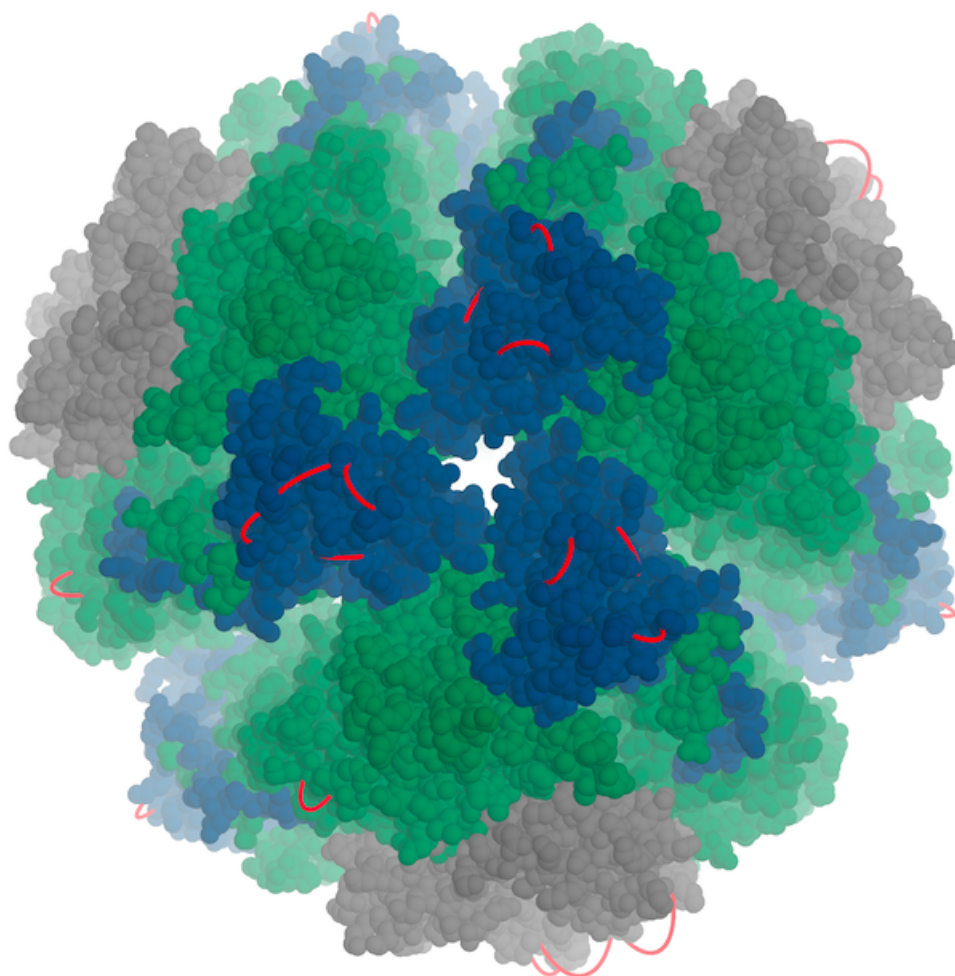


Figure 3.4 Visualization of predicted cross-link locations in a homology model of the urease from *Y. enterocolitica*. The subunits of the urease are colored in blue (alpha subunits), green (beta subunits) and grey (gamma subunits).

out from the surface of the protein. The resulting image (**figure 3.4**) of proteomics cross-links is visually appealing and easily conveys the message that all connections represent intra-, not inter-chain cross-links.

Efficient visualization of the expected outcome allows planning experiments effectively – in this case indicating that experimental proteomics crosslink data will not contain sufficient information to determine the relative orientation and stoichiometry of the components of the urease complex. The OpenStructure script to generate the example is given in supplementary material.

6 Graphical User Interface

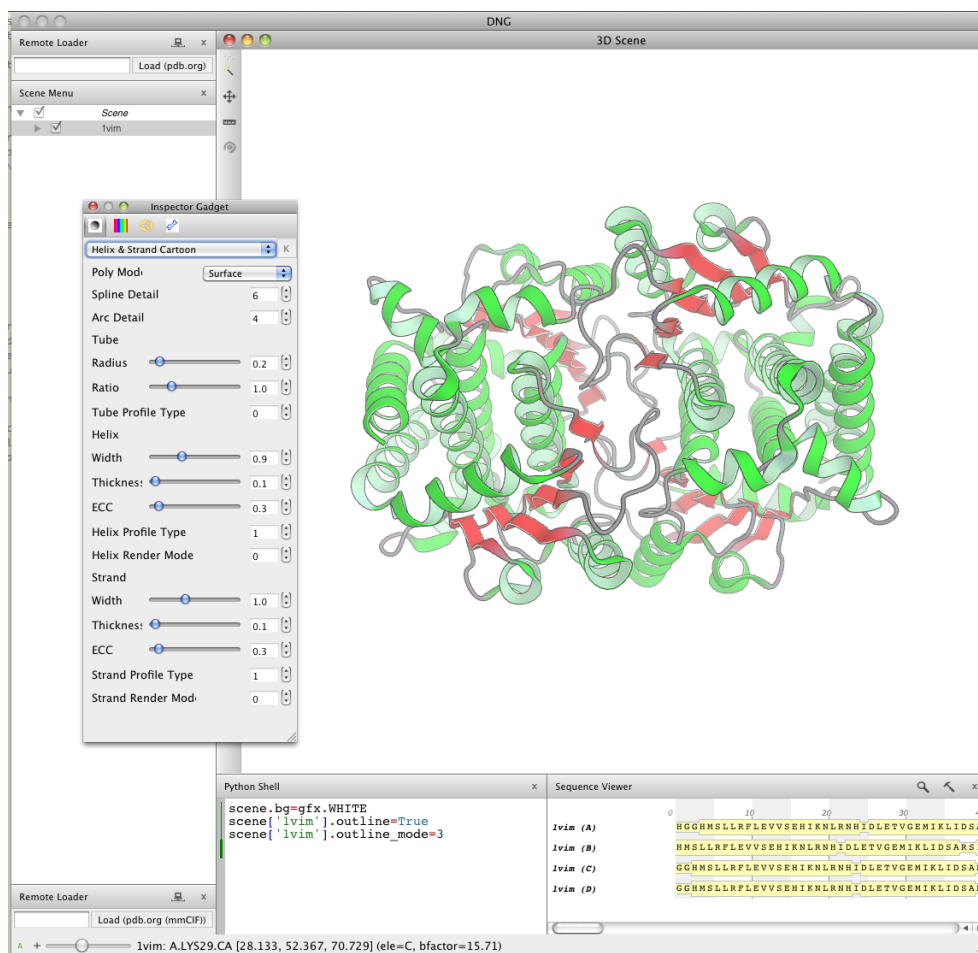


Figure 3.5 Screenshot of the graphical user interface DNG. Controls for data display are organized in a main application window. The majority of the main window by default is taken up by the 3D scene window, which shows a structure rendered in ribbon mode. The user interacts with the scene using the mouse and keyboard shortcuts. On the left side, the currently loaded graphical objects are shown in the scene, a tree view that reflects the structure in the scene graph. The render parameters of graphical objects may be changed with the inspector widget displayed on top of the 3D window. In the bottom-right corner, the sequences of the loaded proteins are shown.

For interactive work, we have developed the graphical user interface DINO/DeepView⁴² Next Generation (DNG). The graphical user interface builds on the visualization and data processing capabilities of the OpenStructure framework and provides controls to interact with macromolecular structures, sequence data and density maps (figure 3.5). A central part of DNG is the Python shell, which allows efficient prototyping and interaction with the loaded data at runtime. Objects may be queried, modified and displayed by using the OpenStructure API. For convenience, the shell supports tab-completion and multi-block editing: complete functions and loops may be pasted into the Python shell.

7 Conclusions

OpenStructure is a software framework tailored towards computational structural biology. Its modular and layered architecture make it an ideal platform for hypothesis driven research and method development, particularly, when density maps, molecular structures and sequence data are to be combined. Together with powerful visualization capabilities, the expressive API allows to implement new algorithms in a very short time. Additionally, through a variety of bindings, third-party applications can be included into the scripts, without worrying about input and output file formats.

OpenStructure has been successfully used as analysis and development platform in several recently published research projects, e.g. QMEAN¹⁴², the local difference distance test⁵³, the identification of two-histidines one-carboxylate binding motifs in proteins amenable to facial coordination to metals^{168–169}, the evaluation of template-based modeling⁵³, the assessment of ligand binding site prediction servers¹⁶⁹ and visualization of very long molecular dynamic simulations^{163–164}.

Local Distance Difference Test - A Robust, Superposition-Free Protein Structure Similarity Measure

This chapter contains a paper manuscript currently in press for Bioinformatics in a revised form. This work has been done in collaboration with Dr. Valerio Mariani, and Dr. Alessandro Barbato. The contributions of VM, AB, TS, and MB were as follows: MB had the idea for the IDDT score, and provided the first implementation used during CASP9. VM implemented the stereo-chemical checks and performed the GDT to IDDT comparison on the CASP9 data. VM and MB implemented the multi-reference code, AB and VM implemented the web-server, MB, VM and TS wrote the manuscript

Mariani, V.*^{1,2}, Biasini, M.*^{1,2}, Barbato, A.^{1,2} and Schwede T.^{1,2} (2013). *The local distance difference test: A robust superposition-free protein structure similarity measure*. Bioinformatics, in press; doi:10.1093/bioinformatics/btt473.

¹ Biozentrum, University of Basel, Klingelbergstrasse 50 / 70, 4056 Basel, Switzerland

² SIB Swiss Institute of Bioinformatics, Basel, Switzerland

* equal contributions

1 Introduction

The knowledge of a protein's three-dimensional structure enables a wide spectrum of techniques in molecular biology, ranging from rational design of mutagenesis experiments to elucidation of a protein's function and to drug design. While the rapid development of DNA sequencing techniques has been providing researchers with a wealth of genomic data, structural biology has not been able to keep the same pace in complementing sequence information with structural data, and the gap between the number of known protein sequences and the number of known protein structures has been growing continuously^{42,50,170–171}. In order to fill this gap, various computational approaches have been developed to predict a protein's structure starting from its amino-acid sequence.

Despite remarkable progress in structure prediction methods, computational models often fall short in accuracy compared to experimental structures. The bi-annual CASP experiment (Critical Assessment of techniques for protein Structure Prediction) provides an independent blind retrospective assessment of the performance of different modeling methods based on the same set of target proteins^{50,172}.

One of the main challenges for the CASP assessors is to define appropriate numerical measures to quantify the accuracy with which a prediction approximates the experimentally determined structure. In the course of the CASP experiment, model comparison techniques have evolved to reflect the current state of the art of prediction techniques: In the first installments of CASP, root mean square deviation (RMSD) between a prediction and the superposed reference structures was used in various forms as the main evaluation criterion^{8,45,173–174}. However, RMSD has several characteristics which limit

its usefulness for structure prediction assessment: the score is dominated by outliers in poorly predicted regions while at the same time it is insensitive to missing parts of the model, and it strongly depends on the superposition of the model with the reference structure.

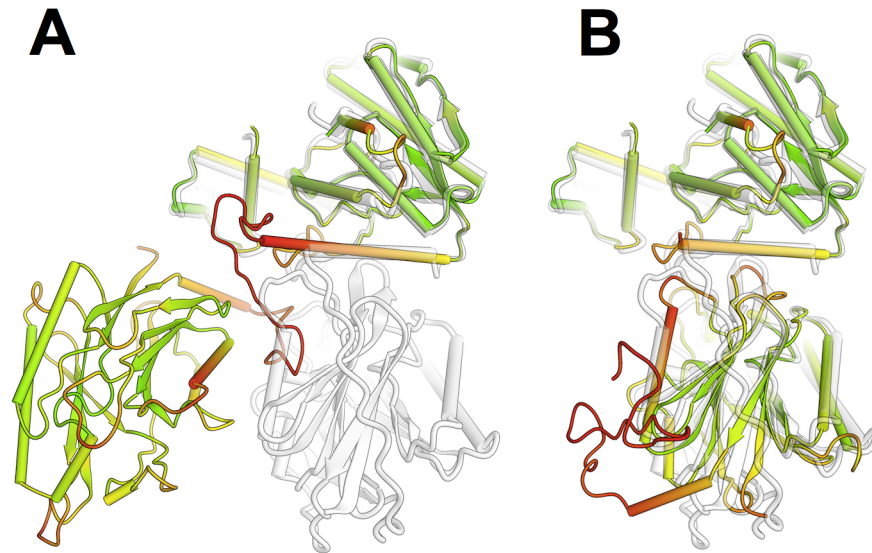


Figure 4.1 Comparison of predicted protein structure model with its reference structure for CASP target T0542. The target structure (shown in grey) consists of two domains. In panel A, a predicted model (TS236, in color) is shown in full length, with the first domain superposed to the target. For graphical illustration, panel B shows the two domains in the prediction separated according to CASP assessment units and superposed individually to the target structure. In both panels, the model is colored according to full-length IDDT scores following a traffic-light-like Red-Yellow-Green gradient, with a red color corresponding to low values of the local IDDT, a green color corresponding to high values, and a yellow color to average values. As superposition-free method, IDDT is insensitive to relative domain orientation and correctly identifies segments in the full length model deviating from the reference structure.

In order to overcome some of the limitations of RMSD in the context of CASP, the Global Distance Test (GDT) was introduced in CASP4^{41,47}. In contrast to RMSD, the GDT is an agreement-based measure, quantifying the number of corresponding atoms in the model that can be superposed within a set of predefined tolerance thresholds to the reference structure. For each threshold, different superpositions are evaluated and the one giving the highest number is selected. The final GDT score is then calculated as the average fraction of atoms that can be superposed over a set of predefined thresholds (0.5, 1, 2 and 4 Å for GDT_HA and 1, 2, 4, 8 Å for GDT_TS, respectively). One of the advantages of GDT is that outliers do not considerably influence the score, while missing segments in the predictions are taken into account. Besides of GDT, several other scores for model comparison have been developed to overcome the limitations of RMSD^{175–177}.

One of the main limitations of measures based on global superposition becomes evident when applied to flexible proteins composed of several domains. The relative orientation of the domains can naturally change. Typically, the global rigid-body superposition is dominated by the largest domain, and as a consequence the smaller domains are not

correctly matched. Artificially high scores are the result. In CASP, the effects of domain movement are reduced by splitting the target into so-called “assessment units” (AU) that are evaluated separately. The definition of assessment units is carried out by visual inspection, and is therefore time consuming. Furthermore, the criteria used to define the AU are often subjective^{178–179}. Grishin and coworkers have proposed an approach to numerically support this decision by analyzing the variability among the predictions for a specific target¹⁷⁹.

Local superposition-free measures based on rotation-invariant properties of a structure are an attractive alternative to overcome several of the shortcomings outlined before. For example, dRMSD - the distance-based equivalent of RMSD - is widely used in cheminformatics to assess differences in ligand poses in binding sites¹⁸⁰. In CASP9, the IDDT score was introduced, assessing how well local atomic interactions in the reference protein structure are reproduced in the prediction⁵³. More recently, other non-superposition-based scores have been proposed, for example CAD score, which is based on residue-residue contact areas as opposed to interatomic distances¹⁴¹.

Initially, most of the scores used in structure prediction assessment aimed at the evaluation of the protein backbone or fold, thereby focusing on C α atom positions. However, with increasing accuracy of prediction methods for template based models, the focus of the assessment has shifted to the evaluation of the atomic details of a model. In CASP7, the first scores based on local atomic interactions were introduced in the form of HBscore, which quantifies the fraction of hydrogen bond interactions in the target protein correctly reproduced in the model^{51,181}. In CASP8, several scores for assessing the local modeling quality were introduced (main chain reality score, hydrogen bond correctness, rotamer correctness, side chain positioning)⁵², as well as an evaluation of the stereo-chemical realism and plausibility of models using the MolProbity score¹⁸². The IDDT score introduced in CASP9 also considers all atoms of a prediction - including all side chain atoms, thereby capturing the accuracy of, for example, the local geometry in a binding site, or the correct packing of a protein’s core.

In this manuscript, we expand the initial concept of local Distance Difference Test (IDDT). We discuss its properties with respect to its low sensitivity to domain movements, and the significance that can be assigned to the absolute score values. Furthermore, we introduce the concept of using multiple reference structures simultaneously, and incorporate stereo-chemical quality checks in its calculation. We finally illustrate how IDDT can be used to highlight regions of low model quality, even in models of multi-domain proteins where domain movements are present.

2 Methods

The Local Distance Difference Test

The local Distance Test (IDDT) score measures how well local interactions in a reference structure are reproduced in a protein model. It is computed over all pairs of atoms in the

reference structure at a distance closer than a predefined threshold R_0 (called inclusion radius), and not belonging to the same residue. These atom pairs define a set of local distances L . A distance is considered preserved in the model M if its length, within a certain tolerance threshold, is the same as the one of the corresponding distance in L . If one or both the atoms defining a distance in the set are not present in M , the distance is considered non-preserved. The fraction of preserved distances as a function of the threshold γ can be expressed as:

$$C(\gamma) = \frac{\sum_{d \in L} c(d, \gamma)}{|L|}$$

Where d is a distance belonging to L , $c(d, g)$ has a value of 1 when the distance d is preserved and 0 when it is not. $|L|$ is the number of elements in L . The final IDDT score is the average of four C values computed using the following thresholds: 0.5 Å, 1 Å, 2 Å, and 4 Å, the same ones used to compute the GDT_HA score^{41,51}:

$$lDDT = \frac{1}{4} [C(0.5) + C(1) + C(2) + C(4)]$$

For partially symmetric residues, where the naming of chemically equivalent atoms can be ambiguous (Glutamic Acid, Aspartic Acid, Valine, Tyrosine, Leucine and Arginine), two local Distance Difference Tests, one for each of the two possible naming schemes, are computed using all non-ambiguous atoms in M as a reference. The naming convention giving the higher score in each case is used for the calculation of the final structure-wide IDDT score.

The IDDT score can be computed using all atoms in the prediction (the default choice), but also using only distances between carbon α atoms, or between backbone atoms. Interactions between adjacent residues can be excluded by specifying a minimum sequence separation parameter. Unless explicitly specified, the calculation of the IDDT scores for all experiments described in this manuscript has been performed using default parameters, i.e. $R_0 = 15\text{\AA}$ using all atoms at zero sequence separation).

Multi-reference Local Distance Difference Test

The local Distance Difference Test can be computed simultaneously against multiple reference structures of the same protein at the same time. The set of reference distances L includes all pairs of corresponding atoms which, in all reference structures lie at a distance closer than the reference threshold R_0 . For each atom pair, the minimum and the maximum distances observed across all the reference structures are compared with the distance between the corresponding atoms in the model M being evaluated. The distance is considered preserved if its length in M falls within the interval defined by the minimum and the maximum reference distances, or if it lies outside of the interval by less than a predefined length threshold. If any of the atoms defining the distance is not present in M , the distance is considered not preserved. The fraction of preserved distances is computed like in the single reference case.

Structure Quality Checks

In order to account for stereo-chemical quality and physical plausibility of the model being evaluated, the calculation of the local Distance Difference test can take violations of structure quality parameters into account. Based on high resolution experimental structures, Engh and Huber have compiled a set of reference average bond lengths and planar angle widths for all standard amino-acids, together with typical standard deviations of their measurements^{84,183}. Here, stereo-chemical violations in the model are defined as bond lengths and angles with values which diverge from the respective average reference value by more than a predefined number of standard deviations (12σ by default). Interatomic distances between pairs of non-bonded atoms in the model are considered clashing if the distance between them is smaller than the sum of their corresponding atomic van der Waals radii¹⁸⁴, within a predefined tolerance threshold (by default 1.5 Å). Tolerance thresholds can be defined for each pair of atomic elements independently.

In case where the side-chain atoms of a residue show stereo-chemical violations or steric clashes, all distances that include any side-chain atom of this residue are considered as not preserved for the IDDT calculation. In case the back-bone atoms are involved stereo-chemical violations or steric clashes, all distances that include any atom of the residue are treated as not preserved.

Determination of the Optimal Inclusion Radius

To determine the optimum value of the inclusion radius parameter R_0 for IDDT, an analysis of all predictions of multi-domain targets evaluated during the CASP9 experiment^{53,179} was carried out (see Table-S1 in supplementary materials for a complete list). GDC-all scores for all predictions were computed based on the Assessment Units (AU) definitions by the CASP9 assessors¹⁷⁹. A weighted whole target GDC-all score was computed for each target as the average GDC-all scores of its AUs weighted by the AU size. GDC-all scores are an all-atom version of GDT with thresholds from 0.5 to 10 in steps of 0.5 Å. GDC-all scores were computed using LGA version 5/2009⁴¹, using the following parameters: -3 -ie -d:4 -sda -swap -o1.

IDDT scores were calculated for the whole targets by including all residues which are covered by any AU. The inclusion radii parameter was varied in the range from 2 Å to 40 Å, and the correlation R^2 score between the distribution of weighted averaged GDC-all scores and the distribution of IDDT scores was computed, and plotted against the value of the inclusion radius (**Figure 4.2** and **Figure 4.3**).

Validation of Baseline Scores for Different Folds

To analyze the influence of the protein fold of the assessed structure on the IDDT score, pseudo-random models were created for different Architectures in the CATH Protein Structure Classification system¹⁸⁵ using the following procedure: Representative domains

longer than 50 residues were selected as evenly as possible amongst the lower level (Topologies) of the CATH classification. For each domain, side chain coordinates were removed and then rebuilt using the SCWRL software package (with default parameters)⁷⁸. Pseudo-random models representing threading errors were then generated by shifting all residues by one alignment position in a backbone only model, and rebuilding the side-chains with SCWRL4, and computing the corresponding IDDT score. This method is loosely based on the approach described in¹⁸⁶. The procedure was repeated iteratively until a threading error of 50 residue positions was reached. Here, we present the results for CATH Architecture entries 1.25 (Alpha Horseshoe) and 2.40 (Beta-barrel), each represented by 60 example structures (**Figure 4.3**).

For estimating IDDT scores of random protein pairs, 200 protein models with wrong fold were generated by selecting pairs of structures with different CATH topologies, generating models by rebuilding side chains on the backbone of the other protein, and computing IDDT scores for these decoy models. The median of the resulting distribution was 0.20 with a 0.04 mean absolute deviation.

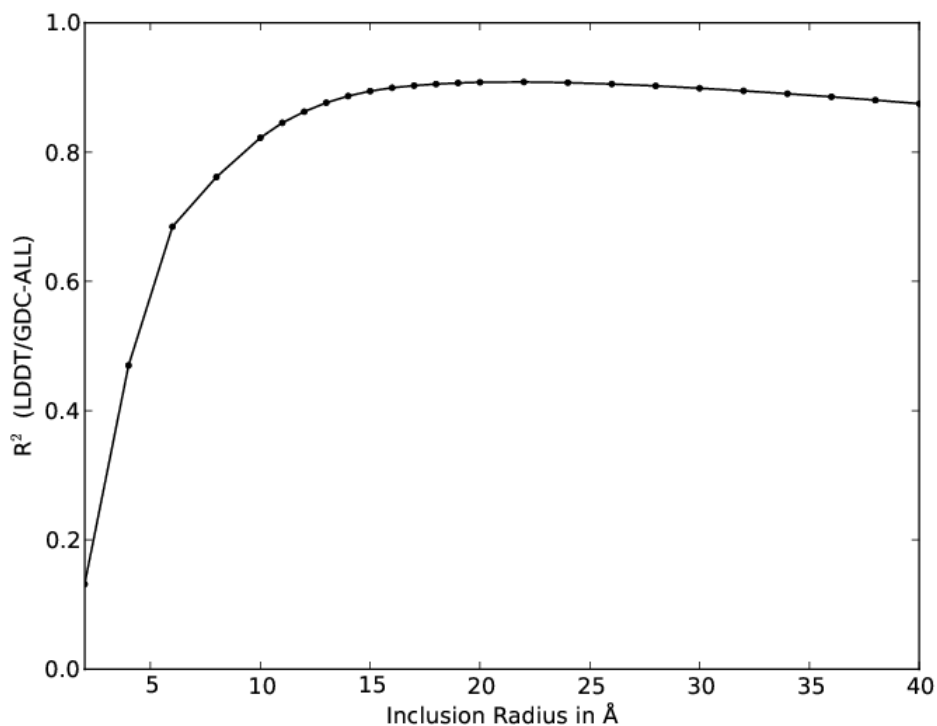


Figure 4.2 Determination of the optimal inclusion radius parameter R_0 . Pearson correlation (R^2) between whole target IDDT scores vs. domain-based weight-averaged GDC-all scores for different values of the inclusion radius parameter R_0 were computed over all CASP9 predictions for multi-domain targets.

Implementation and Availability

IDDT has been implemented using the OpenStructure framework¹³⁵. Source code, stand-alone binaries for Linux and Mac OSX, as well as an interactive web server are available at <http://swissmodel.expasy.org/lddt/>.

3 Results and Discussion

We have developed the local Distance Difference Test (IDDT) as a new superposition free measure for the evaluation of protein structure models with respect to a reference structure. In the following, we will discuss the choice of the optimal inclusion radius parameter R_0 to achieve low sensitivity to domain movements, and analyze the dependence of IDDT on the specific fold architecture. We will discuss the application of IDDT for assessing local correctness in models, including stereo chemical plausibility. Finally, we will present an approach for assessing a model simultaneously against several reference structures, e.g. a structural ensemble generated by NMR. Optimal choice of the inclusion radius parameter R_0 makes IDDT largely insensitive to domain movements

Determination of the optimal inclusion radius.

The nature of the IDDT score is ultimately determined by the choice of the inclusion radius parameter R_0 . For low values of the inclusion radius, only short-range distances are assessed, and only the accuracy of local interactions has a major impact on the final value of the IDDT score. On the other hand, when the value of the inclusion radius parameter is high, the evaluation of long-range atomic interactions gains a bigger contribution in the final score, and the final IDDT score turns into a representation of the global model architecture quality.

For assessing the accuracy of protein models, the inclusion radius should be high enough to give a realistic assessment of the overall quality of the model, but at the same time, the IDDT score should not lose its ability to evaluate the modeling quality of local environments. Especially, scores should not be influenced by changes of domain orientation between the model and the target structures. The optimal value of the inclusion radius parameter R_0 has been determined on a dataset comprising all CASP9 predictions for multi-domain targets, and the corresponding assignment of assessment units (AU) as defined by the CASP9 assessors. Both GDC-all and IDDT scores were computed for all predictions. The IDDT scores were computed against the whole target structures whereas the weighted GDC-all scores were calculated as weighted averages of the AU-based scores (see Methods for details). Hence, the weighted GDC-all scores can be considered to be largely devoid of the influence of domain movements. We computed IDDT scores for a range of R_0 values from 1 to 40 Å, and for each threshold we calculated the correlation with the weight-averaged GDC-all scores for the same predictions. We used

GDC-all (and not the more common $C\alpha$ based GDT) score in order to compare two all-atoms measures on the same set of data. The results are shown in **Figure 4.1**

For small values of the R_0 parameter, the IDDT score essentially reduces to a contact map overlap measure¹⁸⁷ and the correlation with global scores such as GDC-all is rather low. As the inclusion radius increases, longer-range interactions are being evaluated and the correlation shows a steep increase as the IDDT score starts to reflect the global quality of the model. For large values of R_0 , where inter-domain relationships start playing a more significant role and domain movements start to influence the IDDT score, the correlation begins to decrease slowly. However, the slow decrease in correlation for values of the inclusion radius higher than 24 Å (**Figure 4.2**) shows the stability of the IDDT score with respect to the influence of domain movements. Even including all inter-atomic distances in the calculation ($R_0 = \infty$), which maximizes the effect of domain movement, does not significantly lower the correlation with domain-based GDC scores ($R^2 = 0.82$). Based on this analysis, we selected a default value of 15 Å for the inclusion radius R_0 .

Sensitivity analysis vs. relative domain movements.

Proteins consisting of multiple domains can exhibit flexibility between their domains, which can often be experimentally observed in the form of structures with different relative orientations of otherwise rigid domains. In many cases, these relative movements play a functional role. From a modeling assessment perspective, however, the analysis of the relative orientation of the domains must therefore be separated from the assessment of the modeling accuracy of the individual domains.

The insensitivity versus relative domain movement makes the IDDT score an ideal candidate for the unsupervised evaluation of predictions of multi-domain structures, in contrast to scores based on global superposition. To illustrate this behavior, **Figure 4.3** shows IDDT and GDC-all scores computed on full length structures as a function of the AU-based weight-averaged GDC-all scores (x-axis).

As expected, the correlation between the two types of GDC-all scores is rather poor ($R^2 = 0.58$), while the correlation between the AU-based GDC-all scores and the IDDT scores is very good ($R^2 = 0.89$). The hybrid nature of the IDDT score allows it to be global enough to evaluate the modeling quality of the protein domains, but local enough to be only marginally affected by their relative orientations in the compared structures. When using the IDDT score to evaluate predictions, it is not necessary to split the target structure in separate domains, whose identification can be a complex and time consuming procedure.

Validation of IDDT Score Baselines for Different Protein Folds

Since IDDT scores express the percentage of inter-atomic distances present in the target structure that are also preserved in the model, a value of “0” corresponds to zero conserved distances, and “1” to a perfect model. However, these extreme values are in

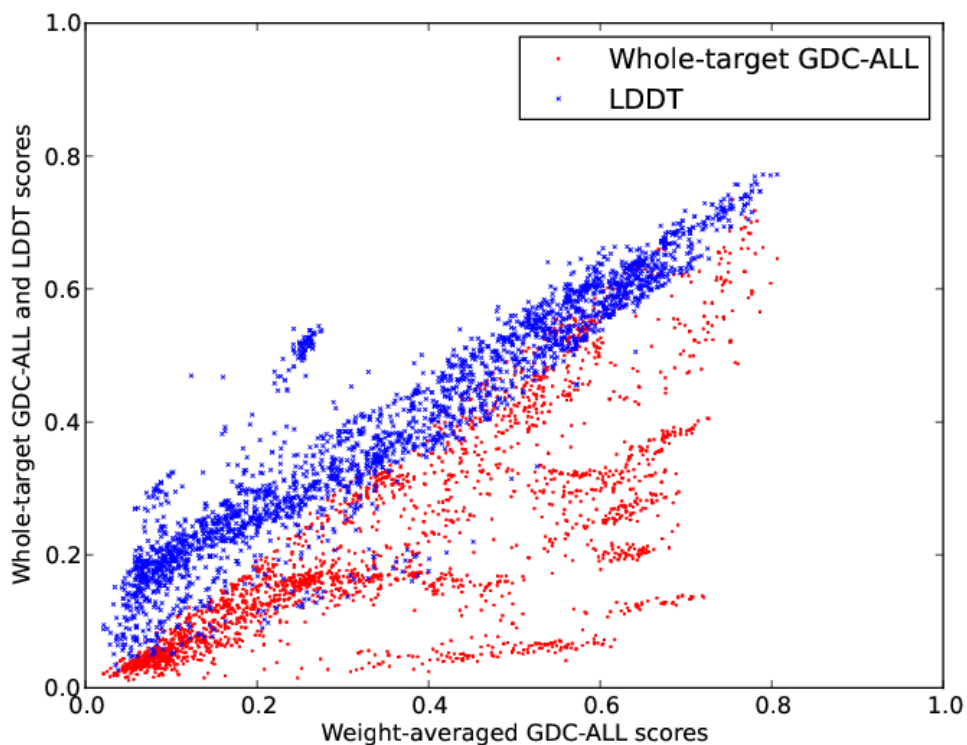


Figure 4.3 Correlation between whole structure GDC and IDDT scores and domain-based weight-averaged GDC scores. For all CASP9 predictions of multi-domain targets, GDC-all scores (red dots), and IDDT scores (blue dots) were computed against the whole unsplit target structures. For the IDDT scores, the default value of 15 Å for the inclusion radius was used.

practice rarely observed, even in extremely high and low quality models. At the high-accuracy end, fluctuations in surface side chain conformations will result in values lower than 1. For very low accuracy models, still some local inter-atomic distances will be preserved if the model has at least a stereo-chemically plausible structure and features some secondary structure elements. In the context of protein model assessment, two types of baseline values are of interest: the expected score when comparing two random structures, and scores for models with correct folds, including alignment errors.

In principle, the first value could be estimated using Flory–Huggins polymer solution theory^{188–189}. However, since protein structures are rich in rigid structural elements like α -helices and β -sheets, where the relative local positions are restricted, they show in general a higher number of preserved local distances than random polymers. Based on these considerations, we decided to empirically derive IDDT baseline scores by comparing a reference structure with a set of well-defined decoy models. The average IDDT score when comparing random structures, i.e. protein models with different architectures (see materials and methods), is 0.20 (+/- 0.04). For estimating the effect of alignment shifts in models with otherwise correct fold and stereochemistry, we created pseudo-models starting from the original protein structure and introducing threading errors of

increasing magnitude for different representative structure architectures from CATH¹⁸⁵. We then compared the pseudo-models to the original structure, computing their IDDT scores against it.

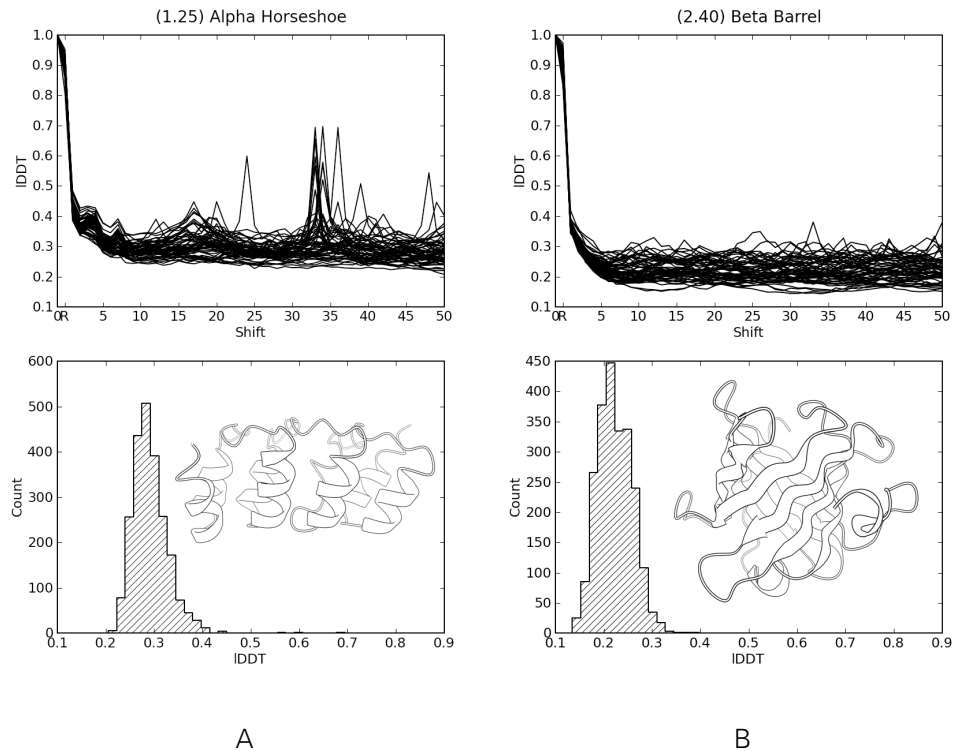


Figure 4.4 Baseline IDDT scores for models with simulated threading errors. IDDT scores of pseudo-models with threading errors for two examples of different CATH Architectures are shown: Alpha Horseshoe (left panel) and Beta Barrel (right panel). The IDDT score is plotted as a function of the introduced threading error (top panels). The histograms (bottom panels) show the distribution of these “baseline” scores for threading error offset > 15 residues for the two architectures. The structure inlays show an example structure of the respective CATH Architecture. Peaks at large off-sets indicate repetitive structural elements with locally correct arrangement.

Here, we show the results for CATH architecture entries 1.25 (Alpha Horseshoe) as example for proteins rich in α -helices, and 2.40 (Beta-barrel) as representative for a β -sheet protein (Figure 4.4). The plots at the top of each panel show the value of the IDDT scores (on the y-axis) for 60 pseudo-models as a function of the magnitude of the threading error (residue offset) on the x-axis. For large threading errors, the IDDT scores converge to a “baseline” range of scores, which appears to be largely independent of the threading error magnitude. We considered scores in this range to be typical IDDT scores for a low quality model with the same architecture as the target structure. For models in the Alpha Horseshoe architecture, the average baseline IDDT score is around 0.28, while for the Beta barrel class the value is much lower around 0.22, showing a clear influence of the architecture of the protein. This indicates that the lower boundary of the IDDT score

can vary as a function of the architecture of the target protein, and direct comparison between absolute raw IDDT scores is only possible when comparing models of the same architecture. This is a common feature of most structure comparison measures.

One interesting feature in **Figure 4.3** are several peaks at larger threading errors in the Alpha Horseshoe architecture example. These peaks correspond to internal repeats in the structure, which give rise to locally correct models when the threading shift coincides with the size of the repeat.

Local Model Accuracy Assessment

Modeling errors are typically not homogeneously distributed over the full lengths of a model, but are localized, e.g. segments in template based models which had to be re-modeled de novo. Residue-based IDDT scores quantify the model quality on the level of a residue's environment, and can reveal regions in the model which have been well-predicted. The low sensitivity of IDDT to relative domain movements also applies to per-residue scores. As shown in **Figure 4.1**, local IDDT scores are not dominated by different domain orientations between the target and the model structures, but correctly reflect the accuracy of the local atomic environment surrounding the residue under investigation in the model. This makes the local IDDT scores ideal to detect local structural divergences in multi-domain structures. **Figure 4.1** shows a superposition of the structure of target T0542 (in transparency) with prediction by group TS236 (colored according to the full-length IDDT score). The models represent each of the two individual domains with high accuracy, but their relative domain orientation does not correspond to the target structure. Superposition-based scores would assign a high score to one of the domains but not to the other, or require scoring based on isolated domain. As illustrated on the right panel (**Figure 4.1**), residues with low IDDT score correspond to regions of large local structural divergence between the two domain structures, irrespective of the domain movement between them. As expected, low local scores can also be detected at the interface between the two domains where the interactions cannot be modeled correctly without knowing their relative orientation in the target.

Stereo-Chemical Realism Assessment

While validation of the stereo-chemical plausibility of protein models is a routine procedure for experimental structure determination, e.g. in X-Ray crystallography⁸⁶, this is not common practice in theoretical modeling. Depending on the applied method, models generated in silico may reveal rather unrealistic stereo-chemical properties. Typically, numerical scores applied in retrospective model assessment compute a measure for the average atomic dislocation between the reference structure and the model, without any concern for the stereo-chemical quality of the latter. Consequently, two models with similar scores may nevertheless differ significantly in their stereo-chemical plausibility, and some models might include atomic arrangements which are physically impossible.

To address this limitation, IDDT incorporates a stereo-chemical plausibility check, which assesses two aspects of model quality: the lengths of chemical bonds and the widths of angles in the model structure. Bond and angle measurements are compared to a set of standard parameters derived from high-resolution crystal structures¹⁸³. A stereo-chemical violation is defined as a parameter deviating from the expected values by more than a specified number of standard deviations (default: 12σ). Inter-atomic distances between non-bonded atoms in the model are compared with the sum of their Van der Waals radii¹⁸⁴, and a violation (“clash”) is assigned if two atoms are closer than the sum of the Van der Waals radii, allowing a certain tolerance (default: 1.5 Å). When calculating the IDDT score, all distances involving side chain atoms of a residue involved in any type of stereo-chemical violations in the model are considered as non-preserved. In cases where backbone atoms are involved in stereo-chemical violations, all distances involving this residue are considered non-preserved. This approach leads to the lowering of the final IDDT score of a model according to the extent of the structure’s stereo-chemical problems (**Figure 4.5**).

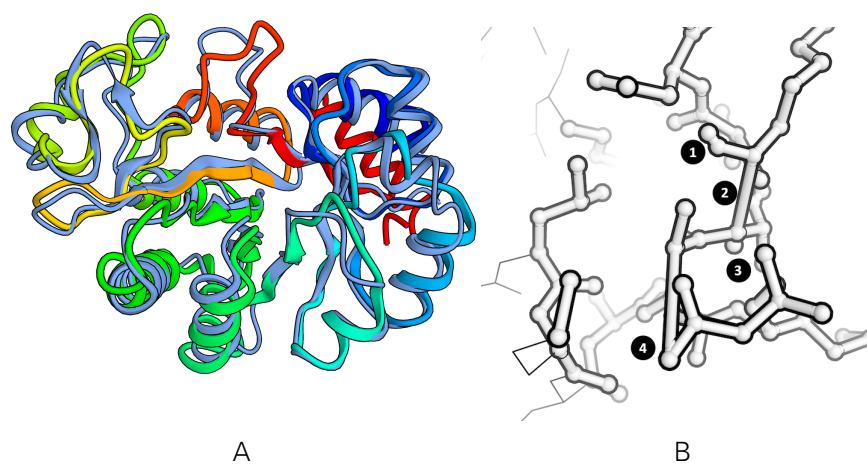


Figure 4.5 Assessing stereo-chemical plausibility. This example illustrates the stereo-chemical quality checks on IDDT score for a model (TS276, left side as ribbon representation) for target T0559-D1 with unrealistic stereochemistry (close up, right panel). Residues with too short (1) or too long (2) chemical bonds, as well as those with close atomic interactions (3) or impossible bond angles (4), result in lower scores during the IDDT computation.

As an example, **Figure 4.5** shows the CASP9 prediction T0570TS276_1 for target T0570-D1. The backbone of the prediction can be superposed accurately to the backbone of the target structure (left panel), and the prediction has indeed a high GDT_HA score (0.814). The analysis using all-atom scores does not immediately reveal the problems, with a GDC-all score of 0.705 and an IDDT score without stereo-chemical checks of 0.682. However, when the IDDT score includes stereo-chemical check, the IDDT score drops to 0.571, highlighting the presence of several problematic residues. Panel B shows a close up of the region around residue Alanine 21, where several stereo-chemical violations are evident.

Multi-Reference Structure Comparison

The typical situation for protein structure prediction assessment is to compare a model against a single reference structure. There are, however, cases where several equivalent reference structures are available, e.g. structural ensembles generated by NMR, crystal structures with multiple copies of the protein in the asymmetric unit (for example, target T603 in CASP9), or independently-determined X-ray structures for the same protein at different experimental conditions. Choosing one of them to be used as a reference for the calculation of model quality scores can only be an arbitrary decision. In all these cases, no structure can be considered more reliable than any other, but the choice of reference for the evaluation score can lead to very different results for models of equal quality. Due to the choice of template, often models have a higher similarity to one or the other.

In case of the local distance difference test, the following approach allows to evaluate a model simultaneously against an ensemble of reference structures: for each pair of atoms, we define an acceptable distance range by taking the minimal and maximal distance observed across all references where the atoms are present. If, in any of the reference structures, the distance is longer than the inclusion radius R_0 , this distance is considered a long-range interaction, and is ignored. For the assessment, the corresponding distance in the model is considered preserved, when it falls inside the acceptable range or outside of it by less than a predefined threshold offset.

One obvious application of the multi-reference IDDT score is the evaluation of models against NMR structure ensembles. For example, in the case of CASP9 target T0559 (PDBID: 2L01), an ensemble of 20 NMR structures has been experimentally determined. Selecting one single chain from the ensemble as reference to evaluate prediction models would be an arbitrary decision, artificially favoring some models which are closer to that specific structure. To estimate the effect of selecting a single reference structure (as must be done for GDT scores), all structures in the ensemble were in turn used as a “model” and evaluated against all the others. Using traditional pairwise comparison with GDC-all scores (Figure 4.6, striped bars), fluctuations of almost 12 GDC points are observed. Ideally, this situation should be avoided, and a prediction should not be rewarded or penalized for being more similar to one member of the ensemble of than to another. The multi-reference version of the IDDT score has been developed to overcome this problem by sampling the conformational space covered by the ensemble and compensating for its variability. Using the same example, the multi-reference IDDT score, which uses one chain as a “model” and all the others together as multi-references, shows a spread of less than 1% (Figure 4.6, dotted bars), indicating its robustness when scoring a model against an ensemble of equivalent reference structures.

4 Conclusions

In this manuscript, we describe the local Distance Difference Test score (IDDT), which combines an agreement-based model quality measure with (optional) stereo-chemical

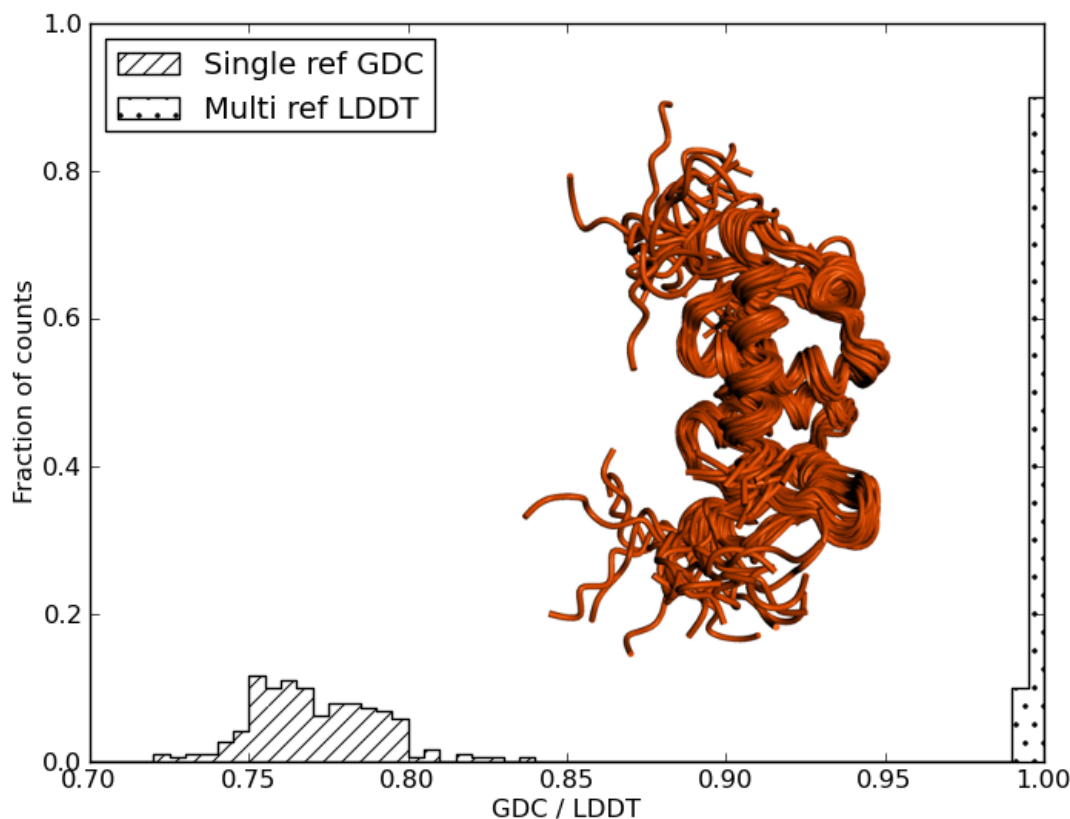


Figure 4.6 Comparing a model against an ensemble of reference structures. The experimental reference structure for CASP target T0559 (human protein BC008182, PDBID:2L01) is an ensemble of NMR structures. The graph shows the effect of selecting a single structure as reference (GDC-all values as striped bars) in contrast to the multi-reference IDDT implementation (dotted bars). For this example, in turn each structure within the ensemble was selected as model and compared to the other members (reference).

plausibility checks. We have demonstrated its low sensitivity with respect to domain movements in case of multi-domain target proteins, which allows for automated assessment without the need for manually splitting targets into assessment units. We also have shown that local atomic interactions are well captured and local IDDT scores faithfully reflect the modeling quality of sub-regions of the prediction. In addition, we present an approach to compare models against multiple reference structures simultaneously without arbitrarily selecting one reference structure for the target, or removing parts which show variability. Additionally, as an agreement-based score, IDDT is robust with respect to outliers.

One disadvantage of the IDDT score is that it does not fulfill the mathematical criteria to be a metric. However, the same is true for most scores commonly applied for structure comparison such as RMSD or GDT. We consider IDDT particularly suited for the evaluation of predictions for the same target protein, for example in the context of the

CASP and CAMEO experiments. For these kind of applications, unlike, for example, for clustering protein structures, we don't see the lack of metric properties as a significant limitation.

FUNDING: We gratefully acknowledge the financial support by the SIB Swiss Institute of Bioinformatics.

Graph-based Constraint Selection for Multi-template Modeling

A critical step of template-based modeling is the selection of suitable template structure information. For well characterized protein families, often 20 or more alternative experimental template structures are available. While all templates may share a similar overall topology, the relative orientation of sub-domains often differs significantly. Such intrinsic movements limit the assignment of consistent structural constraints for the comparative modeling step. Rigid-body superposition and clustering of all templates and their sub-domains is a time consuming and error prone procedure. Here, we propose an efficient and robust procedure to identify stable structural building blocks in ensembles of structures using contact-overlap (COM) map consistency.

1 Introduction

It has been shown repeatedly that combining information of multiple template structures for computational structure prediction is beneficial^{66,190–191}. Skilled human predictors in particular are able to improve models by using information from multiple templates¹⁹². However, the automatic selection of information from multiple sources is still an open challenge. Knowing when to leave a template move to another is non-trivial. Due to the high number of possible combinations it is not tractable to exhaustively enumerate all solutions.

The most successful multi-template modeling programs use extensive sampling and scoring of alternative models to identify suitable combinations of templates^{66,79,82}. For example, in I-TASSER models are assembled on a lattice⁶⁶ and then refined. For proteins of typical size (>200 residues), optimization can require several days. For integration into the SWISS-MODEL web server, such high computational costs are prohibitive and more efficient algorithms for multi-template modeling are required.

Cheng¹⁹⁰ proposed a multi-template selection scheme, where templates are combined in a multiple-sequence alignment, which is then used as input for MODELLER⁷². Starting from a seed template, other templates are added to the multiple sequence alignment if they meet one of two requirements: (a) their sequence alignment E-value is within a certain threshold to templates already added, (b) they cover residues not previously covered in the multiple-sequence alignment. In the algorithm proposed by Cheng, templates are added even if they are structurally inconsistent, e.g. due to ligand induced domain-movement, alignment errors etc. However, it is well-known that the use of inconsistent constraints for multi-template modeling leads to suboptimal models. The models satisfying most of the constraints, often do not represent a biologically relevant conformation, and can even be physically and chemically impossible. Thus, when combining distance constraints from multiple templates, special care needs to be taken to ensure that the used information is consistent. Constraints can be filtered for consistency prior to passing them to MODELLER, e.g. in the case of RaptorX, only templates are added

which have a TMscore⁹ higher than 0.65 with already added templates⁶⁹. This coarse-grained consistency filtering prevents the most detrimental effects of mixing structurally diverse templates, but has limitations, as it enforces consistency in a very global manner. As an alternative, in HHpred server, Söding replaced the standard Gaussian distance constraints in MODELLER by a mixture of two Gaussians, describing correctly and incorrectly aligned residues. The mixture parameters (means, standard deviations and mixture weights) are predicted by a mixture density network (CASP9 abstract booklet). Rather than summing the constraints from multiple templates, the constraints are multiplied. Consistent restraints are reinforced, and contribute significantly more to the molecular PDF than inconsistent constraints. However, the value maximising an individual feature PDF are drawn towards the average, which might not represent a physically plausible conformation. Indeed, the models of HHpred submitted to CAMEO often show physically impossible angles and bond lengths. Additionally, the consistency-enforcement for each constraints is done separately, and does not take the local environment of residues into account.

In our view, consistency should be enforced on the level of a residue's environment. Residues in two structures are consistent, if their local environment is very similar, i.e. if they share contacts to the same sets of residues. The idea of environment conservation was inspired by the structural similarity score IDDT, which quantifies structural similarity by the overlap of conserved local distances. This view is in between the consistency enforcement protocols of RaptorX and HHpred, which either consider similarity on a very global or local scale. In this work, we will describe an algorithm to divide the structures into stable sets, starting from local the environment of residues. Since the sets of consistent residues of the two structures often coincide with biological domains, we call this procedure the Domain-Find (DOMF) algorithm. DOMF shares some ideas with a method to identify RMSD-stable domains developed by Snyder and Montelione for ensembles of NMR structures¹⁹³, but replaces the variance matrix clustering by a graph-based neighbor overlap. First we will discuss the DOMF algorithm and apply it to experimental structures of the same target sequence. The domain-assignment of DOMF is then compared to domain-assignments from the well-established DynDom program¹⁹⁴. In a second step, we will show the application of the DOMF algorithm to the detection of well-modelled regions of homology models. Finally, we will show on a few modeling cases how constraint consistency leads to more accurate models.

2 Materials & Methods

We first describe the core of the DOMF procedure and show an efficient implementation of the DOMF algorithm which uses bitwise operations to speed up the calculation of the neighbor overlap.

The Core of Domain-Find

The method solely relies on distances between pairs of C α atoms to identify consistent sets of residues. The mapping of the C α atoms between the first and second structure is given by a pairwise sequence alignment, e.g. as obtained from BLAST⁵⁸, HHsearch³², or HHblits³³. We denote the Euclidian distance between the i th and j th C α of the first structure as a_{ij} , the corresponding distance in the second structure as b_{ij} . The distances a_{ij} , b_{ij} form distance matrices A and B , respectively.

We define the overlap of two distances a_{ij} and b_{ij} as

$$o_{ij} = \omega(d_{ij}) = \begin{cases} 0 & \text{if } d_{ij} > \tau \\ 1 & \text{if } d_{ij} \leq \tau \end{cases}$$

with $d_{ij} = |a_{ij} - b_{ij}|$. The tolerance parameter τ controls how similar the two distances have to be in order to *agree*. The symmetric matrix C spanned by all o_{ij} is called the contact overlap map (COM). Based on the matrix C , we would now like to find groups of C α atoms that have similar distances in both A and B .

The contact overlap map can also be seen as an adjacency matrix of a graph $G := \{V, E\}$. A vertex v_i is connected (adjacent) to v_j if the C α distances between i th and j th C α in both structures agree, that is $o_{ij} = 1$. We would now like to find clusters of vertices with high connectivity in the group but little connectivity between the groups. This can be achieved with a simple iterative update of edge weights. We note that, if v_i and v_j are part of the same cluster the sets of adjacent vertices n_i and n_j must be very similar. The edge weight o_{ij} between vertex v_i and v_j is iteratively updated as

$$o_{ij} = \begin{cases} 1 & \text{if } u_{ij} \geq \epsilon \\ 0 & \text{if } u_{ij} < \epsilon \end{cases}$$

with

$$u_{ij} = \frac{|n_i \cap n_j|}{|n_i \cup n_j|}$$

ϵ is called the threshold, and is a value between 0 and 1 that describes how much the neighbours of two vertices have to overlap in order to be in the same cluster. This update step essentially draws vertices towards the groups of vertices they most agree with. The algorithm usually converges after 5, sometimes after 10 iterations. The resulting graph only contains edges between vertices of the same cluster, edges of vertices of different clusters are subsequently removed during the iterative update procedure. The building blocks of the structures are then simply the *connected components* of the graph¹⁹⁵: starting from any vertex v_i , all vertices v_j are added to the component of v_i which can be reached from a depth-first or breadth-first search.

The result of DOMF is a partitioning of the residues into domains/clusters (we use the terms domain and cluster interchangeably). We denote each cluster of G as S_i . Vertices which are not consistent in the structures, e.g. which are part of connected components of only one residue are assigned to the *free* (S_f) cluster. Additionally, residues of clusters with less than 16 residues are assigned to the free cluster.

In **figure 5.1**, a few stages of the algorithm are shown for two adenylate kinase structures that undergo a domain rearrangement upon ligand binding.

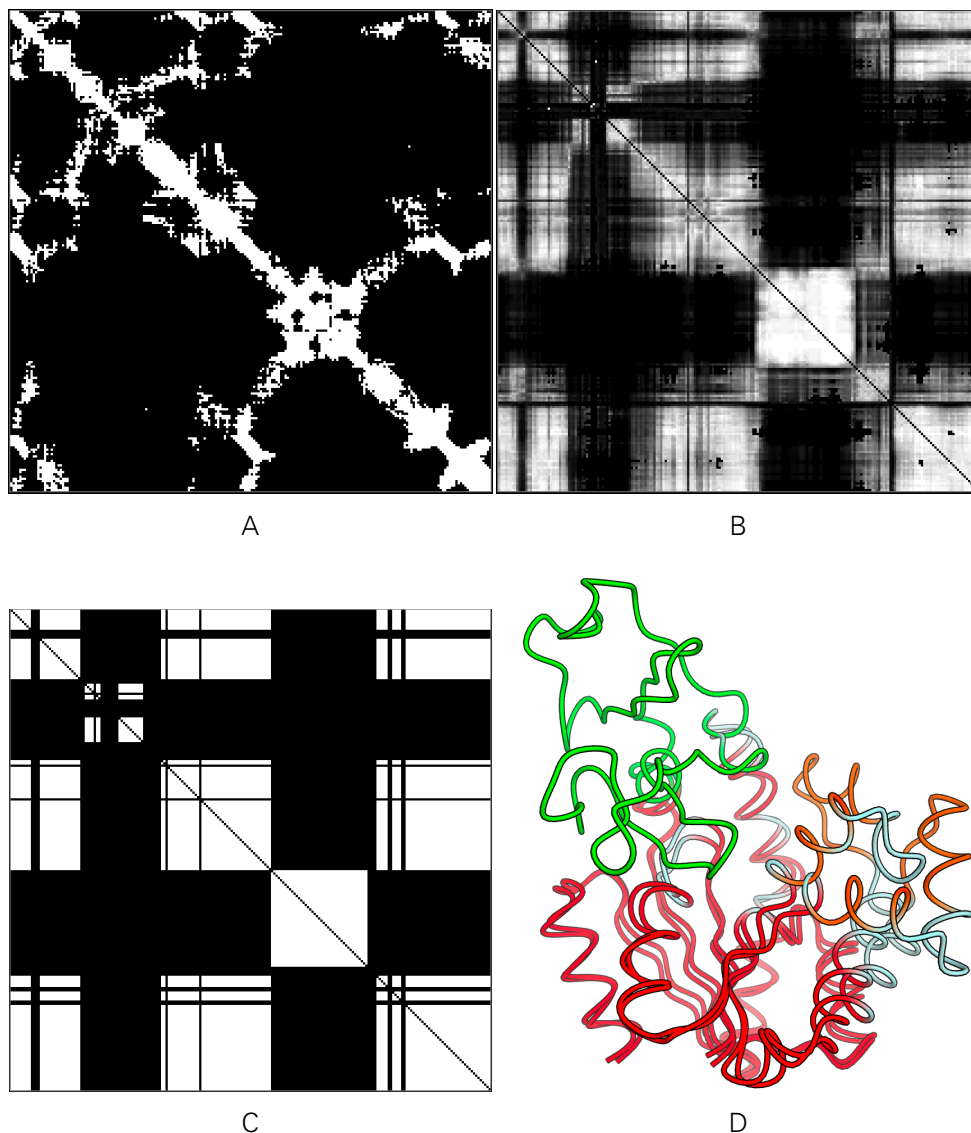


Figure 5.1 Example adjacency matrix and domain assignment for open and closed conformations of the adenylate kinase (PDB identifiers 1ake and 4ake). The adjacency matrix **(A)** after zero iterations, **(B)** after two iteration **(C)** the final assignment after five iterations. White corresponds to an edge weight of one, and black to and edge weight of zero. **(D)** The structures of the open and closed conformation are colored according to the domain assignment. The red, orange and green domains correspond to the Rossman fold, AMP binding site, and lid domain, respectively. The white parts are not structurally conserved and are assigned as unordered.

Global and Local Domain-Find

There are two possible ways to determine consistent sets of $C\alpha$ -atoms:

- Either by considering all $n \times n$ distances of the two structures (global), or

- by only considering distances below a certain distance cutoff, e.g. 15Å (local).

When using a distance cutoff, distances between residues beyond a distance cutoff are ignored by effectively marking the edges as undefined in the adjacency matrix. The first update of the neighborhood similarity is solely based on the local connectivity. During subsequent updates, the previously undefined edge weights become defined based on their neighbor overlap. Since the neighbor overlap is not reliable when calculated from a small number of vertices, we only mark edges as defined when the vertices have a defined connectivity to at least 10 common neighbors. If this criterion is not fulfilled, the edge is marked as undefined. The edge weights might then become defined during subsequent iterations. The local neighborhood update gradually starts to spread to other vertices, until the domains can be read out as the connected components.

In the global scheme, the first iteration is already based on the complete neighbor overlap and has thus a more global character.

To show the difference between the two schemes, consider the output of the DOMF algorithm on two structures of the response regulator PleD (PDB identifiers: 2v0n¹⁹⁶ and 2wb4). PleD has a two-domain architecture with a flexible linker region in the middle. The two structures are overall very similar, but have a slightly different orientation of the two domains. The progress of the DOMF algorithm is shown in **figure 5.2**, for the global and local cases. For the local case, the similarity of the hinge region is the main determinant for the partitioning. The dissimilarity of the hinge region essentially introduces a local region of low connectivity, and, as a result, the similarity between the domains does not propagate through the hinge region. The global case on the other hand is mainly driven by the overall dissimilarity of the two domains, the similarity of the hinge region is less important. The domain partitioning for the local and global DOMF schemes overlap well. A few residues in the vicinity of the hinge region are assigned to the *orange* domain for the global scheme, but are part of the *red* domain for the local case. These residues appear as a white line in the lower-left and upper right corners of the adjacency matrix in **figure 5.2B**. Visually as well as numerically, the convergence of the local domain-find algorithm is faster and *cleaner* for the local case: the connectivity for the global update scheme is smeared out, whereas the local scheme remains sharp and well-defined across all stages.

Fast Update of Edge Weights

Since we consider the edges of all pairs of vertices in the graph, the overall complexity of the algorithm is $O(n^3)$. For small structures of less than 200 amino acids, DOMF converges within fractions of a second. For a structure of 400 residues, the calculation takes 8 times longer. Implementing the inner-most loop in an efficient way, is crucial to keep the running time of the algorithm low, particularly for large n .

The naive and slow implementation operates on the similarity matrix directly. The edge-weight between the i th and j th $C\alpha$ is calculated as

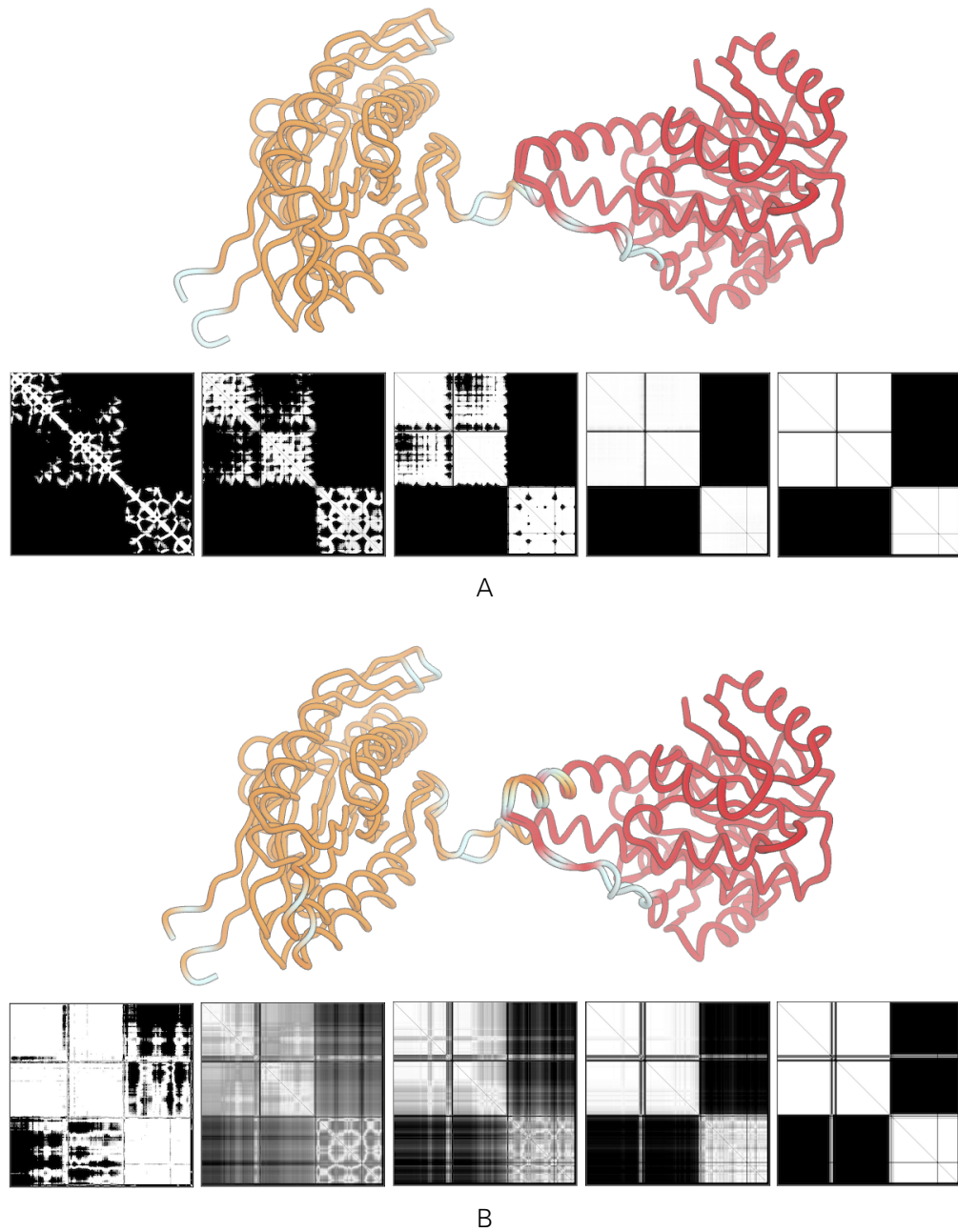


Figure 5.2 State of the adjacency matrix during neighbor overlap updates together with the final domain assignment for two structures of PleD (2v0n, 2wb4). **(A)** the adjacency matrix is constructed from distances which, in either of the two structures are below a cutoff of 15Å, **(B)** the adjacency matrix is constructed from all $n \times n$ distance pairs in the structure. The resulting domain assignment is shown for both the local and global update schemes. The first domain is shown in red, the second in orange. Residues in white are assigned to the free domain.

```

for (w_i, w_j) in (v_i.edge_weights, v_j.edge_weights)
    if not defined(w_i) or not defined(w_j)
        continue
    if w_i > threshold
        denominator += 1
        if w_j > threshold
            nominator += 1
    elif w_j > threshold
        denominator += 1

w_ij_new = nominator / denominator # float division

```

In essence, each edge is in one of 3 states: The edge is undefined, the edge has a value larger than the threshold value, the edge has a weight smaller than the threshold value. These 3 states can efficiently be encoded in two bits: One bit (D) states whether the edge is defined/undefined and one bit (C) states whether the two vertices are connected. The idea is not so much to optimize for space than for finding an efficient representation to update the edge weights. In each byte, we can store 4 neighbouring edge weights, in the format DCDCDCDC. Since we are not interested to know exactly which edges are defined, we can take advantage of bitwise operations to process multiple edges at once when calculating the neighbor overlap.

The intersection is calculated as the bitwise AND of the two bytes. The resulting byte has bits set to one for edges pointing to direct neighbors of v_i and v_j . To quickly identify the number of bits set in a byte, we use a lookup table which maps every integer in the range of 0 to 255 to the number of overlapping edges. For example, the number 206 (binary 11001110) corresponds to two edges set. The calculation of the denominator is a little more involved, since the defined bits and the edge weight bits need to be treated separately: for the edge weights, we calculate the bitwise OR and combine it with a mask for the edge bits to set the defined bits to zero. This result is then combined with the intersection calculated in the previous step with a bitwise OR. The algorithm can be further speed up by performing the bitwise operations on 64 bit integers, which allows to process 32 edges at once. By using this procedure, we have achieved a speed up of a factor of ten, compared to the naive implementation.

Determining the Optimal Threshold Value

The threshold value ϵ has a large influence on the partitioning of the structure into domains. The optimal threshold value is both depending on the overall similarity of the proteins being compared and the topology of the structures. Since certain partitionings are more favourable than others, the threshold values have to be determined for each pair of structures.

To assess the *fitness* of the graph partitioning, we devised an objective function which takes the connectivity of each S_i into account. Intuitively, a good partitioning maximises the connectivity within the domains and the vertices have few *outgoing* connections.

Many different ways have been described in the literature on how to measure the quality of a graph partitioning¹⁹⁷. One possibility is local graph density, which is defined as the fraction of edges in the cluster divided by the number of possible connections. Alternative definitions for graph density exist, e.g. by normalizing by the number of vertices in the cluster.

For our purposes, we take the connectivity within a cluster and between clusters into account. The score for a single domain is defined as

$$\rho(S_i) = 2 \frac{|E(S_i)| - |I(S_i)|}{|S_i| \cdot (|S_i| - 1)}$$

with S_i the vertices of the cluster, $E(S_i)$ edges for which both vertices are in S_i , and $I(S_i)$ are edges for which only one vertex is in S_i . We found that normalization by the number of possible connections in a domain performs substantially better than normalizing by the number of vertices, as it forces the domains to be more compact.

Our objective function combines the local graph density of all clusters as a weighted sum:

$$F(G) = -c \cdot f_f \frac{2 \cdot |E(S_f)|}{|S_f| \cdot (|S_f| - 1)} + \sum_i f_i \rho(S_i)$$

where f_i is the fraction of residues part of S_i , f_f is the fraction of inconsistent residues and c is a constant empirically set to 0.2. The optimal threshold value ϵ is the value which maximizes the objective function. For many of the structure pairs, the objective function is maximised over a relatively wide range of threshold values (**figure 5.3**). Initially, the score remains constant for threshold values between 0 and 0.2, before it starts to drop. A drastic drop of the score occurs around $\epsilon = 0.35$. Here, more and more residues are assigned to the *free* group, even though they have good connectivity. After 0.35, the structures are split into 2 domains, which causes the score to rise again. The score peaks at 0.525, before it decreases again. After $\epsilon = 0.75$, the thresholds are too high and as a result, virtually all residues are assigned to the *free* domain.

A more interesting case is the partitioning of 1lfh.A and 1lfi.A in **figure 5.4**. The partitioning first starts by splitting the structure into two domains ($0.25 \leq \epsilon < 0.7$), then three ($0.7 \leq \epsilon \leq 0.825$). Already at low values of ϵ , DOMF identifies two domains, which is indicative for strong domain orientation movement. The green and red domains, however have almost the same orientation in both structures and it is not before a threshold value of 0.7 that the similarity in the interface between the domains becomes too low to split the structure into three parts. This example illustrates the importance of optimizing the threshold value on a case-by-case basis. In this case, it is helpful that the three domains are very similar in the two structures.

Consistent Constraints for Multi-Template Modeling (MTM)

The main motivation for the development of DOMF has been to create a tool for the automatic detection of consistent structural information in sets of template structures. Here

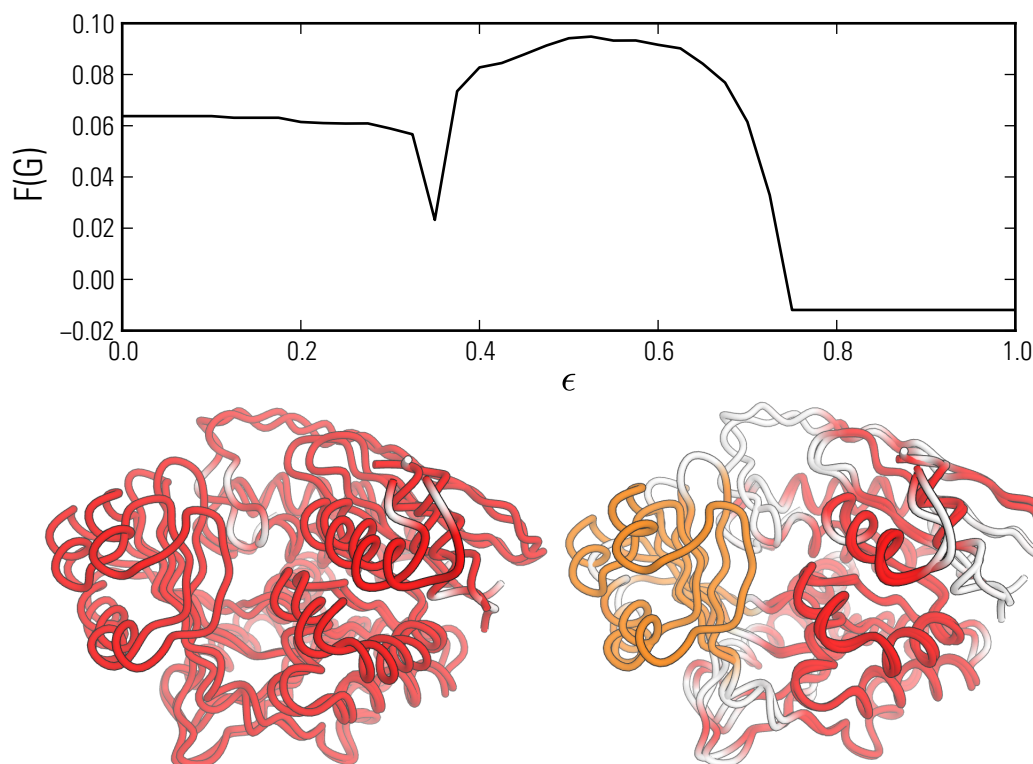


Figure 5.3 $F(G)$ for the partitioning of 1lio.A and 1lii.A into domains. At the bottom, the partitioning of the structures at $\epsilon = 0.2$ and $\epsilon = 0.525$ (the optimal value) is shown. White regions are assigned to the *free* cluster.

we present a very conservative multi-template modeling algorithm which enforces consistency of restraints before they are added to the MODELLER program. For the purpose of constraint selection we are mainly interested in the local conservation of constraints, because MODELLER only generates $C\alpha$ - $C\alpha$ distance constraints below a cutoff of 14\AA (MODELLER reference manual). Hence, in this application it is not important whether long-range interactions are conserved or not and the local update scheme for DOMF will be used.

The MTM algorithm starts with a seed template and gradually extends it with information from other template structures.

1. We start with an initial seed template, e.g. the highest-scoring template from BLAST. The seed template's $n \times n$ $C\alpha$ -distance matrix is extracted, setting distances which are not present as undefined.
2. We identify templates which contain information not covered in the seed distance matrix. If at least $m = 5$ residues are present in the template which are not available in the seed, the DOMF-algorithm is applied to the seed distance matrix and the template structure. The resulting domain partitioning is stored.
3. All of the templates of the previous step are sorted by number of residues which are consistent in both structures (as defined by the domain partitioning). The template

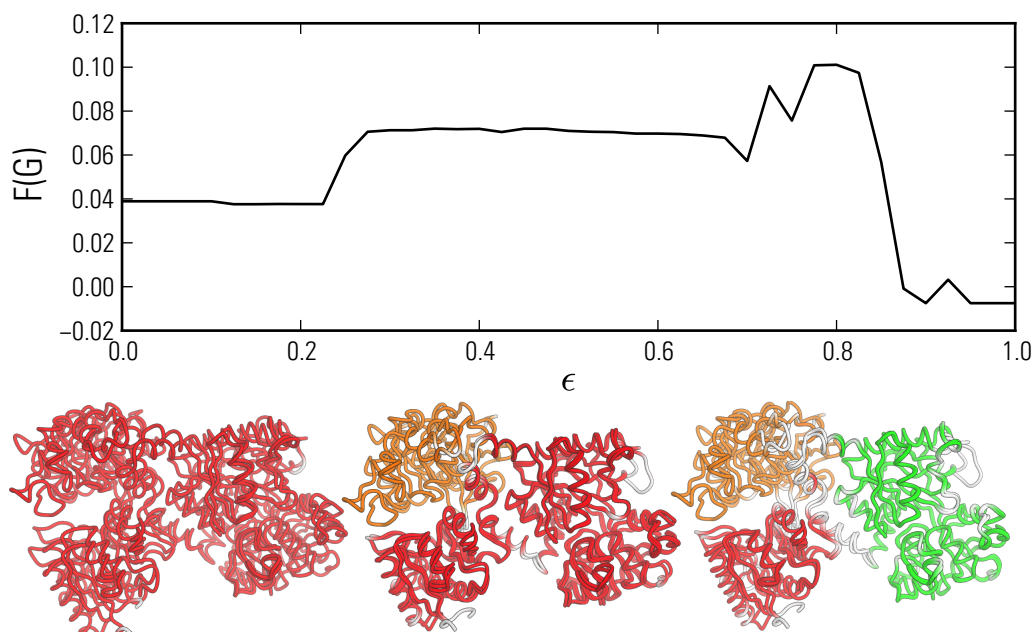


Figure 5.4 $F(G)$ for the partitioning of 1lfi.A into domains. At the bottom, the partitioning of the structures at $\epsilon = 0.2$ and $\epsilon = 0.65$ and $\epsilon = 0.8$ (the optimal value) is shown. White regions are assigned to the *free* cluster.

structure with the highest number of agreeing residues is selected.

4. From the new template structure information which is either consistent with the seed template, or covers residues previously not part of the seed, is added.
5. Steps 2 to 4 are repeated until no template contains new information.
6. The restraints are used as an input to MODELLER.

While more elaborate schemes are certainly feasible, e.g. by mixing and matching parts of multiple templates, these schemes are computationally more demanding and require scoring of alternative solutions. The algorithm we describe here as a proof of principle for constraint selection is very conservative, since improvements are only possible in one of two cases:

1. a template contains additional residues at the N- or C-terminus of the model
2. a template is well-aligned to the target sequence, whereas the seed template contains a deletion

3 Results

Domain-Find on Pairs of Experimental Structures

As a first benchmark, and to understand the behaviour of DOMF in detail, we have applied the DOMF algorithm to pairs of experimental structures. The structures are part of the Protein Structural Change Database (PSCDB)¹⁹⁸, which contains proteins which undergo structural rearrangement upon ligand binding. The domains have been identified using the algorithm of DynDOM^{194,199}. We have applied the DOMF algorithm to all 35 single-chain pairs of the coupled domain motion (CD) category, the other 24 are movements involving more than one protein chain. While such multi-chain movements could also be captured with the DOMF algorithm, analysis of the performance of DOMF on multi-chain proteins is outside of the scope of this work. The structure pairs in the PSCDB offer a good selection of domain movements, from small to large conformational changes.

We have calculated the domain-assignment of DOMF by using the local update scheme with a tolerance of 0.5Å for all 35 structure pairs of the PSCDB dataset. The threshold parameter ϵ has been optimized for each case individually by using the objective function described above. A summary of the results is given in **table 5.1**. For each pair of protein structures, the optimal ϵ , the number of domains identified by DOMF and PSCDB are listed together with the number of residues per domain. The corresponding domains identified by PSCDB and DOMF are determined by pairing the domains with the largest residue overlap. When the numbers of domains are different, a single domain can be paired with more than one domain (e.g. for 1eym.A and 1j4r.A).

For the majority of cases, the optimal threshold value ϵ is between 0.4 and 0.6. For structures which are only split into one domain, the optimal ϵ is below 0.350. Very high threshold values are reached for only two structure pairs: 2qrj.A/2qrl.A and 1lfh.A/1lfi.A which both reach an optimal ϵ at 0.8. Such high values of ϵ are only possible if the internal distances of the two structure are very similar.

For 28 of the 35 structure pairs, the number of domains identified by DOMF and PSCDB are the same. For the remaining seven, DOMF identifies only one domain in six cases, whereas the database lists two. Additionally, there is one case, where DOMF identifies three domains, and PSCDB two. We have looked at these seven cases in more detail. One of these structure pairs is 1yem.A/1j4r.A. **Figure 5.5** shows the alternative domain assignments by DOMF and PSCDB. The two structures are very similar overall ($C\alpha$ -RMSD of 1.147). According to PSCDB, the structures have a larger domain of 78 residues (yellow) and a smaller one of 24 residues (red). DOMF only identifies one domain and marks part of the second domain as unordered. Here, the differences between the two domains are not large enough relative to the intra-domain fluctuation to split the structure into two domains. Similarly, for 1sbq.B and 1u3g.A, two methenyltetrahydrofolate synthetase structures, DOMF assigns an N-terminal helix as unordered, whereas PSCDB defines it as a domain on its own. For 2cgk.B/2uyt.A, DOMF splits the structure into three domains,

unbound	bound	ϵ	N_D	N_P	D_{1D}	D_{1P}	D_{2D}	D_{2P}	D_{3D}	D_{3P}
4ake.A	2eck.A	0.425	3	3	121	133	43	43	17	17
1lfh.A	1lfi.A	0.800	3	3	299	354	136	169	143	143
1n0v.D	1n0u.A	0.400	3	3	451	459	280	282	72	72
2cgg.B	2uyt.A	0.675	3	2	284	257	52	217	32	217
2bjb.A	2o0d.A	0.500	2	2	180	217	191	186	—	—
2ex0.B	2ihz.A	0.475	2	2	233	230	119	145	—	—
1oen.A	2olr.A	0.675	2	2	252	348	113	157	—	—
2gg4.A	2pgc.A	0.500	2	2	212	220	212	219	—	—
1gqz.A	2gke.A	0.700	2	2	144	145	100	125	—	—
1vh3.C	1vh3.B	0.550	2	2	88	122	111	110	—	—
2gca.A	1jbw.A	0.550	2	2	292	292	99	108	—	—
1l5t.A	1lct.A	0.525	2	2	160	163	158	154	—	—
2ghb.A	2gha.A	0.425	2	2	216	213	155	156	—	—
2uvg.A	2uvi.A	0.425	2	2	232	231	166	163	—	—
2c00.A	2vqd.A	0.500	2	2	366	366	65	70	—	—
1zkb.A	1jvy.A	0.450	2	2	207	200	161	158	—	—
1ex6.B	1gky.A	0.725	2	2	93	127	45	55	—	—
3c6q.B	3C6q.D	0.600	2	2	146	154	142	147	—	—
1lio.A	1lii.A	0.525	2	2	193	254	55	67	—	—
2brw.A	1ojp.A	0.725	2	2	490	494	208	223	—	—
1jej.A	1qkj.A	0.725	2	2	178	188	106	159	—	—
1a48.A	2cnq.A	0.675	2	2	141	263	76	30	—	—
2qrl.A	2qrl.A	0.800	2	2	178	191	160	170	—	—
2ous.B	2ouu.A	0.525	2	2	292	298	26	26	—	—
1oid.B	1ho5.A	0.425	2	2	323	329	193	188	—	—
3d8r.A	3d8n.A	0.525	2	2	128	123	121	122	—	—
1zty.A	1zu0.A	0.450	2	2	286	281	242	243	—	—
1z15.A	1z17.A	0.400	2	2	199	199	143	141	—	—
2p0m.A	2p0m.B	0.375	2	2	619	630	18	28	—	—
1eym.A	1j4r.A	0.275	1	2	107	78	107	24	—	—
3cze.A	3czk.A	0.150	1	2	600	363	600	232	—	—
1xgd.A	1ef3.B	0.150	1	2	315	266	315	28	—	—
1meo.A	1rby.A	0.300	1	2	197	99	197	94	—	—
2cbi.A	2j62.A	0.150	1	2	584	456	584	123	—	—
1sbq.B	1u3g.A	0.350	1	2	164	127	164	32	—	—

Table 5.1 Comparison of the domain assignments of PSCDB and DOMF on all 35 single-chain coupled domain movement structures of the PSCDB. For each of the structure pairs, the number of domains identified by DOMF (N_D) and PSCDB (N_P) are listed, together with the domain sizes (D_{1D} to D_{3D} for DOMF, D_{1P} to D_{3P} for PSCDB). ϵ : the threshold value maximising the objective function $F(G)$. Differences in the number of identified domains are highlighted in bold.

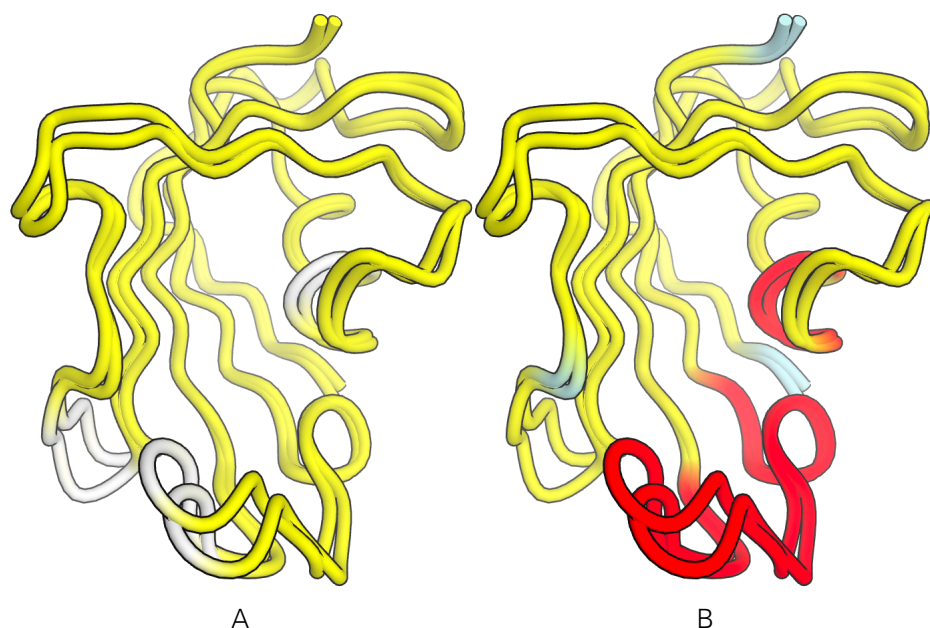


Figure 5.5 Comparison of domains identified by DOMF (A) and PSCDB (B). The domains are shown in red/yellow, disordered parts in grey.

whereas PSCDB lists only two. The largest domain of DOMF overlaps almost perfectly with PSCDB. However, DOMF divides the other domain into two parts: a domain of 32 and 52 residues, respectively. These domains are small compared to a size of 217 for the same residues in PSCDB. The smaller size is mainly caused by an unassigned beta sheet in the interface of the two PSCDB domains. In all 7 cases, where the number of domains found by DOMF does not match the domain definition listed in PSCDB, one of two situations applies: (1) the domain movement is very small and the local environment of these residues does not change enough to cause DOMF to split the structures, (2) the domain listed in PSCDB is very small. Such domains usually have a high surface to volume ratio.

The domain assignment of the two adenylate kinase structures 4ake.A and 2eck.A agrees almost perfectly, with only a handful of residues differing. Similar results are obtained for two crystal structures of elongation factor 2 (1n0v.D and 1n0u.A), the two phosphodiesterases 2ous.B/2ouu.A or the maltotriose binding protein (2ghb.A 2gha.A). In other cases, the domain assignments of DOMF and PSCDB deviate in some details. One such case is the domain assignment for the unbound and bound form of diaminopimelate epimerase (1gqz.A/2gke.A). The first domain agrees rather well. However, the second domain is smaller by around 25 residues. The difference stems mainly from residues at the interface of the two domains which show large deviations, these are assigned as un-ordered by DOMF. Whether these residues should be part of the domain depends on the exact definition of what a domain encompasses and depends on the application at hand. For the purpose of identifying consistent residues in pairs of structure, these residues at the interface are clearly not consistent and the choice made by DOMF is correct.

In general, domains identified by DOMF are smaller than domains from PSCDB. In a sense, the requirements for a domain are more strict and the resulting domains are usu-

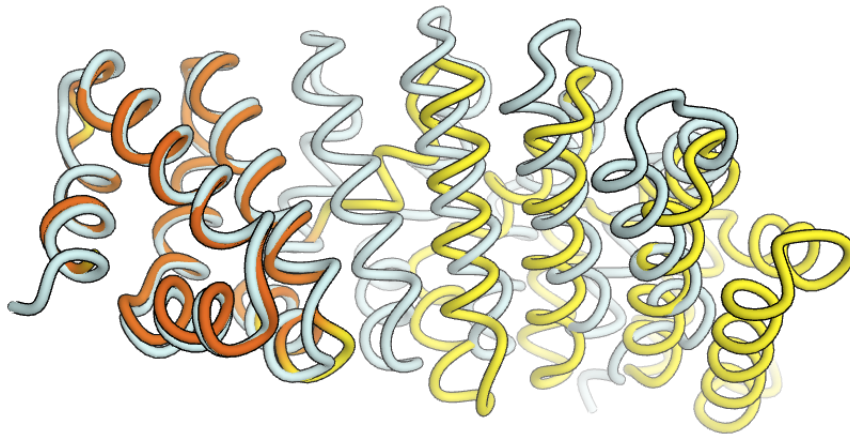
ally more compact. We argue that domains are driven to smaller sizes by the use of internal distances for domain calculation: residues are only part of the same domain when their surrounding is similar. This is in contrast to the method rotation vector analysis performed by PSCDB, which only requires residues to superpose well. Additionally, PSCDB domains tend to be more continuous, i.e. have less unassigned residues in the middle. This makes a direct comparison difficult, as the two algorithms seem to use different domain definitions. To some extent, the differences could be diminished by assigning residues of the free domain to the closest structurally conserved domain.

To summarize the performance of DOMF on pairs of experimental structures, the majority of the difference between PSCDB and DOMF comes from cases where domains are only slightly rotated in space, or when the interface between the domains are very similar in both structures. These small deviations are usually within the tolerance and are thus not detected as concerted movements in DOMF. On the other hand, the clustering of rotation vectors²⁰⁰ underlying PSCDB is able to use the direction-dependence of these structural variations to distinguish between thermal fluctuations and concerted movements. Depending on the applications, the additional accuracy offered by the PSCDB algorithm might be desired.

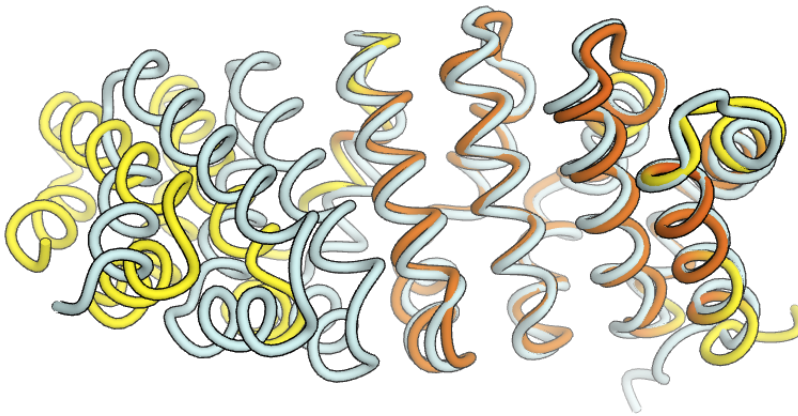
Identifying Correctly Predicted Residues in Models

In the following, we discuss the application of DOMF to identify regions of homology models which have been well predicted. LGA⁴¹, a tool generally applied to identify such regions and superpose structures usually only returns the RT operator which superposes the model onto the largest matching region of the target. However, in presence of domain movements, one would often like to superpose the model onto all regions that have been well predicted, e.g. for visual inspection. For this purpose, we have applied DOMF to model-target pairs. Model regions identified as domains by DOMF are regions which match the target structure and have thus been well predicted. For visual inspection, for each of these well-predicted regions, the model is then superposed onto the target, using atoms of common residues. To illustrate this application of DOMF, we use two examples from the tertiary structure prediction category of the CAMEO web-server⁵⁴.

The sequence for CAMEO target 4HXT_A codes for an engineered, dimeric repeat protein of 252 residues. In **figure 5.6**, the superposition of the prediction from server16 (SWISS-MODEL Next Generation) is shown for both domains that have been identified by DOMF. As can be seen, domain find is able to identify two regions of the model which have been predicted well. However, these domains have been predicted in a different orientation than they are present in the target. Here, DOMF allows us to quickly identify these regions and use them as a seed for superposition. Whether the domain orientation of the model is biologically feasible is a different question. As a second example, we consider the prediction from server11 for target 4HS7_B. Again, two parts of the target have been well-predicted in the model. Moreover, a relatively large fraction of the residues shares little similarity to the target structure, and is thus assigned to the non-consistent domain.



A



B

Figure 5.6 Homology model from server16 for CAMEO target 4HXT_A. The target is shown in white, the residues of the model used for the superposition are highlighted in orange, other residues in yellow.

Constraint Consistency for Multi-Template Modeling

The application of DOMF to the consistent extraction of constraints has been tested on 8 targets of the CAMEO web servers. Some models can be dramatically improved by using information from multiple templates, while others are best predicted using a single template. Which targets can profit from multiple-templates mainly depends on the available template structures. To get an idea how much models can be improved by combining information, we approximate the best-possible IDDT for each residue by using per-residue IDDT of single-template models. Clearly, for each residue, the best possible distance restraints are the ones from the template with the highest per-residue IDDT score. By setting the maximally possible IDDT for each residue to the largest per-residue IDDT in any of the single-template models, we get an approximate idea of the improvements that can be achieved by combining templates. This scheme ignores changes in IDDT which

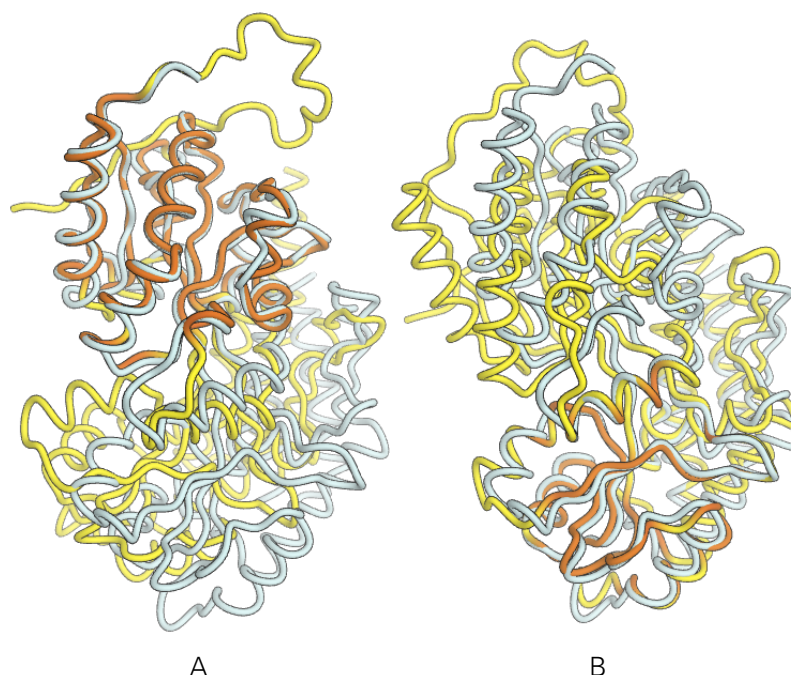


Figure 5.7 Homology model from server16 for CAMEO target 4HS7B_B. The target is shown in white, the residues of the model used for superposition are highlighted in orange, the other residues in yellow.

come from improvements in residues which are nearby.

The maximal per-residue IDDT together with the per-residue IDDT of the model built on the best template have been plotted in **figure 5.8**. The best-possible model is plotted as grey dotted lines, the best single-model template as grey solid lines. For most targets, the best-possible model is substantially better than the best single-template model. For example, for targets 3HKU_A, 3U7R_A, 4DQ2_B, the maximal possible per-residue IDDT is clearly distinguishable from the best single-template model. Other targets, e.g. 4FR9_A and 2LWE_A, can only be marginally improved and it is unlikely that multi-template modeling is of any help. Nevertheless, the latter have been left in the testset as a control: the multi-template modeling algorithm should not lead to significantly less accurate models than the single-template modeling.

Multi-template models have been calculated using the scheme outlined in Materials & Methods. The templates have been sorted by sequence similarity, and the top 20 templates have been used as seeds for the multi-template constraint selection. As a comparison, multi-template models have been calculated which use all constraints of the templates used for the consistent constraint multi-template models. In **table 5.2**, the results for the 8 CAMEO targets are shown. For both single and multi-template models, the models with the highest $C\alpha$ -, all-atom IDDTs and GDT_HA are shown.

For all 8 cases, the $C\alpha$ -IDDT for the consistency-enforced multi-template model is equal or higher than $C\alpha$ -IDDT of the best single-template model; for 6 out of these 8 targets, the $C\alpha$ -IDDT is higher. While for most of them, the improvement is marginal, the consistency-enforced model for 4H1X_A is substantially better. Similar results

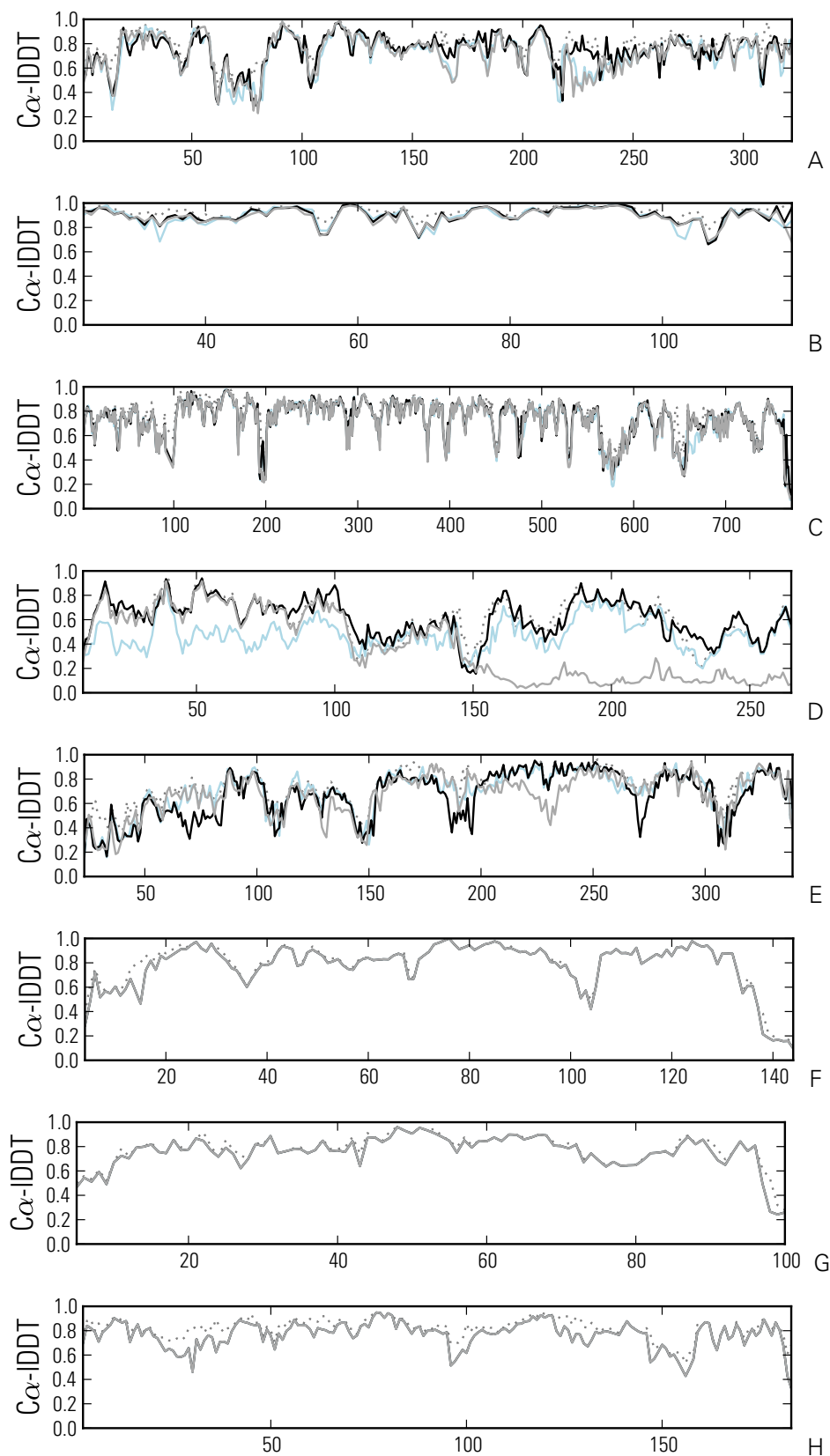


Figure 5.8 Possible improvement over single template model for the CAMEO test cases. (A) 4DQ2_B, (B) 4HL9_A, (C) 4G9I_A, (D) 4H1X_A, (E) 4EV6_E, (F) 4FR9_A, (G) 2LWE_A, (H) 3U7R_D. Per-residue $C\alpha$ -IDDT of best single-template model (grey line), maximal per-residue $C\alpha$ -IDDT of any single-template model (dotted grey line), best multi-template model with constraints consistency (black line), best multi-template model without constraints consistency (light blue line).

are seen for the all-atom IDDT. When compared using GDT_HA, for two targets, the single-template models are slightly better than then multi-template models (4G9I_A and 4HKU_A). The $C\alpha$ -IDDT and all-atom IDDT for the best multi-template model of target 4H1X_A is significantly higher than the single-model counterpart. Here, templates which cover different regions of the target sequence are successfully combined to extend the range of the models. The GDT_HA of the multi-template models are indistinguishable from the single-templates model, indicating that the relative orientation of the second domain has not been properly predicted. Enforcing constraints consistency is important and leads to a model with local accuracy in the N-terminal part comparable to the best single-template model. Combining the templates without enforcing constraints consistency leads to a dramatic loss of local model quality (**figure 5.8D**)

Apart from 4H1X_A, the largest difference is seen for target 4DQ2_B: the best possible single-template model achieves a $C\alpha$ -IDDT of 0.728 (all-atom IDDT 0.618). In comparison, the best multi-template model achieves a $C\alpha$ -IDDT of 0.756 (all-atom IDDT 0.633). The difference can be attributed to higher local per-residue IDDT values for residues around 215-240 (**figure 5.8A**). The remaining residues of the single and multi-template models are almost identical. In this example, constraint consistency improves the resulting model. When using all constraints, the resulting model is worse according to $C\alpha$ -IDDT, IDDT and GDT_HA.

target	IDDT			$C\alpha$ -IDDT			GDT_HA		
	single	cons	multi	single	cons	multi	single	cons	multi
4DQ2_B	0.618	0.633	0.600	0.728	0.756	0.722	0.422	0.476	0.407
4HL9_A	0.751	0.748	0.763	0.896	0.899	0.906	0.718	0.750	0.734
4G9I_A	0.638	0.641	0.630	0.750	0.754	0.748	0.518	0.515	0.503
4EV6_E	0.572	0.572	0.599	0.677	0.682	0.677	0.368	0.368	0.371
4FR9_A	0.657	0.657	0.656	0.787	0.787	0.787	0.566	0.566	0.557
2LWE_A	0.652	0.654	0.644	0.766	0.766	0.766	0.612	0.612	0.612
3U7R_D	0.655	0.655	0.655	0.771	0.771	0.770	0.559	0.565	0.559
4H1X_A	0.323	0.534	0.390	0.369	0.616	0.468	0.363	0.361	0.170

Table 5.2 Modeling accuracy of models built on single and multiple templates with and without consistency enforcement. IDDT, $C\alpha$ -IDDT and GDT_HA of models built on single template (single), multiple templates with (cons) and without (multi) constraints consistency to the target.

4 Conclusions

The DOMF algorithm identifies subsets of residues in ensembles of protein structures whose environment is similar. Starting from an adjacency matrix derived from $C\alpha$ - $C\alpha$ distance agreement, the edge weights between two $C\alpha$ -atoms are iteratively updated until convergence to a well-defined assignment of residues into domains.

DOMF can be applied to identify structural domains in experimental structures. In presence of domain movements, the structure is automatically partitioned into multiple *rigid blocks*. The performance of the algorithm was analysed on structures of the protein structural change database (PSCDB). The number of domains identified by DOMF and PSCDB is identical for the majority of cases. In general, the two methods find a similar domain partitioning. Nevertheless, there are a few notable differences: For a fraction of structures, the domains identified by DOMF are smaller and more compact. In order for two residues to be part of the same domain, their environment is required to be conserved in both structures. In contrast, PSCDB requires the residues only to superpose well.

The algorithm allows to detect well-predicted regions in models. These regions encompass sets of residues whose environment share considerable similarity to the target structure. DOMF finds such well-predicted residues even in presence of large structural deviations, illustrating the stability of the algorithm. This stability is an important feature, as it allows DOMF to be applied to pairs of remote homologs, where alignment errors and intrinsic structural differences in the packing of secondary structure elements are present. We are currently testing DOMF on more CAMEO targets, and, are planning on integrating DOMF into the CAMEO model summary page.

Furthermore, DOMF has been applied to extract consistent sets of constraints for the purpose of multi-template modeling. The constraint selection routine has successfully led to more accurate models. The algorithm outlined here is a proof of principle to illustrate that selection of consistent restraints is (a) feasible and (b) beneficial for generating more accurate models. More work in this direction will be required before turning the algorithm into a solid multi-template modeling program.

Toward the Estimation of the Absolute Quality of Individual Protein Structure Models

This chapter has been published as:

Benkert P.^{1,2}, Biasini M.^{1,2} and Schwede T.^{1,2} (2011). *Toward the estimation of the absolute quality of individual protein structure models*. *Bioinformatics*, 27, 343–350.

¹ Biozentrum, University of Basel, Klingelbergstrasse 50 / 70, 4056 Basel, Switzerland

² SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Author contributions: PB and MB implemented the software, conducted the research, TS, PB and MB wrote the manuscript.

MOTIVATION: Quality assessment of protein structures is an important part of experimental structure validation and plays a crucial role in protein structure prediction, where the predicted models may contain substantial errors. Most current scoring functions are primarily designed to rank alternative models of the same sequence supporting model selection, whereas the prediction of the absolute quality of an individual protein model has received little attention in the field. However, reliable absolute quality estimates are crucial to assess the suitability of a model for specific biomedical applications.

RESULTS: In this work, we present a new absolute measure for the quality of protein models, which provides an estimate of the ‘degree of nativeness’ of the structural features observed in a model and describes the likelihood that a given model is of comparable quality to experimental structures. Model quality estimates based on the QMEAN scoring function were normalized with respect to the number of interactions. The resulting scoring function is independent of the size of the protein and may therefore be used to assess both monomers and entire oligomeric assemblies. Model quality scores for individual models are then expressed as ‘Z-scores’ in comparison to scores obtained for high-resolution crystal structures. We demonstrate the ability of the newly introduced QMEAN Z-score to detect experimentally solved protein structures containing significant errors, as well as to evaluate theoretical protein models.

In a comprehensive QMEAN Z-score analysis of all experimental structures in the PDB, membrane proteins accumulate on one side of the score spectrum and thermostable proteins on the other. Proteins from the thermophilic organism *Thermatoga maritima* received significantly higher QMEAN Z-scores in a pairwise comparison with their homologous mesophilic counterparts, underlining the significance of the QMEAN Z-score as an estimate of protein stability.

AVAILABILITY: The Z-score calculation has been integrated in the QMEAN server available at: <http://swissmodel.expasy.org/qmean>.

1 Introduction

In homology modelling, the quality of a model is largely dictated by the evolutionary distance of the protein of interest (target) to the available template structures. The sensitivity of tools for detecting remote homologues with very low sequence identity has increased significantly in recent years due to the development of sophisticated algorithms^{29,32,201} and growth in sequence databases^{202–203}. However, with decreasing sequence similarity, an increasing amount of structural divergence is observed^{40,67}, and the resulting models

may contain significant inaccuracies, especially models built on distant templates. Typical sources of errors range from misplaced side chains, incorrect loop conformations, backbone distortions, alignment errors, to choice of a template with incorrect fold^{131,204–205}.

Ultimately, the accuracy of a protein model determines its suitability for biomedical applications. However, at the time of modelling the quality of a model is unknown and has to be predicted as well. For this purpose, scoring functions have been developed that evaluate different structural features of protein models in order to generate a quality estimate. Most scoring functions are primarily designed to rank alternative models of the same protein sequence^{63,103,106,108,115,206–211}. However, variability in model quality between different target proteins is typically by far larger than the variability within the ensemble of models generated by different prediction methods for a given protein^{51,205,212}. Therefore, relative ranking of alternative models for a given protein is insufficient for determining its usefulness for biomedical applications such as drug design, mutagenesis experiments, analysis of functional sites, etc. Reliable absolute quality estimates are crucial for the scientist intending to use computational models¹⁷¹.

The prediction of absolute model quality has rarely been addressed in the literature: the pioneering tool ProSA¹⁰⁰ has primarily been developed to evaluate experimental structures and estimates the statistical significance of a structure by comparing its score to random structures with the same sequence. The ProSA Z-score can hardly be used as a measure of absolute model quality as it is highly dependent on the protein size (i.e. the energy gap between the native fold and random decoy structures increases with protein size). Eramian and colleagues²⁰⁶ apply support vector regression to estimate the quality of models based on other modelling cases with similar properties selected from a large database of precompiled structure-model pairs generated by the same method. Wang et al.,²¹³ express the agreement of a model with several structural features predicted from the primary sequence as a reliability measure using the SCRATCH suite¹¹⁶. Most current scoring functions operate on individual protein chains and are not able to deliver quality estimates for biological assemblies.

In this work, we introduce a method for the estimation of the absolute quality of individual protein structure models which is independent of protein size and can be used to both assess isolated chains as well as entire oligomeric assemblies. The absolute quality is estimated by relating the model's structural features to experimental structures of similar size. Based on our recently introduced composite scoring function QMEAN^{108,128}, we analyse different geometrical aspects of proteins. For normalization, the QMEAN score of a model is compared to distributions obtained from high-resolution structures solved by X-ray crystallography. The resulting 'QMEAN Z-score' provides an estimate of the 'degree of nativeness' of the structural features observed in a model and indicates whether the model is of comparable quality to experimental structures. The Z-scores of the individual terms of the scoring function indicate which structural features of a model exhibit significant deviations from the expected 'native' behaviour, e.g. unexpected solvent accessibility, back-bone geometry, inter-atomic packing, etc.

We first describe normalized statistical potential terms and introduce length-corrected QMEAN scores. We then calculate normalized QMEAN scores on all experimental structures from the PDB, and provide an analysis of proteins exhibiting unusually low and

high values. We finally introduce the concept of the QMEAN Z-score and demonstrate the strength of the new score for evaluating both experimental structures and theoretical models.

2 Methods

QMEAN

QMEAN is a scoring function consisting of a linear combination of six structural descriptors as described elsewhere in more detail^{108,128}. In short, two distance-dependent interaction potentials of mean force based on C- β atoms (i.e. residue-level) and on all atom types are used to assess long-range interactions—both are secondary structure dependent; a torsion angle potential over three consecutive amino acids is applied to analyse the local back-bone geometry of the structure and a solvation potential to describe the burial status of the residues; finally, the agreement of predicted and calculated secondary structure and solvent accessibility is included in the form of two agreement terms. Secondary structure prediction is performed by PSIPRED⁸ and solvent accessibility prediction with ACCpro¹¹⁶. The secondary structure and solvent accessibility of the model are calculated by DSSP⁷. While the agreement terms have a significant impact on the performance of QMEAN on theoretical models, they do not add additional information when experimental structures are evaluated. Evaluations on experimental structures are therefore based on the normalized QMEAN4 score (i.e. statistical potential terms only).

The optimization of the weighting factors for the terms contributing to QMEAN has been performed on models from the seventh round of the CASP experiment (CASP7)²¹². To evaluate the performance on an independent dataset, QMEAN has been applied on all server models submitted to CASP8. The length-normalized statistical potentials scores are calculated as follows: the scores of single body potentials (solvation potential and torsion angle potential) are normalized by the number of residues and the scores of the non-bonded interaction potentials (all-atom and C- β potential) are divided by the total number of interactions.

GDT_TS values for the benchmark were parsed from the CASP8 website and quality assessment predictions downloaded from:

http://predictioncenter.org/download_area/CASP8/predictions/QA.tar.gz.

Datasets

PDB TRAINING SET: the statistical potentials were extracted from a non-redundant set of high-resolution structures from the PDB²¹ selected using the PISCES server²¹⁴. A pairwise sequence identity cut-off of 20% is applied and only structures solved by X-ray

crystallography with a resolution better than 2 Å and R-value below 0.25 are included, resulting in a total number of 3544 chains.

CASP7 TRAINING SET: the weighing factors of the QMEAN composite score were optimized based on CASP7 models (human and server)²¹² using the GDT_TS score as target variable⁴¹. From the initial set of 47 214 evaluated models, incomplete models covering <95% of the target sequence or lacking side-chain atoms for >10% of the amino acids were removed. The final CASP7 training set contains 34 322 models from various modelling servers.

CASP8 TEST SET: a total of 31 491 server models from CASP8 were used as an independent test set for the comparison of different implementations of QMEAN and for assessing the performance of the QMEAN Z-score.

PDB REFERENCE SET: a non-redundant reference set of high-resolution PDB structures for the QMEAN Z-score calculation was generated by PISCES using the following criteria: structures longer than 30 amino acids solved by X-ray crystallography, with pairwise sequence identity below 40% and resolution better than 2.5 Å were included, resulting in 9766 structures. Proteins annotated as transmembrane proteins²¹⁵ were excluded. Also, 18 low-scoring outliers showing a normalized QMEANscore (without agreement terms) deviating by more than 3 standard deviations were excluded from the PDB reference set. A complete list of these structures with high scores is provided as Supplementary Data Table S1. The final 'PDB reference set' contains 9451 entries.

BIOLOGICAL UNIT REFERENCE SET: this set contains the biological assemblies of all chains from the PDB reference set. The PISA database²¹⁶ was used to assign the most likely oligomeric state and generate the coordinates of the assembly for all entries of the dataset. The resulting set contains biological units from 9062 unique PDB identifiers—2999 of them are monomers. A 'biological active assembly' may contain multiple chains from the non-redundant chain list.

QMEAN Z-score

To calculate the QMEAN Z-score, the normalized raw scores of a given model (composite QMEAN score and individual mean force potential terms) are compared to scores obtained for a representative set of high-resolution X-ray structures of similar size (number of residues of query proteins $\pm 10\%$). For the analysis of isolated chains, the 'PDB reference set' is used and oligomeric assemblies are evaluated against the 'biological unit reference set'. The same procedure is applied to calculate Z-scores for the agreement terms, i.e. for each structure in the two reference sets PSIPRED and ACCpro have been applied to model the background distribution of expected secondary structure and solvent accessibility prediction accuracy.

The raw QMEAN score and the individual terms have different scales and algebraic signs: QMEAN and agreement terms range from 0 to 1 and the statistical potential terms deliver pseudo energies with negative values for energetically favourable states. In the Z-score calculations, we adjusted the sign of the statistical terms such that higher Z-score consistently relate to favourable states, i.e. higher QMEAN Z-score means better agreement with predicted features and lower mean force potential energy.

Cross-validation

In order to investigate the saturation of the statistics in the QMEAN score calculation and to exclude over-training, a cross-validation experiment has been performed in the form of a leave-1/3-out experiment on the original dataset used to extract the statistical potentials (i.e. the PDB training set). We trained the statistical potentials on 2/3 of the proteins from the original training set and applied the QMEAN score on the remaining 1/3 of the structures. We randomly selected 31 complete SCOP fold classes making up roughly 1/3 of the original set (1523 PDB chains). This results in two sets having no overlap in terms of folds. If the statistics is saturated, the predicted scores of a structure from the test set should not differ considerably between the two potentials implementations, i.e. the one based on the full and the reduced training set. The cross-correlation coefficient between the original QMEAN and the QMEAN score trained on 2/3 of the training set is 0.88 which underlines that the QMEAN score calculation is robust and does not change strongly if applied on folds not used in the generation of the statistical potentials (Figure S7 in the Supplementary Data).

Comparison of predicted protein stability between thermophilic and mesophilic organisms

The dataset of pairs of homologous proteins by Robinson-Rechavi and Godzik²¹⁷ has been used. Three protein structures have in the meantime been superseded by newer entries in the PDB: 1un7 has been replaced by 2vhl, 1nrh by 1u8x and 1jsv by 2afb. One pair of homologues of the original dataset has been excluded (1g6p from *T. maritima*, 1c9o from *Bacillus caldolyticus*) since both are from thermophilic organisms. The final dataset consist of 72 protein pairs.

As in the work of Robinson-Rechavi and Godzik, proteins were shortened according to a structural alignment (FATCAT w/o flexibility,²¹⁸) in order to get homologous protein pairs of similar size.

Implementation

The current version of the QMEAN scoring function has been implemented based on the open source molecular modelling and visualization framework OpenStructure¹³⁵.

3 Results

Normalization of the statistical potentials

Statistical potential scores are calculated as a sum of microstates and therefore have a strong dependence on the size of the assessed protein structure and larger proteins tend

to have lower energies (i.e. higher QMEAN scores). As long as similarly sized models are compared, the strong size dependence does not present an issue. However, it renders the prediction of absolute quality difficult when only looking at single models.

In this work, we introduce normalized statistical potential QMEAN terms. In order to correct for the length dependence of the statistical potentials scores, the scores of single body potentials (solvation potential and torsion angle potential) are normalized by the number of residues and the scores of the two non-bonded interaction potentials (all-atom and C- β potential) are divided by the total number of interactions considered in the calculation. **Figure 6.1** shows the effect of the normalization on the all-atom potential. The all-atom energies of a non-redundant set of 9766 protein structures (single chains) solved by X-ray crystallography are calculated as normalized and non-normalized scores.

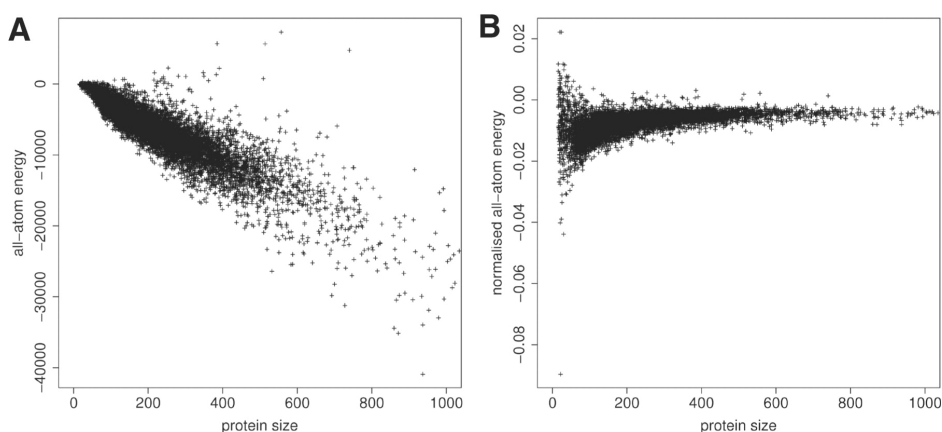


Figure 6.1 Comparison between traditional (A) and normalized all-atom interaction score (B) on a non-redundant set of 9766 high-resolution PDB chains.

A clear correlation with protein size is observed (**Fig. 6.1 left**) for the standard all-atom potential whereas the average energy per interaction of the normalized potential converges to an average value of -0.0058 ± 0.0017 (**Fig. 6.1 right**). This is in accordance with recent results of Thomas et al.²¹⁹ who report an average stability value for protein folds. Smaller proteins adopt a wider range of average per-interaction energies in accordance with the fact that small peptides often exist as a diverse ensemble of conformations or are stabilized in larger complexes. Indeed, the peptides with the highest, i.e. most unfavourable, energies in the dataset are the ribosomal protein THX [PDB:2vqe,²²⁰] and the disordered protein hypocretin presented on a MHC class II protein [PDB:1uvq²²¹]. On the other side of the energy spectrum, we observe three peptide hormones namely hepcidin [PDB:3h0t²²²], endothelin-1 [PDB:1edn²²³] and relaxin [PDB:6rlx²²⁴] with predicted per-interaction energies far below the average value reported above. Energy values and their interpretation are given in Table S2 in the Supplementary Data.

We analysed protein chains with more than 100 amino acids having high predicted interaction energies. The 27 protein chains with highest average interaction values (more precisely those with positive per-interaction energies) all are membrane proteins. These results confirm that the structural features of membrane proteins do not follow the same distribution as proteins in solution, i.e. atomic interactions in membrane proteins and

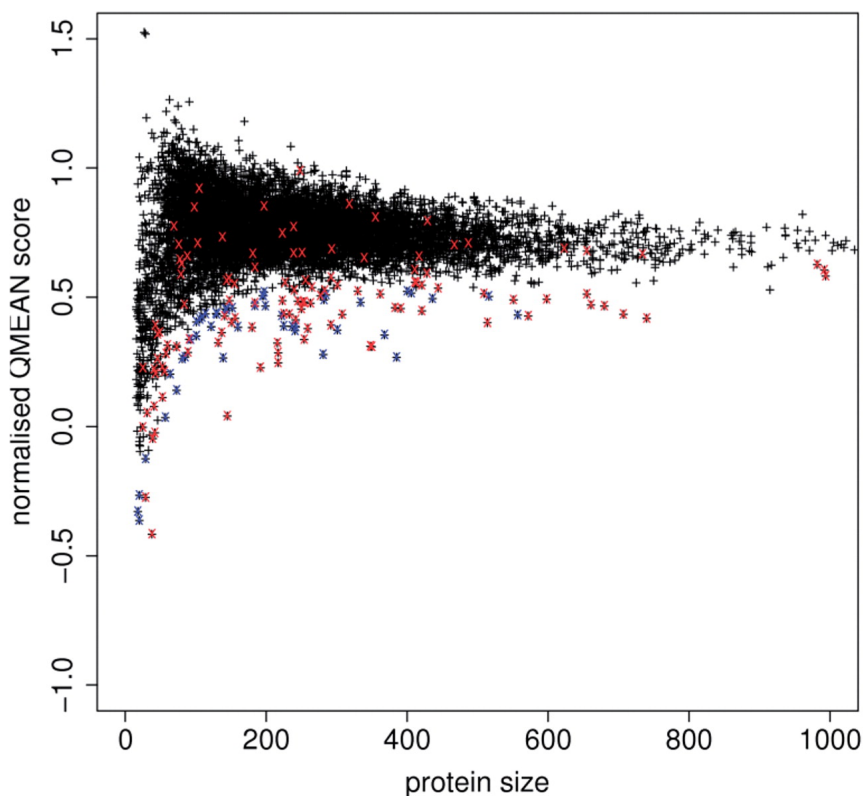


Figure 6.2 Normalized QMEAN score composed of four statistical potential terms (QMEAN4) of 9766 high-resolution structures. Red crosses indicate chains belonging to membrane proteins, blue crosses denote other proteins deviating by more than 3 standard deviations (see Supplementary Table S1 for details).

their solvation properties differ considerably from those found in soluble proteins. We decided that these proteins are better treated in a specialized mean force potential. A variant of the QMEAN score for membrane proteins is currently underdevelopment.

In analogy to the all-atom term, the other three statistical potentials of QMEAN have been normalized and for larger proteins show convergence to an average per residue energy, although with a higher variance (see Figure S1–S3 in the Supplementary Data). The same is true for the composite score of the four statistical potentials scores (QMEAN4, **Fig. 6.2**). In the course of the article, ‘QMEAN’ denotes the complete scoring function consisting of six terms based on normalized potentials. The version of the scoring function based on statistical potentials only is denoted as QMEAN4 in the following.

PDB reference set and QMEAN Z-score concept

In analogy to the average energy per interaction, the average normalized QMEAN4 score is constant over a wide range of protein sizes, i.e. experimental structures adopt a relatively narrow distribution of QMEAN4 scores. While the average normalized score is constant, the variance of the distribution depends on protein size (**Fig. 6.2**).

These observations lead to the idea to use a non-redundant set of protein structures as a reference to evaluate the quality of individual protein structures and models, i.e. the PDB reference set. The dataset contains 9451 non-redundant high-resolution structures, excluding membrane proteins and energetic outliers (highlighted in **Fig. 6.2**). A QMEAN Z-score for a given model is thereby calculated from its normalized QMEAN score by subtracting the average normalized QMEAN score and divided by the standard deviation of the observed distribution. In analogy, Z-scores are calculated for all individual terms of the composite score. In order to facilitate the interpretation, we standardize the algebraic sign of the calculated Z-scores such that higher Z-scores relate to more favourable models.

In the following, we first illustrate the application of QMEAN Z-score for quality estimation on two example proteins, representing a ‘good’ and a ‘bad’ experimental structure. We then extend our analysis on the entire PDB (single chains) and report outliers. The QMEAN Z-score concept is then extended from chains to entire biologically relevant oligomeric assemblies. Finally, we show that the new score can be used as a measure of absolute model quality in the assessment of theoretical models.

QMEAN Z-score analysis of experimental structures

We have applied QMEAN Z-scores to experimental structures from the PDB database²¹. **Table 6.1** and Supplementary Figure S5 show the Z-scores analysis of two experimental structures solved by X-ray diffraction: bacteriophage T4 lysozyme [PDB:2lzm,²²⁵] and Dengue virus NS3 serine protease [PDB:1bef,²²⁶]. The QMEAN Z-score of the lysozyme structure is 0.5, i.e. the score of the structure is clearly within the expected quality range as it deviates less than 1 standard deviation from the mean score in similar sized high-quality proteins from the reference dataset. In contrast, the structure of the NS3 serine protease has a QMEAN score deviating by more than 5 standard deviations indicating that there is clearly something wrong with this structure. Both the composite QMEAN score, as well as all individual terms deviate strongly from expected values (Figure S5, Supplementary Data and **Table 6.1**). Indeed, this structure, as well as several other structures from the same group, have been identified as fabricated and have been retracted (see <http://www.wwpdb.org/UAB.html>). A QMEAN Z-score analysis of all affected structures can be found in Supplementary Table S3.

PDB	QMEAN	C-B	All-atom	Solvation	Torsion
T4 lysozyme, 2lzm	0.5	0.6	1.1	0.7	-0.3
Serine protease, 1bef	-5.5	-3.4	-3.6	-2.7	-4.1

Table 6.1 Z-score analysis of the T4 bacteriophage lysozyme (2lzm, chain A) and the Dengue virus NS3 serine protease (1bef, chain A). Both the QMEAN Z-score as well as the Z-scores of individual statistical potential terms are reported. All structural properties of 1bef deviate significantly from expectation values obtained from high-resolution structures. In the meantime, the structure has been retracted from the PDB.

PDB	Size	QMEAN	C-B	All-atom	Solvation	Torsion
2q97A	354	-1.2	-1.2	-0.5	-0.5	-1.0
2q97T	109	-3.3	-2.2	-1.8	-3.7	-1.1
PISA ^a	926	-1.6	-0.9	-0.6	-1.2	-1.0

Table 6.2 Z-score analysis of the toxofilin/actin both for the isolated chains and as well as the biological assembly defined by PISA. ^a Most probable assembly as proposed by PISA: a tetramer consisting of two copies of the chains A and T. Especially the solvation energy and C- β potential terms exhibit large differences between the Z-score of the isolated toxofilin monomer and the complex with actin.

Group name	Targets	Global r	Mean r	P-value	Mean Δ GDT_TS	P-value
MULTICOM-REFINE	122	0.786	0.729	0.258	0.093	0.936
GS-MetaMQAP	121	0.779	0.708	2.13E-005	0.132	0.004
QMEANnorm	122	0.774	0.738		0.093	
QMEANfamily	107	0.751	0.755	0.0016 ^a	0.089	0.260
QMEAN	121	0.750	0.724	0.017	0.088	0.430
MULTICOM-CMFR	122	0.740	0.759	0.063	0.083	0.227
Bilab-UT	121	0.728	0.693	2.71E-005	0.107	0.239
MULTICOM-RANK	122	0.711	0.708	0.004	0.082	0.178
ModFOLD	122	0.686	0.616	6.24E-022	0.137	0.001
BMF_PP	96	0.683	0.615	9.25E-019	0.197	2.06E-007
SIFT_consensus	117	0.678	0.686	5.16E-007	0.106	0.117
circle	122	0.665	0.712	0.002	0.111	0.143
Pcons_ProQ	122	0.656	0.667	4.55E-010	0.129	0.003
DistillSN	120	0.655	0.476	6.60E-027	0.207	2.76E-012
MUFOLD-QA	122	0.583	0.645	2.14E-010	0.117	0.051
DISTILLF	118	0.581	0.650	2.07E-011	0.141	0.001
MODCHECK-HD	122	0.506	0.304	3.76E-043	0.155	2.40E-006
SELECTpro	122	0.503	0.635	3.22E-014	0.153	6.75E-005
Fiser-QA	121	0.502	0.564	5.24E-019	0.186	4.82E-008
Fiser-QA-COMB	121	0.478	0.521	7.94E-023	0.228	1.31E-010
SIFT_SA	112	0.469	0.636	8.98E-011	0.115	0.086
Fiser-QA-FA	121	0.331	0.524	2.96E-029	0.191	2.91E-007
qa-ms-torda-server	117	0.106	0.058	1.22E-052	0.487	2.92E-034
ProtAnG_s	121	0.081	0.124	4.31E-059	0.139	0.001

Table 6.3 Comparison of normalized QMEAN potentials (QMEANnorm) with single model scoring function of CASP8. Global r: correlation against GDT_TS over all models from all targets; mean r: r averaged over individual targets; mean Δ GDT_TS: average deviation of model with best score and best model. The statistical significance of the difference is measured with a paired t-test on common targets (significantly better performance of QMEAN marked in italic, significance level: 0.05). ^a QMEANfamily is significantly better than QMEAN in ranking models. Performance of the methods described in this work (QMEANnorm) is highlighted in bold.

Scoring Function Term	Wilcoxon	T-test
C- β interaction potential	0.0029	0.0020
All-atom interaction potential	0.0322	0.0293
Solvation potential	0.0031	0.0020
Torsion potential	0.0051	0.0105
QMEAN	0.0001	0.0001

Table 6.4 Analysis of 72 pairs of homologous proteins from *Thermotoga maritima* and corresponding mesophilic organisms²¹⁷. The P-values in two statistical tests (Wilcoxon and t-test) on paired samples are reported. The proteins of thermophilic and mesophilic organisms differ significantly in terms of all QMEAN components.

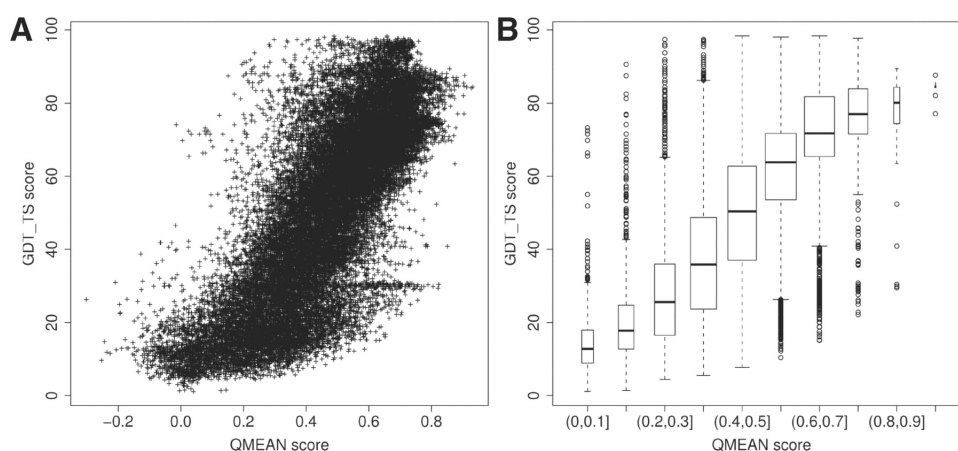


Figure 6.3 Correlation between QMEAN and GDT_TS for all server models of CASP8. **(A)** Scatter plot, **(B)** boxplot.

The ProSA²²⁸ analysis of the two structure can be found in Supplementary Figure S6. The lysozyme structure receives a very low Z-scores of -8.7 . The fabricated structure 1bef, however, also deviates by almost 4 standard deviations from random structures (Z -score = -3.74). In comparison to QMEAN, the score of this model does not differ considerably from many other structures in the PDB. In contrast to QMEAN, the ProSA Z-score shows a clear correlation with protein size which limits its application as an absolute quality measure. We therefore think that a comparison to high-resolution structures instead of random conformations is more meaningful.

We performed the QMEAN Z-score analysis on 144 142 protein chains from the PDB. Of these chains, 134 604 were solved by X-ray diffraction, 7979 by NMR and 1559 by electron microscopy. The Z-score distributions for structures derived by the three different methods show considerable differences (Supplementary Figure S4). The average QMEAN Z-scores are -0.58 for X-ray diffraction, -1.19 for NMR and -2.00 for EM. Among the protein chains solved by X-ray crystallography we observed 1'048 chains (belonging to 417 PDB entries) with a QMEAN Z-score less than -5 . The majority of these proteins were either transmembrane or ribosomal proteins: 61 membrane proteins, 99 oxidoreductases, 109 proteins involved in photosynthesis, 46 transporters and

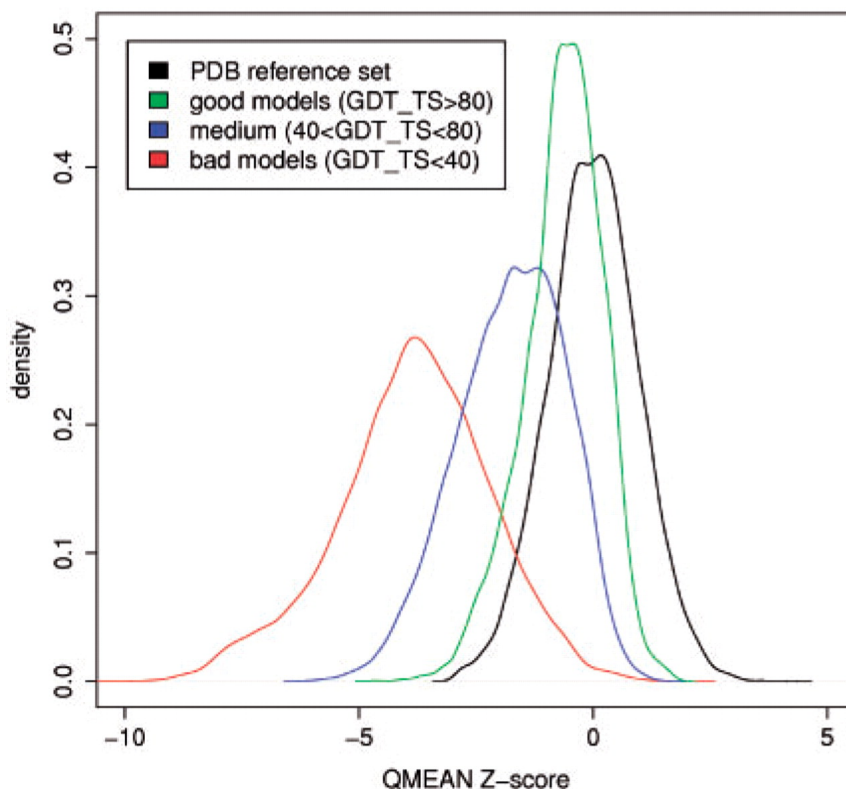


Figure 6.4 Density plot visualizing the QMEAN Z-score distribution of theoretical protein structure models. Z-scores for models from CASP8 are shown in relation to scores of experimental reference structures (black line). The models are split into three quality ranges with low-quality models in red, medium-quality models in blue and good models in green.

55 ribosomal proteins. These numbers underline the importance of a separate treatment of proteins embedded in membranes or bound to RNA, e.g. ribosomes. The remaining 48 proteins with unfavourable QMEAN Z-scores are provided in the Supplementary Data (Table S4). The majority of these structures are of quite low resolution: 79% of the proteins were solved at a resolution $<3 \text{ \AA}$.

To this point, we have applied the Z-score formalism on isolated protein chains. However, many proteins are part of oligomeric complexes and analysing protein stability on the level of isolated chains does not capture the physiologically relevant situation in the cell. We have therefore extended our analysis to complete oligomeric assemblies.

Figure 6.6 illustrates this effect on the example of toxofilin in complex with mammalian actin [PDB:2Q97,²²⁷]. In the complex toxofilin (chain T, blue) adopts a non-globular conformation, which is meaningless in isolation. As expected, the QMEAN Z-score of -3.3 for toxofilin (chain T) in isolation is unfavourable, especially the solvation and the $C-\beta$ interaction terms exhibit large differences between the Z-score of the isolated toxofilin monomer and the complex with actin (**Table 6.2**).

The biological unit reference set contains the most likely biologically relevant oligomeric assembly generated by PISA²¹⁶. **Figure 6.5** shows the QMEAN scores of 9062 oligomeric entries of the biological unit reference set (see Section 2). This dataset is used as a ref-

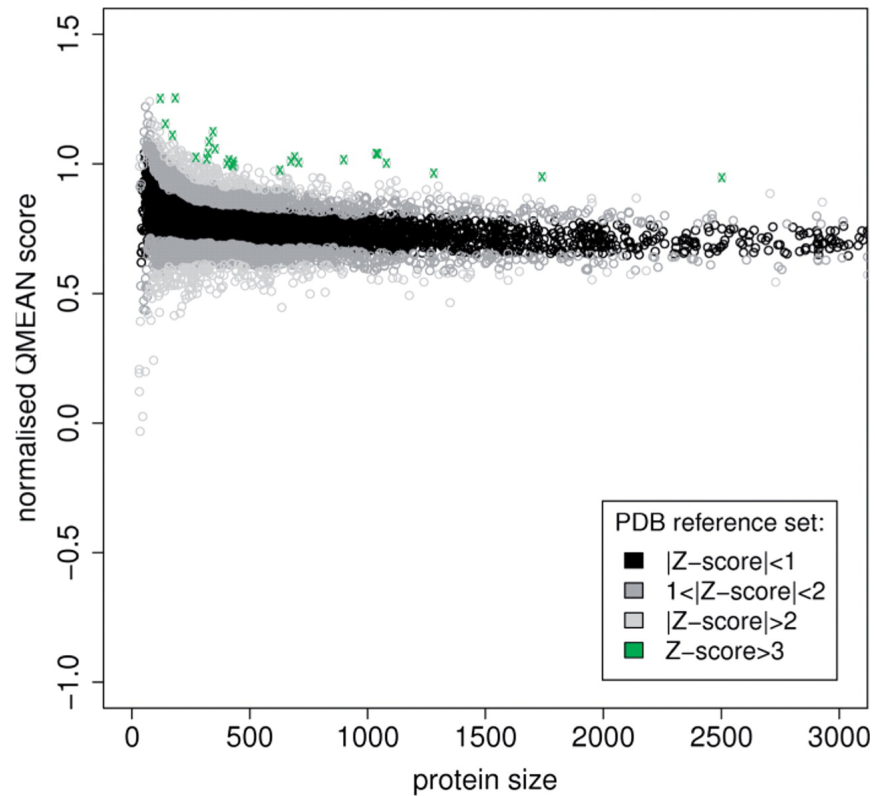


Figure 6.5 QMEAN scores for all structures in the biological unit reference set. Proteins with unusually high QMEAN scores (Z-score >3) marked in green correspond almost exclusively to proteins from thermophilic organisms (see also Supplementary Table S5).

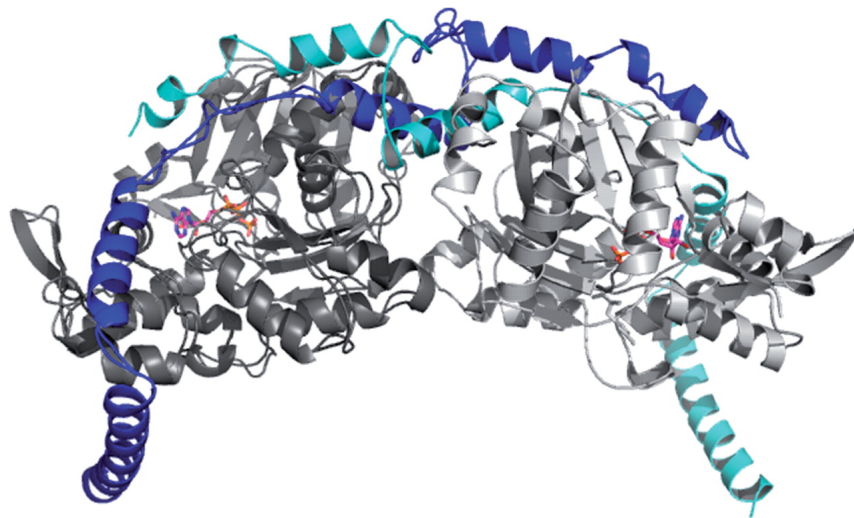


Figure 6.6 Oligomeric complex of mammalian actin (in grey) with toxofilin (chain T, blue) from *Toxoplasma gondii* [PDB:2Q97;²²⁷]. In the complex toxofilin adopts a non-globular conformation, which is meaningless in isolation. As expected, the QMEAN Z-score of -3.3 for toxofilin in isolation is unfavourable (**table 6.2**).

erence set for the assessment of complexes and oligomeric proteins. All structures with extraordinarily high QMEAN scores (Z-score >3 standard deviations; 26 structures) are highlighted with green crosses. Interestingly, 22 out of these 26 are proteins from thermophilic to hyperthermophilic bacteria and archaea, two are designed proteins optimized for stability, and the remaining two are structural genomics targets of unknown function (Supplementary Table S5).

In summary, the proteins at the periphery of the QMEAN score spectrum can be assigned to membrane proteins which exist in a fundamentally different environment compared to soluble proteins and extremely stable proteins found in thermophilic organisms.

Comparison of homologous proteins from thermophilic and mesophilic organisms

The composite scoring function QMEAN seems to capture structural features which distinguish thermostable proteins from proteins in mesophilic organisms. In order to further investigate which terms are most discriminative, we applied QMEAN on a published dataset of pairs of proteins from *Thermatoga maritima* and corresponding homologues from mesophilic organisms²¹⁷. Out of 72 protein pairs, QMEAN assigns in 75% of the cases higher scores to the proteins from *T.maritima*. Over the entire data set, the difference between the QMEAN scores assigned to mesophilic and thermophilic proteins is highly significant ($P = 0.0001$, see **table 6.4**). The comparison is illustrated in form of a diagonal plot in Supplementary Figure S8. These findings indicate that the QMEAN score indeed may be understood as a measure of protein stability.

In agreement with a study on *Thermatoga maritima* in which the authors identified salt bridges and compactness as major determinants of protein stability²²⁹, we observe that the solvation potential and the interaction potentials on residue-level are the most discriminative terms ($P = 0.002$ for both terms in paired t-test).

Analysis of theoretical models using normalized QMEAN scores and QMEAN Z-scores

In the following, the normalized QMEAN scoring function is applied on theoretical models from CASP8 and its performance is compared to other methods. We demonstrate the value of the QMEAN Z-score as a statistically well-founded measure of absolute quality and end with a critical discussion of the limitations of this approach for predicting absolute local per-residue errors.

CASP data is a good testing ground for scoring functions since it includes models spanning a wide range of quality generated by a variety of different modelling algorithms. Figure ?? shows the global correlation between the size-normalized QMEAN score and the GDT_TS distance to the native structure of all CASP8 server models. A global correlation coefficient of 0.77 overall CASP8 models is obtained. QMEAN6 scores perform significantly better than QMEAN4 to estimate the quality of predicted structures (correlation on CASP8 data was 0.77 versus 0.66). While for assessing experimental structures,

the agreement terms do not provide additional value, these terms are especially effective in the medium to low model quality range²³⁰.

Table 6.3 shows a comparison of the normalized QMEAN scoring function (denoted as QMEANnorm) with methods in the quality estimation category of CASP8²³¹. Only scoring functions operating on individual models are used (i.e. no consensus methods and methods using structural information from homologous proteins). Compared to the original QMEAN scoring function, the normalized QMEAN shows a considerably better global correlation to GDT_TS which forms the basis for absolute quality predictions. The new version is also significantly better in ranking the models ($P = 0.017$) while the difference in picking the good models (mean delta GDT_TS of selected and best model) is not significant ($P = 0.43$). MetaMQAP and MULTICOM-REFINE have a slightly better global r but the former performs significantly worse in model ranking/selection. In terms of global correlation, the three methods perform equally well on easy targets (mean GDT_TS of top 5 models greater than 50) but QMEAN performs worse on the harder ones (see Supplementary Tables S6 and S7). The performance of QMEAN with respect to other state-of-the-art methods such as ProSA¹⁰⁰ and DFIRE¹⁰⁶ has also been recently assessed in an independent study²³⁰. QMEAN was found to be the best performing method in terms of the selecting the best model.

The robustness of the QMEAN Z-score on experimental structures lead us to apply the same concept to describe the absolute quality of theoretical protein structure models. Large deviations from expected values of experimental reference structures may be an indicator for modelling errors. The significance of the deviation as expressed by the QMEAN Z-score provides a quantitative and statistically well-founded measure of model reliability and therefore represents an absolute quality estimate of the model. (Note that the Z-score formalism does not affect QMEAN's ability to rank and select models.)

Figure 6.4 visualizes the differences in the QMEAN6 Z-score distributions between experimental structures of the PDB reference set (black line) and the CASP8 server models coloured according to model quality ranges (i.e. the GDT_TS distance to the native structure). The Z-score distribution of low-quality models with GDT_TS below 40 is clearly shifted towards lower Z-scores compared to experimental structures (mean Z-score = -3.85). Only a small overlap of the distributions is observed: 85% of the bad models with a Z-score above -2 are small structures below 150 residues. As can be seen in **Figure 6.2**, the variance of the QMEAN score increases with decreasing size and as a consequence the separation between good and bad structures becomes less pronounced (see also Supplementary Figure S9). Another reason for the overlap of the distributions is that 36% of the overlapping bad models are incomplete with $<80\%$ residues resolved which lowers the GDT_TS score but not the normalized QMEAN score. The 'good' models depicted in green reach QMEAN Z-scores comparable to experimental structures (mean Z-score = -0.65) and the 'medium' quality models (in blue) are located in between (mean Z-score = -1.75). A clear correlation between the GDT_TS distance of the model to target structure and the QMEAN Z-score for all CASP8 server models larger than 150 residues is observed underlining the suitability of the QMEAN Z-score as an estimate of model quality (**Fig. 6.3** and Supplementary Fig. S10).

The prediction of local (per-residue) error estimates is an active field of research. For

our previously introduced local QMEAN score (QMEANlocal)¹²⁸, normalized interaction potentials lead to a slight performance increase (data not shown). However, the precision of current local scoring functions applied on single models is not sufficient as reliable absolute quality estimate. Nevertheless, a distinction between more and less deviating regions is still possible. Supplementary Figures S11 (boxplot) and S12 (ROC analysis) show the performance of QMEANlocal in estimating per-residue errors on all CASP8 models. Only a weak correlation between local score and C- α deviation exists. The ROC analysis shows that QMEANlocal is able to enrich residues from the models with low deviation from the native structure. More than half of the residues with a calculated C- α deviation below 2.5 Å are identified among the 10% best scoring residues.

4 Conclusions

In this work, we present a new method for estimating the absolute quality of a single protein structure, i.e. without including additional information from other models or alternative template structures. The measure is based on the composite scoring function QMEAN which evaluates several structural features of proteins. The absolute quality estimate of a model is expressed in terms of how well the model score agrees with the expected values from a representative set of high resolution experimental structures. The resulting QMEAN Z-score is a measure of the ‘degree of nativeness’ of a given protein structure. The Z-scores of the individual components of the composite QMEAN score point to structural descriptors that contribute most to the final score, and thereby indicate potential reasons for ‘bad’ models.

A large-scale benchmark of experimental structures revealed two groups of proteins on the periphery of the QMEAN score distribution: on one side there are membrane proteins whose structural integrity is maintained by the lipid bilayer and as a consequence their physico-chemical properties differ considerably from those of soluble proteins. On the other side of the QMEAN score spectrum, proteins from thermophilic organisms are predominant. In a direct comparison of pairs of homologous proteins, proteins from thermophilic organisms receive significantly higher QMEAN scores compared to their mesophilic counterparts.

Finally, we show that the QMEAN Z-score is a useful measure for the description of the absolute quality of theoretical models and is a valuable measure for identifying experimental structures with significant errors. Compared to most existing scoring functions, QMEAN Z-scores can be both applied on isolated chains or biological assemblies.

The QMEAN Z-score calculation has been integrated in the QMEAN server¹²⁸^A, and the ‘Structure Assessment’ tools of SWISS-MODEL Workspace^{130,232}^B. A stand-alone version is available on request from the authors.

^A <http://swissmodel.expasy.org/qmean>

^B <http://swissmodel.expasy.org/workspace/>

5 Acknowledgements

The authors thank Florian Kiefer for his support with generating the oligomeric assemblies from PISA and all members of the group for fruitful discussions.

FUNDING: The development of QMEAN Z-score has been supported by the SIB Swiss Institute of Bioinformatics; Biozentrum der Universität Basel, Switzerland.

CONFLICT OF INTEREST: none declared.

QMEANdist - QMEAN Enhanced with Distance Restraints from Alignments

Quality estimation of protein structure models is important for many stages of homology modeling, e.g. to rank alternative models for the same sequence, or to judge the quality of a model on an absolute scale. Previously, we have introduced QMEAN, a composite scoring function combining potentials of mean force with terms measuring the agreement of observed and predicted sequence features. During CASP8, we have found QMEAN to assign low energies to models of certain groups, even if the models share little similarity to available template structures. These models were heavily optimized with potential of mean force and QMEAN is unable to judge their quality with confidence. In this work, we present QMEANdist, a variant of QMEAN which tries to address this limitation by incorporating information from evolutionary related protein structures. The structures define an C α -C α distance propensity, by which models are scored. QMEANdist has been extensively validated during the CASP9 experiment as part of quality estimation category. We show, that the method delivers performance comparable to consensus scoring functions for ranking and model selection.

1 Introduction

In theoretical modeling protein structure models are inferred from sequence in absence of experimental data²⁰⁷. The structural energy landscape of a protein is explored by various means, e.g. by starting from a related template structure, or ab-initio folding techniques. Since the usefulness of a model directly correlates with its quality, the ability to judge how good a model is not only of academic but also of biological interest¹⁷¹. Model quality assessment is thus an important aspect of every modeling endeavour.

CASP (critical assessment of techniques in structure prediction) is a biennial double-blind experiment to objectively compare competing approaches in the theoretical modeling field^{45–53}. Model quality assessment servers are evaluated separately in the model quality assessments (QA) category. QA methods at CASP can be broadly categorized into 3 groups: consensus-based, single-model and quasi-single model methods.

Physics- and knowledge-based quality estimation programs employ potentials of mean force, force fields, and agreement terms of predicted sequence features^{100,103,106,120,142} to estimate the quality of models. These methods are referred to as single-model quality estimation programs, as they evaluate the quality of each model separately.

Previously, we have developed the single-model MQA program QMEAN^{108,142}. It uses a linear combination of potential of mean force terms and two terms comparing predicted and observed sequence features. QMEAN has been shown to be one of the top-performing single-model quality estimation programs at CASP8. While single-model quality estimation have their benefits and are especially appealing from an academic perspective, the employed scoring schemes suffer from the same limitations as traditional force fields and scoring functions: They are often unable to distinguish between near-native and non-native conformations^{81–83}. During CASP8, we have found QMEAN to

systematically overpredict the quality of models from certain structure prediction servers. For example, RBO-Proteus uses potentials of mean force to optimize the conformation of the models. From the perspective of a potential of mean force, these models were *convincing*, even though they were far from the native state.

Pcons¹²⁰ and similar methods have been fairly successful at predicting the quality of models using consensus features extracted from all models submitted for a target sequence^{108,120,233}. Since structural features more commonly found in the models are more likely to be correct, the average structure is assigned the highest score. The top-performing consensus servers at CASP additionally include pre-filtering steps to select a fraction of the models for clustering¹⁰⁸. In the context of CASP, consensus methods are by far the most successful in terms of performance. Two reasons contribute to the success: First, the number of models per target exceeds 300 and as such sufficiently large to obtain statistics of structural features. Second, the models at CASP are built by different modeling methodologies. Errors associated with these methods, e.g. alignment errors, structural sampling biases, tend to cancel out. In an ad-hoc experiment, Kryshtafovych *et al.* have shown that the performance of consensus methods for quality estimation rapidly decreases when less models are available¹¹⁸. For many homology modeling projects, it is not feasible to obtain such large number of models for the same target sequence, especially with different modeling programs. Thus, the conditions under which clustering methods perform best are rarely met for real modeling projects.

Between single-model QA programs and consensus approaches, a new set of approaches has emerged at CASP7 and CASP8, which combine information from templates with traditional scoring functions^{234–235}. Since similar sequence implies structural similarity⁴⁰, models that are closer to related experimental structures are thought to be better models for that target sequences. These scoring functions assess each model separately, but rely on the availability of structural information, hence the name quasi single-model estimation programs.

Paluszewski and Karplus²³⁵ extract distance constraints from alignments identified by their threading method SAM. For each pair of residues in the model, they determine a desired distances, a weighted average of observed distances in the templates. Distances from template sequence closer to the target sequence are assigned a higher weight. The weights are therefore interpreted as a confidence value for the distance constraints. Often, the distances observed for a residue pair show considerable variation. Thus, the weighted average can result in a desired distance far away from any distance observed in the templates.

In this work, we explore the combination of structural features of templates with the QMEAN scoring function. Distances between residues in templates are combined by expressing each distance as a Gaussian function, whose weight is proportional to the evolutionary distance between target and template and the standard deviation is scaled by the uncertainty of the restraint. Models are then scored by agreement with these restraints. The scoring function has participated in CASP9's QA category as QMEANdist. After showing results on CASP8 data, we will thoroughly discuss the performance of QMEANdist using the official results obtained during the prediction season.

2 Materials & Methods

Datasets

CASP8 | The models of all CASP8 servers were used as a training set for QMEAN-dist. For each week of CASP, we have created a timestamped PDB snapshot to be used as template library. Template searches were performed against these snapshots. The GDT_HA, GDT_TS⁴¹ scores for all models were calculated on the whole target structure with TM-score²³⁶. C α -IDDT scores (chapter 'Local Distance Difference Test - A Robust, Superposition-Free Similarity Measure for Protein Structures') were calculated without clash-filters and stereo-chemical checks.

CASP9 | QMEANdist participated in the QA server category of CASP9. Predictions were submitted for all 116 targets of CASP9. The predictions and GDT_TS scores for all servers have been downloaded from the prediction center website. IDDT scores and C α -IDDT scores for the full-length targets have been calculated without clash-filters and stereo-chemical checks.

Model Quality Assessment Method

The method for model quality assessment consists of the following steps, which are described in detail in the following sections:

1. Identification of templates using BLAST and HHsearch
2. Clustering of the identified templates sequences by sequence similarity
3. Extraction of distance constraints from the identified templates
4. Scoring of models by the extracted constraints
5. Scoring of the models using the QMEAN6 scoring functions
6. Combination of the QMEAN6 and distance constraints scores

Template Identification

For each target, templates were identified using BLAST⁵⁸ and HHsearch³². BLAST was run with the default parameters against all chains of the SWISS-MODEL template library. The HHsearch profile was built by using buildali.pl from the HHsearch package using default parameters and 3 PSI-BLAST iterations. The profile was then searched against a 70% clustered database of profiles of PDB chains. Then, the search was extended to all PDB profiles whose clusters have been identified during the first step.

Clustering of Template Sequences

Due to the uneven distribution of sequences in the PDB, the set of identified templates often contains multiple template sequences with identity above 90%. Since our structural scoring algorithm combines information from all identified templates, information from these sequences would be overrepresented. We account for the presence of multiple, very similar template sequences by downweighting the contributions of these structures accordingly (see 'Scaling of Constraints from Structures with Similar Sequences').

To group templates with highly similar sequences, the sequences are clustered using a greedy sequence identity clustering algorithm: The templates are first sorted by decreasing target sequence coverage. The cluster list is initialized to an empty list. Each template is then compared against the centroids of the existing clusters. When the sequence identity is higher than a defined threshold, the template is added to that cluster, otherwise the sequence is compared to the next cluster. If the template sequence does not share considerable identity to any of the existing clusters, a new cluster is added, setting the template as the centroid.

For the final scoring function, we have used a sequence identity threshold of 90%. While the clustering step in general did improve the performance of the scoring functions, we found that the threshold could be varied between 70% and 90% without impact on the performance.

Distance Constraints

Each template defines distance constraints for the interaction of residues in the final model. The residues in the template structure are mapped onto the target sequence using the target-template alignment. For each pair of $C\alpha$ atoms closer than a distance threshold D and further than s residues apart in sequence, a restraint is added. Setting the sequence separation to 0 includes contacts from neighbouring residues, whereas a larger separation increases the importance of interactions between distant residues. While a larger sequence separation avoids trivial nearest-neighbours contacts, we have found that setting the sequence separation to large values above 6 together with a small distance cutoff is problematic for short proteins: For these small structures, local interactions are an important contributor to stability. For the final scoring function, we have identified an optimal value of 4 for sequence separation and 15Å for the distance cutoff.

We have also experimented with using $C\beta$ - $C\beta$ distances instead of $C\alpha$ - $C\alpha$. However, we did not find an increase in performance on any of the testsets and hence decided to stay with $C\alpha$ - $C\alpha$ distances.

Each pair of $C\alpha$ atoms (i, j) meeting the requirements of sequence separation and distance cutoff described above is represented as a Gaussian function:

$$c_{ijk}(d) = \frac{A_{ijk}}{\sigma_{ijk}\sqrt{2\pi}} \exp\left[-\frac{(d - \mu_{ijk})^2}{\sigma_{ijk}^2}\right]$$

μ_{ijk} is the mean of the Gaussian function and is set to the distance found in the template, σ_{ijk} and A_{ijk} are standard deviation and scaling factors, respectively. They are calculated as described in the next two sections.

CALCULATION OF SIGMA | σ_{ijk} is set to the average root mean displacement of the two residues defining the restraint. The root mean displacement is calculated from the average atomic B-factors of the two residues b_i as $\sigma_{ijk} = \sqrt{3b_i/8\pi^2}$.

CALCULATION OF SCALING FACTOR | A_{ijk} balances the contributions of structural consensus and evolutionary distance between target and template. By expressing A_{ijk} as a function of the sequence similarity between target and template, sequences closer to the target are given a stronger weight. Setting A_{ijk} to a constant for all templates would weight all restraints the same, irrespective of the evolutionary distance between the target and the template. After experimenting with a variety of measure for the evolutionary relatedness of target and template, a measure based on the BLOSUM62²⁸ scoring matrix was found to perform best. It was substantially superior to sequence identity, E-value and raw alignment scores.

The question remains on how A_{ijk} should depend on the sequence similarity measures. By defining A_{ijk} as a linear function of sequence similarity, the consensus part of the scoring algorithm was too influential. For many targets, constraints from distant templates were driving the best-scoring template away from the optimum. To increase the importance of closely-related sequences, we express A_{ijk} as an exponential of sequence similarity, e.g. $A_{ijk} = \exp[c \cdot sim]$. The coefficient c leading to optimal results was found to be 16.

Scaling of Constraints from Structures with Similar Sequences

The uneven sampling of sequence space in the PDB leads to an over-representation of certain structures. Typically, structural bias is avoided by choosing a representative structure for a cluster of sequences. However, the structures within a cluster might differ considerably, due to domain rearrangements, crystallization conditions or ligand-induced fit. Thus, it is not possible to select a single representative structure, since the target sequence might be in one or the other conformation. In addition, different homology detection algorithms can produce alternative alignments. They differ in the location of insertions, deletions, and overall coverage of the target sequence. It is clear that structural constraints which are consistent in alternative alignments are more reliable as they are robust against fluctuations of alignment algorithms.

To account for alternative alignments and structures of highly similar sequences, constraints are scaled by how often the particular $C\alpha$ - $C\alpha$ pair is observed in the template cluster. To illustrate, in **figure 7.1** a sample cluster with two templates is shown. The first template covers the N-terminal and central region of the target sequence, whereas the second template covers the C-terminal and central regions. Distances between residues of the C-terminal part can only be observed in the second template. However, distances

between C α atoms in the central part can be observed twice. Distance between C- and N-terminal residues are never observed.

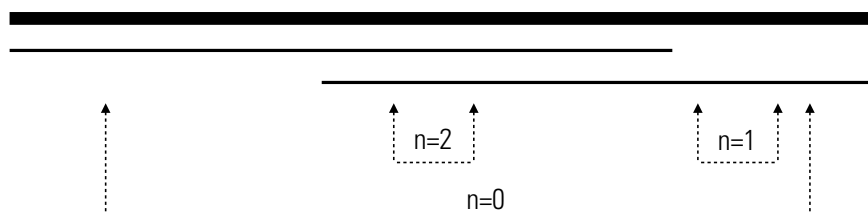


Figure 7.1 Constraints are downweighted by the number of times the pairs of residues appears in a cluster. The thick black line denotes the target sequence, the thin black lines the templates structures.

Depending on the scoring scheme of the alignment method, the insertions and deletions can happen a few residues earlier or later. When determining the number of observations from raw alignment counts, constraints from residues which appear as a deletion in one sequence, but not in the other would not get downweighted. When determining which pairs of C α atoms can be observed in the structure, we consider short deletions below 5 residues in the template sequences as aligned. Thus, small relative shifts of insertions and deletions effectively leads to a downweighting of constraints.

The additivity of the Gaussian constraints ensures that the score of distances which are identical in all structures of a cluster result in the same score as if there would only be one structure. When the distances differ considerably, the Gaussian constraints represent two alternative conformations for the residue pair.

Scoring of Models

After the distance constraints have been extracted from the templates, the models are scored. For all pairs of C α atoms (i, j) at a distance below the distance threshold in the model, the agreement of the distance with the propensity function is calculated:

$$C_{ij}(d_{ij}) = \sum_k c_{ijk}(d_{ij})$$

, where d_{ij} is the distance between C α atoms i and j and the sum runs over all distance restraints defined between the two residues. The final score for the model is the sum of the individual distance agreements, **figure 7.2**.

The formulation of distance constraints as sum of Gaussian functions is employed by the MODELLER programs as well⁷². It treats each observed distance as an alternative conformation, without enforcing any consistency on the restraints. Multiple, alternative conformations for the model can coexist.

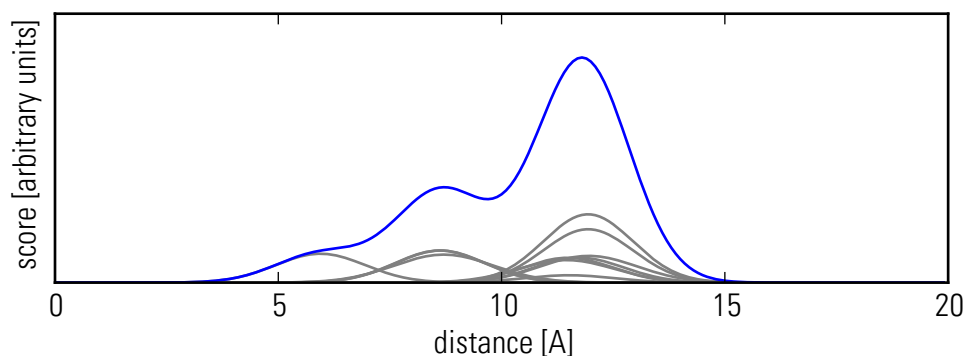


Figure 7.2 Example propensity for distance constraints between two residues. In grey, the individual Gaussian functions are shown, one for each identified template. The blue line shows the sum of all Gaussian functions.

The magnitude of these scores depends on several factors: First the number of restraints considered in the calculation, second, the similarity of the templates to the target sequence, and third, the variation of the restraints. Since these factors all depend on the target, the distance scores can only be understood as a relative ranking, which is valid for models of a given sequence. These scores are helpful in identifying the best model among a set of models, but do not predict the quality of the model on a global scale. However, global quality estimates are important in determining the usefulness of a model. Additionally, as soon as the distance scores are combined with other scoring functions, e.g. QMEAN, a sufficiently high global correlation is required to fix the relative importance of the individual scoring functions.

To convert the relative scores into scores on an absolute scale, we use the sequence similarity and structural score of the highest scoring template. The score value is related to GDT_HA by comparing to GDT_HA values obtained for target-template pairs with similar sequence similarity. Since GDT is clearly bound by the coverage of the target sequence, the obtained GDT estimate is multiplied with the target sequence coverage of the highest ranking template.

Combination of Distance Score with QMEAN

The QMEAN6 norm scores for each model are calculated after Benkert¹⁴². The predicted solvent accessibility and secondary structure for the QMEAN agreement terms are calculated with SSpro4¹¹⁶ (version 4.03) and PSIPRED⁸ (version 2.61), respectively.

The distance score is linearly combined with QMEAN6. When templates could be identified, the relative scaling factors of distance score and QMEAN6 have been set to 0.8 and 0.2, respectively. When no templates are identified, the importance of QMEAN6 is set to one. The scaling factors have been determined empirically on the CASP8 training set.

Implementation

The scoring function has been implemented as part of SWISS-MODEL Next Generation using the OpenStructure framework¹³⁵. The low-level calculations are implemented in C++. Most of the logic that deals with setup/initialisation/output is implemented in Python.

3 Results & Discussion

Results on CASP8 data

Before turning our attention to the CASP9 predictions, we would like to have a closer look at the contributions of the individual terms of QMEANdist. We will discuss the influence of QMEAN and the distance score on template selection, model ranking and absolute quality prediction of models. **Table 7.1** summarizes the performance of the individual terms of QMEANdist.

	QMEAN		dist		QMEANdist	
	GDT_HA	C α -IDDT	GDT_HA	C α -IDDT	GDT_HA	C α -IDDT
<i>r</i> per-target	0.732	0.805	0.844	0.901	0.854	0.917
loss	0.057	0.065	0.050	0.050	0.048	0.039
<i>r</i> pooled	0.766	0.849	0.830	0.857	0.848	0.889

Table 7.1 Performance of QMEAN, the distance score (dist) and QMEANdist on all server predictions of CASP8. For each of the structural similarity measures (GDT_HA and C α -IDDT) the loss (difference between best and best-scoring model), average Pearson coefficient and Pearson correlation of all pooled predictions are listed.

MODEL RANKING | Per-target correlation measures how well quality estimation programs are able to rank the models of a given target. Compared to the QMEAN scoring function, the distance constraints substantially improve the ability to rank models (**figure 7.3**). On average, QMEANdist achieves a Pearson correlation to GDT_HA of 0.85, compared to $r = 0.732$ for QMEAN. With the exception of 6 targets which have marginally lower Pearson correlations than QMEAN alone, the per-target correlations are higher in all cases. In some cases, the correlations increase from 0.6 to almost 0.95 for QMEANdist. The incorporation of QMEAN slightly lowers the correlation coefficient for *easy* targets. In a sense, the distance constraints from the templates are informative enough and QMEAN is adding noise to the ranking. Still, the incorporation of QMEAN into the scoring is beneficial and leads to a significant increase of performance. The difference is especially noticeable for medium to difficult targets, where multiple equally plausible templates are available.

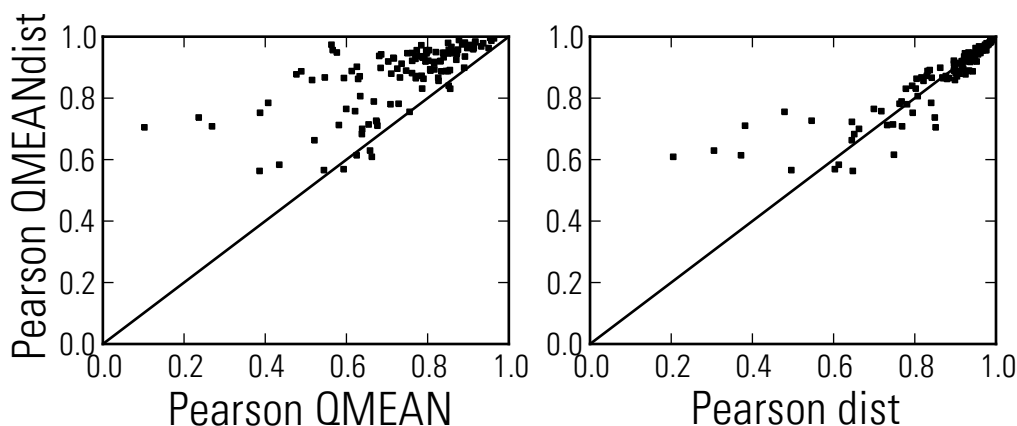


Figure 7.3 Pairwise comparison of per-target correlation coefficients for QMEANdist, the distance score (dist) and QMEAN to GDT_HA on CASP8 predictions.

Similar results are obtained for correlation to $C\alpha$ -IDDT. An increase in performance is observed for both distance scoring and QMEANdist. For all 3 scoring schemes, the correlations to $C\alpha$ -IDDT are substantially higher than the corresponding correlations to GDT_HA. Two reasons contribute to this: first the distance scoring schemes and potentials of mean force are based on internal distances and are thus more closely related to $C\alpha$ -IDDT. Second, lower correlations to GDT_HA are observed for targets which exhibit domain movement. The effect of domain movement is less strongly pronounced for $C\alpha$ -IDDT, which makes the correlations appear higher.

MODEL SELECTION | While relative ranking performance depends on the ordering of all models, model selection measures the capability of a scoring function to pick the best model. We compare the loss of GDT_HA, i.e. the difference between the GDT_HA of the best model in the set minus the GDT_HA score of the highest scoring model, for each scoring function.

As with model ranking, the combination of distance scores with QMEAN is beneficial, **figure 7.4**. QMEAN alone selects on average a model with a 5.7 lower GDT_HA than the best model. The distance score alone achieves a loss of GDT_HA of 5.0, slightly lower than QMEAN. The combination of QMEAN and distance constraints has a positive effect for model selection. Even though the improvement is less pronounced than for per-target and global correlations, the average loss of GDT_HA decreases to 4.8 GDT_HA points.

In analogy to loss of GDT_HA, we define the $C\alpha$ -IDDT loss as the difference between the model with the highest $C\alpha$ -IDDT and the $C\alpha$ -IDDT of the selected model. The loss of $C\alpha$ -IDDT also profits from incorporation of distance constraints. Already the distance scoring alone is substantially better than QMEAN (0.065 vs. 0.050). Another boost in performance is observed when the distance score and QMEAN are combined (0.039).

ABSOLUTE QUALITY PREDICTION | The performance of QMEANdist to estimate model quality on an absolute scale was measured by pooling the models of all targets (**figure 7.5**) and calculating the correlations between QMEANdist and the GDT_HA/ $C\alpha$ -IDDT scores,

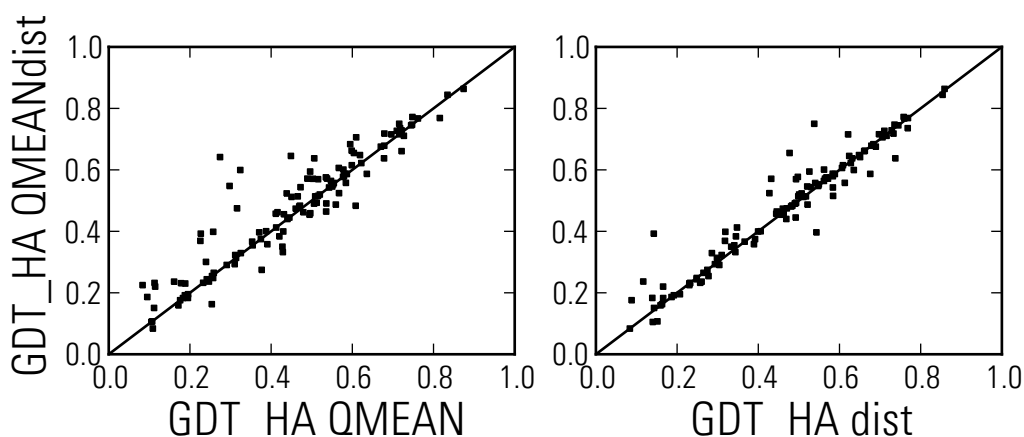


Figure 7.4 GDT_HA of selected models of QMEANDist in comparison to the distance score (dist) and QMEAN on the CASP8 training set. Points above the diagonal indicate targets for which the selected model of QMEANDist had a higher GDT_HA than dist/QMEAN.

respectively. The correlation of QMEANDist to GDT_HA was found to be 0.85 compared to 0.75 for QMEAN6. Correlation to $C\alpha$ -IDDT was improved slightly as well, from 0.85 to 0.89. Although the correlations to $C\alpha$ -IDDT are slightly higher, the absolute difference between the predicted quality and the $C\alpha$ -IDDT scores are bigger. There is one-to-one correspondence between QMEANDist scores and GDT. However, for $C\alpha$ -IDDT, the predicted scores are clearly off-diagonal and the QMEANDist score heavily under-estimate the $C\alpha$ -IDDT. This is a direct effect of the training of the global QMEANDist scores on GDT. For better absolute quality estimates with respect to $C\alpha$ -IDDT, the data would need to be retrained on $C\alpha$ -IDDT.

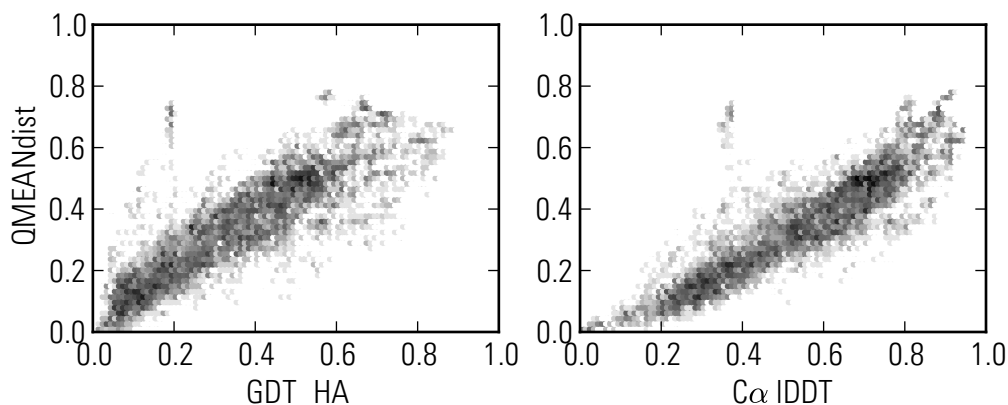


Figure 7.5 Global correlation of QMEANDist score to GDT_HA and $C\alpha$ -IDDT of all pooled CASP8 targets. The plot shows the logarithmic point density from white to black.

In the upper left corner of **figure 7.5**, a group of models are assigned a QMEANDist score between 0.4 and 0.75, even though they have a GDT_HA of only 0.2. These structures are all models of target T0498, a designed protein with a sequence identity of 95% to T0499. T0498 adopts a 3α -fold, whereas T0499 adopts an $\alpha\beta$ -fold^{66,237}. Such targets are very challenging for automated methods as the usual assumptions of sequence and struc-

ture relationship do not necessarily hold for designed proteins: A single point-mutation having a dramatic effect on the overall structure would most likely lead to a loss of function, and thus be deleterious in a living organism. But these sequences were designed in absence of typical evolutionary pressure, and were allowed to undergo massive conformational changes.

Regardless of the structural similarity measure, the spread between predicted and observed similarity is relatively large. For example, a QMEANDist value of 0.6 corresponds to GDT_HA values between 0.4 and 0.7, the difference between a fold-level accurate and a prediction of medium quality. There is substantial room for improvement to make the quality estimates more accurate.

Results on CASP9 data

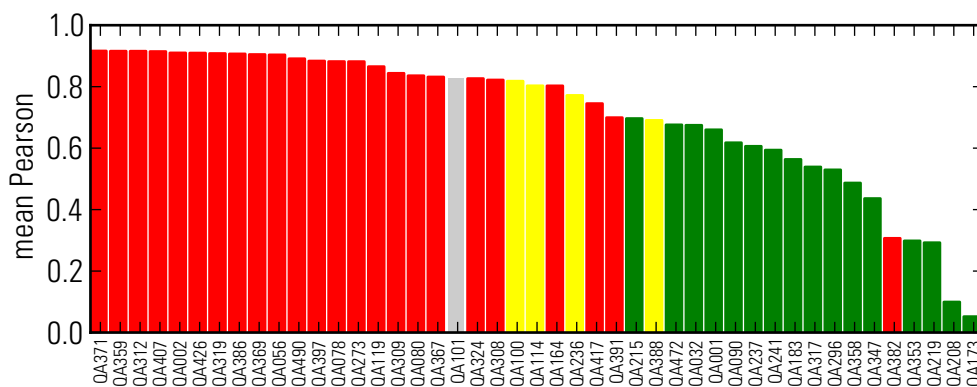
In the following, the results of QMEANDist at the CASP9 experiment are discussed. The discussion occasionally touches on the performance of other quality estimation servers but is not intended to be a comprehensive evaluation of the CASP9 QA category. For an in-depth discussion of all QA servers, the official QA assessment¹¹⁸ is to be consulted.

Unlike the official QA assessment which was based on GDT_TS scores, GDT_HA is used herein, which is more discriminative for high-quality models: the four distance cutoffs of GDT_TS (1Å, 2Å, 4Å, and 8Å) are replaced by 0.5Å, 1Å, 2Å, and 4Å. Additionally, quality prediction servers are evaluated against $C\alpha$ -IDDT, a measure which is more robust in presence of domain movements.

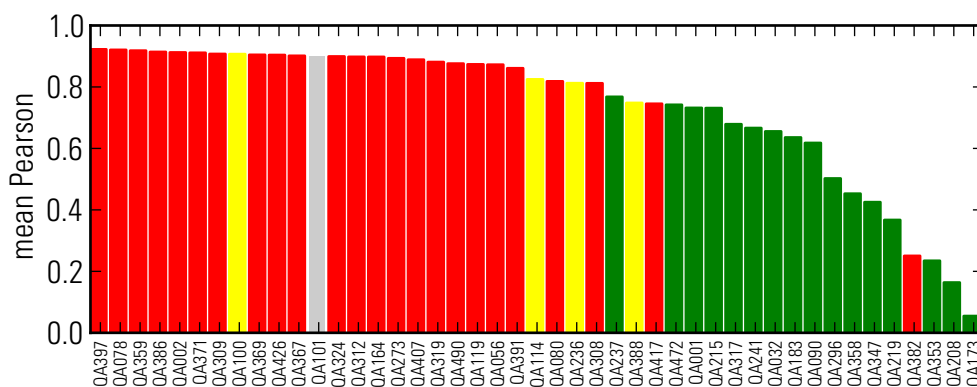
Since QMEANDist was still in development during the prediction period of CASP9, the set of parameters used differs from the final and optimal set. For the CASP9 predictions, the following parameters were used: the weights are calculated as $\exp(sim)$ (coefficient set to 1), a sequence separation of $s = 8$, and a cutoff distance of $D = 12\text{\AA}$.

PER-TARGET CORRELATION | For GDT_HA, the consensus-based methods achieve the highest per-target correlations (**figure 7.6**). They are followed by QMEANDist (QA101, $r = 0.850$) and the quasi-single model method Splicer_QA (QA100, $r = 0.837$). While the performance gap between consensus and quasi-single model methods is small, there is a large gap to the first single-model MQA programs: MULTICOM-NOVEL (QA215) and QMEAN (QA427) achieve an average Pearson correlation of 0.71 and 0.69, respectively. For $C\alpha$ -IDDT, QMEANDist ($r = 0.900$) moves to rank 12, statistically indistinguishable from Splicer_QA ($r = 0.907$) on the 8th rank.

As with CASP8 server predictions, quality estimates tend to correlate better with $C\alpha$ -IDDT than GDT_HA. Similarly, the top-performing consensus servers for GDT_HA have a higher correlation to $C\alpha$ -IDDT, although the correlation improvement is less pronounced for these methods. The differences in correlation gain is partially explained by the dependence of the top-performing clustering methods on rigid-body superpositions to compare the model structures. Methods based on inter-atomic distances fail to mimic the behavior of GDT in presence of domain movements. This causes the correlations to be smaller. The top-performing consensus methods for $C\alpha$ -IDDT, ModFOLDclust2



A



B

Figure 7.6 Mean Pearson correlations of predicted quality of all CASP9 QA servers and structural similarity to target. (A): GDT_HA, (B): $C\alpha$ -IDDT. QMEANDist (QA101) in grey, consensus groups in red, quasi-single model in yellow, single model servers in green.

(QA397) and IntFOLD-QA (QA078), are indeed methods that use inter-atomic distances instead of global superpositions to cluster the models¹¹⁹.

While for 75% of the CASP targets, the per-target correlations of QMEANDist were above 0.88, QMEANDist failed to properly rank the models for a small fraction of targets. The majority of these targets have been assigned to the free-modeling category. Since QMEANDist was unable to obtain any evolutionary restraints for these targets, the quality predictions are solely based on the QMEAN6 scoring function. Lower correlations can thus be expected.

MODEL SELECTION | **Figure 7.7** shows the GDT_HA and $C\alpha$ -IDDT losses for all QA servers of CASP9. Again, for GDT_HA the typical separation between clustering, quasi-single and single-model MQA programs is visible, albeit less pronounced than for per-target correlations. For both GDT_HA and $C\alpha$ -IDDT, QMEANDist was ranked as the first non-consensus method, followed by Splicer_QA (QA100). The loss of QMEANDist was indistinguishable from the majority of consensus methods.

	better		worse		tie		delta	
	GDT	IDDT	GDT	IDDT	GDT	IDDT	GDT	IDDT
QA002	36	44	49	36	31	36	0.011	0.000
QA312	31	40	48	35	37	41	0.013	0.000
QA407	36	44	51	41	29	31	0.010	0.003
QA273	35	34	47	39	34	43	0.005	0.003
QA472	56	58	18	15	42	43	-0.038	-0.037

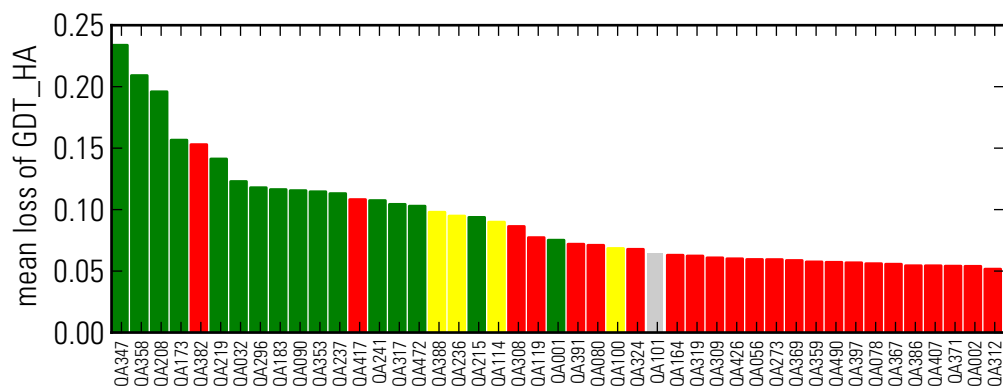
Table 7.2 Head-to head comparison of QMEANdist and the two top performing servers for $C\alpha$ -IDDT (QA407, QA273), GDT_HA (QA002, QA312), and QMEAN (QA472). The table lists the number of targets where QMEANdist was better, worse or equally good at model selection. The delta column lists the difference in loss of GDT_HA and $C\alpha$ -IDDT between the methods. Positive numbers denote higher losses for QMEANdist.

In order to understand the differences between QMEANdist, QMEAN, and the top-performing servers, we have performed head-to-head comparisons for GDT_HA and $C\alpha$ -IDDT, **table 7.2**. For each of the servers, the number of wins (QMEANdist selects a better model), losses (QMEANdist selects a less-accurate model), ties (the selected models are within 0.01 GDT_HA/ $C\alpha$ -IDDT), and the average difference in accuracy have been calculated. Overall, the two top-performing servers for GDT_HA (QA002/QA312) selected a better model in 49/48 cases, whereas the model of QMEANdist is better for 36/31 targets. Here, the number of targets where QMEANdist selects a better model is significantly lower. Compared on $C\alpha$ -IDDT, the same two servers are virtually indistinguishable from QMEANdist, both in terms of number of wins and average difference of selected models. For the two top-performing servers for $C\alpha$ -IDDT (QA407/QA273), the number of wins and losses is more balanced.

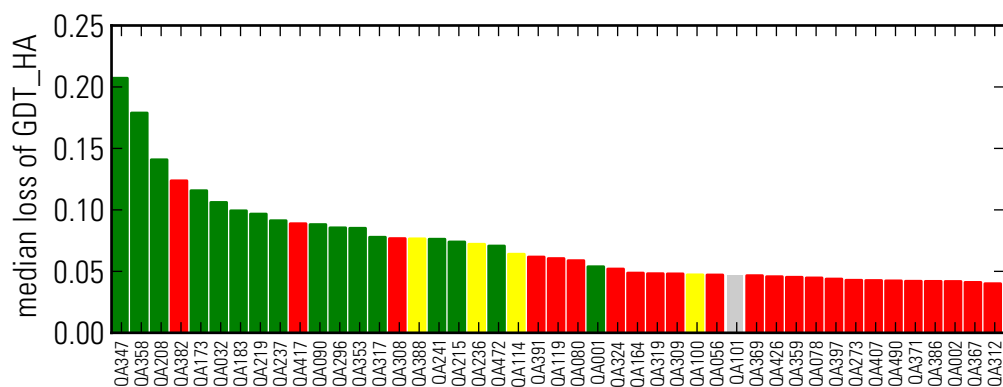
A substantial increase in performance in model selection is seen in comparison to QMEAN (QA472). QMEANdist selects *worse* models for 18 targets. With the exception of five targets, the losses are smaller than 5 GDT_HA points. The wins, on the other hand exceed 5 GDT_HA points in 34 cases.

ABSOLUTE QUALITY PREDICTION | The ability of the MQA methods to predict the quality of models on an absolute scale is shown in **figure 7.9**. The correlations have been calculated both for IDDT and GDT_HA. Correlation of absolute quality to GDT_HA is dominated by consensus methods. Splicer_QA, however, outperforms all other QA methods when correlating against $C\alpha$ -IDDT. The perceived dominance of consensus methods at CASP for absolute quality predictions is not due to intrinsic limitations of quasi-single model methods, but due to the choice of the structural similarity measure. Reliable absolute quality predictions are feasible by only relying on information from available template structures and empirical energy functions. No clustering of submitted models is required.

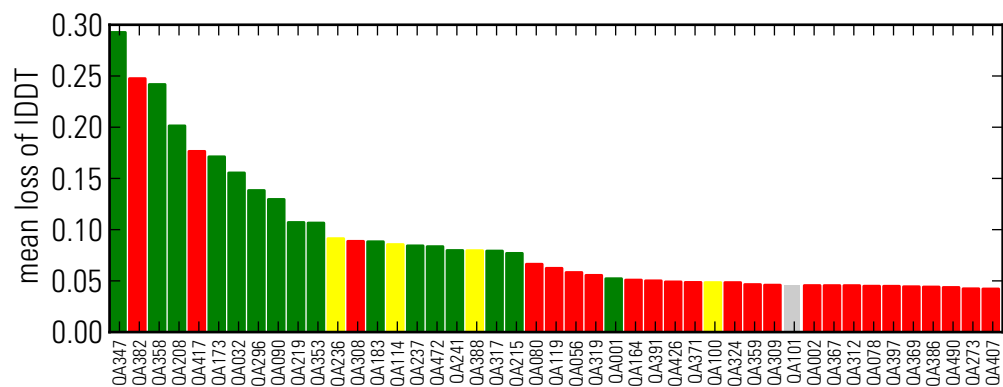
QMEANdist performed poorly at predicting the global quality of the CASP9 dataset. The predictions are only slightly better than the predictions of QMEAN alone. The relations between scores and predicted quality of the targets exhibit a different slope, which causes them to not superpose well and spread out. Partially, the lack of improvement



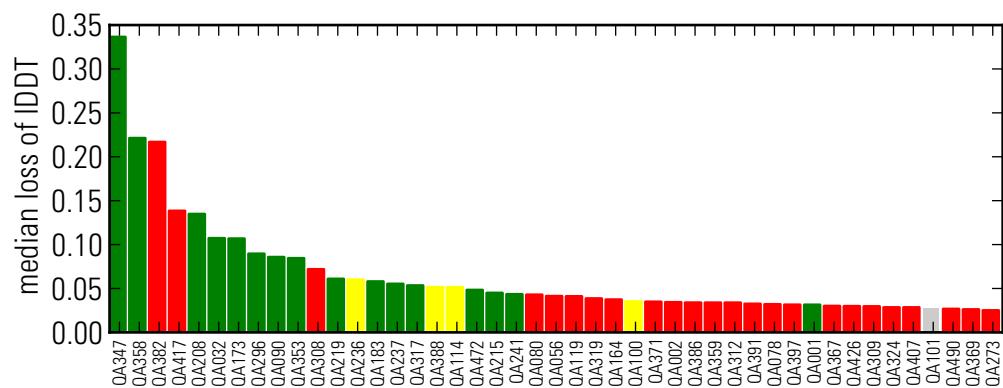
A



B



C



D

Figure 7.7 Loss of GDT_HA and $C\alpha$ -IDDT on CASP9 targets. (A) mean GDT_HA loss, (B) median GDT_HA loss, (C) mean $C\alpha$ -IDDT loss, (D) median $C\alpha$ -IDDT loss. QMEANDist (QA101) in grey, consensus groups in red, quasi-single model in yellow, single model servers in green.

over QMEAN is explained by the use of a smaller coefficient for the distance constraint weighting. A smaller weighting causes constraints from distant templates to be more influential and the magnitude of the scores is more dependent on the number of Gaussian constraints per residue pair. A larger coefficient on the other hand, reduces the number of significant constraints. When using the improved parametrization, the correlation to GDT_HA increases from 0.76 to 0.83.

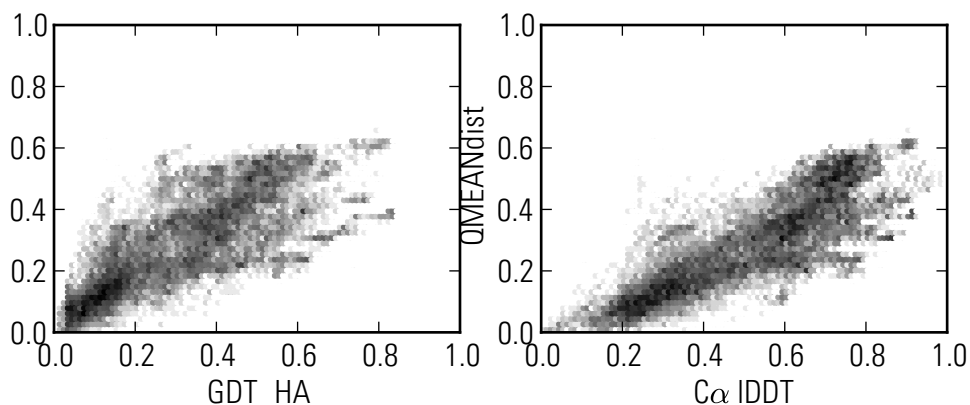
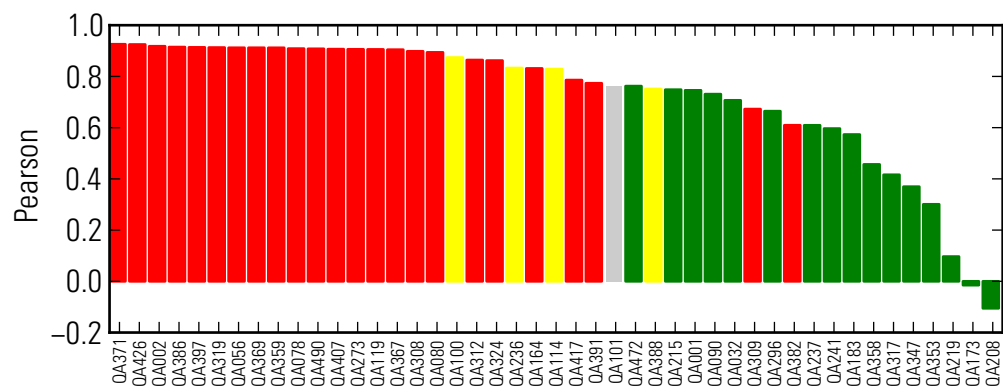


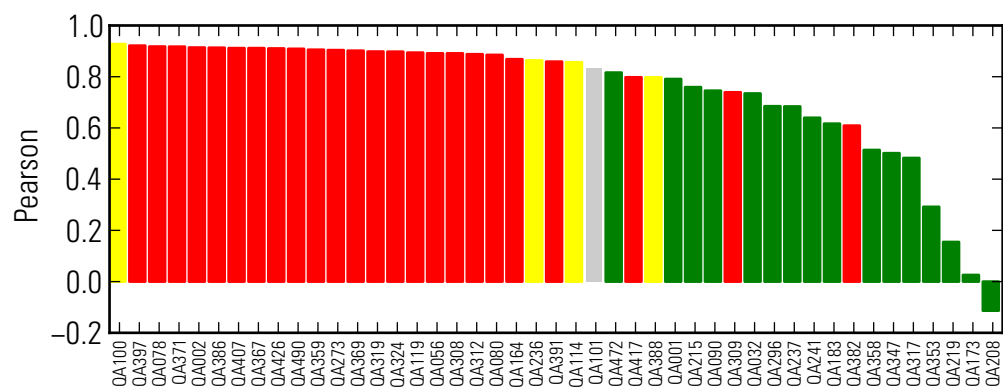
Figure 7.8 Absolute quality prediction of QMEANdist on CASP9 targets

The Importance of the Similarity Measure

As described above, there seems to be a clear bias of methods towards either $C\alpha$ -IDDT or GDT_HA. When one or the other structural similarity measure is used, the final ranking of the methods changes. When using GDT_HA, methods based on global superposition of models and templates perform better, whereas methods using internal distances perform better when compared against $C\alpha$ -IDDT. For most targets, the difference between the two similarity measures are small. However, the choice of one similarity measure goes beyond method preferences. Especially in presence of domain movements, superposition-based measures do not lead to meaningful results. For example, CASP target T0542 has a length of 590 residues and exhibits a two domain architecture. There are two sets of templates available which differ in the relative orientation of the domains. Some of the models submitted by predictors are based on the first, some on the second group of templates. For the assessment of the TBM category, the T0542 has been split into two assessment units. However, for the QA category the quality estimation programs are compared against the full models. It is clear that the whole-target GDT_HA scores are largely dominated by the domain movement and not by how well the predictors modeled the individual domains. For example, when evaluated against GDT_HA the correlation of the QMEANdist score is only 0.640, even though the method is perfectly capable of ranking the models for $C\alpha$ -IDDT ($r = 0.972$). In this particular case, GDT_HA scores do not capture local model quality well-enough. This reinforces the importance of superposition-free scores in presence of domain movements.



A



B

Figure 7.9 Absolute quality prediction of all QA servers expressed as Pearson correlation coefficient between score and similarity measure. (A): GDT_HA, (B) C α -IDDT. QMEANdist (QA101) in grey, consensus groups in red, quasi-single model in yellow, single model servers in green.

4 Conclusions

In this work, we presented QMEANdist, a variant of QMEAN augmented with constraints from related template structures. Compared to QMEAN alone, the method's ability to rank and select models is greatly improved. In addition, the absolute quality estimates are more reliable. The analysis of the QA results from CASP9 has shown that our quasi-single model method QMEANdist produces per-target quality estimates comparable to consensus programs. The ability to rank models for a given target of QMEANdist was indistinguishable from many of the clustering methods participating at CASP9. The main differences are due to the presence of free-modeling targets, where little to no template information is available. The ability of QA methods to select accurate models is an important performance figure. Performance is less dominated by low-quality models,

and focuses more on models with highest quality for each target. During CASP9, QMEANDist was one of the top-performing methods for model selection: for median loss of IDDT, QMEANDist ranks as the fourth-best method. These results, together with a detailed comparison to the two best-performing QA methods according to $C\alpha$ -IDDT and GDT_HA losses, underline the stability and viability QMEANDist.

The results of CASP9 have shown substantial room for improvement of QMEANDist in two areas: absolute quality prediction and local, per-residue quality estimates. Consensus programs perform substantially better than QMEANDist at absolute estimates. As illustrated by absolute quality prediction estimates of Splicer_QA, more accurate absolute quality estimation are feasible for template-based MQA programs. The weighting scheme employed by QMEANDist is based on the sequence similarity of target-template alignment alone. There is potential for improving the quality estimates of QMEANDist by incorporating more target-template alignment properties into the weighting. The use of properties based on profile-profile scores, predicted secondary structure features will most likely have a positive effect on model selection and absolute quality estimation.

Likewise, the local quality estimates of QMEANDist can only be understood as a failed experiment. Correlating the propensity scores to local residue quality has proven to be difficult, as both the number of constraints and the magnitude of the propensity functions vary greatly. More work is required to obtain reliable local quality estimates.

Additionally, it was shown that the choice of similarity measure alters the relative ranking of the groups. Servers using inter-atomic distances for clustering perform better on $C\alpha$ -IDDT, whereas methods based on a superposition of models perform better for GDT_HA and GDT_TS. As shown for target T0542, the evaluation of model quality servers on GDT measures in presence of domain movements is not meaningful. The results are dominated by the effect of relative domain orientation changes and not the ability to predict the accuracy of the individual domains. Domains should be split into assessment units, or evaluated against $C\alpha$ -IDDT, as done in this work.

Automated Modeling in SWISS-MODEL Next Generation

An automated homology modeling pipeline for inclusion into the SWISS-MODEL web server is outlined. The method applies state-of-the-art homology detection programs to identify evolutionary related template structures. The quality of each template is estimated using a probabilistic approach: each property of the target-template pair, e.g. sequence identity, HHsearch alignment score, or agreement between predicted solvent accessibility acts as a predictor of the template's quality. The quality estimates are combined by considering the uncertainty of each prediction, effectively increasing the importance of accurate predictors. Models are then built for the 30 top-scoring templates, and their quality assessed with the composite scoring function QMEAN. Finally, models are scored by consensus to identified template structures.

It will be demonstrated, that the combination of template properties for quality estimation greatly improves template selection performance for all targets difficulties. The performance greatly benefits from inclusion of both QMEAN and structural consensus as well. For an objective comparison to other existing structure prediction servers, the complete pipeline is assessed as part of the CAMEO live benchmark, which reveals strengths and weaknesses of the presented approach.

1 Introduction

The SWISS-MODEL web-server for protein homology modeling is a widely-used service to predict the structure of proteins from their primary amino acid sequence^{42,131}. Structures are modelled by homology to experimentally determined protein structures, the templates. SWISS-MODEL uses as much of the available template information as possible when building models, without resorting to extensive sampling to guide the model building process. Models only deviate from the templates where it is unavoidable, e.g. due to insertions and deletions, or non-conserved sidechains.

The current version of SWISS-MODEL builds models on a single template. When closely-related template structures are available, the models are locally very accurate, and are often closer to the target structure than models from multi-template modeling programs. Nevertheless, multiple template modeling provides some benefits: in the Domain-Find chapter, an analysis was performed to quantify how much models can be improved by including information from multiple templates. Models with an overall low IDDT score often contain stretches of residues with a higher per-residue IDDT than the best single-template model. The chemical environment of these residues more closely resembles the target structure. Would restraints from these residues be selected, more accurate models could be built. In practice, it turns out to be challenging to identify the optimal combination of restraints without resorting to computationally demanding sampling protocols^{66,79,82}. Improvements seen for multi-template models are often due to model extension: models built on a template covering only a part of the target sequence are extended with structural information from other available templates. By that, the

effective coverage of the target sequence is increased. For many applications, local prediction accuracy is more important than complete coverage of the target sequence^{204,238}. Thus, local accuracy is to be retained while extending the coverage of the built models.

Regardless whether models are built on single or multiple templates, the templates identified by homology detection programs need to be ranked according to their quality prior to model building. In this work, a template ranking scheme is developed that allows to accurately predict the quality of a template. The template ranking serves as the basis for a complete homology modeling pipeline for inclusions into the SWISS-MODEL web server. The ranking scheme is optimized for a single-template modeling pipeline, but has been designed to facilitate the transition from pure single-template modeling to a multi-template modeling approach. First, the automated modeling pipeline will be outlined in detail. Then, the template selection and model building steps are analyzed on training sets, giving insight into the factors that contribute the most to template selection performance. Finally, to assess the overall viability of the approach, the method is compared to existing protein structure prediction servers registered in the CAMEO live benchmark.

2 Materials & Methods

Method Overview

The modeling pipeline comprises the following steps, which are described below in more detail:

- Templates for the query sequence are identified using HHsearch/HHblits and BLAST.
- The template's quality is estimated from its properties.
- The templates are ranked according to the estimated quality
- Models are built for the 30 top-ranking templates using PROMOD-II and MODELLER
- The model quality is assessed with QMEAN and combined with the quality estimates on the template level.
- Models are further scored by structural agreement with the top-ranking templates.

Datasets

TARGET SEQUENCES | A set of target sequences from high-resolution X-ray structures has been derived from the PISCES web-server²³⁹. The following parameters have been used to generate the test set: mutual sequence identity less than 15%, R-factor of less than 0.3, resolution < 1.5 Å, and a minimal length of 60 amino acids. Chains where the atom sequence covers less than 50% of the SEQRES sequence have been removed from the set. The resulting test set contained 1304 protein chains.

BLAST TRAINING SET | For each target sequence of TARGET SEQUENCES, BLAST²⁹ (version 2.2.16) has been run against all sequences of the SMTL with default parameters (substitution matrix: BLOSUM62, gap opening penalty: 11, gap extension penalty: 1) and an E-value cutoff of 0.0001. This resulted in a total of 8890 target-templates alignments.

HHSEARCH TRAINING SET | For each target sequence of TARGET SEQUENCES, a profile was built with the buildali.pl script from the HHsearch package³² (version 2.0.13) using three PSI-BLAST iterations. The profile was searched against a 70% clustered database of SEQRES profiles. Then, the search was extended to all sequences of the SEQRES by creating a temporary database containing sequences of all clusters identified in the first search step. HHsearch was used with the default parameters. To avoid overweighting of target sequences with many templates, the identified templates have been clustered, and only the centroids of the 90% clusters have been used in the evaluation. The final HHsearch Training Set contained a total of 151347 target-template alignments.

HHBLITS TRAINING SET | This set was created like the HHsearch Training Set, except that the profiles have been built with HHblits³³ (version 2.0.13) against a 20% non-redundant database (NR20) with one iteration. The profile is then searched against all HHblits profiles of the SMTL. The final HHblits Training Set contained a total of 141362 target-template alignments.

ARTIFICIAL TEMPLATE SELECTION TEST SET | Starting from the HHsearch Training Set, artificial template selection tests have been created by removing templates above certain sequence identity, irrespective of the template-target coverage. Sequence identity thresholds of 80%, 50%, 30%, 25%, 20%, and 15% were chosen. These test sets have then been used to evaluate the template selection performance. The test sets comprised the following number of targets:

Test set	15%	20%	25%	30%	50%	80%
Targets	797	917	958	980	1022	1035

CAMEO TEST SET A | The CAMEO test set was derived from CAMEO targets that were submitted to structure prediction servers between 13th of January 2012 and 20th of April 2012⁵⁴. NMR structures and structures with lower than 75% coverage of the submitted SEQRES sequence were removed from the set. The final set contained 205 targets of all difficulty levels. Pseudo-models were built by copying conserved coordinates from the templates to the model, ignoring insertions and deletions. To measure the structural similarity between the pseudo-model and available experimental structures with matching sequence, the $C\alpha$ -I-DDTs were calculated for the pseudo-models. Since the SMTL might contain multiple polypeptide structures with matching sequence, the $C\alpha$ -I-DDT of each model was set to the highest $C\alpha$ -I-DDT obtained against any of the chains in the template library.

CAMEO TEST SET B | The CAMEO test set was derived from CAMEO targets submitted to structure prediction servers between 12th of October 2012 and 14th of December 2012⁵⁴. The resulting set contained 191 targets.

Template Identification

To identify evolutionary related protein structures, we apply three well-established homology detection programs: HHsearch, HHblits and BLAST. These sequence and profile-profile search programs have repeatedly performed well at the CASP experiment^{52–53}. A detailed comparison of HHsearch and BLAST has been carried out by Sadowski and Jones²⁴⁰. They find that it is beneficial to apply both methods as they target different sequence identity regimes: for closely-related templates, the scoring model underlying BLAST is more accurate for creating alignments, whereas the HHsearch scoring model is taking over for more distant sequences that do not share considerable identity to each other.

Template Properties

Several ranking schemes have been applied to template selection: feed-forward neuronal networks²⁴¹, alignment E-values¹⁹⁰, profile-profile alignment scores²⁴², selection by sequence identity²⁴³, etc. The template selection described herein sees each of the template properties, e.g. sequence identity, secondary structure agreement, as predictors for the template's quality q . The higher a template's quality the more closely the model built on the template resembles the actual target structure. When the target structure is known, e.g. for training purposes, q can be calculated from a pairwise structure comparison. In other cases, q is to be estimated from available template properties.

The following template properties have been used to estimate q : sequence identity, sequence similarity, agreement between predicted and observed secondary structure, agreement between predicted and observed secondary solvent accessibility, QMEAN4 score of the model, and HHsearch raw alignment score (**table 8.1**).

Predictor	Description
$cov \times id$	sequence identity
$cov \times sim$	sequence similarity
$cov \times acc$	solvent accessibility agreement
$cov \times sse$	secondary structure agreement
$cov \times hh_score$	normalized HHblits/HHsearch score
$cov \times QMEAN4$	QMEAN4 norm score of the model.*

Table 8.1 Available predictors for template/model quality q . * only available after model has been built.

In the following, precise definitions for these properties are given:

SEQUENCE IDENTITY/SIMILARITY | The sequence identity is calculated as the fraction of identical residues divided by the number total of aligned residues in the target-template alignment, ignoring gaps. The sequence similarity measure is based on a linear transformation of the BLOSUM62 substitution matrix³⁵. Consider an alignment of sequences A and B . The score $M(a, b)$ for a column in the pairwise alignment of A and B consisting of amino acid a and b is given as:

$$M(a, b) = \begin{cases} \frac{m(a,b) - \min(m)}{\max(m) - \min(m)}, & \text{if } a \neq \text{gap and } b \neq \text{gap} \\ 0, & \text{otherwise} \end{cases}$$

where $m(a, b)$ are the scores from the BLOSUM62 substitution matrix²⁸, $\min(m)$ and $\max(m)$ are minimal and maximal substitution scores found in the matrix. The similarity between the sequences A and B then defined as

$$\text{sim}(A, B) = \frac{1}{L} \times \sum_{i=1}^l M(a_i, b_i)$$

where L is the number of aligned columns not containing any gaps. Unlike sequence identity, when a sequence is aligned to itself, the score is not always 100%, but depends on the amino acid composition of the sequence. For a sequence with a typical amino acid composition, the sequence similarity is around 0.62, for sequences with a bias towards rare amino acids, the similarity of the sequences can reach higher values.

QMEAN4_NORM SCORE | After ranking the templates, models are built for the top 30 templates with PROMOD-II²⁴⁴. The all-atom, $C\beta$, solvation and torsion potential energies for the models are calculated according to Benkert *et al.*¹⁴². The potentials are then combined using weights trained on a large set of models built by PROMOD-II.

HHSEARCH/HHBLITS SCORES | The HHsearch and HHblits scores were retrieved from HHsearch and HHblits for each identified template. The scores are based on the raw alignment scores from the maximum accuracy algorithm of the HMM-HMM alignment step. Additionally, clusters of highly conserved residues are rewarded using a local auto-correlation function. More details on the HHsearch and HHblits score are given in the HHsearch paper³².

AVERAGE PROFILE COLUMN ENTROPY | The column entropies of the target profiles have been calculated as

$$H = - \sum_a p_a \log p_a$$

where the sum runs over all amino acids a in the column, p_a is the frequency of occurrence of that amino acid in the column as defined by the HHsearch profile of the target sequence. The average column entropy of the target profile is then calculated as the arithmetic mean of the individual column entropies.

SOLVENT ACCESSIBILITY/SECONDARY STRUCTURE AGREEMENT | Solvent accessibility and secondary structure are predicted with SSpro4¹¹⁶ (version 4.03) and PSIPRED⁸ (version

2.61), respectively, by using the alignment obtained from HHsearch/HHblits. The predicted burial/secondary structure states of the target are compared with the template. The agreement is defined as the matching fraction of solvent accessibility/secondary structure states of the aligned residues.

Choice of Structural Similarity Measure

In this work, q is defined as the IDDT of the pseudo-model built on the template. This is in contrast to established quality estimation programs where the quality is expressed either as global distance test GDT (global) or S-score (local). However, as discussed in the IDDT chapter, structural similarity measures operating on global superpositions are weak measures for the similarity of protein structures in presence of domain movements. Especially, for large training sets, the splitting of targets into assessment units is not feasible. We have thus chosen IDDT as the measure for template quality, as it is largely independent of domain movements, while maintaining a high correlation to GDT-style measures⁴¹.

The choice of the underlying similarity measure determines the balance between coverage and local reliability of the selected template/model. When template selection is optimized for RMSD, shorter models often lead to smaller root mean square deviations and are preferred. Local reliability of the model is the major driving force. For agreement-based measures, such as GDT and IDDT, higher coverage of the target sequence typically implies higher similarity. There are two counter-acting forces that play a role for template selection: To achieve higher scores, either a model with higher coverage, or a shorter, but more reliable model is selected. Local reliability has to be traded for coverage. Identifying the optimal template according to GDT or IDDT thus drives template selection towards longer alignments.

Derivation of IDDT PDFs for Properties

For templates which share considerable sequence identity with the target, predicting the template quality from sequence identity alone leads to very accurate results. Including secondary structure agreement does not improve template selection, since sequence identity is already discriminative enough. However, for templates in the twilight zone³², sequence identity is no longer a good measure for evolutionary divergence. The inclusion of structural information from predicted features can significantly improve template selection and even threading performance^{68–69}. The relative importance of the template properties for template selection clearly depends on the sequence identity regime. A linear combination of template properties is unlikely to perform well, since the linear regression model assumes that the relative importance of the properties remains constant.

Another way of looking at it, is to consider the uncertainty with which each property m_i predicts q . The smaller the range of observed qualities for a given m_i , the more accurate the prediction is. The reliability of each m_i can be estimated from large sets of

target-template pairs with known q : assuming that the data in the training set is a representative sample of target-template alignments, the probability of the template having a certain q , is given by the IDDT density.

There are many possibilities to determine the reliability of a property in predicting the IDDT. They all require approximating the 2-dimensional training data for each property by a probability density function. For simplicity, we have chosen to approximate each property by N equi-distant IDDT distribution fits. The distribution for a certain m_i is then obtained by linear interpolation between the distributions. Each of the N equi-distant distributions is calculated by estimating the IDDT density using a non-parametric approach. More specifically, we use density estimation (KDE)²⁴⁵ to approximate the IDDT histograms. The PDF of a sample $X = \{x_1, x_2, \dots, x_n\}$ is approximated by a sum of kernel functions, one placed on each x_i :

$$f_h(x) = \sum_i^n \frac{1}{n} K_h(x - x_i) = \sum_i^n \frac{1}{nh} K\left(\frac{x - x_i}{h}\right)$$

Kernel functions are required to be symmetric and integrate to one. The bandwidth parameter h influences the smoothness of the approximation. We have used a Gaussian kernel with a bandwidth of $h = 1$ for all properties. The density estimation has been performed with SciPy¹⁴⁸.

Since the IDDT is coverage dependent and the above properties are calculated on the aligned part only, the property values are multiplied by the coverage of the target-template alignment. This automatically downweights short alignments, while rewarding longer alignments.

IDDT Prediction

For each property, a probability density function of the template having a quality q , given the property is obtained. This is written as $P(q|m_i)$. The IDDT value maximising the PDF for properties m_i is the most probable IDDT according to the data. By assuming that the properties are statistically independent, the joint probability distribution is obtained by multiplication of the individual PDFs. Thus, the IDDT q for a template, given its properties M are given by:

$$\arg \max_q P(q|M) = \arg \max_q \prod_i P(q|m_i)$$

Model Building

Models are built for the top- n templates. Above a sequence similarity of 0.4, the models are built with PROMOD-II. If PROMOD-II fails to built a model, a second model is built using MODELLER. Below 0.4 sequence similarity, models are built with both PROMOD-II and MODELLER. QMEAN4 is run on the two models, and whichever model gives the higher score is selected.

Scoring of Models with QMEAN and QMEANDist

After the models have been built, the models are further assessed using two scoring functions: First, the quality of the models is assessed using the QMEAN4 scoring function. The predicted quality of the model is set to the value of q maximising the PDF of the template, multiplied by the PDF of the QMEAN4 score.

For structural scoring of the built models, we use two variants of the constraint scoring from the QMEANDist scoring function: distance constraints from the top 20 templates are extracted and weighted by

- an exponential of sequence similarity (variant A)
- an exponential of the predicted quality q of the model (variant B)

3 Results & Discussion

Coverage-Dependence

The agreement-based nature of the IDDT implies a strong dependence on coverage of the target-template alignment. In fact, for structures with an even distribution of local contacts, the IDDT is clearly bounded by the coverage of the template to the target. Since all properties are averages over the aligned residues, the properties themselves are coverage-independent. A short target-template alignment has the same average sequence similarity as a longer alignment, even though the two are clearly different, and a longer alignment is most likely superior in terms of IDDT. To simulate the coverage-dependence of IDDT, each of the properties is multiplied by the coverage, prior to correlating it to IDDT. In the following, the predictors are referred to as $cov \times x$, e.g. $cov \times id$ for the sequence identity predictor, to indicate inclusion of coverage dependence.

The strong correlation between coverage and IDDT of templates identified by HHsearch can be seen in **figure 8.1A**. Similar behaviour has been found for templates identified by HHblits and BLAST. The theoretical, coverage-imposed upper limit of IDDT values is only violated by a few target-template pairs. For low coverage values, the predicted IDDT values are more accurate. With increasing coverage, the spread increases and coverage is a less reliable predictor for template quality. Still, the correlation between q and coverage is very strong, much stronger than expected for random protein pairs. Clearly, the prediction power of coverage is heavily influenced by the fact that the plots only contain target-template pairs identified as significant hits by HHsearch. The length of these alignments, and thus the coverage, is driven by the underlying scoring model. Longer hits are only possible when they share considerable similarity to the query profile.

Discussion of the IDDT Predictors

The quality estimate of each property for the template is represented as a probability density function (PDF) of the pseudo-model having a quality q . The most probable q , according to the property, is defined as the value of q maximising the PDF. The higher-order descriptors of the PDF give additional information on the reliability of the quality estimate. Some predictors, are more accurate for low property values, whereas others are more accurate for high property values. The relations between predicted quality and the predictors of HHblits templates are shown in **figure 8.1**. $cov \times id$ is a prime example of a predictor for which the prediction accuracy changes as a function of the input. For values between 0 and 20, the spread of IDDT values for a given sequence identity is large (up to 0.6-0.7). In fact, in this region, the signal is mostly dominated by the coverage; sequence identity itself is very noisy. Above 20, the prediction accuracy increases and between 40 and 100, $cov \times id$ is a very accurate predictor for template quality. $cov \times sim$ behaves very similarly to $cov \times id$, but delivers better results over a larger range of input values. The sharp bend present in the sequence identity predictor is also visible in for sequence similarity, albeit less pronounced. Both $cov \times acc$ and $cov \times sse$ show a considerable spread of IDDT values in the upper value range. Nevertheless, when combined with the other predictors, they contribute significantly to better template selection performance.

Normalization of HH Scores

Many alignment properties are target dependent. Prominent examples of such properties include E-values, P-values, and raw alignment scores. Their scale depends both on the target length as well as the amino acid composition. As a result, it is not possible to compare E-values for different target sequences. Similar difficulties arise when comparing the potentials of mean force energy of structures with different size¹⁴².

Prior to calculating a *global* quality for the template, the properties need to be transformed to remove target-specific effects and make them comparable on a global scale. For sequence identity, sequence similarity and the agreement terms, normalization by length and coverage leads to accurate predictors. However, the HHsearch and HHblits alignment scores require a more in-depth treatment. In addition to the length-dependence, the magnitude of the alignment scores depends on the multiple sequence alignment used to derive the profile. More specifically, it depends on the number of highly conserved profile columns in the hidden Markov model: **Figure 8.2** shows that the magnitude of the average per-column score of profiles aligned to themselves increases with decreasing profile entropy. Profile entropy is a direct effect of the multiple sequence alignment used to derive the profile. For profiles derived from alignments with a high number of effective sequences, the evolutionary rates for the amino acid positions are well approximated by the sequences in the alignment. Important amino acids appear as conserved, whereas the non-conserved profile columns exhibit a high column entropy. For profiles derived from alignment with a low number of effective sequences, however, many of the amino acid columns appear as low-entropy, even though they are non-conserved. The

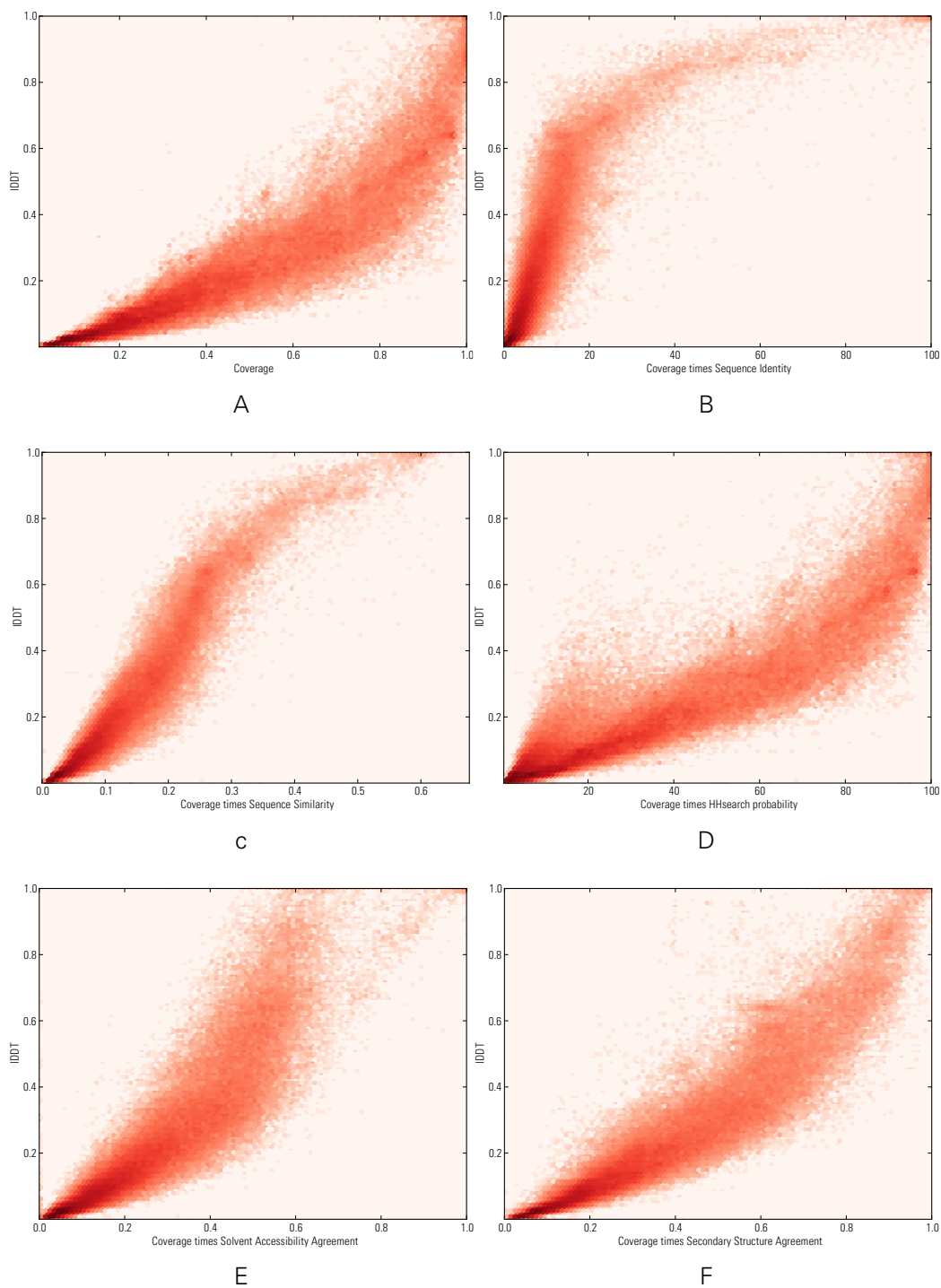


Figure 8.1 Relation between template properties and C_{α} -IDDT of the resulting pseudo-model. The plots are colored according to logarithmic point density from white to red. The following properties are shown: (A) coverage, (B) coverage times sequence identity, (C) coverage times sequence similarity, (D) coverage times HHsearch probability, (E) coverage times overlap of predicted solvent accessibility between target and template, (F) overlap predicted secondary structure between target and template

column scores for two similar, highly-conserved profile columns have a larger magnitude than scores for two similar, but less conserved profile columns, the reason being that the conserved columns are very much different from the background null model. In summary, the difference in the effective number of sequences leads to two distinct, but related effects: Target-template alignments from profiles with low average column entropy are less accurate since the profiles do not model the evolutionary events well enough. These alignments are less reliable, hence the quality of the template can less accurately be predicted (figure 8.3). For these reasons, the HHsearch score data has been split according to profile entropy and fitted individually.

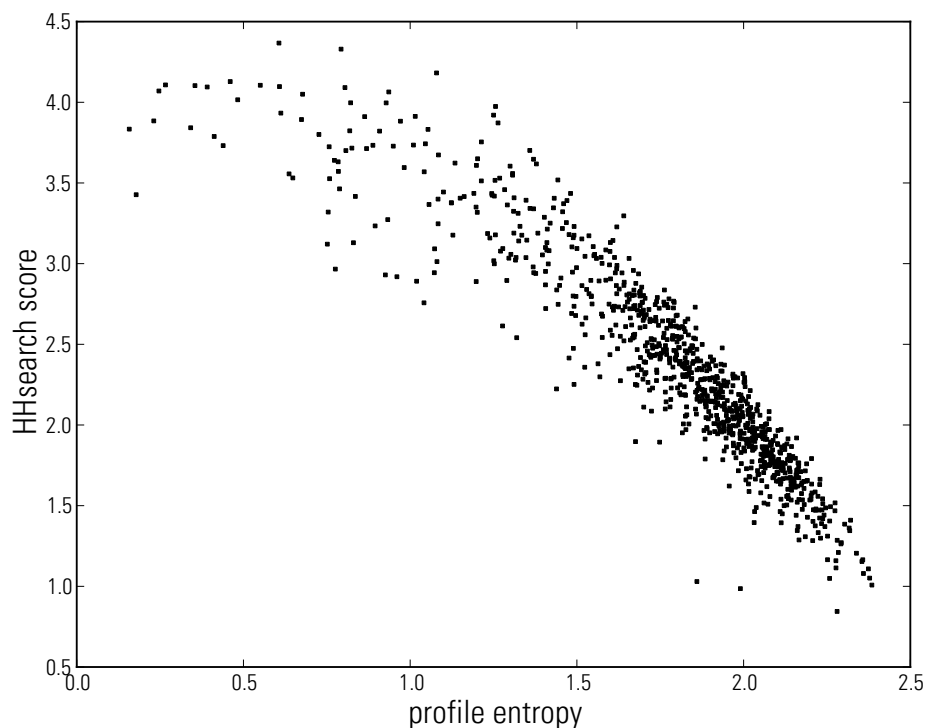


Figure 8.2 Dependency of HHsearch alignment scores on profile entropy. The HHsearch alignment scores divided by length of the alignment for the target profile aligned to itself are plotted as a function of average profile column entropy.

Choice of Bandwidth Parameter

The bandwidth parameter h used in this work for kernel density estimation is considerably larger than what would be recommended by Scott's or Silverman's rule^{246–247}. As a result, the data is only weakly approximated and the PDF is a heavily over-smoothed representation of the data. We justify the use of such large value for the bandwidth by the avoidance of overfitting of the data and the way with which the resulting probability density functions are combined. Since the final PDF is the product of the property PDF's low values of a single PDF have a strong influence on the IDDT value maximising the combined distribution. By opting for smoother distributions, a single value has in general a less drastic effect on the final IDDT, which leads to more robust estimates.

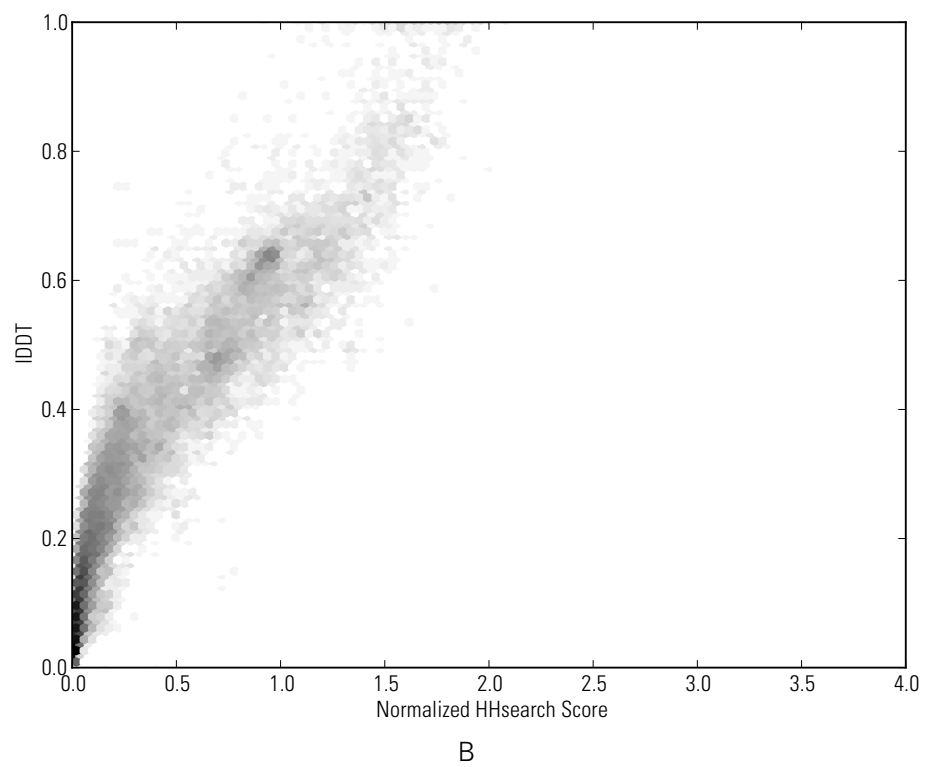
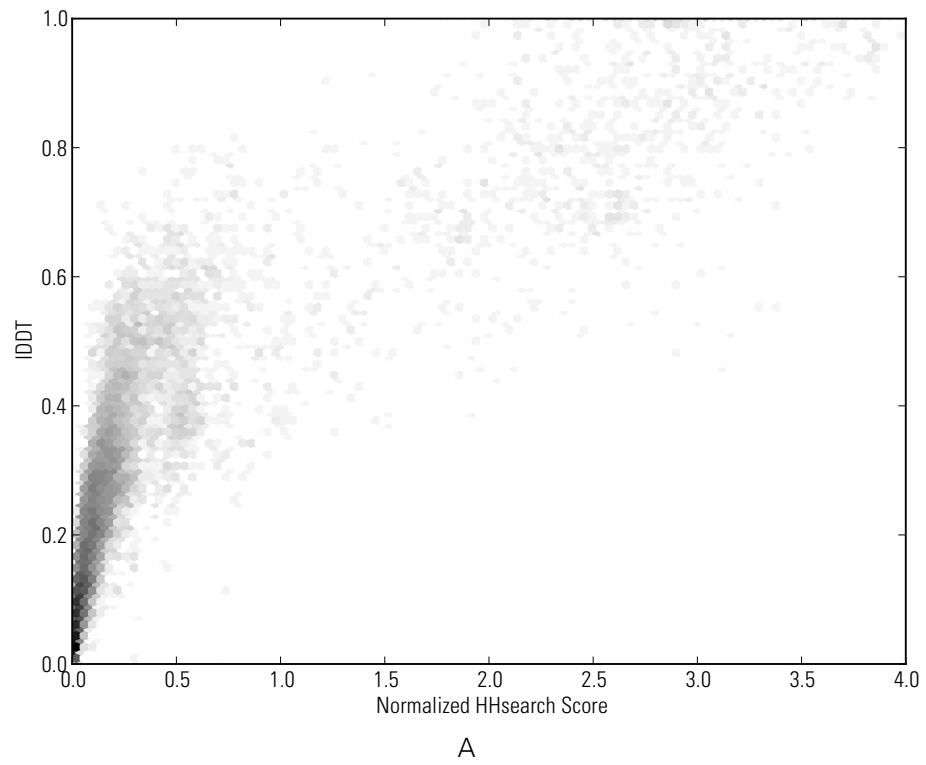


Figure 8.3 Effect of target profile entropy on HHsearch score. The normalized HHsearch score is plotted for pairs with (A) average column entropy between 1.0 and 1.6 and (B) average column entropy between 2.1 and 2.7)

Discussion of Template Selection Performance on Training Sets

From the set of targets used for training, we have derived 5 test sets that simulate template selection in all sequence similarity regimes. For each of these test sets, we have removed templates above certain sequence identity thresholds. These test sets are only approximations of typical template selection scenarios. However, they are useful in identifying which measures are most accurate selectors across difficulty levels. The test sets have been separately compiled from templates for each template identification method. In the following, we discuss the performance on the HHsearch test set. Similar behaviour was observed on the HHblits and BLAST test sets.

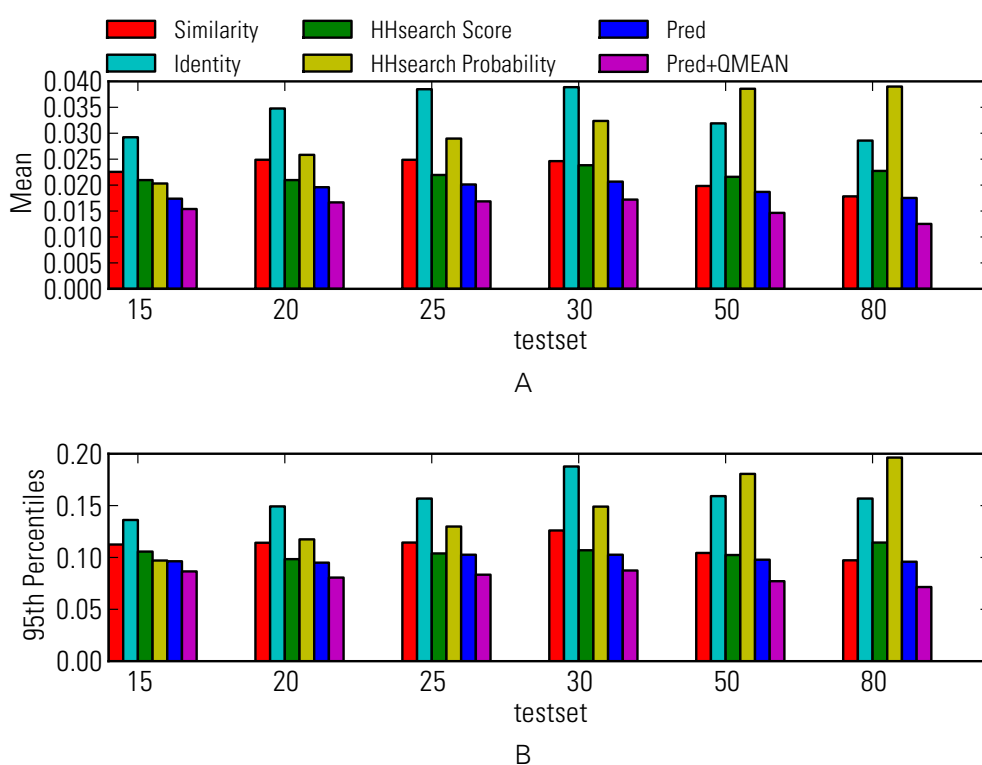


Figure 8.4 Loss of IDDT on test sets derived from training data. From left to right, the test sets culled at 15%, 20%, 25%, 30%, 50% and 80%. (A) mean loss of IDDT, (B) 95th percentiles of IDDT losses

We have found the *loss of IDDT* to be a useful measure when evaluating the template selection performance of multiple methods: The loss expresses the difference between the template with the highest IDDT in the set and the selected template. For targets, where templates above 50% sequence identity are available, sequence similarity is a better approximation of the evolutionary distance between the target and the template than HHsearch (**figure 8.4**); the templates selected by sequence achieve on average a lower loss of IDDT. The profile-profile scores smudge the sharply-peaked sequence similarity signal and make it less informative. However, below 50% percent, HHsearch score is a more accurate predictor. This illustrates that predictors based on a single property are

not likely to produce accurate results across all target difficulties. Likewise, a linear combination of properties is suboptimal as the underlying regression model assumes that the relative importance of every term remains constant. The probabilistic combination of the template's properties ($pred$) on the other hand delivers almost constant performance over the complete range of test set difficulties. The mean loss of IDDT numbers are substantially smaller than for the predictors based on a single property. The probabilistic combination of the individual properties implicitly includes the uncertainty of the properties in predicting the IDDT. Depending on the target difficulty, and available templates, template selection is driven by another set of properties. For example, the quality scores of distant templates are driven by the normalized HHsearch score, whereas for closer templates, the predicted quality is mainly determined by sequence similarity and sequence identity. The relative contribution of each property m_i is linked to the entropy of $P(q|m_i)$ (figure 8.5). When the entropy is small, the spread of expected IDDT scores decreases and contributes more to the template selection. For both the normalized HHsearch and sequence similarity, the entropy decreases with increasing similarity between the target and template. However, the entropy decrease is more pronounced for sequence similarity. For more closely related targets, the relative importance of sequence similarity for template selection thus increases.

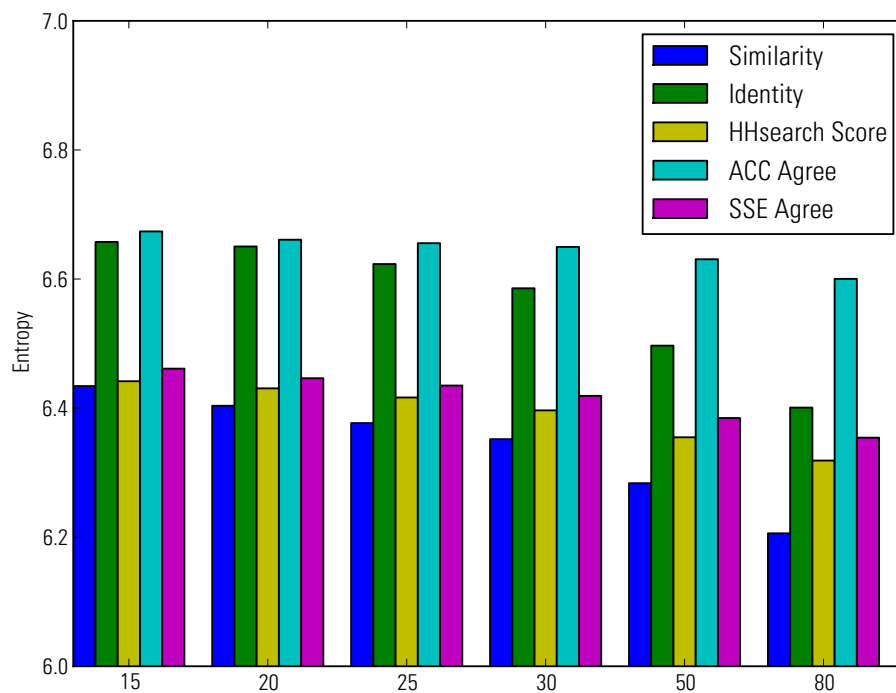


Figure 8.5 Average entropy of $P(q|m_i)$ for the template selection test sets capped at 80%, 50%, 30%, 25%, 20% and 15% sequence identity. To avoid influence of coverage, only templates with more than 90% coverage of the target sequence have been used.

Effect of QMEAN

In our tests, QMEAN contributes significantly to the template selection performance (figure 8.4). We have found that the QMEAN score is especially helpful in detecting large alignment errors and wrongly assigned folds. Models built on a misaligned template, often appear as elongated structures with their hydrophobic core turned towards the solvent. By considering sequence features alone, the misaligned templates might achieve higher scores, due to higher sequence similarity, secondary structure agreement and HH-search scores, even though it is significantly worse than other available templates. This highlights a key feature of a modeling pipelines: It is important not to prejudge the validity of sequence-structure relationships from sequence properties alone^{233,248}. Many errors are only detectable when taking the structure of the resulting model into account.

An example, where sequence feature alone are not sufficient to judge the quality of the resulting model is target 2v9vA, part of the 20% test set. The sequence has a length of 145 residues and codes for SelB of *Moorella thermatica*²⁴⁹. The models selected by sequence properties alone (built on template 2xv4A, figure 8.6A) and the model selected by sequence properties combined with QMEAN (built on template 3qphA, figure 8.6B) differ by more than 18 IDDT points (table 8.2). Both of the models receive negative QMEAN Z-scores. Still, there are substantial differences between the two models: While solvation scores are within 1σ of the expected QMEAN scores of X-ray structures for the model built on 3qphA, the solvation score for 2xv4A deviates by 4σ . Likewise, the all-atom interaction potential, $C\beta$ -interaction and torsion potentials deviate considerably from what would be expected for a protein-like conformation. This clearly indicates that the models exposes some of the hydrophobic core to the solvent. The QMEAN scores for the model built on 3qphA are, with the exception of the torsion energy, within 1σ of X-ray structure of comparable size. And indeed, the template selected by *pred + QMEAN* is the template with the highest IDDT.

Discussion of Template Selection Performance on CAMEO Test set A

The performance of template selection methods was further evaluated on a test set composed of 10 weeks of CAMEO targets. This set includes 205 targets of all difficulty levels. Since the identified targets for each template are fixed, this test measures raw template selection performance. Improvements arising from choice of threading or fold recognition algorithms are beyond the scope of this test. A complete evaluation of performance of the complete modeling pipeline, including template identification and model building steps, are described in the section *Performance on CAMEO test set B*.

The performance of the template selection methods on CAMEO test set A is summarized in table 8.3. Sequence similarity multiplied by coverage selects the best template for almost half the targets. On average, a template with 2.5 IDDT points less than the best template is selected. For the lower sequence identity range, sequence similarity is still able to select the majority of structures within 2 IDDT points, albeit the number of cases it fails increase (data not shown). Using template quality estimation (*pred*), results

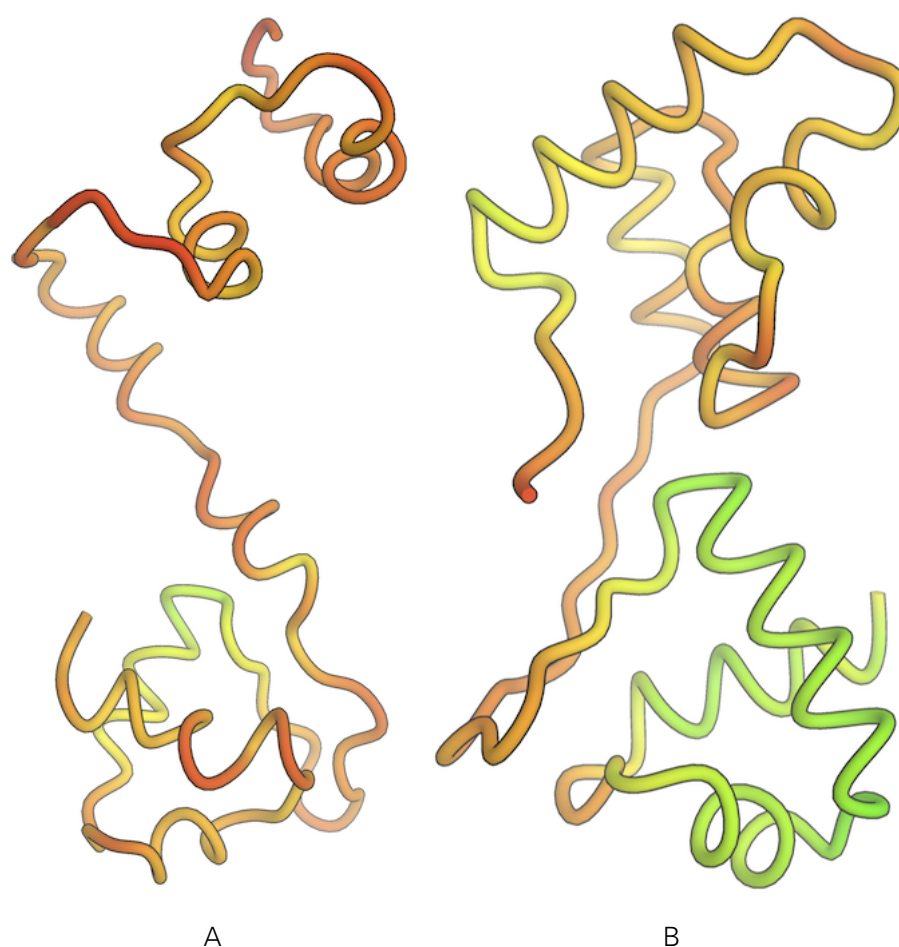


Figure 8.6 Comparison of $C\alpha$ -IDDT of models selected by *pred* (A) and *pred+QMEAN* (B). The structures are colored according to the $C\alpha$ IDDT to the target structure with a gradient from red to yellow to green.

Template	IDDT	ID	Sim	SSE	ACC	Cov
2vx4A	0.273	14.0	0.287	0.675	0.675	0.844
3qphA	0.457	11.4	0.259	0.610	0.610	0.911
	All-Atom	C- β	Torsion	Solv	QMEAN	
2vx4A	-3.09	-2.159	-2.92	-4.01	-5.32	
3qphA	-0.68	-0.98	-2.17	-0.72	-2.22	

Table 8.2 Comparison of sequence properties and QMEAN scores of models selected by *pred* (2vx4A) and *pred+QMEAN* (3qphA), respectively. ID: sequence identity, Sim: sequence similarity, SSE: secondary structure agreement, ACC: solvent accessibility agreement, Cov: coverage of template to target, All-Atom, C- β , Torsion, Solv, QMEAN: QMEAN Z-scores calculated for the models

in a lower loss of IDDT. The improvement is reflected in both lower average and median

IDDT losses. Ranking of templates by structural similarity performs equally well. On average, a template which is worse by 1.9 IDDT points is selected. Adding QMEAN to *pred* slightly lowers the loss of IDDT. The biggest improvement is achieved by using the predicted IDDT score of *pred+QMEAN* as weights for the structural scoring. Here, the loss of IDDT drops to just above one IDDT point.

Method	Mean Loss	Median Loss	Max Loss
sim	0.0266	0.0049	0.340
pred	0.0189	0.0044	0.227
struct	0.0190	0.0030	0.433
pred+QMEAN	0.0175	0.0053	0.223
pred+QMEAN+struct	0.0105	0.0036	0.170

Table 8.3 Loss of $C\alpha$ -IDDT on 10 weeks of CAMEO targets for the template selection methods evaluated in this work. *sim* selection by sequence similarity times coverage, *pred* selection by predicted quality q , *struct* selection by structural scoring with QMEANdist, *pred+QMEAN* selection by predicted quality including QMEAN4 score of model, *pred+QMEAN+struct* selection by structural scoring using QMEANdist, but replacing the sequence similarity weighting with *pred+QMEAN*.

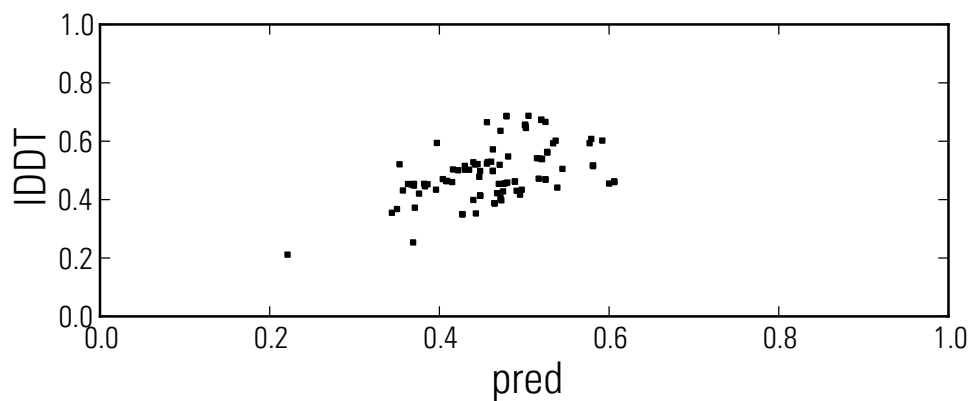
Global IDDT Prediction

One of the goals for template selection procedure presented in this work is to be able to predict the quality of the models on a global scale. To understand how accurate these predictions are on an absolute scale, we have plotted the predicted IDDT by *pred+QMEAN* for all selected templates of the 50% test set against the actual IDDT of the selected template (**figure 8.8**). Prediction and actual IDDTs achieve a Spearman rank correlation of 0.91, and a Pearson's r of 0.94. The predictions are most accurate above an IDDT of 0.6. Below, the methods tends to over-predict the IDDT. The over-prediction results in a mean difference between predicted and actual IDDT of 4.5 IDDT points.

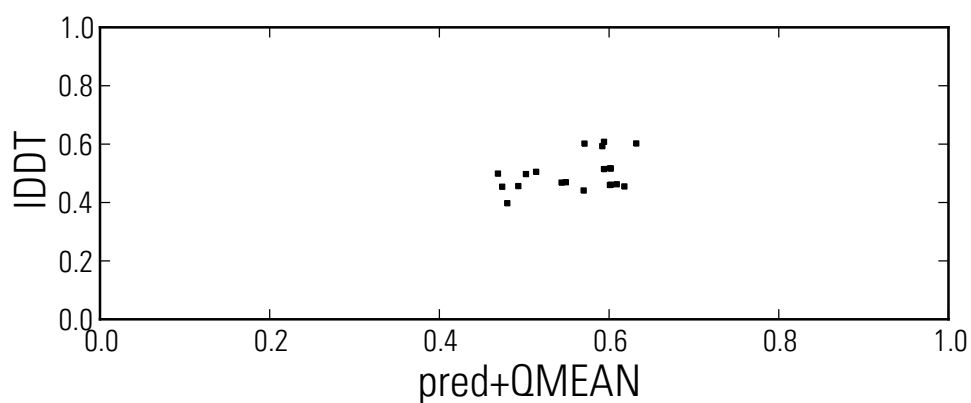
Comparison of MODELLER and PROMOD-II

Once a suitable template has been identified, the alignment is turned into a 3-dimensional model using a coordinate modeling engine. We have investigated the use of two well-established modeling programs, MODELLER⁷², and the modeling engine behind SWISS-MODEL (PROMOD-II)⁴².

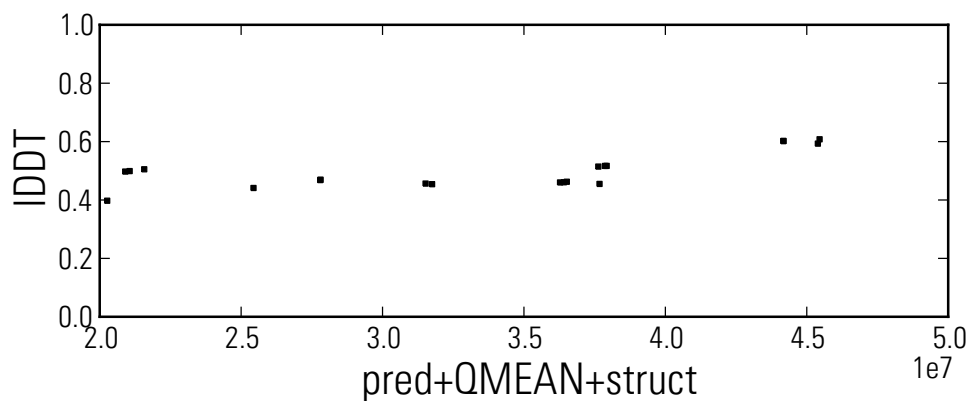
To asses which of the two approaches gives better results for single template modeling, we have built models with both MODELLER and PROMOD-II in parallel. To minimize effects from handling of non-standard residues, the target/template alignment was first turned into a pseudo-model by copying backbone and sidechains of conserved residues. A fake 100% alignment between the pseudo-model and the target sequence was then



A



B



C

Figure 8.7 An example where integrating structural information into the scoring scheme is beneficial for template selection. The 3 plots show the predicted versus actual IDDT of the template for CAMEO target 3vbp_A **(A)** Predicted quality from sequence features alone, **(B)** predicted quality including QMEAN4 score, **(C)** predicted quality including sequence features, QMEAN4 score and structural clustering. The number of points decreases from **(A)** to **(B)**, since PROMOD-II was unable to build models for the remaining cases.

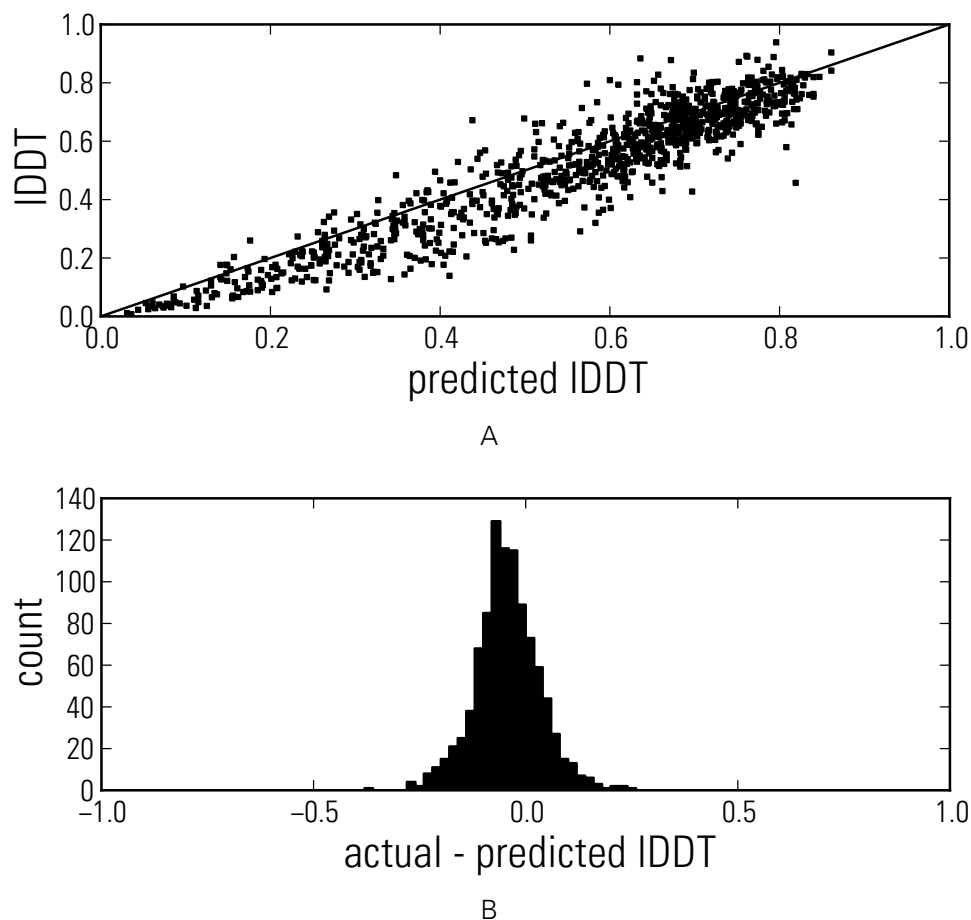


Figure 8.8 Predicted vs. actual IDDT of all selected templates of the 30% test set. (A): Scatter plot of predicted vs actual IDDT, (B) histogram of the differences between actual and predicted. Negative numbers indicate that the IDDT has been over-predicted, e.g. is lower in reality.

constructed and passed to MODELLER and PROMOD-II. The performance of MODELLER and PROMOD-II was tested on a 281 target-template alignments of various difficulty levels. Since PROMOD-II has failed to build models for 27 targets, the final comparison was performed on 254 models.

The models built by MODELLER and PROMOD-II have virtually the same IDDT scores when evaluated on the $C\alpha$ level. They cover the complete range of IDDT scores, from 0.3 to almost perfect models with IDDT scores close to one. Regardless of the target difficulty, both modeling engines produce very similar model structures. In **figure 8.9**, the all-atom IDDT scores of the models built by PROMOD-II and MODELLER are plotted against each other. Most points cluster around the diagonal. For targets with IDDT scores below 0.6, both modeling engines on average produce models of the same quality. For a few models, MODELLER is significantly better. Above IDDT scores of 0.7, the better models are mostly generated by PROMOD-II. The difference between the two modeling engines increases with higher IDDT scores.

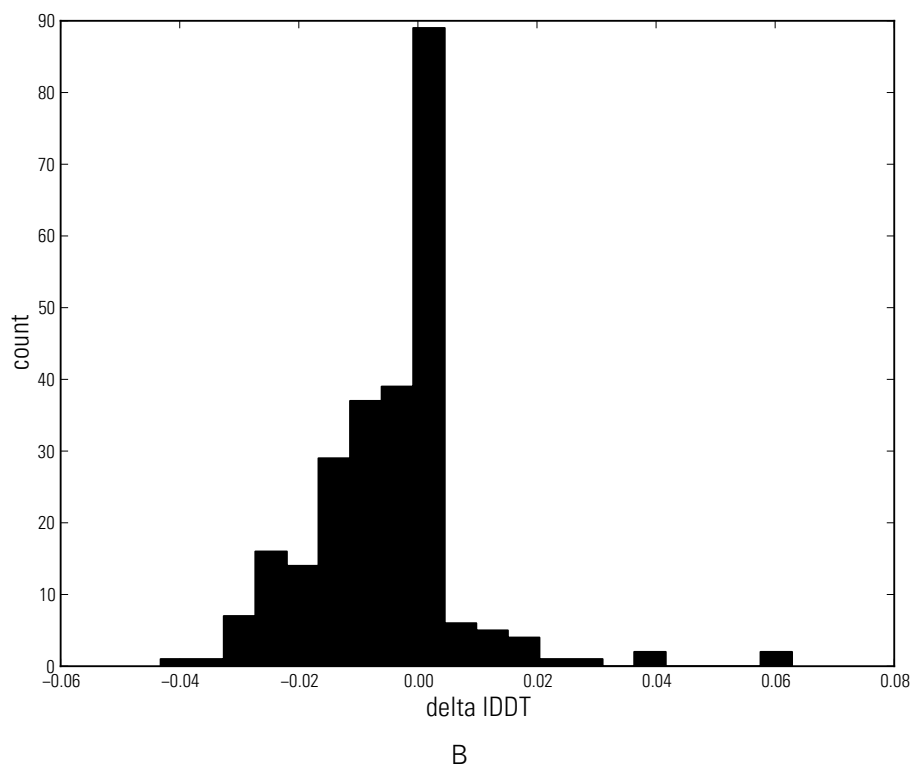
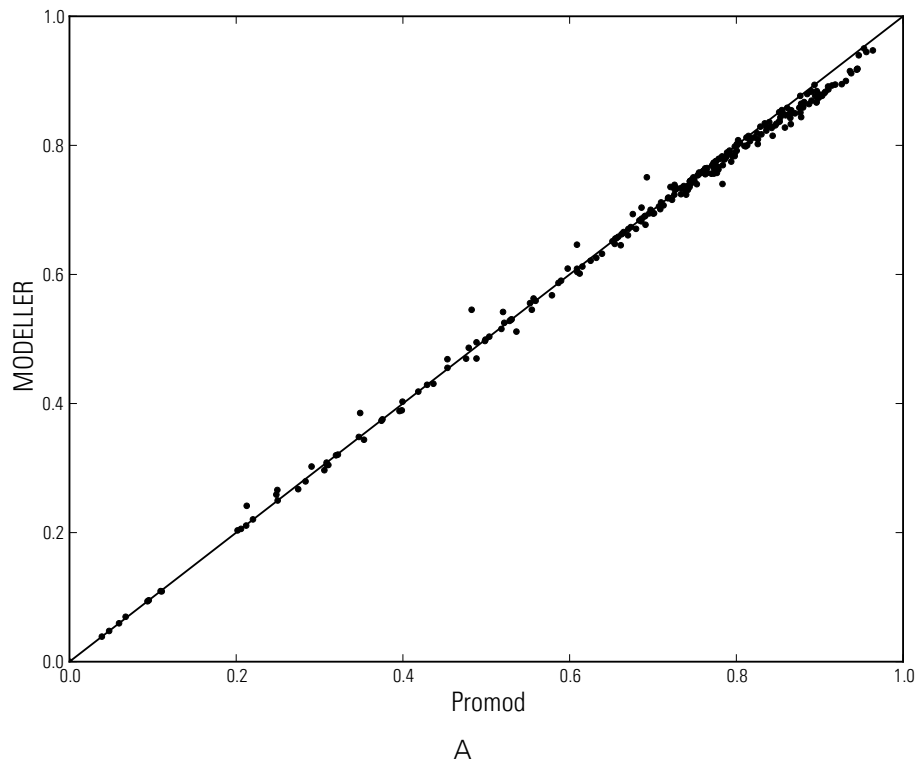


Figure 8.9 IDDT of models built on the same target-template alignment by MODELLER and PROMOD-II, respectively. **(A)** all-atom IDDT, **(B)** histogram of IDDT differences between model built by MODELLER and PROMOD-II. Negative differences mean the model built by PROMOD-II is better.

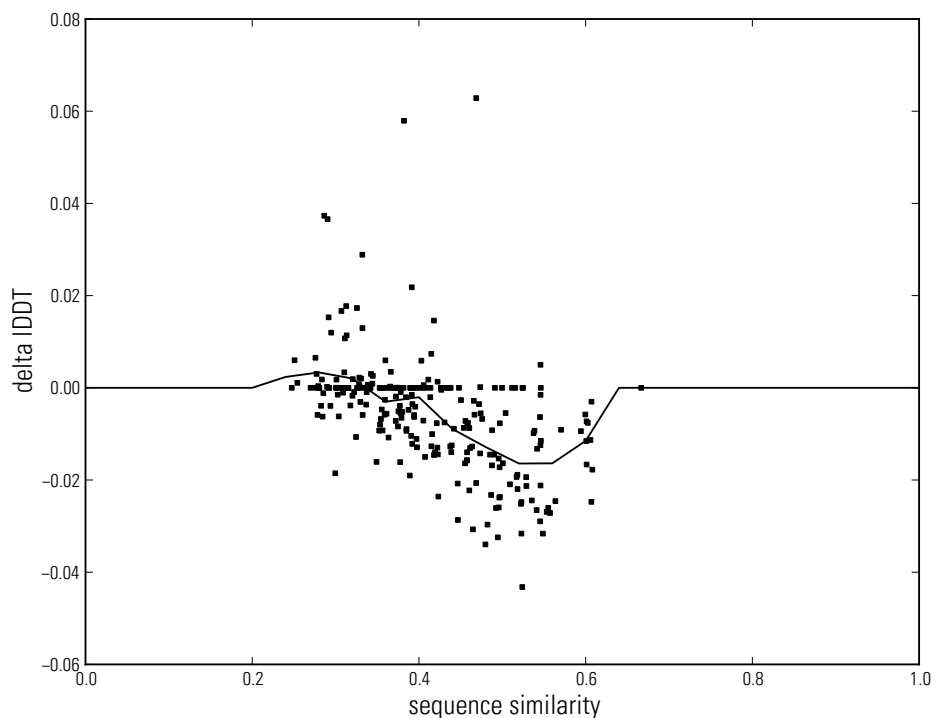


Figure 8.10 IDDT difference of models built by PROMOD-II and MODELLER plotted against the sequence similarity between target and template. Negative differences means the model built by PROMOD-II is more accurate than the model from MODELLER. The black line is a sliding average.

Since the target difficulty and thus the expected IDDT value for a certain model strongly correlates with the evolutionary distance between the target and the template, one would expect that most high IDDT models are built on templates that are relatively close to the target structure and share a high sequence similarity to the target. In **figure 8.10**, the all-atom IDDT difference between the model produced by PROMOD-II and MODELLER are plotted against the sequence similarity between the target and the template the models were built on. The same trend that can be seen in **figure 8.9** can also be seen in this figure. With increasing sequence similarity, PROMOD-II is able to consistently build more accurate models than MODELLER. Above a sequence similarity of 0.4, the PROMOD-II model is, with one exception, always at least as good as the MODELLER model. The black line on the plot shows the average difference between the MODELLER model and the PROMOD-II model. Down to a sequence similarity of 0.35, it is on average better to use PROMOD-II models. Below that, MODELLER produces slightly better models.

To analyze the reasons for differences between the models from PROMOD-II and MODELLER, the local per-residue IDDT scores from all residues of the 254 targets were calculated. To test the hypothesis, that the main differences are caused by the way the two programs approach the modeling of conserved sidechains, we have plotted the per-residue IDDT of conserved residues from MODELLER and PROMOD-II against each other. Similarly, we have plotted the non-conserved residue IDDT scores against each

other (figure 8.11). For both the conserved and non-conserved residues, the majority of the residue IDDT scores are above 0.5. The non-conserved IDDT scores overall align very well on the diagonal. The spread seems to increase for higher IDDT scores, even though this might be caused by the fact that there are more points in the high IDDT range. On average, the sidechains by MODELLER achieve a higher score by 0.001 IDDT. The difference is significant (p-value of $1e - 28$), albeit small. For the non-conserved scores, on the other hand, the differences between MODELLER and PROMOD-II are more pronounced. It can be clearly seen that the per-residue IDDT scores of PROMOD-II have an advantage over the IDDT scores from MODELLER. The effect is most prominent in the high IDDT range. The points move off-diagonal and shift towards the lower right corner. On average, the PROMOD-II residues achieve a IDDT score that is higher by 0.015 IDDT points.

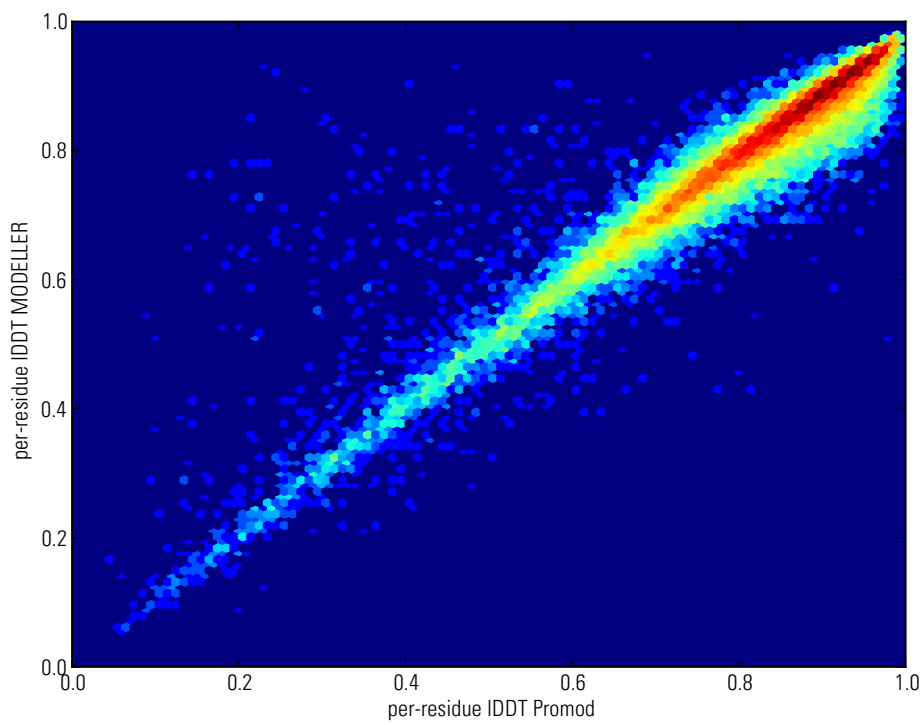
The sidechain modeling of PROMOD-II for conserved residues is very conservative. Sidechains atoms are copied *as is* from the template to the model. MODELLER, on the other hand, rebuilds the sidechains. The dihedral angle and distance constraints from the sidechains in the template are combined with prior knowledge of sidechain orientations from a large set of PDB structures⁷². Overall, the sidechains of PROMOD-II remain closer to the template structure, which in turns makes them more accurate. The differences between the PROMOD-II and MODELLER models can be explained by PROMOD-II's ability to more accurately model conserved sidechains.

COMPARISON OF LOOP MODELING | To compare the ability of the modeling programs to build insertions, we compared the IDDT scores of all residues part of an insertion. The spread between the IDDT scores of PROMOD-II and MODELLER is relatively large, suggesting that the local environment of loops built by the two program are rather different. On average, PROMOD-II builds more accurate loop residues (mean IDDT difference: 0.0091, median: 0.0018). However, no clear trend can be seen, as the spread between the IDDT scores is high.

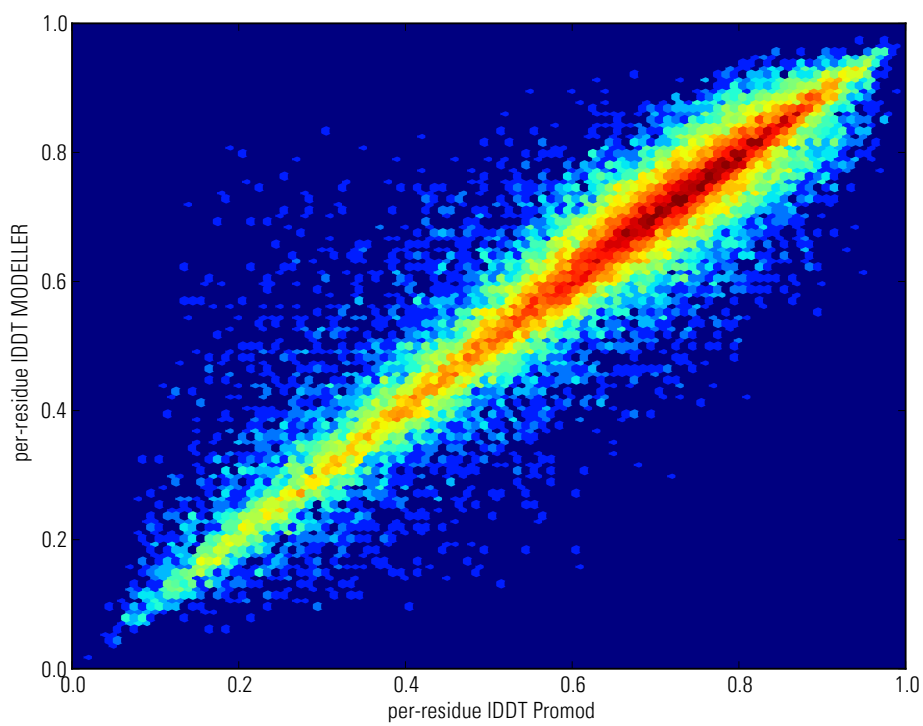
Performance on CAMEO test set B

For comparison of the pipeline to other structure prediction methods, we registered as a server in the structure prediction category of CAMEO. The registered servers differ in many aspects, e.g. template libraries, homology detection algorithms, template selection, structural modeling and sampling etc. All these aspects affect performance of a method in some way or another. This comparison gives the most complete evaluation of a pipeline's performance and assesses it as a whole.

While every server receives the same target sequences, models are not sent back for all sequences, e.g. due to timeout, server maintenance, or difficulties in building a model. Thus, a direct comparison of the servers on all targets is not possible. Instead, we will perform head-to-head comparisons of our pipeline to all participating servers. Servers will only be evaluated on common targets, e.g. targets where both servers returned at least one model.



A



B

Figure 8.11 Per-residue IDDT of models built on the same target-template alignment by MODELLER and PROMOD-II. **(A)** conserved residues only, **(B)** non-conserved residues only.

server	better	worse	tie	total	mean	median
server0 (SWISS-MODEL)	73	44	17	134	0.075	0.007
server4	89	50	14	153	0.026	0.024
server6	78	41	15	134	0.018	0.011
server9 (SMNG baseline)	78	30	43	151	0.017	0.005
server7	72	46	18	136	0.016	0.006
server13 (M4T)	36	57	8	101	0.013	-0.007
server8	68	47	15	130	0.008	0.006
server5 (IntFOLD-TS)	52	74	9	135	-0.008	-0.011
server12	50	101	8	159	-0.032	-0.015
server11 (Robetta)	33	115	10	158	-0.067	-0.046

Figure 8.12 Head-to-head comparison of the pipeline presented in this work and servers registered in the CAMEO structure prediction category. For each server, the number of models where our pipeline returns a better model, worse model equally good model (within 0.5 IDDT points), mean and median IDDT differences are listed. Names of publicly available servers are given in parenthesis.

Table 8.12 shows the results of the head-to-head comparisons. For 7 out of 10 servers, our pipeline on average produces more accurate models, according to median difference in IDDT in 6 out of 10. Similarly for 6 out of 10 servers, the number of better models is higher. SMNG is considerably outperformed by server12 and server 11 (Robetta). Here the fraction of targets, where SMNG returns a better model is small. Additionally, the average IDDT difference is 3.0 and 7.5, respectively.

To our knowledge, with the exception of server4, the servers which are on average less accurate than the method presented here (server0, server6, server7, and server8) are all single-template modeling servers. This clearly illustrates the limits of single-template modeling. For better performance, using information from multiple templates is inevitable. The only multi-template modeling server which on average produces less accurate models than our method is server4: on average the returned models are less accurate by 2 IDDT points. The bad performance of server4 can be attributed to the newly introduced stereo-chemical filtering step in IDDT. The models of server4 contain substantial stereo-chemical problems. a substantial fraction atoms is involved in distorted bonds and angles deviating considerably from the optimal values. When calculating the IDDT, interaction from these atoms are considered to be non-conserved, hence the lower scores.

The current version of SWISS-MODEL is participating as server0 in the CAMEO structure-prediction category. For many targets, SWISS-MODEL does not return any model, e.g. due to problems with loop modeling or identification of templates. On common targets, the average difference in IDDT is substantial (>7 IDDT points), whereas the median difference is close to zero. The large mean IDDT difference is mainly attributed to a few targets, where SMNG produces significantly better models (**figure 8.13**). For a few of the targets, the difference is a matter of interpretation what a good model is. For SWISS-MODEL, a shorter but more accurate model is thought to be better, whereas in SMNG coverage and local accuracy are balanced. In general, the results of SWISS-

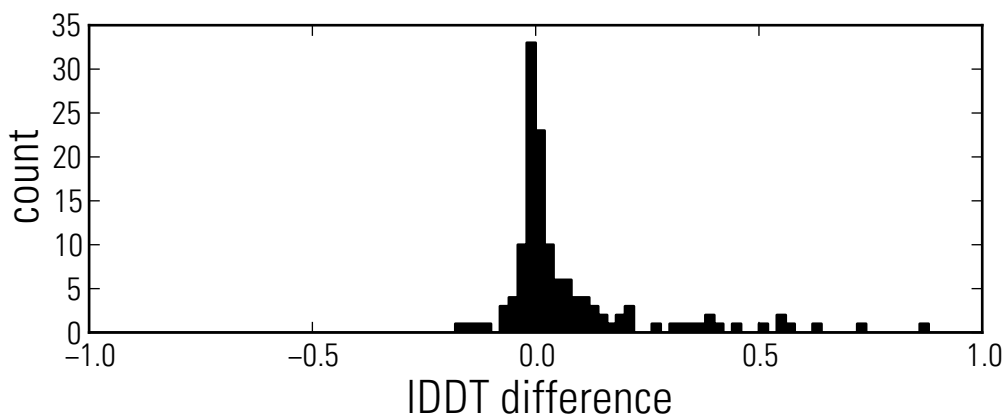


Figure 8.13 Comparison of common models of SWISS-MODEL and SMNG. Negative values denote models for which the model of SWISS-MODEL is more accurate.

MODEL tend to get better when templates with high sequence identity to the target are available. Here, the applied template selection protocol, which is primarily based on sequence identity is superior. The selected template in SMNG is sometimes drawn away from the one with the highest sequence identity by the structural clustering step. It should be investigated whether structural clustering should be disabled for high-accuracy targets in SMNG.

4 Conclusions

In this work, we have outlined the automated modeling pipeline of SWISS-MODEL Next Generation. The idea to formulate the template selection as a quality estimation problem has been used by other automated modeling pipelines, e.g. IntFOLD-TS²³³ and HH-pred³². However, to our knowledge, this is the first time that quality estimates are combined using a probabilistic approach. The method presented in this work is both fast and intuitive to understand. The influence of each measure on the final template selection is proportional to the uncertainty of the feature. Unreliable properties are down-weighted and contribute less to the final score. The reliability of each feature is a function of the property itself and is adapted based on fits to target-template alignments of known quality. Scaling of uncertainty is an important feature of the template selection and crucial for good performance in all target difficulty regimes.

We have shown that the reliability of the IDDT prediction from HHsearch and HHblits scores heavily depends on the average column entropy of the target profile. The accuracy of IDDT estimation is improved by fitting the normalized HHsearch/HHblits scores for low and high entropy profiles separately. Since the solvent-accessibility and secondary structure predictions are calculated from the HHsearch and HHblits profiles, the performance might also depend on the average column entropy. It remains to be seen to what extent the performance degrades when only few sequences are available. Separate fits for low and high entropy profiles might be beneficial as well.

On the in-house test sets, we have shown that the performance of selecting templates using a combination of properties is significantly better than a single feature, such as sequence similarity. Additionally, we have shown the benefit of incorporating QMEAN into the scoring of models. QMEAN reduces the number of wrongly-aligned templates considerably. On average, QMEAN is also beneficial in reducing alignment errors in models.

Scoring the models structurally by consensus of local contacts tremendously lowers the loss of IDDT. Local contacts present in multiple models are reinforced. Models satisfying more of the *consensus contacts* are ranked higher. This drives the selection of the top-ranking model towards the *average template structure*, a desirable property, since alignment errors tend to cancel out. As noted by Peng and Xu⁶⁹, structural information can have negative effects for alignment accuracy and template selection performance in the high sequence identity regime. It shall be investigated whether it is beneficial to disable the structural consensus scoring for simple modeling cases, e.g. templates with a high predicted IDDT.

When pre-multiplying the properties with coverage, two independent properties of the target-template alignment are combined into one number. A template with 100% sequence identity and 50% coverage will get the same predicted quality as a template with 50% sequence identity and 100% coverage. However, the 100% sequence identity template is clearly preferable as is much more closely related to the target and should get a higher score. For modeling on single templates, combining coverage and properties is a viable solution. However, in the context of multi-template modeling, smaller coverage of a template structure can be compensated by extending the model with another template. Instead, the IDDT predictors could be redesigned to predict a coverage-corrected IDDT, and take the length of the target-template alignment into account.

The comparison to existing TBM servers on the CAMEO live benchmark has shown that for many targets single-template modeling delivers better performance than multi-template modeling approaches. However, substantial room for improvement is possible, particularly when templates only cover parts of the target sequence.

SMNG Web Interface

In his chapter, we give a short overview of the newly developed user interface for SWISS-MODEL. This work has been performed together with Konstantin Arnold, Stefan Bienert, Tobias Schmidt and Andrew Waterhouse. KA has been setting up the productive web-server, SB has implemented the backend, AW implemented the frontend, TS has implemented the backend of the ligand annotation, MB has been coordinating and managing the project.

1 Introduction

SWISS-MODEL is one of the most widely-used web services for homology modeling^{42,130–131,232,244}. Around 2000 models are built every day by scientists around the world. The modeling is primarily motivated by a particular biological research question, and serves as the basis for the design of mutagenesis experiments, structure-based drug design²⁵⁰, binding site studies^{171,204}, binding site prediction¹⁶⁹ etc. The biology is not a side-show of the model building but a primary concern. One of the goals of homology-modeling pipelines thus has to be to provide models that are directly applicable to particular research questions. Several factors contribute to usefulness of models: First and foremost, models have to be accurate, as the applicability of models strongly correlates with quality²⁰⁴. In addition, the biological context of the model is important. This includes modeling of the protein in its correct oligomeric state, and the inclusion of biologically relevant ligands. Last but not least, reliable local quality estimates give the researcher an idea in which areas a model can be trusted.

In this chapter, the user interface of SWISS-MODEL Next Generation and the template library are outlined. For a scientific description of the modeling pipeline itself, we refer to the 'Automated Modeling in SWISS-MODEL Next Generation' chapter on [page 119](#).

2 Implementation

The backend of SMNG is written in Python²⁵¹ and uses the Django web framework²⁵². Python seamlessly integrates with OpenStructure¹³⁵ for input validation of sequences, alignments and structure files. Input validation is performed directly in the HTTP request handler. More demanding calculations, such as sequence searches and model building are delegated to the cluster using the Sun Grid Engine. The web-interface then polls for completion of the calculation.

On the client-side, SMNG uses the jQuery JavaScript library²⁵³ for cross-browser compatibility. For rendering of vector graphics, the frontend uses raphael.js²⁵⁴ which provides a convenient interface on top of SVG (or SGML for older versions of Internet Explorer).

The frontend communicates with the backend asynchronously. The data is requested when used the first time and sent back as a JSON (JavaScript Object Notation) stream. The use of asynchronous requests is important for reducing the latency in an interactive interface. Once the page has been loaded, only parts of the website which change need to be transferred from the server.

3 Template Library

Since homology modeling pipelines draw most of their strength from available experimental structures, ready and fast access to this information is key. Even more so, when biological information has to be available at every step of the homology modeling pipeline. In SWISS-MODEL the template information is stored in a newly designed template library (SMTL). It is a large database of structural and biological information for experimentally determined protein structures derived from the PDB^{18,21}.

Biological Units

The smallest unit of structural information in the SMTL are the biologically functional units, short biounits. It is the SMTL's counterpart of biological assemblies of the PDB. Since there might be multiple assemblies for a given PDB entry, there is a 1 to n mapping between PDB and SMTL entries. The biounits consist of one or more entities, i.e. types of molecules present in the structure. For example, a typical homo-tetramer consists of four polypeptide chains and solvent molecules. In the biounit, the four polypeptide chains are grouped as one *polypeptidic* entity, the water in a second entity. For a hetero-oligomer, the peptide are grouped by their SEQRES sequence, which results in one entity per unique sequence found in the structure.

When new structures have been released by the PDB, they are typically added to the SMTL within 2 weeks. Information for the entries, such as deposited sequences, coordinates, primary citations etc. are read from the mmCIF files and converted to a SMTL-specific format. Annotations on the residue level, such as predicted secondary structure, solvent accessibility, and DSSP states for the residues are calculated and stored alongside the structures. Likewise, HHsearch and HHblits profiles are calculated if one of the deposited structures contains a sequence that was not previously part of the SMTL.

Deprecated entries are kept in the database for backward compatibility, since they might still be referenced by some projects. However, the chains are removed from the indices and will not be available for a sequence/profile search performed after deprecation.

Sequence and Profile Databases

Sequences and HMMs of the SEQRES sequences of all chains in the template library are stored in databases that can be directly searched with BLAST⁵⁸, HHsearch³² and HH-

blits³³. Since there are usually multiple polypeptide mapping to the same SEQRES sequence, we normalize the sequences by storing each sequence only once. The sequences are referenced by their MD5, and denormalized upon use.

The following databases exist:

- SMTL100 is a non-redundant sequence database containing all SEQRES chains currently in the SMTL. SMTL90 and SMTL70 are like the SMTL100 a BLASTable sequence databases but culled at 90% and 70% sequence identity, respectively.
- HMM70 is an HHsearch profile library containing the centroids of sequence clusters at 70%. Profiles for individual SEQRES sequences are kept in the profiles subdirectory. Typically, homology detection with HHsearch first runs the query profile against the 70% clustered database and then creates a temporary database containing all profiles of centroids identified as hits. This two-layer approach has proven to be effective in reducing the memory consumption and runtime of the HHsearch alignment step.
- HHBLITSDB contains the profiles for HHblits. Since database creation for HHblits databases is much more time-consuming than for HHsearch databases, database creation time becomes limiting and searching against the 70% cluster centroids and extension of the searches to all sequences of these clusters is not beneficial for search performance. Instead all sequences are kept in one database.

Annotation of Cloning Artifacts

A large number of proteins crystallized today are purified using purification tags such as poly-histidine tags, TAP tags etc²⁵⁵. While of importance to the purification itself, these tags are not relevant to the biological function of the protein. Purification tags pose challenges for sequence searches. When the target sequence contains expression tags, often, templates containing the same tag as the target sequence show up as high-identity hits, even though the protein itself shares little sequence similarity with the target sequence. Thus, expression tags should be excluded from the alignments when selecting templates.

Deposited entries contain partial purification tag annotations. While the number of well-annotated sequences increases, there is still a large fraction of entries which is not properly annotated. For these cases, we rely on heuristics to determine expression tags. The complete annotation procedure is as follows:

- Sequence regions containing the word "TAG" are annotated as expression tags. In addition, sequence regions with the annotation 'cloning' in different spelling variations (clonning, clonong, etc) are annotated as tags.
- We have found that many of the regions identified in the first step usually only encompass the actual purification tag and do not include linker regions. Thus, we add residues next to regions identified as tags in the first step, when they could not be mapped to any sequence in reference sequence databases such as UNIPROT.
- Regions of sequences containing four histidines in a row and not related to any sequence in sequence databases (not including PDB) are also annotated as tags.

Web Access to the Template Library

Virtually all information from the template library entries, .e.g. secondary structure annotations, calculated profiles, coordinates, are available in a web interface. A typical SMTL entry, the methyl transferase from the Dengue virus in complex with S-adenosyl homocysteine and ribavirin 5 (1r6a.1), is shown in **figure 9.1**. In the top-left, the structure is rendered with OpenAstexViewer²⁵⁶. The top right lists primary citations, ligands and the polypeptide chains present in the biological assembly. The bottom part shows the sequences of the chains present in the structure aligned to the SEQRES entries. It can be immediately seen that 5 residues at the N-terminus and more than 30 residues at the C-terminus of chain A are not resolved in the structure. The 3-dimensional structure viewer is synchronized with the sequences displayed in the bottom part. Moving the mouse over a residues will highlight it in the 3D view. Likewise, clicking on a secondary structure element centers the 3D display on that element.

The screenshot displays the SWISS-MODEL Template Library web interface for entry 1r6a.1. The top navigation bar includes the SWISS-MODEL logo and the entry ID. The main content area is divided into two columns. The left column features a 3D molecular model of the protein structure, with a 'Reset' button and a 'Genome polyprotein' section. The right column contains a detailed information panel for the entry, including the title, method, oligo state, ligands, links, citation, and chains. Below the information panel, the genome polyprotein sequence is displayed with a color key and alignment markers. The bottom of the page includes the Swiss Institute of Bioinformatics logo and a 'Back to the Top' link.

Figure 9.1 Entries from the SMTL are accessible from the web interface.

The web-interface also features an online annotation system for ligands. Ligands can be annotated as buffer, post-translational modifications, biologically relevant or synthetic binders etc. The annotations are stored for each biounit, which allows to mark potassium ions, which are typically part of the solvent, as biologically relevant in an potassium channel. The system has been developed for the CAMEO ligand binding category, and allows users to decide which ligands should be included in the evaluation. The same system is used to decide which ligands are biologically relevant when building homology models. In **figure 9.1**, sulfate ions are annotated as buffer, whereas S-adenosyl-L-homocysteine is annotated as biological ligand and ribavirin as a synthetic molecule. The ligand annotation system is described in more detail in the PhD thesis of Tobias Schmidt²⁵⁷.

4 Modeling Interface Walkthrough

Input

Typically, a modeling project starts with a target sequence for the protein of interest. The sequence can be entered as FASTA, ClustalW, ProMod, PIR, plain string, or in the form of a UNIPROT accession code. As soon as the sequence is entered, the input is sent to the server for validation. If the input is a valid sequence, the input form is hidden and replaced by a rendered display of the target sequence. The immediate feedback is important, as input boxes are a major source of confusion for users. The validation procedure informs the user immediately whether the input was understood by the web server (figure 9.2) or not.

The screenshot shows the SWISS-MODEL workspace interface. At the top, there is a navigation bar with the SWISS-MODEL logo and the text 'SWISS-MODEL | Workspace'. Below this, there is a header section with the SIB logo and 'BIOZENTRUM Universität Basel The Center for Molecular Life Sciences'. The main content area is titled 'Template Library | Modelling'. It features a 'Target Sequence:' input field with a file upload icon and a 'Cancel' button. Below the input field, there are three lines of target sequence data, each with a 'Target' label and a corresponding sequence and residue count. The first line is 'MVFVAVCVINGDAKGTVFFEQESSGTFVKVSGEVCGLAKGLHGFHVHFGDNTNGCMSSG 60', the second is 'PHFNYPYGKEHGAPVDENRHLGDLGNIEATGDCPTKVNITDSKITLFGADSIIGRTVVVHA 120', and the third is 'DADDLGQGGHELKSTGNAGARIGCGVIGIAKV 153'. Below the sequences, there are input fields for 'Project Title' (containing 'SODC_DROME P61851 Superoxide dismutase [Cu-Zn] (1.15.1.1)'), 'Email' (containing 'marco.biasini@unibas.ch'), and 'Template Search Program(s):' with checkboxes for 'Blast' and 'HHblits'. At the bottom, there are two buttons: 'Search For Templates' and 'Build Model'. A footer message reads 'You are currently not logged in - to take advantage of the workspace, please log in.' The bottom navigation bar includes 'Swiss Institute of Bioinformatics', 'Contact Us', and 'Back to the Top'.

Figure 9.2 A new modeling project is typically started by entering a target sequence or UNIPROT accession code. The input is immediately validated when entered.

In addition to single sequences, a SWISS-MODEL projects can be started by

- a target/template alignment, allowing the user to specify the mapping between target and template residues. Since the sequence of the user's template might be different from the sequence in the SMTL (non-standard residues, atom vs. SEQRES etc.), the template sequence is aligned to the corresponding entry in the SMTL. Correspondence is checked by testing various pattern of PDB and SMTL identifiers. The user's

template sequence is required to share at least 80% sequence identity with the SMTL sequence.

- a SWISS PDB Viewer project file.
- uploading a template structure and a target sequence. This option is useful when the user is in possession of an unpublished, experimentally determined structure.

Once the input has been validated, the user can choose to start a new project using either manual or automated mode. In manual mode, a template search is performed and the user selects suitable templates by hand. Automated mode uses the automated template selection procedure outlined in more detail in the 'Automated Modeling in SWISS-MODEL Next Generation' chapter and directly returns models.

For this walkthrough, we will use the Superoxide dismutase from *D. melanogaster* (UNIPROT accession code: P61851).

Template Search Results

In manual modeling mode, the user is redirected to the template results page when the template search completes. The page serves both as an overview of available templates as well as an interactive template selection tool.

The top part of the screen contains a summary of the top-ranking templates identified by the template search methods (**figure 9.3**). Three types of views are available: a template summary table, listing all templates in tabular form, and two sequence/structural similarity plots which show the templates in relation to each other. Templates for a subsequent modeling step can be selected in any of the three views. Since the selections between the views are synchronized, selection of a template in the table, automatically selects the template in the two similarity plots and vice versa. The alignment of all selected templates is shown in the lower part of the screen.

ALIGNMENT VIEWER | For the purpose of sequence and alignment visualisation the SMNG web interface contains an interactive alignment viewer (**figure 9.3**, bottom). The alignment viewer is used in all parts of the website, since it is central to the template selection process, it is described here in more detail. The alignment viewer supports a variety of display styles and coloring modes, which can be changed dynamically. For example, the ClustalW coloring scheme groups amino acids by property and colors each group separately. Other options are to color by amino acid property, B-factor of the template structure, solvent accessibility or local QMEAN score of the models. Secondary structure, predicted or calculated by DSSP, can be shown as an SVG overlay: helices and extended segments are drawn as boxes and arrows around the sequences, respectively. In addition, the alignment viewer is synchronized with a 3D structure view. Hovering over a residue in the alignment view automatically highlights the residues in the 3D display. Likewise, the coloring schemes are shared between the alignment viewer and the 3D display.

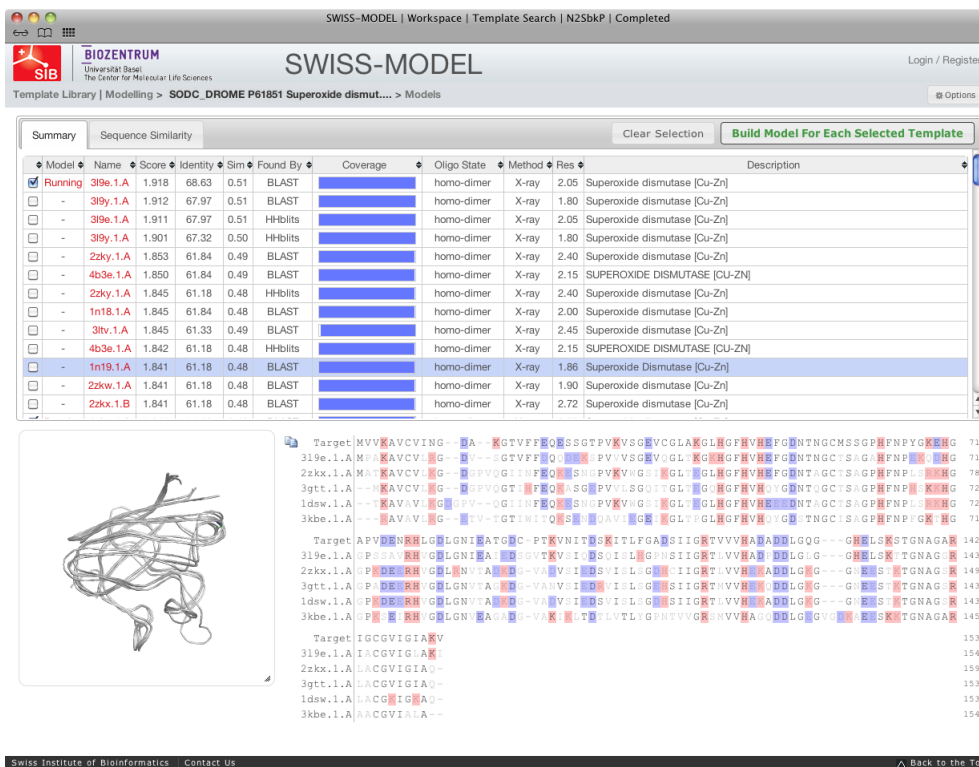


Figure 9.3 Templates identified for the superoxide dismutase from *D. melanogaster*. The top part lists the identified templates in tabular form. At the bottom, the selected templates are shown and superposed together with the alignment. Charged amino acids in the target and template sequences are colored in red and blue. Mismatching residues in the alignment are shown in grey. The predicted secondary structure for the templates and the target sequence is displayed using an SVG overlay. The sequence alignment and structure views are synchronized, hovering over one of the residues in the alignment will highlight it in the structure.

SEQUENCE SIMILARITY PLOT | As an alternative form of display, the template search results page contains a graph, which displays the templates in relation to each other. This complements the table of identified templates with a focus on the relations between templates. The target sequence is depicted as a red dot together with each template shown as a circle. The distances between the templates are proportional to the pairwise similarity between the templates. An example of two sequence similarity plots is shown in **figure 9.4**. The target-sequence coverage is shown in a thick line around the circle. For example, a template which covers the N-terminal half of the template will have a thick border from top to bottom in clockwise direction. Templates which share a high sequence identity, can be immediately identified, as they group together on the screen. We have found the sequence similarity plot to be a very helpful tool in understanding the available templates.

The layout of the template plot is performed by the *neato* program of the GraphViz package²⁵⁸, which performs a dimensionality reduction, from the high-dimensional sequence and structure space to the two dimensions of a computer screen. The dimensionality reduction inevitably leads to a loss of information, and two templates sometimes

appear close on the computer screen, even though they are far apart in sequence space.

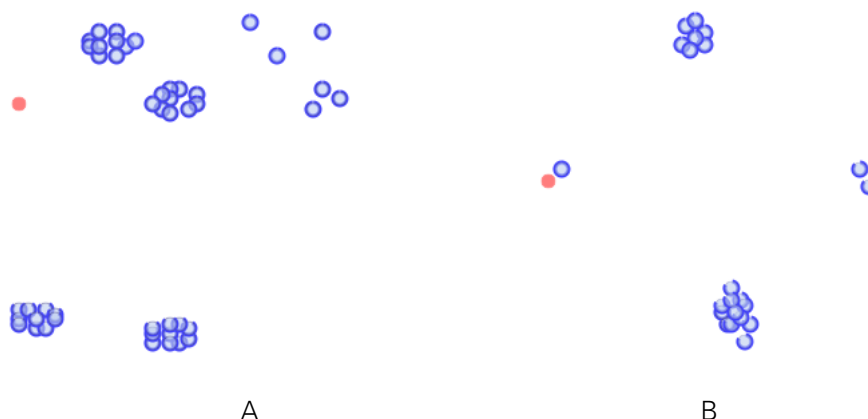


Figure 9.4 Two example sequence similarity plots of identified templates. (A) ATP synthase subunit beta (P991122), (B) Flavodoxin (P00321). Similar templates cluster together and appear in *grape*-like arrangements.

ORIENTING THE TEMPLATES | For visual comparison of the identified template structures, it is important that the template structures are shown in the same orientation. Since the orientation of the structures in the PDB is arbitrary, the structures need to be aligned accordingly before displaying them. For this purpose, we transform all template structures into the same frame of reference according to a sequence-dependent superposition. The alignment of all structures is performed as follows:

- The sequences of all templates are clustered at 90% sequence identity.
- The template structures within the cluster are structurally aligned onto the centroid of the cluster (the template with highest coverage of the target sequence). The result of the superposition is a rotation matrix and a translation vector which optimally superposed the template structures onto the cluster centroid.
- We apply the **DOMF** algorithm (chapter 'Graph-based constraint selection for multi-template modeling') to all pairs of centroids and recursively group the centroids with the largest number of residues part of domains. The larger of the two structures is chosen as the representative and the smaller superposed onto it. The merging continues until all structures are merged.

The result of this procedure is a tree of transformations. To transform all structures into the same frame of reference, we traverse the tree and apply the transformation matrices to the structures.

The superposition of template structures makes it possible to directly compare the templates by visual inspection. Flexibility and variability in the structures become immediately apparent. While the majority of images created with the procedure outlined above are close to perfect, in a few instances the images are suboptimal. One of the problems arises through the use of *molscript* to generate static images. The viewer's position is automatically adjusted to fill the available space on the image. As a result, the distance

of the camera to the protein is affected by protein size. For smaller proteins, the viewer moves closer to the protein, whereas larger proteins require the viewer to move back. The differences in the viewer's position are especially noticeable when open and closed conformations of a protein are available. The positioning of the camera would need to be fixed for all images, e.g. by calculating the bounding box of all template structures and using that to position the camera.

Modeling Results

Once models are built with manual or automated modeling mode, they are available on the template results page. A short summary of the template, global model quality¹⁴², the included ligands, and oligomeric state is displayed in small boxes (figure 9.5). To explore different aspects of the models, the user can switch between various coloring modes. Typical tasks include inspection of local model quality or modeled ligands. For example, since insertions and deletions are major sources for uncertainty in models, it is usually a good idea to check the locations of insertions and deletions, and to assess if these loops play a role in the binding site. This task is facilitated by coloring the locations of insertions and deletions in the model. Alternatively, as indicator of local model quality, the structures can be colored by local QMEAN scores.

SWISS-MODEL | Workspace | Model Results | N25bkP : 07

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SWISS-MODEL

Template Library > Modelling > SODC_DROME P61851 Superoxide dismut... > Models > 07

Project: SODC_DROME P61851 Superoxide dismutase [Cu-Zn] (1.15.1.1); Created: 2013-02-07. Report Archive

Logs

2D 3D detach Reset

Model Information

Coordinates: [img]

Residue Range: [img] (A:2-153) [img] (B:1-153)

Oligo State: homo-dimer

Ligands: 2 x ZN, 2 x CU

Template Information

Template: 319y.1.A

Title: Superoxide dismutase [Cu-Zn]

Oligo state: homo-dimer

Seq Id (%): 67.97

Ligands: 2 x ZN, 2 x CU

Resolution: X-ray, 1.80Å

Seq Sim: 0.51

Coverage: 1.00

Found By: BLAST

Model Quality

Qmean4: [img] -0.80

CBeta: [img] 0.59

All Atom: [img] -0.92

Solvation: [img] -2.89

Torsion: [img] 0.97

Model-Template Alignment: [-]

Model:A	M	V	V	K	A	V	C	V	I	N	G	D	A	K	T	V	F	F	E	Q	E	S	S	G	T	P	Y	K	V	S	G	E	V	C	L	A	K	L	G	L	H	G	F	H	V	H	E	F	G	D	N	T	N	G	C	M	S	S	G	P	H	F	N	P	Y	G	K	E	H	G	A	P	V	D	E	N	R	H	L	80
Model:B	M	V	V	K	A	V	C	V	I	N	G	D	A	K	T	V	F	F	E	Q	E	S	S	G	T	P	Y	K	V	S	G	E	V	C	L	A	K	L	G	L	H	G	F	H	V	H	E	F	G	D	N	T	N	G	C	M	S	S	G	P	H	F	N	P	Y	G	K	E	H	G	A	P	V	D	E	N	R	H	L	80
319y.1.A	M	P	A	K	A	V	C	V	L	R	G	D	V	S	G	T	V	F	F	Q	D	E	K	S	P	V	V	S	G	E	V	Q	G	L	T	K	G	R	H	G	F	H	V	H	E	F	G	D	N	T	N	G	C	T	S	A	G	A	H	F	N	P	E	K	Q	D	H	G	C	P	S	S	A	V	R	H	V	80		
Model:A	C	D	L	G	N	I	E	A	T	G	D	C	P	T	K	V	N	I	T	D	S	K	I	T	L	F	G	A	D	S	I	I	G	R	T	V	V	H	A	D	A	D	L	G	Q	G	G	H	E	L	S	K	T	G	N	A	G	A	R	I	C	G	V	I	G	I	A	K	V	153										
Model:B	C	D	L	G	N	I	E	A	T	G	D	C	P	T	K	V	N	I	T	D	S	K	I	T	L	F	G	A	D	S	I	I	G	R	T	V	V	H	A	D	A	D	L	G	Q	G	H	E	L	S	K	T	G	N	A	G	A	R	I	C	G	V	I	G	I	A	K	V	153											
319y.1.A	G	D	L	G	N	I	E	A	T	E	D	A	G	V	T	K	V	S	I	Q	D	S	Q	I	S	L	H	G	P	N	S	I	I	G	R	T	L	V	V	H	A	D	P	D	D	L	G	G	N	E	L	S	K	T	T	G	N	A	G	G	R	I	A	C	G	V	I	G	I	A	K	I	154							

Swiss Institute of Bioinformatics | Contact Us | Back to the Top

Figure 9.5 Information for built models is shown on the model results page.

Ligands are added to the models from the templates, provided the binding site is conserved and the ligands are marked as biologically relevant in the SMTL. Additionally, at

the user's choice, the oligomeric state of the model can be based on the template. Work is on the way to include the oligomeric state prediction developed by Florian Kiefer during his PhD²⁵⁹. When the oligomeric state of the template and the model is predicted to be the same, the model is built as a homo-oligomer.

For archiving and later retrieval, the results of a SWISS-MODEL homology modeling project are available as a downloadable archive containing all relevant files, including the coordinates of the models and QMEAN scores. In addition, a summary of the modeling project is available in a project report. The report describes the applied methods in text-form and lists versions of databases and programs used. The primary motivation of the project report has been to create a standardized text document which can be copied and pasted into a Materials & Method section of a paper.

5 Conclusions

The current developments in SWISS-MODEL have been presented with a focus on the completely overhauled user interface. The interface leverages on JavaScript capabilities of modern web browsers to transform homology modeling into an interactive experience. The version of SWISS-MODEL presented here builds on a newly developed template library, the SMTL, for structural information. It is essentially a cleaned view on structural data from the PDB with a focus on the biology of proteins. For example, ligands are annotated using an online ligand annotation system, which allows to distinguish solvent molecules from biological and synthetic binders. Additionally, to facilitate modeling, the SMTL includes HHblits and HHsearch profiles, predicted secondary structure features, solvent accessibility etc. Data in the SMTL can be easily browsed from a web interface, making it possible to quickly look through large numbers of structures and compare their biology. Plans are on the way to extend the annotation system to complete biounits, e.g. to mark certain biological assemblies as biologically irrelevant, or to distinguish between wild-type proteins and crystallized mutants.

A newly designed template selection step allows to quickly compare features of the templates, including their structure, and choose an appropriate template for the subsequent modeling step. Templates can either be selected in a tabular view or sequence similarity graphs, which allow to compare the templates to each other. Additionally, the template structures are superposed onto each other to facilitate structural comparison of the templates. The manual template selection step is targeted at intermediate to experienced users who would like to use biological knowledge to select templates.

Whenever possible, models are built including biologically relevant ligands in the correct oligomeric state. This places the models into biological context and is an important prerequisite for further use of the models in mutagenesis, and ligand binding site studies.

While others, e.g. ModBase²⁴⁸ provide tools to perform interactive modeling inside of UCSF Chimera, we would like to completely push the modeling away from the desktop into the browser. As one of the most important additions to support this goal, the alignment viewer should be extended to handle alignment editing. Upon shifting in-

sertions and deletions, the changes on the structure of the model should be visible in real-time. Experienced users can directly observe the structural impact of an alignment shift. Moreover, real-time calculations of local quality scores would help users to see if the new alignment makes sense on the structural level.

Knowledge-Based Extension of Fragmented Models at Low Resolution in ARP/wARP

This work has been performed in collaboration with Tim Wiegels from the EMBL Hamburg. TW implemented the sequence docking, and performed the performance evaluation, MB implemented the fragment library, and evaluated sampling performance. TW and MB wrote the manuscript.

Wiegels T.*¹, Biasini M*^{2,3}, Lamzin V.S.¹, Schwede T.^{2,3} (2013). *Knowledge-based extension of fragmented models at low resolution in ARP/wARP*. Manuscript in Preparation

¹ European Molecular Biology Laboratory (EMBL) - Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

² Biozentrum, University of Basel, Klingelbergstrasse 50 / 70, 4056 Basel, Switzerland

³ SIB Swiss Institute of Bioinformatics, Basel, Switzerland

* equal contributions

SYNOPSIS: During automatic protein model building, chain breaks between partially built fragments are automatically detected and filled using structural information from the PDB in order to achieve higher completeness and accuracy of automatically built protein structures at medium-to-low resolutions.

ABSTRACT: During automated protein model building from medium-to-low resolution crystallographic data, the density between two built chain fragments is often too poorly defined to accurately model any protein chain. This is especially true during early stages of model building, and is the case not only for loops but also for helices or strands. A novel method is presented for the automatic detection of breaks in partially built protein chains during model building that makes use of electron density information, secondary structure predictions and statistical descriptions of the relationship between gap length and missing residues. Structural information obtained from the PDB is used to fill these structural gaps, with experimental data guiding the scoring of the candidate fragments. The obtained structure models are up to 20% more complete and thus less fragmented, specifically at crystallographic resolutions between 3.0 and 3.8 Å. The described method has been incorporated into the ARP/wARP package for crystallographic model building.

1 Introduction

Macromolecular structures of proteins, DNA RNA or complexes thereof, are a focus of attention in structural biology. This can be attributed to their high biomedical significance and their role as major players in the key processes of life. For a deep understanding of their function, it is crucial to have complete knowledge of the spatial arrangement of their constituent atomic blocks. Three-dimensional macromolecular structures find application in diverse areas of pharmaceutical and biotechnological industry and research.

Macromolecular crystallography (MX) has been the primary technique for the determination of structures of biomolecules at an atomic level of detail. MX has provided over 85% of all entries in the Protein Data Bank^{21,260} and over 90% of those complexes of proteins that are larger than 80 amino acids. The continuous exponential growth in the num-

ber of deposited PDB entries demonstrates the increasing demand for crystallographic three-dimensional (3D) structural information on biological macromolecules.

Many challenging structure determination projects come to a halt at a certain point. In particular, crystals of large proteins and their complexes may not diffract to a resolution where an atomic model can be straightforwardly constructed. Indeed, even after semi-high-throughput sample screening, the crystals of currently studied projects diffract on average to about 4 Å resolution on synchrotron beamlines²⁶¹, and only a small fraction of the measured X-ray data results in a structure being deposited in the PDB. It has been estimated that the gap between collected data sets and published structures is about 50 to 1²⁶².

The apparent problem with low-resolution X-ray diffraction is that the amount of collected data is insufficient for calculation of an electron density map that is detailed enough for atomic modelling and subsequent structural refinement²⁶³. For example, for a protein crystal with 55% solvent content that diffracts to a resolution of 2 Å, there are 8 reflections per atom. If 4 atomic parameters (for example xyzB) are needed to be refined, the observation-to-parameter ratio is 2, and the task is numerically overdetermined. However, for the same structure at a resolution of 4 Å there is only 1 observation per atom, which is insufficient to refine several atomic parameters²⁶⁴. This lack of observations at reduced data resolutions requires the use of additional parameters in the form of constraints or restraints. It also causes smoothing of density maps and a loss of detectable atomic features. The development of automated structure determination methods in MX has been predominantly focused on high-resolution data, where bonded or at least angle-bonded atoms are resolved. Reduction of model completeness at medium-to-low resolution implies an increase in the number of shorter, unconnected fragments built. Thus, the determination of low-resolution structures is usually beyond the normal operational range of crystallographic software and involves a large, if not excessive, amount of manual intervention.

Recently, impressive results have been reported for low-resolution structure determination, although rarely can a complete structure be built without user intervention. For example, the Buccaneer software can build up to 80% of the model at a resolution down to 3.2 Å provided that the initial map correlation is higher than 0.6²⁶⁵. Using the PHENIX AutoBuild wizard²⁶⁶, it was shown that structures with data extending to resolutions around 2.8 Å could be built automatically to a completeness of higher than 80%. At a resolution of 3.3 Å, the model completeness dropped to 60%. A comparable performance is obtained for model building with ARP/wARP¹⁶⁰, version 7.3. Estimates from the ARP/wARP remote model-building web service (July to October, 2012) suggest that structures at a resolution of 2.6 Å are typically built to a completeness of 85%. At 3 Å resolution, the model completeness decreases to 75%, and for cases with a resolution of 3.5 Å, one typically obtains a structure with only 70% model completeness. Here, the recently incorporated automatic detection of non-crystallographic symmetry into both the model building and refinement stages of ARP/wARP²⁶⁷ leads to an average of 7% more model completeness for cases between 2.3 and 3.2 Å.

Meanwhile, the theoretical modelling field has made progress in predicting the structure of proteins from the primary amino acid sequence. These methodologies operate

in absence of experimental data and use evolutionary relationships to guide the modelling process^{32,68,268}. Over last few years, the methods have advanced to a point, where accurate predictions are possible, even when only distant templates are available. For example, during the CASP9 instalment, the Zhou group was able to predict the structure of target T0523 (120 residues) down to an all-atom RMSD of 2.2Å, based on a template that shares only 15% sequence identity to the target⁵³.

Despite their increasing scope, algorithmic methods and databases from structural bioinformatics are rarely exploited in an integrative manner to aid macromolecular structure determination. It has been shown that employing a coordinated use of structural bioinformatics and modern X-ray data interpretation software can lead to impressive results, although existing methods are not yet fully applicable to everyday use^{139,269–270}. To fully take advantage of the technical possibilities of both experimental and theoretical methods, novel, sophisticated software solutions are required. Initial efforts in this direction include the application of homology modelling in molecular replacement^{72,271}, and the modelling of loops into smeared and non-interpreted electron density aided by the incorporation of database-derived information. Examples of methods exploiting bioinformatics for such loop modelling include the Loopy module of ARP/wARP²⁷², XPLEO²⁷³, LAFIRE²⁷⁴ and phenix.fit_loops from the PHENIX suite²⁷⁵. All of the noted softwares have shortcomings, however. Either, the anchoring residues and the length of the missing structural fragment are required as user input or the protein chain fragments need to be sequence-assigned to identify loop regions – and thus they are not suitable for incorporation into automated model building procedures.

Here we introduce the FittOFF (Fitting OF Fragments) method, which identifies chain breaks between partially built fragments from ARP/wARP intermediate models and uses structural information obtained from the PDB to complete these structural gaps, thereby increasing the structural information available for the next iteration of modelling. These structural gaps are defined by the stem residues that anchor the loop region to the intermediate model and the number of missing residues therein. As opposed to loop-building approaches commonly used in MX, the identification of structural gaps does not require the anchoring fragments to be sequence-assigned. Hence, the method becomes especially applicable to low-resolution cases, for which automatic sequence assignments rarely function well.

In a number of selected examples we demonstrate how the developed methodology is implemented in the ARP/wARP package as a dedicated and efficient module.

2 Methods

The FittOFF method is based on the analysis of the partially built protein chain fragments produced during the ARP/wARP model building protocol, and most specifically, whether or not such fragments should be connected by a single protein fragment of variable length. Within ARP/wARP, an initial protein model, a set of ‘free atoms’ with no chemical identity, or a mixture of both – the hybrid model – undergo iterative refinement. In each building cycle some ‘free atoms’ gain chemical identity and are recognised

as part of a protein chain fragment, others remain free. The evolving hybrid model combines two sources of information: it incorporates chemical knowledge from the partially built model and the free atoms aid the further interpretation of the electron density in areas where no model is yet available. The FittOFF method itself can be divided into two essential parts:

- The identification of the stem residues (residues anchoring a structural gap) and estimation of the number of residues necessary to close structural gaps in intermediate models obtained from the ARP/wARP model building process and
- The sampling of a large number of backbone conformations from a structural database to find the best suitable fragment to be placed in the structural gap (called the FRAGRA method in the following).

Identification of Gaps in Intermediate Models

At high resolution, the mutual location of chain fragments and the number of enclosed missing residues can easily be derived from docking the modelled structure to the sequence. However, at resolutions lower than 2.5 Å the sequence docking algorithm in ARP/wARP and similar approaches, all based on the recognition of the side chain electron density, which becomes very weak beyond this point, do not work sufficiently well. As a result, the fragment's position in the protein sequence remains largely unknown. Determining which chain fragments of partially built models are consecutive is a challenging task, and the designation of which residues are stem residues for a particular gap can be unclear, as shown in **figure 10.1A**. In addition to identifying which fragments to connect, the number of residues between the chain fragments has to be determined, **figure 10.1B**.

To position chain fragments in the protein sequence, we dock them according to their secondary structure. Since there are often multiple possible matches, the best three docking positions are stored for each fragment. Gaps that connect two consecutive fragments are then identified. Potential gap lengths are filtered for false-positives using a knowledge-based approach relating the number of residues contained in a gap to the distances between the C α atoms of the stem residues. As a final validation criterion, the density between each pair of partially built protein chains that are deemed to enclose a structural gap is examined. Only structural gaps that are, at least partially, supported by density will be put forward for fragment fitting with FRAGRA.

SECONDARY STRUCTURE DOCKING | To obtain results that are similar to sequence docking, even at low resolution, we developed a method that detects the best agreement between a segment of secondary structure assigned to a chain fragment and a secondary structure predicted from the corresponding amino acid sequence. We used the approach by Zhang and Skolnick²⁷⁶ to assign a secondary structure state for each residue of the partially-built model based on the C α -coordinates of five neighbouring residues. Their method is applicable to highly fragmented models as it only considers the neighbouring local

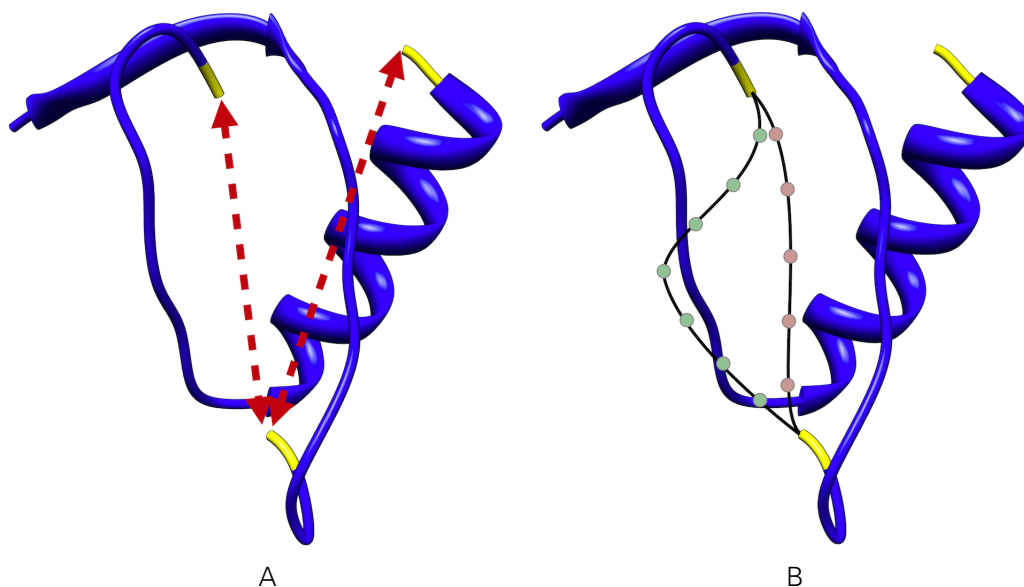


Figure 10.1 Difficulties of defining stem residues and gap length. The question of which stem residues should be connected is depicted in (A). Additionally, there is the question of how many residues are missing in the potential gap, shown in (B).

geometry of fragments. Due to the required number of neighbouring $C\alpha$ atoms only chain fragments of at least seven residues receive an assignment.

The secondary structure of the whole protein meanwhile, is derived from the input amino acid sequence using either PSIPRED^{8,277–278} or SSPro4^{116,279}. The assigned secondary structure of the chain fragment (the pattern) is then slid over the secondary structure prediction (the template) and at each offset, the number of matches between pattern and template at each position is computed and evaluated (see **figure 10.2**). The alignment with the highest number of matches denotes the best fitting position. A predefined number of best fits (three in the current implementation) are kept for each chain fragment. For each fit the percentage of matches with the secondary structure prediction is stored. All docked chain fragments are compared to each other and their relative positions are analysed. In short, if one chain fragment ($frag_i$) has been docked from position 0 to 14 in the sequence and another one ($frag_j$) from position 19 to 34, a gap of length four is assumed between the last $C\alpha$ atom of $frag_i$ and the first $C\alpha$ atom of $frag_j$. If some chain fragments have been indeed sequence-assigned by ARP/wARP, this information is taken into account to assign further structural gaps and validate the results from secondary structure docking.

RELATING GAP LENGTH TO THE DISTANCE BETWEEN STEM RESIDUES | Saving the three best positions for each chain fragment following secondary structure docking can, in the worst case, lead to nine different lengths for the same structural gap. To decide which gap length is most likely, the predicted secondary structure content and distance between the terminal $C\alpha$ atoms is compared to experimental data from the PDB, assigning a probability value to each of the possible lengths.



Figure 10.2 Schematic overview of the sliding window method for sequence and secondary structure docking.

Statistical data on secondary structure composition, enclosed residues and distances between $C\alpha$ atoms was collected from a large set of experimental structures. The analysis was performed on a non-redundant set of 6,613 protein chains obtained from the PDB (PDB50, Release: January 2011, solved by MX at resolution of 2.0 Å or better). This subset was generated with CD-Hit²⁸⁰ by clustering all protein chains of at least 20 amino acids at 50% sequence identity.

The probability of observing a certain gap-length at distances between stem residues was calculated for gaps with a certain secondary structure composition, e.g. 50%, 75% or 100% helix, sheet or coil. An overview of occurrences for gaps with two, four, six, or eight residues at distances between 0 Å to 20 Å, ignoring the secondary structure content for visibility, is given in **figure 10.3**.

For each possible gap length identified by secondary structure docking, the probability of that gap being ‘true’ can be calculated from the distance between the $C\alpha$ atoms of the stem residues and the secondary structure content. Gaps below a certain probability threshold are ignored for further processing.

FRAGRA

FRAGRA is a knowledge-based method to remodel structural gaps implemented within the OpenStructure framework¹³⁵. The method was originally designed to produce reasonable models of short loops that could be used to complete homology models in seconds rather than hours and to close structural gaps up to 14 residues in length. FRAGRA uses a large database of backbone conformations (hereafter called fragment database). FRAGRA differs from related methods such as XPLEO²⁷³, using existing fragments from the PDB instead of rebuilding gaps with physical approaches. This results in a drastic decrease in required computation time (under a minute for FRAGRA compared to up to two hours with XPLEO for a gap of 12 residues). In principal, FRAGRA follows the concept of knowledge-based loop modelling described by others²⁸¹. In the context of FittOFF, FRAGRA is used to find candidate backbone conformations that fit a structural gap with the stem residues and number of enclosed residues identified as described above, **figure 10.4A**.

FRAGMENT DATABASE | The fragment database has been constructed from about 60,000

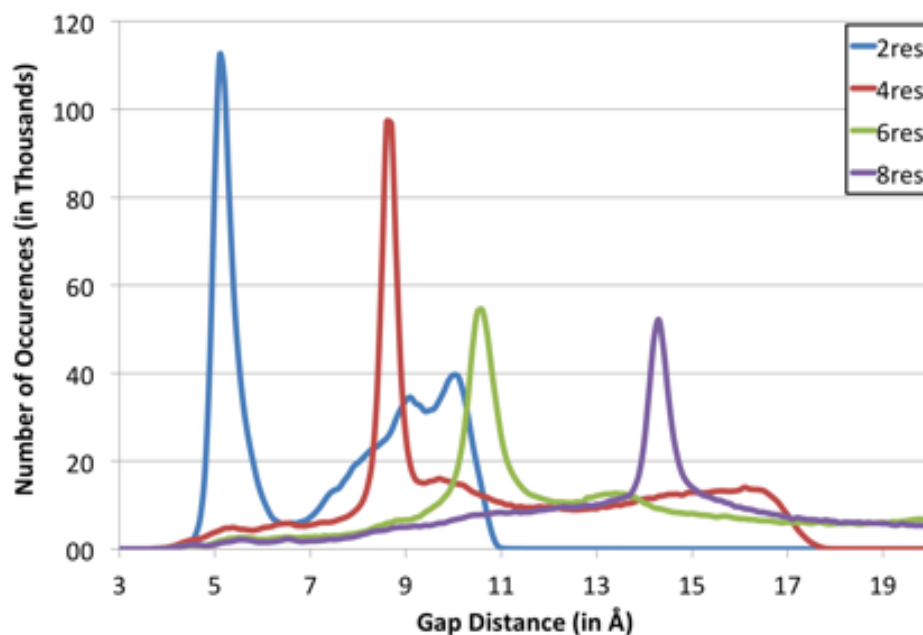


Figure 10.3 Relations between gap length and distance between stem residues for gaps of 2, 4, 6, and 8 residues. Occurrences of a certain distance (in Å) are shown. Distinctive peaks are evident, for example, for gaps of two residues containing helices at 5.1 Å a gap of two residues containing a beta-sheet at 10.0 Å and a gap of eight residues forming an alpha helix at 14.3 Å.

protein chains solved by X-ray crystallography. Only structures with experimental data extending to a resolution of 2.2 Å or better were included in order to provide a good trade-off between the quality of the backbone models and the number of chains included in the database. The database uses a hash generated from the geometry of the two residues lining each fragment, the so-called stem geometry, to quickly look up possible backbone conformations for the structural gap. Here, $C\alpha - C\alpha$ distances as well as the angle between the $C\alpha - C\alpha$ and the planes formed by N - $C\alpha$ - C of the N-terminal residue and the $C\alpha - C - O$ plane of the C-terminal residue are used as descriptors. For each stem geometry, the fragments in the database are structurally non-redundant, that is, not to fragments is closer than a certain RMSD to another. The RMSD threshold is determined as a function of the fragment length.

SAMPLING OF BACKBONE FRAGMENTS | Possible backbone conformations are sampled from the fragment database by using the stem geometry of the structural gap as a lookup key. Usually, the conformation from the database do not perfectly connect the two stem residues. Therefore, to improve the fit at the stems, small fragments with a length of three residues are used to bridge between the residue before and after the stem residue. Optionally, the geometry of the backbone is additionally optimised using a steepest decent optimisation of bonded terms.

SCORING AND FILTERING THE CANDIDATES | The list of candidates found during backbone

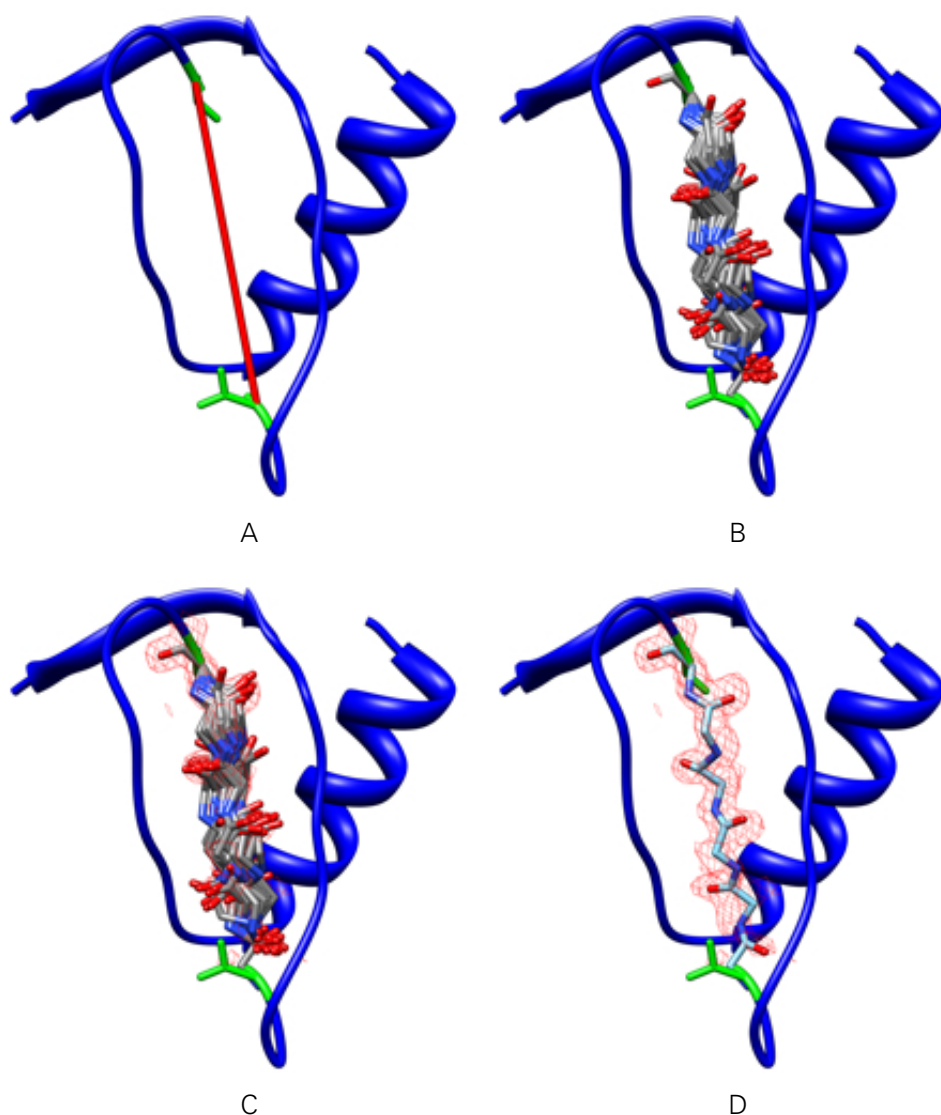


Figure 10.4 Overview of FRAGRA in the FittOFF method. An identified structural gap (A), stem residues are marked in green, gap length indicated by a red pseudo bond, is used for fragment fitting in FRAGRA, the initial database search results in a large number of candidates (B). By taking the residual density into account (C), it is possible to rank the candidates by map correlation and define one, or a few, top scoring results (D). The map correlation can further be used to identify regions that have been incorrectly identified as gaps.

sampling contains 1500 fragments on average, **figure 10.4B**. To decide which one fits the gap best, a scoring scheme is applied. At first, fragments that clash with the already built protein structure are filtered out. A finer ranking is achieved by spatially correlating the candidates to the residual density, **figure 10.4C**. The expected density is computed by placing a Gaussian sphere of density at each atom and the real-space correlation to the experimental density is calculated as described by DiMaio¹³⁶. The fragments are then output ranked by the real-space correlation values, **figure 10.4D**. If the gap length is not

determined precisely by FittOFF, different lengths are used in FRAGRA and the real gap length is identified by analysing the map correlations for different trials.

APPLICATION OF FITTED FRAGMENTS TO MODEL BUILDING | Within the ARP/wARP workflow, FittOFF is applied to the intermediate model as depicted in **figure 10.5**. Specifically, the obtained best fitting fragments are added as $C\alpha$ seed points to the current hybrid model. This hybrid model is used for subsequent tracing of protein chains in the main-chain building block. The method is invoked by default if the resolution of the data is poorer than 2.5 \AA , as this is a typical resolution at which the standard sequence docking algorithm does not work sufficiently anymore.

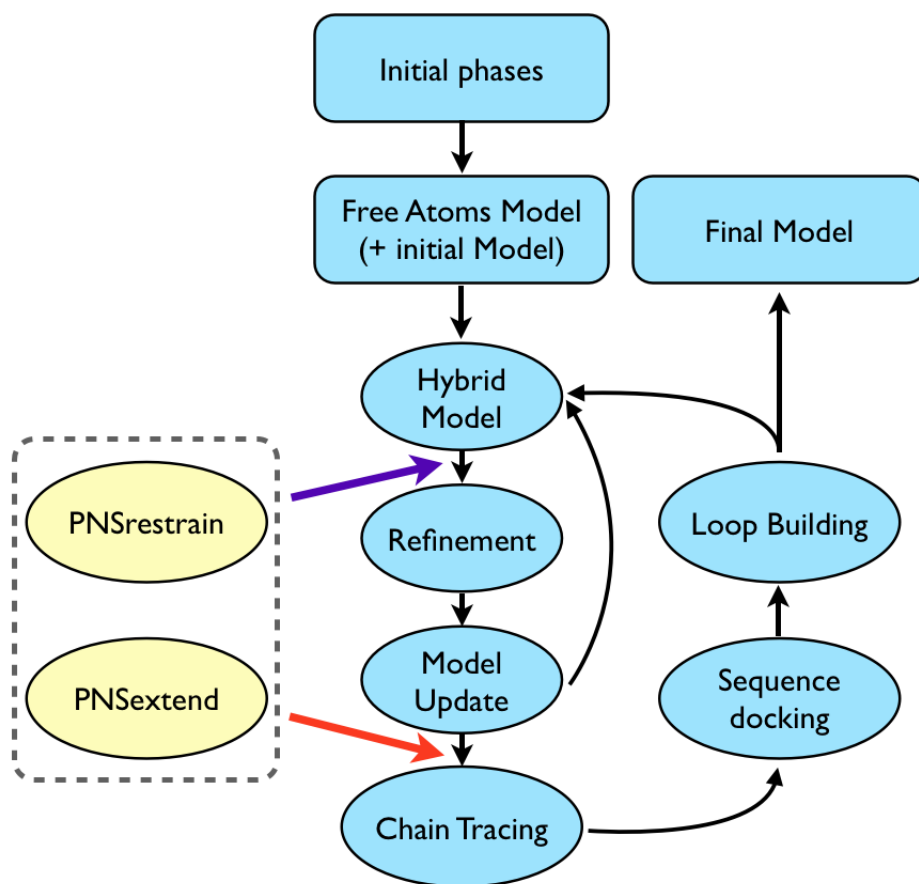


Figure 10.5 Flowchart of the ARP/wARP protein model building, including automatic detection and closing of structural gaps.

3 Results

Data

The method was initially developed on high resolution test cases obtained from the PDB. A good representative example is the 1.6 Å structure of a B subunit of a mutated shiga-like toxin, PDB ID 1c48²⁸². The molecule is arranged as a homo-pentamer, with each subunit composed of 69 residues.

For the initial proof of principle, 1c48 was fragmented to obtain six gaps of various length and secondary structure content (between two and nine residues, with mainly helical, mainly sheet, mainly loop and mixed content). This “gapped” structure was then submitted to FitOFF-based loop modelling.

Computational tests for subsequent examination of the method’s effectiveness were carried out using ten randomly chosen structures from the PDB that were determined at resolutions ranging from 3.0 to 3.8 Å. Choosing this resolution regime further ensured that the tested structures could have been used for the generation of the fragment database used in FRAGRA. Only structures with low molecular weight (15 - 25 kDa) characterised by various secondary structural content were selected. For all of these test cases, secondary structure predictions were generated with PSIPRED version 3.0 and SSPro version 4.1. Additionally, secondary structure assignments of the structure from the PDB were generated with the Zhang algorithm.

Application of the Method in Absence of Coordinate Error

To assess its effectiveness, the FittOFF method was used to improve the completeness of the artificially “gapped” structure of 1c48. The identification of gaps and the fitting of the highest-ranking fragments have been evaluated independently. Of the six gaps in the test structure, five could be successfully detected using secondary structure docking. The sixth gap was anchored to a fragment of five residues, thus preventing assignment of secondary structure. Two different protocols were tested for gap retrieval. Initially, a rigid gap filtering was attempted, necessitating a secondary structure docking of the anchoring fragments with at least 60% confidence (referred to as *confdock*) and a probability for the suggested number of residues missing in the gap of at least 50% (*conf_{pvec}*). Starting from the gap candidates and corresponding lengths identified by secondary structure docking, all with values beyond the limits of the distance databases (*gapdistance* > 40 Å, *gaplength* > 14 residues) of physically impossible ($C\alpha - C\alpha$ distance > 4.5 Å) were immediately discarded. After filtering using relations between gap distance and the number of missing residues and check for uninterpreted density, three of the five gaps were automatically detected without any mistakes. By applying a “loose protocol” that required a lower value of *confpvec* (10%), all five gaps could be identified. As might be expected, loosening the

filtering cutoff led to more than one potential gap length being suggested for three of the five gaps. An overview of the number of gap candidates and corresponding lengths filtered out in both protocols is given in **table 10.1**.

For testing the fitting of the fragments, the gaps obtained from the loose protocol were fed into FRAGRA. By applying the map correlation as a ranking criterion, all false-positive solutions could be eliminated, leaving only the best-ranked fragments for the expected gaps and gap lengths. To obtain a good estimate of the validity of these fitted fragments, they were superposed against the corresponding areas in the reference structure, **figure 10.6**. Although, poorer results are achieved for long gaps, **figure 10.6B**, even for larger deviations the fitted fragments follow a path very similar to the protein backbone in the reference structure.

	Rigid protocol		Loose protocol	
Detected by SS-docking	15	(45)	15	(45)
Beyond database limits	4	(12)	4	(12)
Physically impossible	1	(5)	1	(5)
Filtered by distance check	7	(25)	2	(15)
Filtered out by density check	0	(0)	3	(4)
Final results	3	(3)	5	(12)

Table 10.1 Validation test of FittOFF - ‘artificial gapping’. The table shows the number of gap candidates (corresponding gap lengths given in brackets) filtered out by the applied protocols.

Incorporation of FittOFF into ARP/wARP Protein Model Building

The application of the FittOFF method to standard ARP/wARP protein model building was investigated using three protocols each with differing parameters. The first two protocols are the rigid and the loose ones described above in the last section. In these protocols, the average map correlation over all highest-ranking fragments issued by FRAGRA was calculated and only fragments with a map correlation higher than the average were admitted to ARP/wARP model building as $C\alpha$ seeds. In addition, a third protocol (also loose in regard to filtering) that fed back all fitted fragments into ARP/wARP was used, denoted as the loosest protocol. Every protocol was executed with two different secondary structure predictions, generated with PSIPRED and SSPro4. Additionally, secondary structure assignments from the Zhang algorithm were also tested as secondary structure information, resulting in nine test environments for each structure. In almost all cases, we observed higher model completeness with less fragments. The improvement for the best cases is shown in **figure 10.7**. Up to 25% more residues could be built for resolution as low as 3.8 Å (**figure 10.7A** and **figure 10.7B**). In addition, decreases in R-factor by 4% and doubling of the average length of fragments were observed in several cases (**figure 10.7C**). The best results for extra residues could be obtained following the application of the loose protocol, although the best improvement in fragmentation was

seen following application of the rigid protocol as shown in [table 10.2](#). Regarding the source of secondary structure information, it was found that the prediction methods delivered results comparable to those obtained with the Zhang assignment of the deposited structure. Nevertheless, slightly better results were achieved using the Zhang assignment ([table 10.3](#)). No apparent relation was detected between the secondary structure content of a structure and the extent of improvement. On average, the loose protocol coupled with a secondary structure prediction from SSPro4 provided the best improvements with 9% more residues built and 22% longer fragments.

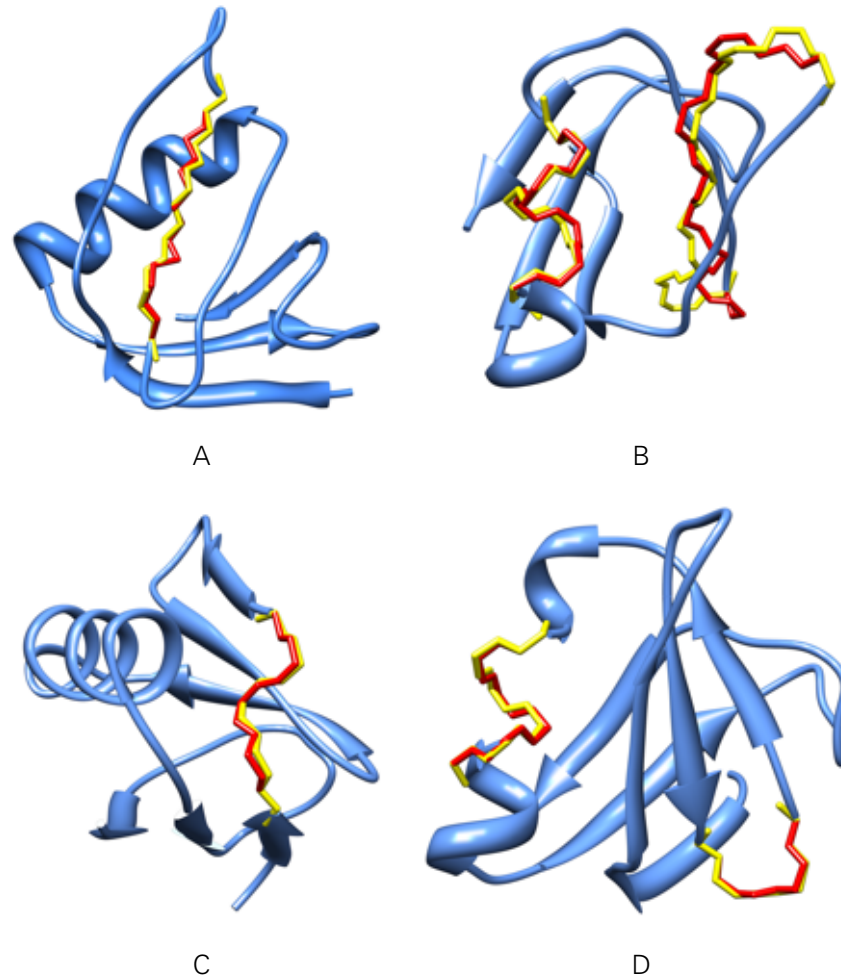


Figure 10.6 Validation test of FittOFF method with artificially “gapped” test case 1c48. Part (A) to (D) show the different structural gaps, with (D) showing the gap that could not be automatically detected due to short anchoring fragments. Fitted fragments are shown in stick representation for the minimal backbone. The fitted fragment is colored in red, the reference structure is shown in yellow. The biggest deviation can be seen in (B), for a gap including parts of a beta-sheet and a loop.

For testing of FittOFF, only cases from the PDB were used. This made it possible to compare the models built by the standard ARP/wARP with and without the FittOFF protocol

to the crystal structure from the PDB. The results for the superpositions of three models are shown in **table 10.4**. It is shown that all models built to a higher extent with FittOFF also had a smaller $rmsd$ and $rmsd_{adj}$ ($rmsd$ scaled to the cube root of the number of aligned residues²⁸³) to the reference structure.

Protocol	Residues built	Model completeness	Fragment length	R-Factor
Rigid	+19%	+11%	+98%	-4.3%
Loose	+25%	+12%	+78%	-3.8%
Loosest	+21%	+11%	+78%	-3.8%
Overall best	+25%	+12%	+98%	-4.3%

Table 10.2 Comparison of the best result obtained by model building with FittOFF for each of three different protocols for gap filtering.

Secondary Structure	Residues built	Model completeness	Fragment length	R-Factor
PSIPRED	+18%	+11%	+78%	-3.8%
SSPro4	+19%	+10%	+60%	-3.4%
Zhang assignment	+25%	+12%	+98%	-4.3%
Overall best	+25%	+12%	+98%	-4.3%

Table 10.3 Comparison of the best result obtained by model building with FittOFF for each of three different sources of secondary structure information.

Testcase	Size	Resolution (Å)	Aligned Residues		rmsd		rmsd _{adj}	
			ARP	FittOFF	ARP	FittOFF	ARP	FittOFF
1PLR	258	3.0	225	240	0.80	0.75	0.13	0.12
2QSR	173	3.1	119	138	0.93	0.84	0.19	0.16
2AJ2	208	3.2	144	165	0.67	0.67	0.13	0.12

Table 10.4 Best results obtained with FittOFF; results of the structural alignments of the model built with the standard ARP/wARP protocol and the protocol incorporating FittOFF and the reference structure. The $rmsd$ has been calculated over all C atoms in each model.

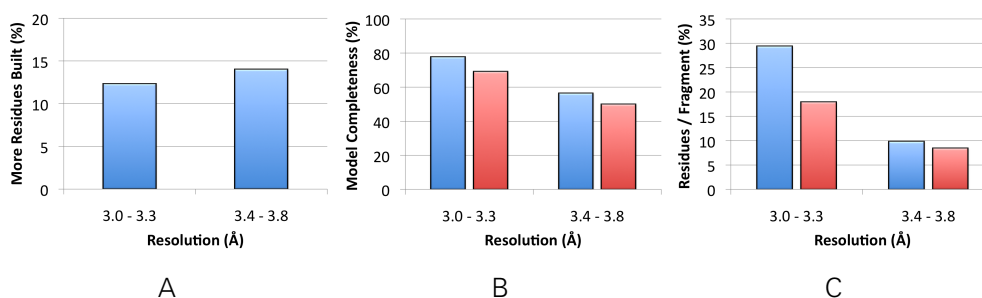


Figure 10.7 FittOFF applied to ARP/wARP. Best results are shown for tests with different protocols and sources of secondary structure information. The red columns denote the values obtained with the standard ARP/wARP model building protocol, the blue column with incorporating FittOFF. (A) The percentage of extra residues built compared to the standard ARP/wARP protocol; (B) Average completeness of the built model; (C) Average length of built fragments.

4 Conclusions

The obtained results support assertion that application of methods from theoretical modelling to automatic protein model building in macromolecular crystallography can be beneficial. The application of FittOFF gave rise to notable improvements in crystallographic model building at resolutions between 3.0 and 3.8 Å, pushing model building for some case studies towards 90% completeness and a significantly better rmsd_{adj} to the crystal structure from the PDB (**table 10.4**). The method imposes negligible overhead on the computation time required by the standard ARP/wARP protein model building protocol and is thus applicable in general use. Further optimisation of the regime of secondary structure prediction, and thus docking, will certainly provide additional enhancement. Due to the planned invocation of FittOFF in ARP/wARP web-based model building, a continuous evaluation on a wide variety of cases will be performed automatically.

In the “artificial gapping” test scenario, the chain fragments are free of phase dependent coordinate error. There are also no mistakes in traced chain fragments such as route shortcuts or spurious loops. Moreover, the electron density was of high quality (1.6 Å) and the correct secondary structure information was used, meaning the test cases were somewhat idealised. It was shown that all gaps surrounded by fragments docked into the secondary structure could be identified without introducing any false-positive gaps. Furthermore, wrongly recognised gap lengths could be eliminated using a threshold applied to the map correlation of the fitted fragment to the residual density. Although the deviation from the reference structure rises with longer fitted fragments, they are generally highly similar to the path taken by the protein backbone.

The way the fragments fitted by FRAGRA are used in the protein model building protocol has certain advantages over other possible approaches. For example, plain use of the coordinates of the fragments fitted into identified structural gaps may not be the best option as it introduces a certain degree of model bias. In addition, some parts of the new loop could be out of density, causing the accurately modelled parts to be removed during refinement. In our implementation all fitted fragments are only used as potential $C\alpha$ seeds (suggestions) to ARP/wARP for subsequent building of longer chain fragments. Therefore, the method is not expected to build parts of the structure that lack support for coordinate placement in terms of electron density and plausible stereochemistry. If the additional $C\alpha$ atoms admitted to further chain tracing by FittOFF are in agreement with the density, the structural gaps will be closed. If, on the contrary, the suggestions do not match the density they will not be used for building a chain. This is especially important for the additional information derived from FittOFF, which has been shown to be less accurate for longer gaps. There may still be small decreases in model completeness or higher fragmentation for some cases. This may occur for models that are built only to a modest extent and can be caused by different paths that will be followed during chain tracing. Following an incorrect path could always lead the tracing to areas of low density and thus prevent the building of longer chains. Going forward, it may be worth investigating the placement of all protein atoms as seeds for model building as opposed to

merely as $C\alpha$ -candidates. Thus, FittOFF could also be applied after ARP/wARP's final model building cycle.

The accuracy of the method depends predominantly on the degree of fragmentation of the initial model and its coordinate accuracy. For the identification of structural gaps correctly built and long chain fragments with defined secondary structure are of benefit. The use of fragment fitting in model building is always advantageous, but the degree of improvement depends even more strongly on the completeness, fragmentation and correctness of the model, which all in turn depend on the quality of the initial phases and the data. More specifically, for a model consisting only of chain fragments shorter than seven residues, no chain fragments can be docked to the secondary structure prediction. Additionally, for such a model, there is also a high probability that most, if not all, chain fragments are modelled incorrectly or with high positional error.

The performance could be improved by using the Zhang assignment of the final model instead of the predicted secondary structure by PSIPRED and SSPro4. This shows that the method presented here is sensitive to the quality of the predicted secondary structure. Still, the predicted secondary structure by PSIPRED and SSPro4 is in many cases sufficient to dock the partially modeled fragments with high confidence into the sequence. Fitting fragments into long gaps might result in only marginally reliable fragments (Figure 3.1b), which might not lead to the closure of a structural gap. However, the host fragments may be partially extended, thus leading to a shorter gap and resulting in a more reliably fitted fragment in the next iteration.

The possibility to deduce the correct location of structural gaps implies that the anchoring fragments have been docked into right position in sequence. Thus, the secondary structure docking should also be tested for its application to aid the sequence assignment in ARP/wARP at medium-to-low resolution. Highly ambiguous dockings could be further improved by a combination of secondary structure docking and identification of large side chains in the density.

The next step in the development of the FittOFF method must be its release within the next version of the ARP/wARP software suite. Once this has been achieved, it will also be necessary to evaluate protocols combining fragment fitting and already implemented NCS-detection and extension²⁶⁷ on an array of test cases containing NCS-relations to see if their combined effectiveness is more than the effects of either addition alone.

Software Availability

The FittOFF method is currently being incorporated into a future release of ARP/wARP.

Acknowledgements

We would like to thank the European Molecular Biology Laboratory (EMBL) for the funding of the PhD fellowship to Tim Wiegels. Further we would like to thank Ciaran Carolan, Philipp Heuser and Andres Gruber for their help and enlightening discussions.

FRAGRA - Knowledge-based Backbone Conformation Sampling

What follows is a description of the FRAGRA method for sampling of backbone conformations. It is essentially an extended version of the Materials and Methods part together with some results in the "Knowledge-based extension of fragmented models at low resolution in ARP/wARP" manuscript. We first give some methodological background before touching on sampling performance of the FRAGRA method.

1 Introduction

Traditionally, there have been two different approaches to filling structural gaps in protein structures. On one end of the spectrum, *ab initio* methods sample loop conformations based on prior knowledge of backbone conformations and an energy function to guide the sampling process^{73,284}. On the other side are knowledge-based methods that make use of large libraries of fragments from experimentally solved structures^{74,285}. The main argument of *ab initio* versus knowledge-based is one of coverage of structural space vs. efficiency. *Ab initio* methods have good coverage of the structural space, but due to their rigorous sampling tend to be slow, especially for long loop lengths. Knowledge-based methods are very fast, mainly due to the fact that they only look at backbone conformations that are feasible in nature, but have the problem that only a fraction of the possible backbone conformations are available in structural databases. For gaps of 15 residues and more, fragment-based methods are not viable anymore, since the structural coverage is too low. However, performance starts to decay already for shorter loop²⁸⁵. For longer loops, only *ab initio* sampling is applicable. Recently, hybrid methods have been developed that combine the speed of fragment-based (knowledge-based) with the rigorous sampling of *ab initio* methods⁷⁵. For an extensive comparison of available loop sampling and scoring methods, see Rossi^{76,286}.

In the following, we describe a knowledge-based method to remodel structural gaps in the context of ARP/wARP model building. It was developed with the following goals in mind: the method should produce accurate results for short loops, finish the remodeling in seconds rather than hours and be able to close structural gaps up to 14 residues in length.

We start by describing how the fragment library is built, then go on to a benchmark of the raw sampling performance in comparison to other available programs. Then, we describe the backbone geometry regularizer, which improves accuracy of the sampling results. Finally, we show clash filters help to reduce the number of false-positive fragment candidates.

2 Fragment Database

The fragment database has been constructed from roughly 60'000 protein chains solved by X-ray crystallography. Only structures solved at 2.2Å or higher resolution have been included into the database. The cutoff of 2.2Å resolution provided a good tradeoff between the quality of the backbone models and number of chains included in the database.

When sampling fragments for a given structural gap, the two residues lining the gap region constrain the geometry of the possible backbone candidates. These two residues are coined *stem residues* as they are anchoring the backbone of the missing part to the rest of the structure. Only fragments that are more or less matching the stem geometry are suitable backbone candidates. The geometry of the stem, thus provides a good filtering opportunity to reduce the computational cost. We use the $C\alpha/C\alpha$ distance as well as the angle between the $C\alpha/C\alpha$ vector and the planes formed by N- $C\alpha$ -C of the N-terminal residue and the $C\alpha$ -C-O plane of the C-terminal residue as descriptors. Together with the length of the fragment, they form the stem pair geometry (**figure 11.1**). The distance and angles are binned. Fragments that fall into the same bin are called a stem group. Structural redundancy of the candidates in one stem group is removed: only fragments that differ more than a certain RMSD are included. This makes sure we are not wasting any CPU cycles on recomputing energies of similar fragments but rather spend computation on a more diverse set of candidate fragments.

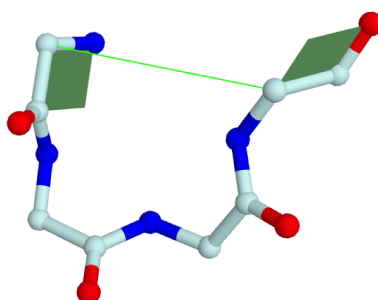


Figure 11.1 The stem geometry is calculated from the $C\alpha/C\alpha$ distance as well as the angle between the $C\alpha/C\alpha$ vector and the planes formed by N- $C\alpha$ -C of the N-terminal residue and the $C\alpha$ -C-O plane of the C-terminal residue.

The internal layout of the database has been optimized for fast fragment retrieval as well as fast loading from disk (see **figure 11.2**). The whole database can be loaded into memory, even on low-end consumer hardware. Thus, no access of the file system is required during sampling. These two requirements had a large impact on the internal database layout and lead to the design of a very compact, and efficient storage scheme that is described in more detail in the next section.

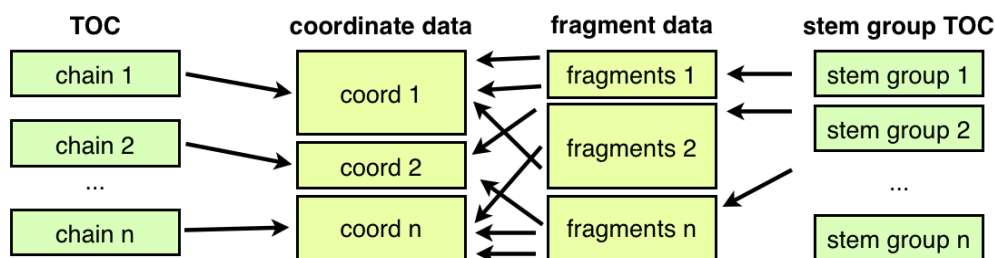


Figure 11.2 Data layout of the fragment library. See text for a description.

Instead of storing the coordinates of the backbone per fragment, the database is divided into 2 parts, very similar to the tables in a relational database. The first part is called the coordinate section and contains the actual backbone coordinates of the protein structures. This part is divided into a data section and a table of contents (TOC). The table of contents is implemented with a quadratically probed hash-table with an open addressing scheme and maps the entry id (PDB identifier plus chain name) to the offset in the data section. The data section itself is a compressed representation of the backbone coordinates of the protein chains. The chains have been translated such that their minimal position of the axis-aligned bounding box coincides with the coordinate origin. Each position takes 48bits, 16bit per dimension in a fixed-precision format. This leaves us with a precision of 0.01\AA per dimension and a maximal chain extent of 635.36 along each of the x , y and z axes, more than enough for all of the protein chains that have been crystallized so far. The second section stores the fragments for a given stem pair geometry. A hash table maps the stem pair geometry to the block of fragment entries of the stem pair geom. Each fragment entry has the size of one int (32bit) and maps back to the structure section. While this data layout comes at the expense of cache efficiency as the coordinates per stem pair geometry are spread out over the whole coordinate section and there is not way to fit the whole structure section into the cache, the backbone coordinates can be reused for overlapping fragments and fragments of different length. This reduces the space required considerably.

As a post-processing step to the database creation, unused coordinates are removed, e.g. only coordinates which are accessed by at least one fragment are kept in the database.

3 Sampling Procedure

The task of the sampling is to produce a list of fragments that are suitable candidates to remodel the backbone of the structural gap. This list is called the candidate list. As input, the sampling routine requires 6 coordinates (the N , $C\alpha$, C atom positions of the N -terminal stem and the $C\alpha$, C , O atom positions of the C terminal stem) and the loop length. The stem pair geometry is calculated from these 6 coordinates. Each fragment in the stem group is then superposed onto the stem atoms by minimizing the squared errors using singular value decomposition²⁸⁷ and kept if the RMSD is lower than a certain threshold. While this step generally yields good candidates, they do not fit perfectly

at the stems. Inserting them as is would produce suboptimal backbone geometries with bond lengths and angles that are far from being physically/chemically feasible. We have two different "protocols" to optimize the stems. The first one is to use small fragments of length 3 that are used to bridge between the stem atoms and atoms on the backbone fragment to get better fits at the stems. The second one is use to use a backbone optimization scheme with fixed stem atoms (see below).

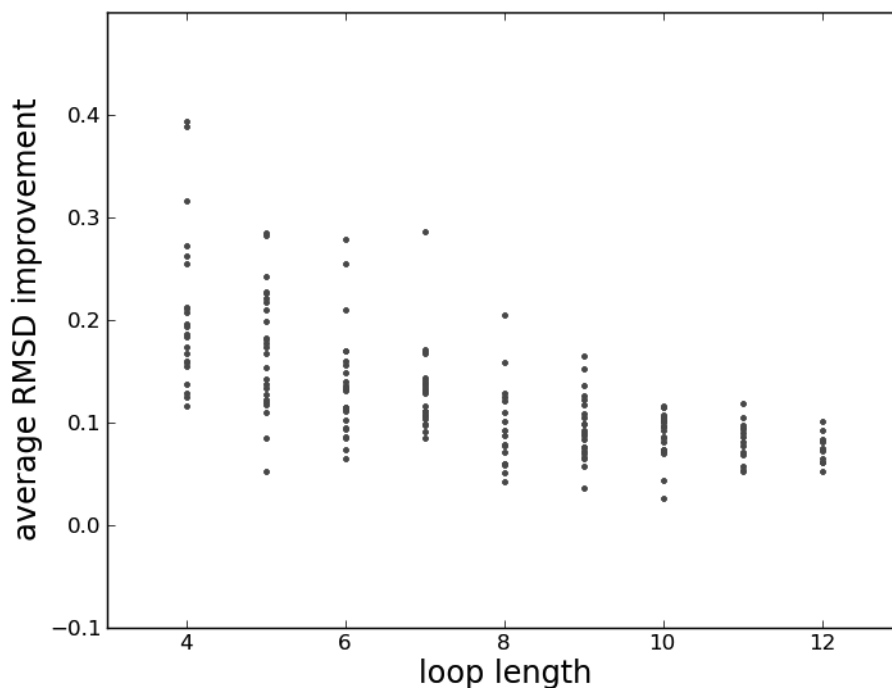


Figure 11.3 Effect of healing and regularisation of fragment geometry on RMSD for loops of different lengths. The best decrease in RMSD is observed fro small loops. For longer loops, the improvement is still non-zero, but less pronounced The best decrease in RMSD is observed fro small loops. For longer loops, the improvement is still non-zero, but less pronounced..

When used separately, both the healing and the backbone optimization have a positive effect on the average RMSD of the backbone candidates to the native loop. For loops of length 4, the backbone optimizer improves the backbone RMSD by 0.17\AA , whereas the healing protocol improves the backbone RMSD by 0.12\AA (**figure 11.3**). The combination of healing and backbone optimizer improves by an average of 0.21\AA RMSD. For both the backbone optimizer and the healing, the effect in RMSD difference becomes smaller with increasing loop length. This is not surprising as the contribution of the residues close the stem atoms becomes smaller with increasing loop length. For a better comparison, one should look at the RMSD improvement per residue position. For the backbone optimization, one would expect that the RMSD improvement is biggest for the terminal loop residues and becomes smaller for residues in the middle.

In this section we compare the performance of our loop-sampling method to other available loop modeling programs. This comparison is based on the loop test set and analysis by Rossi²⁸⁶. The test set consists of high-resolution protein structures. For each

loop length between 4 and 12, certain regions of the structure are removed and need to be remodeled. The obtained ensemble of backbone conformations are then compared to the native structure and an RMSD value is calculated. For each loop length, the average of the lowest RMSD for each modeling case is calculated and plotted in **figure 11.4**. The solid lines are the results for sampling alone, whereas the dashed lines are the results obtained from sampling and scoring. As can be seen, the gap between the best possible loop in the generated ensemble of loops and the one that has been selected as the best one (energy-wise) is considerably large. The method that performs best in selecting the loops in the ensemble is the ab initio method Prime.

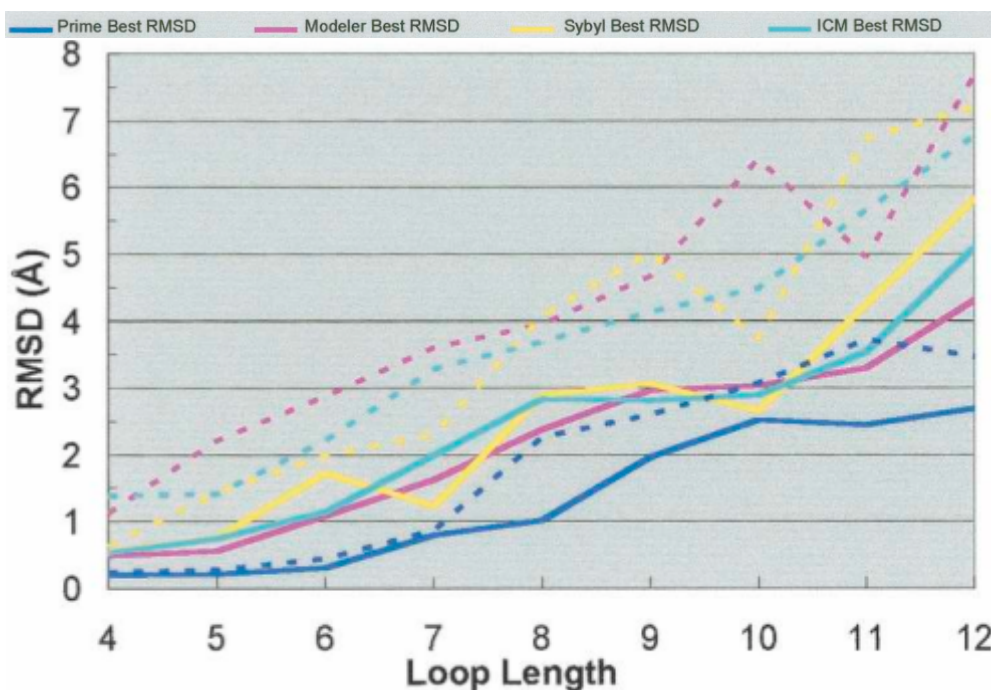


Figure 11.4 Performance of various loop sampling packages on the Rossi testset²⁸⁶. The dashed-line represent the best-scoring loop candidates, the solid lines are the loop candidate with the smallest RMSD to the native conformation.

To compare the sampling performance of our method, we repeated the calculations done by Rossi²⁸⁶ with our method on the same test set (see **figure 11.5**). For knowledge-based methods, Rossi and coworkers removed fragments from structures that share a sequence identity of 90% or higher to the structure containing the gap to be modeled. We do the same for our fragment database. As can be seen, on average our method produces a backbone fragment below 2.0 up to loop lengths of 10 residues. For longer loops, the performance starts to decay. The best method from the set tested by Rossi, Prime, performs very similarly. They tend to produce slightly better result at short loop lengths, whereas we find loops closer to the native loop conformations for loops of lengths of 10. However, one should not over-interpret these result, as the comparison is not completely fair. We are only assessing the sampling of our method and do not pay any attention to the impact the sampling has on selecting a good backbone fragment. The sampling re-

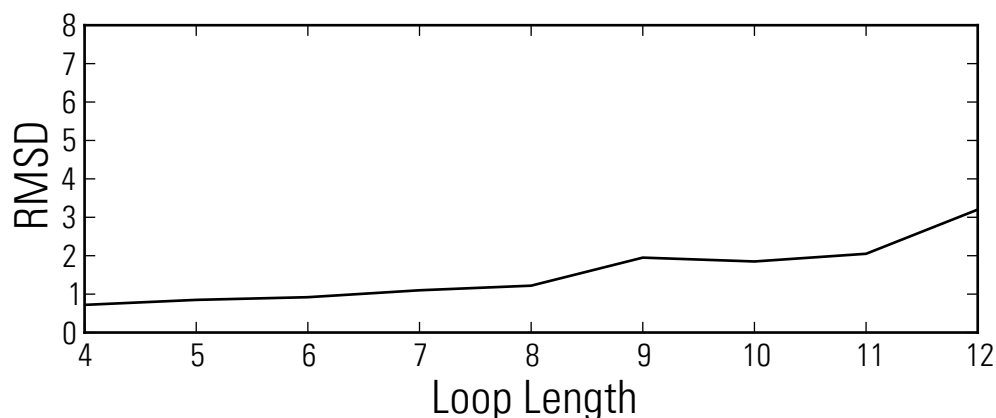


Figure 11.5 Performance of the loop sampling method presented here on the Rossi testset. X-axis: loop length, Y-axis: average RMSD value of best candidate for each target.

sult of Prime, on the other hand is the result of a highly tuned interplay of scoring and sampling.

Optimizing the Backbone Geometry

As input, the backbone optimizer takes a set of backbone coordinates. The N, C α and C atoms of the first amino acid and the C, C α , O atom of the last amino acid are fixed to the coordinates of the stem; their position remains constant throughout the minimization procedure. This is required to fit the backbone candidate to the stem residues where we would like to create perfect geometry where the backbone candidate connects to the backbone of the protein.

The backbone optimizer considers only energies coming from bonded terms. There is no term capturing van-der-Waals and Coulomb forces. A harmonic potential is used to represent bonds:

$$V_b(b) = \frac{1}{2}K(b - b_0)^2$$

The bond-length b is allowed to oscillate around an equilibrium length. K is the spring constant and determines how strongly the bond fluctuates. The spring constant is higher for double bonds, e.g. the C-O bond than it is for single bonds, e.g. C α -C. The equilibrium lengths/angles and spring constants are taken from the CHARMM force field (`par_all27_prot_na.prm`). Analogously, angles between 3 consecutive atoms are described by a harmonic potential that is parametrized on the angle:

$$V_a(\gamma) = \frac{1}{2}K(\gamma - \gamma_0)^2$$

The potential for dihedral angles is

$$V_d(\gamma) = K[1 + \cos(n \cdot \gamma - \delta)]$$

Here, n is the multiplicity of the dihedral angle, δ determines the location of the minima, K is a constant that scales the energy appropriately. For all involved backbone dihedrals,

the multiplicity is 2 or 1, meaning that the potential has two or one minima, respectively. To account for the energetic asymmetry of the omega cis and trans conformations, two potentials are overlaid on top of each other. One with multiplicity one, and one with multiplicity two. Together these potentials make the trans-peptide conformation energetically more favourable. Although this it does not occur in the case of backbone dihedrals, the multiplicity may also be 3, for example for the CHI1 torsion of lysine.

At each step, the backbone optimizer accumulates forces resulting from the bonded terms. The forces are calculated as the derivatives of the potential functions given above. The positions are updated by moving a tiny fraction along the resulting force vector. This procedure is repeated iteratively until a fixed number of steps is reached. The minimum of the bonded term potential is very far away from the normal loop conformation and optimizing for many steps would drive the fragments away from their current position. This is not desirable as the minimization step should merely serve to remove bonds and angle that are not chemically feasible.

4 Loop Ranking

The candidate list contains between 100 and 4000 backbone candidates with an average number of fragments around 1500. For ARP/wARP $C\alpha$ -seeding, only a few loops are to be selected. Naturally, we would like the selected loop to be as close to the native structure as possible. We use a combination of density map correlation and clash filtering to select suitable loop candidates.

Density Correlation

The conformation of the fragment is converted to a density map following the approach by DiMaio¹³⁶. At each atom position, a Gaussian sphere of density is placed, with magnitude proportional to the atomic weight. The resulting density map is then correlated to the experimental density map with a real-spatial cross-correlation.

Filtering Clashing Backbone Candidates

Very often the backbone candidates sterically clash with the protein. While slight steric overlap can be corrected by energy optimization procedures, candidates that are strongly overlapping with the rest of the protein structure can be safely discarded: It would be futile to further process them as for sure they are not good candidates. A first step is to calculate a clash score and remove clashing loops from the candidate list. This provides a very powerful filter to remove true negatives very early on.

We use a very simple steric energy function⁷⁷ that is zero for atoms that are not clashing and increases linearly to a value of up to 10 for clashing atoms. The used cutoff radii (r_i

for the first, r_j for the second atom) are based on the element of the atom. We use 1.5 for carbon, 1.3 for nitrogen and oxygen and 1.7 for sulphur atoms. The energy is defined as:

$$E(d_{ij}, r_i, r_j) = \begin{cases} 0, & \text{if } d_{ij} > r_i + r_j \\ 10, & \text{if } d_{ij} < 0.8254 \cdot (r_i + r_j) \\ 57.273 \cdot \left(1 - \frac{d_{ij}}{r_i + r_j}\right), & \text{otherwise} \end{cases}$$

Below are two typical cases where the RMSD of the backbone candidate is plotted against the clash score (**figure 11.6**). On the left side, the filtering works very nicely and we keep the candidates close to the solution while removing loops of high RMSD. In the second case, we see, that loops with very diverse RMSD have a steric energy close to zero. Several false-positives remain after the filtering step. The latter is a typical scenario for loops that are on the protein surface, whereas the first is a typical case when the loop is on the inside of the protein: The conformational space is very confined and candidates that are not close to the solution are clashing with the protein.

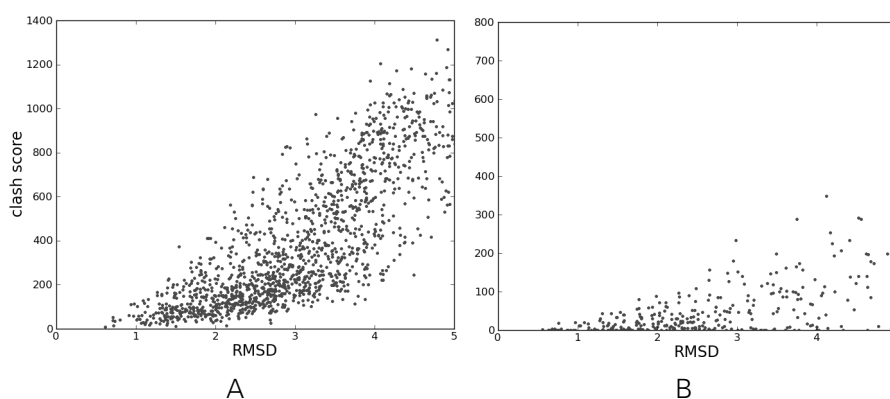


Figure 11.6 Two typical examples of RMSD to native loop plotted against clash score. (A) internal loop of a protein for which the space is very confined. Any loop deviating from the native will inevitably clash with other parts of the protein. (B) Surface loop. Conformations with large deviations from the native loop are possible without clashes.

Ranking with Statistical Potentials

We have looked into the application of statistical potentials as a filter for backbone fragments. More specifically, a torsion potential over 3 consecutive phi/psi pairs and a solvation potential from QMEAN. In addition, a residue-level potential parametrized on the $C\alpha$ - $C\alpha$ distance and the angle between the two $C\alpha$ - $C\beta$ vectors has been implemented. We have tested the performance of the potentials of mean force on a few artificial test sets. The performance of these potentials as filters was found to be very limited. For ARP/wARP model building, scoring of loops by traditional energy functions is likely to be even less efficient, since the surrounding of the loop is not complete and thus lacking many stabilizing interactions. In addition, the sampling produces backbone conformations without sidechains and scoring would be limited to backbone atoms only. Many

loop-stabilizing interactions are mediated by sidechains atoms which will not be included in the scoring.

5 Conclusions

The fragment sampling protocol presented is a very efficient procedure to sample backbone conformations. On the Rossi testset, the procedure is able to identify loop conformations which are on average within 2Å RMSD to the native structure for loops up to 10 residues in length. This performance is comparable to the other available loop sampling programs described by Rossi, but at a fraction of the computational cost. For a detailed analysis of the fragment sampling protocol in the context of ARP/wARP model building, we refer to the 'Knowledge-Based Extension of Fragmented Models at Low Resolution in ARP/wARP' chapter.

Summary and Outlook

In this work, we have presented novel methods for protein structure comparison, protein model quality assessment, automated homology modeling, and X-ray density map interpretation. These methods have been developed based on the OpenStructure framework, introduced in the thesis. The framework has made it possible to seamlessly integrate the presented algorithms with existing tools. Moreover, the powerful visualisation capabilities of the framework have been indispensable for development, and understanding the molecular aspects of proteins.

First, we have described the local distance difference test (IDDT) for comparison of protein structures, the measure has been used as one of the official criteria for the CASP9 and CASP10 tertiary structure prediction evaluation. The stereo-chemical and clash filters have helped to identify physically impossible models and driven participating servers to generate more realistic models. In addition, IDDT has been successfully applied in the CAMEO live benchmark to assess the quality of structure prediction servers in an unsupervised manner. Here, the true power of IDDT comes into play, as the score accurately describes structural similarity even in presence of domain movements. The splitting of structures into assessment units, as it is currently done in the CASP experiment, is rendered unnecessary.

With DOMF, we have presented a graph-based method to identify common building blocks in protein structures. The algorithm uses an iterative neighborhood update to partition the residues into domains. On a test set of proteins which undergo structural rearrangements, we have successfully used DOMF to identify structural stable sets of residues. A multi-template modeling algorithm based on DOMF has been presented, which shows that it is beneficial to filter constraints for consistency at the residue level prior to passing them to the modeling program.

For assessment of model quality we have developed two extensions to the QMEAN scoring function. The first, QMEAN Z-score, relates the energy of models to that of experimental structures of similar size. By that, the quality estimates are placed on an absolute scale. The score is expressed as a Z-score of the model being of X-ray like quality. The work on QMEAN Z-score has identified the requirement for specialized scoring functions for particular classes of proteins. For example, due to the different physico-chemical properties of biological membranes, interactions which are favourable inside a membrane are energetically prohibited in aqueous solution. When calculating the potential of mean force energy of membrane proteins with potentials trained on soluble proteins, the energies are often higher than for soluble proteins of similar size. Therefore, specialized scoring functions targeted at the characteristics of membrane proteins need to be developed.

QMEANdist complements the potential of mean force terms with constraints from evolutionary related protein structures. In addition to scoring the models with QMEAN, the models are scored by agreement with a local $C\alpha$ - $C\alpha$ atom distance propensity obtained from templates. These restraints are very helpful in distinguishing correct from

incorrect folds, and successfully filter models with over-optimized potentials of mean force terms. QMEANDist has been extensively benchmarked during the CASP9 experiment and been shown to perform very well for ranking and selection of models. The CASP9 benchmark identified room for improvement for local, per-residue quality prediction. The magnitude of local scores depend on the number of restraints, their amplitude, and spread and it has proven difficult to correlate them to a structural similarity measure. The Q-score employed by McGuffin and Roche¹¹⁹, or the filtered constraints by Paluszewski²³⁵ seem to be better suited for local quality prediction.

Motivated by results of QMEANDist in the CASP9 QA category, QMEANDist was integrated into the automated modeling pipeline of the next generation of SWISS-MODEL. The sequence similarity weighting scheme employed during CASP9 was replaced by a probabilistic template quality estimation. Properties of the target-template alignment and predicted features (secondary structure and solvent accessibility agreement) act as IDDT predictors and are combined in a probabilistic manner. Our template selection scheme ensures good performance in all sequence identity regimes. Depending on the evolutionary distance between the target and the template, the relative importance of the predictors is adapted. In the high-sequence identity regime, template selection is mainly driven by sequence similarity, whereas in the low-sequence identity regime, the importance of HHsearch/HHblits scores and other predictors increases. The new pipeline represents a substantial improvement compared to the current version of SWISS-MODEL. On average, the models are more accurate, and the number of target sequences where a model can be built is increased. Nevertheless, there is gap to top-performing structure prediction servers in CAMEO. Further work on the modeling pipeline in two areas is therefore required: first, model accuracy and coverage would tremendously benefit from extension of single-template models. For these cases, a multi-template modeling pipeline based on DOMF should be implemented. Second, it should be investigated whether the template quality predictor parametrisation can be optimized. Currently, the properties are multiplied by coverage to account for the coverage-dependence of IDDT. The predictors could be modified to estimate a coverage-corrected IDDT, i.e. an IDDT calculated on residues present in the both the template and the target. These predictors could additionally be parametrized on the alignment length, to account for the length-dependence of the properties.

The newly designed web-interface for SWISS-MODEL Next Generation is a large step towards providing a more interactive modeling environment for non-experts. The completely overhauled template selection and comparison page, allows to compare both the biology and structure of identified templates at a glance. By building on this solid foundation, new interactive modeling applications are within reach. For example, tools to analyze bindings sites of proteins or an alignment editor with live update of protein structure models.

Acknowledgement

I would like to thank *Torsten Schwede* for his constant support and wise words. It has been a fantastic few years with many lessons learned. I am especially grateful for his ability to put the work we do into bigger perspective. Being reminded that it's all about biology is a good thing once in a while... Also, I would like to thank *Andrew Torda* for being part of my thesis committee and agreeing to serve as the co-referee.

Many people have made the years unforgettable. Most of this work would not have been possible without their constant support and feedback. In particular, I am greatly indebted to *Pascal Benkert*. He has been a fantastic mentor during my *early years*. Working late to get the CASP9 server up and running was an experience I will look back to, when I am old and bald. Also, I am grateful for all the evenings at the Linde, where we could have sworn we just had 4 glass of beer but the waitress made us to pay for 5. *Ansgar Philippsen* for introducing me to that little piece of software that later evolved into OpenStructure. I still remember the moment as if it were yesterday. A huge 'thank you', for making the graphics look that great and all the discussions on software design. A huge 'thank you' also goes to Tim Wiegels from the EMBL in Hamburg. Working together on ARP/wARP didn't feel like work at all. *Tobias Schmidt* for the many hours of coffee drinking, the scientific and technical discussions, and getting Drug The Bug from a decent product to the nifty gadget it is now. *Juergen Haas* for always finding another five minutes to help and read a manuscript, and keeping me motivated. *Andrew Waterhouse* for your fantastic work on SWISS-MODEL and for teaching me that normal people use the backspace key to delete characters, *Bienchen* for all the work on OpenStructure, SWISS-MODEL, and the after-work beer at the Cargo Bar. *Valerio Mariani* for the work on IDDT and his burning passion to bundle, *Lorenza Bordoli* for being the good soul of the group, *Andreas Schenk* for all the help and patience to explain things. *Gabriel Studer* for being a fun office mate and his unbelievable enthusiasm for all things scientific. *Konstantin Arnold* for all his help in setting up web services and managing the BC2 computing infrastructure. Last but not least I would like to thank *Andreas Bergner* for carefully reading my thesis and giving very valuable feedback.

References

- 1 Dugas, H. (1999). *Bioorganic Chemistry: A Chemical Approach to Enzyme Action*. Springer.
- 2 Pal, D. and Chakrabarti, P. (1999). Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations. *Journal of Molecular Biology*, 294(1), 271 - 288.
- 3 Ramachandran, G. and Mitra, A. K. (1976). An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. *Journal of Molecular Biology*, 107(1), 85 - 92.
- 4 Berg, J., Tymoczko, J. and Stryer, L. (2002). *Biochemistry*, volume Teile 1-34 of *Biochemistry*. W.H. Freeman.
- 5 Baldwin, R. L. (2007). Energetics of protein folding. *Journal of Molecular Biology*, 371(2), 283-301.
- 6 Eisenberg, D. (2003). The discovery of the α -helix and β -sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), 11207-11210.
- 7 Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577-2637.
- 8 Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2), 195-202.
- 9 Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, 33, 2302-2309.
- 10 Bourne, P. and Weissig, H. (2003). *Structural Bioinformatics*. John Wiley & Sons.
- 11 Goodsell, D. S. and Olson, A. J. (2000). Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure*, 29(1), 105-153.
- 12 Hilser, V. J., Wrabl, J. O. and Motlagh, H. N. (2012). Structural and Energetic Basis of Allostery. In Rees, DC, editor, *Annual Review of Biophysics*, number 41 in Annual Review of Biophysics, pages 585-609. .
- 13 Rhodes, G. (2010). *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Elsevier Science.
- 14 Keeler, J. (2011). *Understanding NMR Spectroscopy*. Wiley.
- 15 Glaeser, R. (2007). *Electron crystallography of biological macromolecules*. Oxford University Press.
- 16 Adams, P. D., Afonine, P. V., Grosse-Kunstleve, R. W., Read, R. J. and Richardson, J. S. et al. (2009). Recent developments in phasing and structure refinement for macromolecular crystallography. *Current Opinion in Structural Biology*, 19(5), 566-572.
- 17 Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V. and Stark, H. (2010). Ribosome dynamics and tRNA movement by time-resolved electron cryo-microscopy. *Nature*, 466(7304), 329-333.

- 18 Bernstein, F. C., Koetzle, T., Williams, G., Meyer, E. and Brice, M. et al. (1977). Protein Data Bank - Computer-Based Archival file for Macromolecular Structures. *Journal of Molecular Biology*, 112(3), 535-542.
- 19 Chandonia, J. and Brenner, S. (2006). The impact of structural genomics: Expectations and outcomes. *Science*, 311(5759), 347-351.
- 20 Berman, H., Henrick, K. and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structure Biology*, 10, 980.
- 21 Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G. and Bhat, T. N. et al. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235-242.
- 22 Velankar, S., Alhroub, Y., Best, C., Caboche, S. and Conroy, M. J. et al. (2012). PDBe: Protein Data Bank in Europe. *Nucleic Acids Research*, 40(D1), D445-D452.
- 23 Kinjo, A. R., Suzuki, H., Yamashita, R., Ikegawa, Y. and Kudou, T. et al. (2012). Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, 40(D1), D453-D460.
- 24 Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H. and Berglund, A.-C. et al. (2009). PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *Journal of Applied Crystallography*, 42, 376-384.
- 25 Sanderson, K. (2009). New protein structures replace the old. *Nature*, 459(7250), 1038-1039.
- 26 Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T. and Swindells, M. B. et al. (1997). Cath—a hierarchic classification of protein domain structures. *Structure*, 5(8), 1093-108.
- 27 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. *Journal of Molecular Biology*, 247(4), 536-540.
- 28 Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22), 10915-10919.
- 29 Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J. and Zhang, Z. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-3402.
- 30 Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755-763.
- 31 Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, 7(10).
- 32 Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7), 951-960.
- 33 Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(December), 173-5.
- 34 Eddy, S. (2004). Where did the BLOSUM62 alignment score matrix come from?. *Nature Biotechnology*, 22(8), 1035-1036.
- 35 Valdar, W. S. (2002). Scoring residue conservation. *Proteins-Structure Function and Bioinformatics*, 48(2), 227-241.

- 36 Tan, Y. H., Huang, H. and Kihara, D. (2006). Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. *Proteins-Structure Function and Bioinformatics*, 64(3), 587-600.
- 37 Finn, R. D., Clements, J. and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(2), W29-W37.
- 38 Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96), 223-230.
- 39 Vendruscolo, M. (2012). Proteome folding and aggregation. *Current Opinion in Structural Biology*, 22(2), 138 - 143. Theory and simulation/Macromolecular assemblages.
- 40 Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 5, 823-826.
- 41 Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13), 3370-3374.
- 42 Guex, N., Peitsch, M. C. and Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis*, 30(S1), S162-S173.
- 43 Dessailly, B. H., Nair, R., Jaroszewski, L., Fajardo, J. E. and Kouranov, A. et al. (2009). PSI-2: Structural Genomics to Cover Protein Domain Family Space. *Structure*, 17(6), 869-881.
- 44 Zhang, Y., Hubner, I., Arakaki, A., Shakhnovich, E. and Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Science of the United States of America*, 103(8), 2605-2610.
- 45 Martin, A. C., MacArthur, M. W. and Thornton, J. M. (1997). Assessment of comparative modeling in casp2. *Proteins, Suppl 1*, 14-28.
- 46 Alwyn Jones, T. and Kleywegt, G. J. (1999). CASP3 comparative modeling evaluation. *Proteins, Suppl 3*, 30-46.
- 47 Zemla, A., Venclovas, Moul, J. and Fidelis, K. (2001). Processing and evaluation of predictions in CASP4. *Proteins, Suppl 5*, 13-21.
- 48 Tramontano, A. and Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins, 53 Suppl 6*, 352-68.
- 49 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.
- 50 Moul, J. (2005). A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3), 285 - 289.
- 51 Battey, J. N. D., Kopp, J., Bordoli, L., Read, R. J. and Clarke, N. D. et al. (2007). Automated server predictions in CASP7. *Proteins-Structure Function and Bioinformatics*, 69(8), 68-82. 7th Meeting on Critical Assessment of Techniques for Protein Structure Prediction, Pacific Grove, CA, NOV 26-30, 2006.
- 52 Keedy, D. A., Williams, C. J., Headd, J. J., Arendall III, W. B. and Chen, V. B. et al. (2009). The other 90% of the protein: Assessment beyond the C alpha s for CASP8 template-based and high-accuracy models. *Proteins-Structure Function and Bioinformatics*, 77, 29-49.

- 53 Mariani, V., Kiefer, F., Schmidt, T., Haas, J. and Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins-Structure Function and Bioinformatics*, 79(10), 37-58.
- 54 Haas, J., Schmidt, T., Biasini, M., Arnold, K. and Waterhouse, A. et al. (in preparation). CAMEO - Continuous Automated Modeling Evaluation. .
- 55 Chou, P. and Fasman, G. (1974). Prediction of Protein Conformation. *Biochemistry*, 13(2), 222-245.
- 56 Rost, B. and Eyrich, V. A. (2001). EVA: Large-scale analysis of secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 45(S5), 192-199.
- 57 Schwede, T. and Peitsch, M. (2008). *Computational Structural Biology: Methods and Applications*. World Scientific.
- 58 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403-410.
- 59 Marti-Renom, M., Madhusudhan, M. and Sali, A. (2004). Alignment of protein sequences by their profiles. *Protein Science*, 13(4), 1071-1087.
- 60 Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10), 846-856.
- 61 Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J. E. and Kelley, L. A. (2008). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins-Structure Function and Bioinformatics*, 70(3), 611-625.
- 62 Wu, S. and Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10), 3375-3382.
- 63 McGuffin, L. J. (2008). The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics*, 24(4), 586-587.
- 64 Xu, J., Li, M., Kim, D. and Xu, Y. (2003). RAPTOR: optimal protein threading by linear programming. *Journal of Bioinformatics and Computational Biology*, 1(1), 95-117.
- 65 Xu, J., Li, M. and Xu, Y. (2004). Protein threading by linear programming: Theoretical analysis and computational results. *Journal of Combinatorial Optimization*, 8(4), 403-418.
- 66 Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins-Structure Function and Bioinformatics*, 77(S9), 100-113.
- 67 Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2), 85-94.
- 68 Peng, J. and Xu, J. (2010). Low-homology protein threading. *Bioinformatics*, 26(12), i294-i300.
- 69 Peng, J. and Xu, J. (2011). RaptorX: Exploiting structure information for protein alignment by statistical inference. *Proteins-Structure Function and Bioinformatics*, 79(S10), 161-171.
- 70 Blundell, T., Sibanda, B., Sternberg, M. and Thornton, J. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111), 347-352.
- 71 Greer, J. (1980). Structure of haptoglobin heavy-chain and other serine protease homologs by comparative model-building. *Biophysical Journal*, 32(1), 218-219.

- 72 Sali, A. and Blundell, T. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3), 779 - 815.
- 73 Canutescu, A. A. and Dunbrack, R. L. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12, 963–972.
- 74 Jamroz, M. and Kolinski, A. (2010). Modeling of loops in proteins: a multi-method approach. *BMC Struct. Biol.*, 10, 5.
- 75 Lee, J., Lee, D., Park, H., Coutsiyas, E. A. and Seok, C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins*, 78, 3428–3436.
- 76 Rossi, K. A., Nayeem, A., Weigelt, C. A. and Krystek, S. R. (2009). Closing the side-chain gap in protein loop modeling. *J. Comput. Aided Mol. Des.*, 23, 411–418.
- 77 Canutescu, A. A., Shelenkov, A. A. and Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12, 2001–2014.
- 78 Krivov, G. G., Shapovalov, M. V. and Dunbrack R. L., J. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4), 778–95.
- 79 Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F. and Thompson, J. et al. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. In Johnson, ML and Brand, L, editor, *Methods in Enzymology, Vol 487 : Computer Methods, Pt C Methods in Enzymology*, pages 545-574. .
- 80 Kinch, L., Shi, S. Y., Cong, Q., Cheng, H. and Liao, Y. et al. (2011). CASP9 assessment of free modeling target predictions. *Proteins-Structure Function and Bioinformatics*, 79(10), 59-73.
- 81 Kim, D. E., Blum, B., Bradley, P. and Baker, D. (2009). Sampling Bottlenecks in De novo Protein Structure Prediction. *Journal of Molecular Biology*, 393(1), 249-260.
- 82 Zhang, J., Liang, Y. and Zhang, Y. (2011). Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure*, 19(12), 1784-1795.
- 83 Zhu, J., Fan, H., Periole, X., Honig, B. and Mark, A. E. (2008). Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins-Structure Function and Bioinformatics*, 72(4), 1171-1188.
- 84 Engh, R. and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Crystallographica Section A*, 47(Part 4), 392-400.
- 85 Laskowski, R., MacArthur, M., Moss, D. and Thornton, J. (1993). Procheck - A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(Part 2), 283-291.
- 86 Read, R. J., Adams, P. D., Arendall III, W. B., Brunger, A. T. and Emsley, P. et al. (2011). A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure*, 19(10), 1395-1412.
- 87 Lazaridis, T. and Karplus, M. (2000). Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, 10(2), 139-145.
- 88 Ivetac, A. and Sansom, M. S. P. (2008). Molecular dynamics simulations and membrane protein structure quality. *European Biophysics Journal with Biophysics Letters*, 37(4), 403-409.
- 89 DeRonne, K. W. and Karypis, G. (2009). Improved estimation of structure predictor quality. *BMC Structural Biology*, 9.

- 90 Zhang, J., Zhang, J., Wang, Q., Shang, Y. and Xu, D. et al. (2011). Quality Assessment of Predicted Protein Structures by Using Molecular Dynamic Simulations. *Biophysics Journal*, 100(3, 1), 214.
- 91 Zhang, J., Wang, Q., Vantasin, K., Zhang, J. and He, Z. et al. (2011). A multilayer evaluation approach for protein structure prediction and model quality assessment. *Proteins-Structure Function and Bioinformatics*, 79(10), 172-184.
- 92 Solis, A. and Rackovsky, S. (2006). Improvement of statistical potentials and threading score functions using information maximization. *Proteins-Structure Function and Bioinformatics*, 62(4), 892-908.
- 93 Solis, A. and Rackovsky, S. (2002). Optimally informative backbone structural propensities in proteins. *Proteins-Structure Function and Genetics*, 48(3), 463-486.
- 94 Solis, A. and Rackovsky, S. (2000). Optimized representations and maximal information in proteins. *Proteins-Structure Function and Genetics*, 38(2), 149-164.
- 95 Fitzgerald, J. E., Jha, A. K., Colubri, A., Sosnick, T. R. and Freed, K. F. (2007). Reduced C-beta statistical potentials can outperform all-atom potentials in decoy identification. *Protein Science*, 16(10), 2123-2139.
- 96 Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27.
- 97 Deng, H., Jia, Y., Wei, Y. and Zhang, Y. (2012). What is the best reference state for designing statistical atomic potentials in protein structure prediction?. *Proteins-Structure Function and Bioinformatics*, 80(9), 2311-2322.
- 98 Dill, K. A. (1997). Additivity Principles in Biochemistry. *Journal of Biological Chemistry*, 272(2), 701-704.
- 99 Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force - An approach to the knowledge-based prediction of local structures in globular-proteins. *Journal of Molecular Biology*, 213(4), 859-883.
- 100 Sippl, M. (1993). Recognition of errors in 3-dimensional structures of proteins. *Proteins-Structure Functions and Genetics*, 17(4), 355-362.
- 101 Bahar, I. and Jernigan, R. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *Journal of Molecular Biology*, 266(1), 195-214.
- 102 Melo, F. and Feytmans, E. (1997). Novel knowledge-based mean force potential at atomic level. *Journal of Molecular Biology*, 267(1), 207-222.
- 103 Samudrala, R. and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, 275(5), 895-916.
- 104 Tobi, D., Shafran, G., Linial, N. and Elber, R. (2000). On the design and analysis of protein folding potentials. *Proteins-Structure Function and Genetics*, 40(1), 71-85.
- 105 Lu, H. and Skolnick, J. (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins-Structure Function and Genetics*, 44(3), 223-232.
- 106 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 Science*, 11(11), 2714-2726.

- 107 Shen, M.-Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11), 2507-2524.
- 108 Benkert, P., Tosatto, S. C. E. and Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins-Structure Function and Bioinformatics*, 71(1), 261-277.
- 109 Zhou, H. and Skolnick, J. (2011). GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophysical Journal*, 101(8), 2043-2052.
- 110 Jones, D., Taylor, W. and Thornton, J. (1992). A new approach to Protein Fold Recognition. *Nature*, 358(6381), 86-89.
- 111 Holm, L. and Sander, C. (1992). Evaluation of Protein Models by Atomic Solvation Preference. *Journal of Molecular Biology*, 225(1), 93-105.
- 112 Albiero, A. and Tosatto, S. (2006). Fine-grained statistical torsion angle potentials are effective in discriminating native protein structures. *Current drug discovery technologies*, 3, 75-81.
- 113 Betancourt, M. and Skolnick, J. (2004). Local propensities and statistical potentials of backbone dihedral angles in proteins. *Journal of Molecular Biology*, 342(2), 635-649.
- 114 Kocher, J., Rooman, M. and S.J., W. (1994). Factors Influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *Journal of Molecular Biology*, 235(5), 1598-1613.
- 115 Tosatto, S. (2005). The Victor/FRST function for model quality estimation. *Journal of Computational Biology*, 12(10), 1316-1327.
- 116 Cheng, J., Randall, A. Z., Sweredoski, M. J. and Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33(Suppl 2), W72-W76.
- 117 Benkert, P., Schwede, T. and Tosatto, S. C. E. (2009a). QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. *BMC Structural Biology*, 9.
- 118 Kryshchuk, A., Fidelis, K. and Tramontano, A. (2011). Evaluation of model quality predictions in CASP9. *Proteins-Structure Function and Bioinformatics*, 79(10), 91-106.
- 119 McGuffin, L. J. and Roche, D. B. (2010). Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, 26(2), 182-188.
- 120 Larsson, P., Skwark, M. J., Wallner, B. and Elofsson, A. (2009). Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins-Structure Function and Bioinformatics*, 77(9), 167-172.
- 121 Hinsen, K. and Sadron, R. C. (2000). The molecular modeling toolkit: a new approach to molecular simulations. *J. Comput. Chem*, 21, 79-85.
- 122 Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D-Biological Crystallography*, 66(Part 4), 486-501.

- 123 Canutescu, A. and Dunbrack, R. (2005). MolIDE: a homology modeling framework you can click with. *Bioinformatics*, 21(12), 2914-2916.
- 124 Eswar, N., Eramian, D., Webb, B., Shen, M. and Sali, A. (2006). *Protein structure modeling with MODELLER*, pages 145-159. Humana Press Inc.
- 125 Kohlbacher, O. and Lenhof, H. (2000). BALL - rapid software prototyping in computational molecular biology. *Bioinformatics*, 16(9), 815-824.
- 126 Gruenberg, R., Nilges, M. and Leckner, J. (2007). Biskit - A software platform for structural bioinformatics. *Bioinformatics*, 23(6), 769-770.
- 127 Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14, 33-38.
- 128 Benkert, P., Kuenzli, M. and Schwede, T. (2009b). QMEAN server for protein model quality estimation. *Nucleic Acids Research*, 37, W510-W514.
- 129 Arnold, K., Kiefer, F., Kopp J. Battey, J. N. D., Podvinec, M. and Westbrook, J. D. et al. (2009). The Protein Model Portal. *Journal of Structural and Functional Genomics*, pp. 1-8.
- 130 Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006). The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2), 195-201.
- 131 Bordoli, L., Kiefer, F., Arnold, K., Benkert, P. and Battey, J. et al. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols*, 4(1), 1-13.
- 132 Philippsen, A., Schenk, A. D., Signorell, G. A., Mariani, V. and Berneche, S. et al. (2007). Collaborative EM image processing with the IPLT image processing library and toolbox. *Journal of Structural Biology*, 157(1), 28-37.
- 133 Sanner, M., Olson, A. and Spehner, J. (1996). Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3), 305-320.
- 134 DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P. and Plake, H. R. et al. (2009). Thermodynamic and Structural Effects of Conformational Constraints in Protein-Ligand Interactions. Entropic Paradoxy Associated with Ligand Preorganization. *Journal of the American Chemical Society*, 131(46), 16758-16770.
- 135 Biasini, M., Mariani, V., Haas, J., Scheuber, S. and Schenk, A. D. et al. (2010). OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics*, 26, 2626-2628.
- 136 DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. and Baker, D. (2009). Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta. *Journal of Molecular Biology*, 392(1), 181-190.
- 137 Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W. and Kipper, J. et al. (2007). Determining the architectures of macromolecular assemblies. *Nature*, 450(7170), 683-694.
- 138 Trabuco, L. G., Villa, E., Mitra, K., Frank, J. and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 16(5), 673-683.

- 139 Rigden, D. J., Keegan, R. M. and Winn, M. D. (2008). Molecular replacement using ab initio polyaniline models generated with ROSETTA. *Acta Crystallographica Section D-Biological Crystallography*, 64(Part 12), 1288-1291.
- 140 Holm, L. and Sander, C. (1993). Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology*, 233(1), 123 - 138.
- 141 Olechnovic, K., Kulberkyte, E. and Venclovas, C. (2012). Cad-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins-Structure Function and Bioinformatics*, pp. n/a–n/a.
- 142 Benkert, P., Biasini, M. and Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27, 343–350.
- 143 Frigo, M., Steven and Johnson, G. (2005). The design and implementation of fftw3. In *Proceedings of the IEEE*, pages 216–231.
- 144 Guennebaud, G., Jacob, B. and others (2010). Eigen v3. <http://eigen.tuxfamily.org>.
- 145 Schroeder, W., Martin, K. and Lorensen, B. (2004). *The Visualization Toolkit, Third Edition*. Kitware Inc.
- 146 Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B. and Echols, N. et al. (2011). The Phenix software for automated determination of macromolecular structures. *Methods*, 55(1), 94-106.
- 147 Dubois, P., Hinsén, K. and Hugunin, J. (1996). Numerical Python. *Computers in Physics*, 10(3).
- 148 Jones, E., Oliphant, T., Peterson, P. and others (2001). SciPy: Open source scientific tools for Python. <http://www.scipy.org/>.
- 149 Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering*, 9(3), 90–95.
- 150 Sukumaran, J. and Holder, M. T. (2010). DendroPy. *Bioinformatics*, 26(12), 1569–1571.
- 151 Chaudhury, Lyskov, S. and Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*.
- 152 Gamma, E., Helm, R., Johnson, R. and Vlissides, J. (1995). *Design Patterns*. Reading, MA: Addison Wesley.
- 153 Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.3r1.
- 154 Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197.
- 155 Needleman, S. G. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443-453.
- 156 Armon, A., Graur, D. and Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology*, 307(1), 447 - 463.
- 157 Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R. and McGettigan, P. A. et al. (2007). Clustal W and Clustal X version 2.0.. *Bioinformatics (Oxford, England)*, 23(21), 2947–2948.

- 158 Egloff, M., Benarroch, D., Selisko, B., Romette, J. and Canard, B. (2002). An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *EMBO Journal*, 21(11), 2757-2768.
- 159 Halstead, S. B. (2007). Dengue. *Lancet*, 370(9599), 1644-1652.
- 160 Langer, G., Cohen, S. X., Lamzin, V. S. and Perrakis, A. (2008). Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols*, 3, 1171-1179.
- 161 Philippsen, A., Schenk, A., Stahlberg, H. and Engel, A. (2003). IPLT-image processing library and toolkit for the electron microscopy community. *Journal of Structural Biology*, 144(1-2), 4-12.
- 162 Mariani, V., Schenk, A. D., Philippsen, A. and Engel, A. (2011). Simulation and correction of electron images of tilted planar weak-phase samples. *Journal of Structural Biology*, 174(2), 259-268.
- 163 Yang, M. H., Nickerson, S., Kim, E. T., Liot, C. and Laurent, G. et al. (2012). Regulation of RAS oncogenicity by acetylation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(27), 10843-10848.
- 164 Shan, Y., Eastwood, M. P., Zhang, X., Kim, E. T. and Arkhipov, A. et al. (2012). Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell*, 149(4), 860-870.
- 165 Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F. and Rinner, O. et al. (2010). Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. *Molecular & Cellular Proteomics*, 9(8), 1634-1649.
- 166 Ha, N., Oh, S., Sung, J., Cha, K. and Lee, M. et al. (2001). Supramolecular assembly and acid resistance of *Helicobacter pylori* urease. *Nature Structural Biology*, 8(6), 505-509.
- 167 Kahraman, A., Malmstrom, L. and Aebersold, R. (2011). Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics*, 27(15), 2163-2164.
- 168 Amrein, B., Schmid, M., Collet, G., Cuniasse, P. and Gilardoni, F. et al. (2012). Identification of two-histidines one-carboxylate binding motifs in proteins amenable to facial coordination to metals. *Metallomics*, 4(4), 379-388.
- 169 Mariani, V., Kiefer, F., Schmidt, T., Haas, J. and Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins-Structure Function and Bioinformatics*, 79(10), 37-58.
- 170 Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27), 11079-11084.
- 171 Schwede, T., Sali, A., Honig, B., Levitt, M. and Berman, H. M. et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. *Structure*, 17, 151-159.
- 172 Moulton, J., Pedersen, J., Judson, R. and Fidelis, K. (1995). A large-scale experiment to assess protein-structure prediction methods. *Proteins-Structure Function and Genetics*, 23(3), R2-R4.
- 173 Hubbard, T. J. (1999). Rms/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins, Suppl 3*, 15-21.

- 174 Mosimann, S., Meleshko, R. and James, M. N. (1995). A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins*, 23(3), 301-17.
- 175 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), 776-85.
- 176 Sippl, M. J. (2008). On distance and similarity in fold space. *Bioinformatics*, 24(6), 872-873.
- 177 Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins-Structure Function and Bioinformatics*, 57(4), 702-710.
- 178 Clarke, N. D., Ezkurdia, I., Kopp, J., Read, R. J. and Schwede, T. et al. (2007). Domain definition and target classification for CASP7. *Proteins*, 69 Suppl 8, 10-8.
- 179 Kinch, L. N., Shi, S., Cheng, H., Cong, Q. and Pei, J. et al. (2011). CASP9 target classification. *Proteins-Structure Function and Bioinformatics*, 79(10), 21-36.
- 180 Bordogna, A., Pandini, A. and Bonati, L. (2011). Predicting the Accuracy of Protein-Ligand Docking on Homology Models. *Journal of Computational Chemistry*, 32(1), 81-98.
- 181 Kopp, J., Bordoli, L., Battey, J. N. D., Kiefer, F. and Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins-Structure Function and Bioinformatics*, 69(8), 38-56.
- 182 Chen, V. B., Arendall W. B., r., Headd, J. J., Keedy, D. A. and Immormino, R. M. et al. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, 66(Pt 1), 12-21.
- 183 Engh, R. A. and Huber, R. (2006). *Structure quality and target parameters* John Wiley & Sons, Ltd.
- 184 Allen, F. H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B*, 58(Pt 3 Pt 1), 380-8.
- 185 Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B. and Rentzsch, R. et al. (2011). Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research*, 39(1), D420-D426.
- 186 Shi, S., Pei, J., Sadreyev, R. I., Kinch, L. N. and Majumdar, I. et al. (2009). Analysis of CASP8 targets, predictions and assessment methods. *Database (Oxford)*, 2009, bap003.
- 187 Vendruscolo, M., Subramanian, B., Kanter, I., Domany, E. and Lebowitz., J. (1999). *Statistical Properties of Contact Maps*, volume 59. College Park, MD, ETATS-UNIS: American Physical Society.
- 188 Flory, P. (1969). *Statistical mechanics of chain molecules*. Interscience Publishers.
- 189 Huggins, M. (1958). *Physical chemistry of high polymers*. Wiley.
- 190 Cheng, J. (2008). A multi-template combination algorithm for protein comparative modeling. *BMC Structural Biology*, 8.
- 191 Chakravarty, S., Godbole, S., Zhang, B., Berger, S. and Sanchez, R. (2008). Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC Structural Biology*, 8(1), 31.

- 192 Venclovas, C. and Margelevicius, M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins-Structure Function and Bioinformatics*, 61(7), 99-105.
- 193 Snyder, D. A. and Montelione, G. T. (2005). Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins-Structure Function and Bioinformatics*, 59(4), 673-686.
- 194 Qi, G., Lee, R. and Hayward, S. (2005). A comprehensive and non-redundant database of protein domain movements. *Bioinformatics*, 21(12), 2832-2838.
- 195 Hopcroft, J. and Tarjan, R. (1973). Efficient algorithms for graph manipulation. *Commun. ACM*, 16(6), 372-378.
- 196 Wassmann, P., Chan, C., Paul, R., Beck, A. and Heerklotz, H. et al. (2007). Structure of BeF₃-modified response regulator PleD: Implications for diguanylate cyclase activation, catalysis, and feedback inhibition. *Structure*, 15(8), 915-927.
- 197 Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1), 27-64.
- 198 Amemiya, T., Koike, R., Kidera, A. and Ota, M. (2012). PSCDB: a database for protein structural change upon ligand binding. *Nucleic Acids Research*, 40(D1), D554-D558.
- 199 Amemiya, T., Koike, R., Fuchigami, S., Ikeguchi, M. and Kidera, A. (2011). Classification and Annotation of the Relationship between Protein Structural Change and Ligand Binding. *Journal of Molecular Biology*, 408(3), 568-584.
- 200 Poornam, G. P., Matsumoto, A., Ishida, H. and Hayward, S. (2009). A method for the analysis of domain movements in large biomolecular complexes. *Proteins-Structure Function and Bioinformatics*, 76(1), 201-212.
- 201 Dunbrack Jr., R. L. (2006). Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology*, 16(3), 374-384.
- 202 Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C. and Boeckmann, B. et al. (2005). The universal protein resource (uniprot). *Nucleic Acids Research*, 33(Suppl 1), D154-D159.
- 203 Tramontano, A. and Morea, V. (2003). Exploiting evolutionary relationships for predicting protein structures. *Biotechnology and Bioengineering*, 84(7), 756-762.
- 204 Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540), 93-96.
- 205 Koh, I. Y. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D. and Madhusudhan, M. S. et al. (2003). Eva: evaluation of protein structure prediction servers. *Nucleic Acids Research*, 31(13), 3311-3315.
- 206 Eramian, D., Eswar, N., Shen, M. Y. and Sali, A. (2008). How well can the accuracy of comparative protein structure models be predicted?. *Protein Science*, 17, 1881-1893.
- 207 Marti-Renom, M., Stuart, A., Fiser, A., Sanchez, R. and Melo, F. et al. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29, 291-325.
- 208 Melo, F. and Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. *Journal of Molecular Biology*, 277(5), 1141-1152.

- 209 Pettitt, C. S., McGuffin, L. J. and Jones, D. T. Improving sequence-based fold recognition by using 3d model quality assessment. *Bioinformatics*, 21(17), 3509-3515.
- 210 Randall, A. and Baldi, P. (2008). Selectpro: effective protein model selection using a structure-based energy function resistant to blunders. *BMC Structural Biology*, 8(1), 52.
- 211 Wallner, B. and Elofsson, A. (2003). Can correct protein models be identified?. *Protein Science*, 12(5), 1073-1086.
- 212 Moulton, J., Fidelis, K., Kryshtafovych, A., Rost, B. and Hubbard, T. et al. (2007). Critical assessment of methods of protein structure prediction - Round VII. *Proteins-Structure Function and Bioinformatics*, 69(8), 3-9.
- 213 Wang, Z., Tegge, A. N. and Cheng, J. (2009). Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins-Structure Function and Bioinformatics*, 75(3), 638-647.
- 214 Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, 19(12), 1589-1591.
- 215 White, S. H. (2009). Biophysical dissection of membrane proteins. *Nature*, 459(7245), 344-346.
- 216 Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3), 774-797.
- 217 Robinson-Rechavi, M. and Godzik, A. (2005). Structural genomics of *thermotoga maritima* proteins shows that contact order is a major determinant of protein thermostability. *Structure*, 13(6), 857 - 860.
- 218 Ye, Y. and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19(Suppl 2), ii246-ii255.
- 219 Thomas, A., Joris, B. and Brasseur, R. (2010). Standardized evaluation of protein stability. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, 1804(6), 1265 - 1271.
- 220 Kurata, S., Weixlbaumer, A., Ohtsuki, T., Shimazaki, T. and Wada, T. et al. (2008). Modified Uridines with C5-methylene Substituents at the First Position of the tRNA Anticodon Stabilize U-G Wobble Pairing during Decoding. *Journal of Biological Chemistry*, 283(27), 18801-18811.
- 221 Siebold, C., Hansen, B. E., Wyer, J. R., Harlos, K. and Esnouf, R. E. et al. (2004). Crystal structure of hla-dq0602 that protects against type 1 diabetes and confers strong susceptibility to narcolepsy. *Proceedings of the National Academy of Sciences of the United States of America*, 101(7), 1999-2004.
- 222 Jordan, J. B., Poppe, L., Haniu, M., Arvedson, T. and Syed, R. et al. (2009). Hepcidin Revisited, Disulfide Connectivity, Dynamics, and Structure. *Journal of Biological Chemistry*, 284(36), 24155-24167.
- 223 Janes, R., DH, P. and Wallace, B. (1994). The crystal-structure of human endothelin. *Nature Structural Biology*, 1(5), 311-319.
- 224 Eigenbrot, C., Randal, M., Quan, C., Burnier, J. and O'Connell, L. et al. (1991). X-ray structure of the human relaxin at 1.5 Å - Comparison to insulin and implications for receptor-binding determinants. *Journal of Molecular Biology*, 221(1), 15-21.

- 225 Weaver, L. and Matthews, B. (1987). Structure of Bacteriophage-T4 Lysozyme refined at 1.7 Å resolution. *Journal of Molecular Biology*, 193(1), 189-199.
- 226 Murthy, H. M. K., Clum, S. and Padmanabhan, R. (1999). Dengue virus ns3 serine protease. *Journal of Biological Chemistry*, 274(9), 5573-5580.
- 227 Lee, S. H., Hayes, D. B., Rebowksi, G., Tardieux, I. and Dominguez, R. (2007). Toxofilin from *Toxoplasma gondii* forms a ternary complex with an antiparallel actin dimer. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41), 16122-16127.
- 228 Wiederstein, M. and Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, 35(Suppl 2), W407-W410.
- 229 Robinson-Rechavi, M., Alibes, A. and Godzik, A. (2006). Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: Lessons from structural genomics of *Thermotoga maritima*. *Journal of Molecular Biology*, 356(2), 547-557.
- 230 Rykunov, D. and Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, 11(1), 128.
- 231 Cozzetto, D., Kryshtafovych, A. and Tramontano, A. (2009). Evaluation of CASP8 model quality predictions. *Proteins-Structure Function and Bioinformatics*, 77, 157-166.
- 232 Schwede, T., Kopp, J., Guex, N. and Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, 31(13), 3381-3385.
- 233 McGuffin, L. J. and Roche, D. B. (2011). Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins-Structure Function and Bioinformatics*, 79(10), 137-146.
- 234 Zhou, H. and Skolnick, J. (2008). Protein model quality assessment prediction by combining fragment comparisons and a consensus C-alpha contact potential. *Proteins-Structure Function and Bioinformatics*, 71(3), 1211-1218.
- 235 Paluszewski, M. and Karplus, K. (2009). Model quality assessment using distance constraints from alignments. *Proteins*, 75, 540-549.
- 236 Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4), 702-10.
- 237 He, Y., Chen, Y., Alexander, P., Bryan, P. N. and Orban, J. (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38), 14412-14417.
- 238 Fiser, A. (2010). Template-Based Protein Structure Modeling. In Fenyo, D, editor, *Computational Biology*, number 673 in *Methods in Molecular Biology*, pages 73-94. Humana Press Inc.
- 239 Wang, G. and Dunbrack, R. L. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33(Suppl 2), W94-W98.

- 240 Sadowski, M. I. and Jones, D. T. (2007). Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *Proteins-Structure Function and Bioinformatics*, 69(3), 476–485.
- 241 Hildebrand, A., Remmert, M., Biegert, A. and Soeding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *Proteins-Structure Function and Bioinformatics*, 77, 128-132.
- 242 Kelley, L. A. and Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols*, 4(3), 363-371.
- 243 Kiefer, F., Arnold, K., Kuenzli, M., Bordoli, L. and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, 37, D387-D392.
- 244 Schwede, T., Kopp, J., Guex, N. and Peitsch, M. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, 31(13), 3381-3385.
- 245 Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3), 832–837.
- 246 Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.
- 247 Silverman, B. W. (1986). *Density estimation: for statistics and data analysis*. London.
- 248 Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D. and Schlessinger, A. et al. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, 39(1), D465-D474.
- 249 Ganichkin, O. and Wahl, M. C. (2007). Conformational switches in winged-helix domains 1 and 2 of bacterial translation elongation factor SelB. *Acta Crystallographica Section D*, 63(10), 1075–1081.
- 250 Fan, H., Irwin, J. J., Webb, B. M., Klebe, G. and Shoichet, B. K. et al. (2009). Molecular Docking Screens Using Comparative Models of Proteins. *Journal of Chemical Information and Modeling*, 49(11), 2512-2527.
- 251 The Python Software Foundation (1990-2013). The Python Programming Language. <http://www.python.org/>.
- 252 Django Software Foundation (2013). Django - The Web framework for perfectionists with deadlines. <http://www.djangoproject.com>.
- 253 John, R., Corey, F., Katz, Y. and Dan, H. (2013). jQuery JavaScript Library. <http://jquery.org>.
- 254 Dimitry, B. (2013). Raphael.js JavaScript Library. <http://raphaeljs.org>.
- 255 Kim, Y., Babnigg, G., Jedrzejczak, R., Eschenfeldt, W. H. and Li, H. et al. (2011). High-throughput protein purification and quality assessment for crystallization. *Methods*, 55(1), 12-28.
- 256 Hartshorn, M. (2013). OpenAstexViewer - Software for molecular visualisation. <http://openastexviewer.net>.
- 257 Schmidt, T. (2012). *Computational Approaches for Investigating Protein-Ligand Interactions - Towards an in-depth Understanding of the Dengue Virus Methyltransferase*. PhD thesis, Biozentrum, University of Basel.

- 258 Ellson, J., Gansner, E., Koutsofios, L., North, S. and Woodhull, G. (2001). Graphviz — Open Source Graph Drawing Tools. In *Lecture Notes in Computer Science*, pages 483–484.
- 259 Kiefer, F. (2012). *Modeling of tertiary and quaternary protein structures by homology*. PhD thesis, Biozentrum, University of Basel.
- 260 Rose, P. W., Beran, B., Bi, C., Bluhm, W. F. and Dimitropoulos, D. et al. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, 39(1), D392-D401.
- 261 Holton, J. M. (2005). Abstract W0308. In *Am. Crystallogr. Assoc. Annual Meeting*.
- 262 Stroud, R. M., Choe, S., Holton, J., Kaback, H. R. and Kwiatkowski, W. et al. (2007). Annual progress report synopsis of the Center for Structures of Membrane Proteins. *Journal of Structural and Functional Genomics*, 10, 193-208.
- 263 Dyda, F. (2010). Developments in low-resolution biological X-ray crystallography. *F1000 biology reports*, 2, 80.
- 264 Morris, R., Perrakis, A. and Lamzin, V. (2007). *Getting a macromolecular model: model building, refinement, and validation*. Oxford Oxford University Press.
- 265 Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallographica Section D-Biological Crystallography*, 62(Part 9), 1002-1011.
- 266 Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W. and Zwart, P. H. et al. (2008). Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D-Biological Crystallography*, 64(Part 1), 61-69.
- 267 Wiegels, T. and Lamzin, V. S. (2012). Use of noncrystallographic symmetry for automated model building at medium to low resolution. *Acta Crystallographica Section D-Biological Crystallography*, 68(Part 4), 446-453.
- 268 Simons, K., Bonneau, R., Ruczinski, I. and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins-Structure Function and Genetics*, pp. 171-176.
- 269 Qian, B., Raman, S., Das, R., Bradley, P. and McCoy, A. J. et al. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, 450(7167), 259-U7.
- 270 Terwilliger, T., DiMaio, F., Read, R., Baker, D. and Bunkoczi, G. et al. (2012). phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. *Journal of Structural and Functional Genomics*, 13, 81-90.
- 271 Claude, J., Suhre, K., Notredame, C., Claverie, J. and Abergel, C. (2004). CaspR: a web server for automated molecular replacement using homology modelling. *Nucleic Acids Research*, 32(2), W606-W609.
- 272 Joosten, K., Cohen, S. X., Emsley, P., Mooij, W. and Lamzin, V. S. et al. (2008). A knowledge-driven approach for crystallographic protein model completion. *Acta Crystallographica Section D-Biological Crystallography*, 64(Part 4), 416-424.
- 273 van den Bedem, H., Lotan, I., Latombe, J. and Deacon, A. (2005). Real-space protein-model completion: an inverse-kinematics approach. *Acta Crystallographica Section D-Biological Crystallography*, 61(Part 1), 2-13.

- 274 Yao, M., Zhou, Y. and Tanaka, I. (2006). LAFIRE: software for automating the refinement process of protein-structure analysis. *Acta Crystallographica Section D-Biological Crystallography*, 62(Part 2), 189-196.
- 275 Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B. and Davis, I. W. et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D-Biological Crystallography*, 66(Part 2), 213-221.
- 276 Zhang, Y. and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1029-1034.
- 277 Bryson, K., McGuffin, L., Marsden, R., Ward, J. and Sodhi, J. et al. (2005). Protein structure prediction servers at University College London. *Nucleic Acids Research*, 33(2), W36-W38.
- 278 McGuffin, L., Bryson, K. and Jones, D. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4), 404-405.
- 279 Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins-Structure Function and Genetics*, 47(2), 228-235.
- 280 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.
- 281 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-Structure Function and Genetics*, 54(3), 583-595.
- 282 Ling, H., Boodhoo, A., Hazes, B., Cummings, M. and Armstrong, G. et al. (1998). Structure of the Shiga-like toxin I B-pentamer complexed with an analogue of its receptor Gb(3). *Biochemistry*, 37(7), 1777-1788.
- 283 Maiorov, V. and Crippen, G. (1995). Size-Independent Comparison of Protein 3-dimensional Structures. *Proteins-Structure Function and Genetics*, 22(3), 273-283.
- 284 Deane, C. and Blundell, T. (2000). A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins-Structure Function and Genetics*, 40(1), 135-144.
- 285 Baeten, L., Reumers, J., Tur, V., Stricher, F. and Lenaerts, T. et al. (2008). Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLoS Computational Biology*, 4, e1000083.
- 286 Rossi, K. A., Weigelt, C. A., Nayeem, A. and Krystek, S. R. (2007). Loopholes and missing links in protein modeling. *Protein Science*, 16, 1999-2012.
- 287 Kabsch, W. (1976). Solution for the best rotation to relate 2 sets of vectors. *Acta Crystallographica Section A*, 32(SEP1), 922-923.

OpenStructure — Technical Annex

This chapter contains bits and pieces of OpenStructure — implementation details that didn't fit into the general OpenStructure overview chapters and papers.

1 The Query Language

Entity views are a fundamental part of the OpenStructure framework. They are conveniently created using a dedicated mini-language, called the query language. The query language allows to define binary predicates against which parts of the structure, e.g. atoms, residues and chains are matched. Parts, which fulfil all of the criteria are selected, and returned in a new entity view.

Here, we will describe how the query language syntax and its implementation in detail.

Features of the Query Language

The predicates may use any of the available built-in properties defined for the atoms, residues, and chains. Examples include the atom name (*aname*), residue number (*rnum*), chain name (*cname*) or atom element (*ele*). Typically, properties of chains are prefixed with *c*, properties on residues with *r* and atom properties with *a*. A complete list of built-in properties is given in the OpenStructure documentation. In addition, the predicates may refer to user-defined properties declared using the *generic properties system* (see below).

Selecting atom in proximity of another atom or point is achieved with the *within-operator* of the query language: To select all atoms within 5 Å of the origin of the reference system, the query `5 <> {0,0,0}` may be used. The `<>` operator is called the 'within' operator. Instead of a point, the within statements can also be used to return a view containing all chains, residues and atoms within a radius of another selection statement applied to the same entity. Square brackets are used to delimit the inner query statement: For example, the statement `5 <> [rname=HEM and ele=C] and rname!=HEM` selects all non-HEM atoms within 5 Å of carbon HEM atoms.

Since selection statements both can be applied to *EntityHandles* and *EntityViews*, complex selections can be carried out by chaining the selection statements. For example, chaining is particularly useful in combination with within-statements and whole-residue matching, as it allows to select atoms within a certain radius of a ligand, extend the selection to residues and then select the sidechains thereof.

Implementation of the Queries

The conversion of the query string into an internal representation consists of 3 stages.

First, the selection string is converted into a stream of tokens, that is a stream of identifiers, operators and values. Second, this stream is consumed by a recursive decent parser and transformed into an abstract syntax tree (AST). Each node of the tree denotes a construct occurring in query string. The syntax is *abstract* in the sense that it does not represent every detail that appears in the real syntax. For instance, grouping parentheses and operator precedence are implicit in the tree structure. Additionally, convenience shortcuts such as range selection are encoded as composite sub-trees. An example is the numeric range selection $rnum=1:100$ that is stored as $rnum>=1$ and $rnum<=100$ in the tree. The leaf nodes of the syntax tree are formed by predicates, e.g. $rname=ASN$, the internal nodes are boolean operators. Their left and right child nodes are either predicates or other boolean operators. In **figure A.1** are two examples that show the AST for two typical queries.

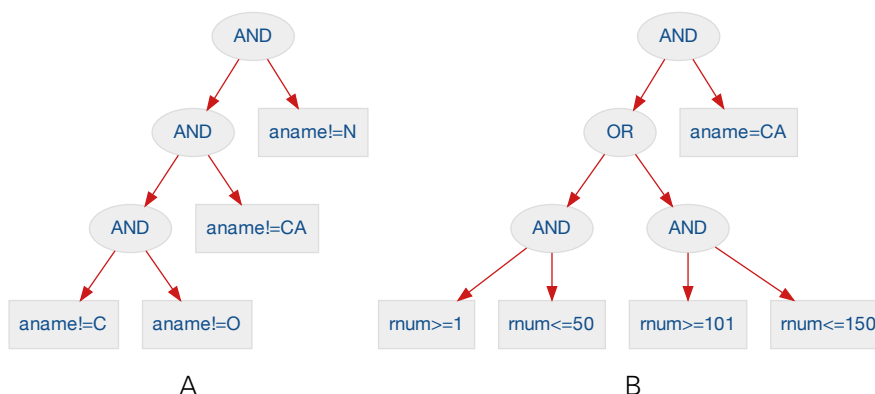


Figure A.1 Abstract syntax trees (ASTs) for two selection statements: **(A)** 'aname!=N,CA,C,O', **(B)** 'aname=CA and rnum=1:50,101:150'

Some of the predicates operate on the level of the residue, some on the level of the chain, some at the level of the atom. This means that each predicate assumes one of 3 values, true, false, maybe. The *maybe* state introduces opportunities for optimized query execution. When several predicates are combined, the complete query can evaluate to true or false, even though some of the predicates are in maybe state. For example, atom predicates are in a state *maybe* when evaluating the chain or residue predicates, but assume states *true* or *false* when considering a certain atom.

The AST is split into 3 evaluation contexts, one for each level of the chain, residue, atom hierarchy. The contexts consists of a set of instructions in reverse polish notation (RPN) that are executed on the tri-states of the predicates. The value of the predicates are shared between the 3 contexts. Starting from the chain, the chain, residue atom tree (CRA-tree) is traversed in breadth-first order. For many queries, the context does not need to be evaluated down to the level of the atom, since the result is already *defined* at the residue or chain level. For example, when evaluating the query statement $rname=GLY$ and $aname=CA$, the tree traversal can stop at the level of residues, except for GLY residues: the expression *false AND maybe* can never be true, whereas *true AND maybe* evaluates to maybe.

When at any point of the query execution, the expression evaluates to true, all sub

elements of CRA-tree are included; when the expression evaluates to false, execution continues with the next sibling; when the outcome is maybe, execution has to decent to the child nodes.

2 Crystallographic Density Maps in OpenStructure

Crystal lattice

A crystal is an object of apparent infinite extent that is created by replicating a *unit cell* along the major axes of the crystal. In the simplest case, the unit cell consists of a single point; the resulting crystal is a point lattice. Since the crystal is infinite, the points in the lattice can not be distinguished by their distance to the border. These points are absolutely identical and any of these points is able to reproduce the whole lattice. In the more general case, unit cells are not infinitely small points but have a physical extent. Points within a unit cell are usually represented as fractional coordinates f . Any point p in the unit cell can be written as

$$p = f_x \cdot \vec{a} + f_y \cdot \vec{b} + f_z \cdot \vec{c}$$

\vec{a} , \vec{b} and \vec{c} are the basis vectors of the unit cell. For a point within a unit cell, u , v and w are limited to the half-closed interval $[0, 1)$. For any point p in the crystal, the fractional part of f , represents the location in the unit cell, whereas the integral part describes the unit cell repeat the point falls into. The above equation can also conveniently be described by a matrix vector product:

$$p = \begin{pmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{pmatrix} \cdot \begin{pmatrix} f_x \\ f_y \\ f_z \end{pmatrix} = U \cdot f$$

Fractional coordinates have the advantage to be independent of the unit cell parameters and can be used to represents points for any unit cell. The inverse of the above equation generates fractional coordinates from a point:

$$f = U^{-1} \cdot p$$

The size and shape of the unit cell are completely described by the length of 3 base vectors \vec{a} , \vec{b} and \vec{c} and the angles α , β , γ between them. The matrix U can be constructed from these parameters. Assuming \vec{a} is along the x-axis and \vec{b} lies in the XY-plane, the matrix can be calculated as:

$$U = \begin{pmatrix} 1 & \cos \gamma & \cos \beta \\ 0 & \sin \gamma & -\sin \beta \cdot s \\ 0 & 0 & \sin \beta \cdot \sqrt{1 - s^2} \end{pmatrix} \cdot \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$

with

$$s = \frac{\cos \gamma \cos \beta - \cos \alpha}{\sin \beta \sin \gamma}$$

Space Groups and Symmetry Operators

The crystal is formed by repeating the unit cells along all 3 major axes of the unit cell. Using the fractional coordinate notation from before, a point with fractional coordinates f_0 , will be mapped to points $f_n = f_0 + (l, m, n)$ with $l, m, n \in N$. These points are equivalent, since any of these points is able to reproduce the infinite lattice. The points within one unit cell can have symmetry relations themselves. These relations are defined by the space group of the crystal. The smallest set of points of the crystal that is not symmetry related is called the asymmetric unit of the crystal. A unit cell can be reconstructed by applying all symmetry operators of the space group to the points in the asymmetric unit. The simplest of the space groups is P 1 that has only one symmetry operator, the identity operator. In the following, we introduce a notation for symmetry operators that describes the mapping to be applied to each component of a fractional coordinate. The mappings for the components are separated by comma. Let's first start with the identity operator:

$$X, Y, Z$$

The above operator maps all points onto themselves. In matrix notation, this can be written as the 3x3 identity matrix:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \vec{f}$$

Multiplying this matrix with a column-vector yields the transformed coordinates. A rotation around Y by 90 degrees, can be written as

$$Z, Y, -X$$

It can be seen that the point (1, 0, 0) gets mapped onto (0, 0, 1), whereas a point on the y axis gets mapped onto itself. Again, in matrix notation

$$\begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \cdot \vec{f}$$

Symmetry operators can also introduce shifts:

$$\frac{1}{2} + X, \frac{1}{2} + Y, Z$$

Here the shifts are specified as fractions of the unit cell. This operator can not be described by a 3x3 matrix alone, but requires a matrix plus a translation vector that is applied after the matrix transformation:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \vec{f} + \begin{pmatrix} 1/2 \\ 1/2 \\ 0 \end{pmatrix}$$

Since the symmetry operators of the space group form a closed set, for every symmetry operator there is a second symmetry operator that reverses its effect (the inverse). As an example, for the symmetry operator $\frac{1}{2} + X, \frac{1}{2} + Y, Z$, it is $\frac{1}{2} + X, \frac{1}{2} + Y, Z$ that produces the same set of points.

Implementation of Crystallographic Density Maps in OpenStructure

OpenStructure supports X-ray density maps through the XtalMap class, available in the non-orthogonal-maps branch. Internally the class only stores a little more than an asymmetric unit and uses the symmetry operators of the space group and cell repeats to reconstruct any region of the crystal. This is called symmetry extension. While the theory behind symmetry extension of a density map is trivial, implementing in a fast way requires intelligent bookkeeping. The idea used for the XtalMap class is largely based on Kevin Cowtan's clipper library.

An XtalMap is either created manually or loaded from a CCP4 or MRC file. These files contain the voxel values for one asymmetric unit. To reconstruct the whole unit cell of a P 1 21 1 space group with symmetry operators

$$X, Y, Z \text{ and } -X, Y, -Z$$

only half the pixel of the unit cell size are required in both X and Z, whereas all voxels along Y are required. For simple space groups, the asymmetric unit will fill the whole 3D voxel array. More complex space groups will not fill the whole cuboid. Maps however, are required to store a cuboid of data and thus there will be redundancy in the data.

For simplicity let's call the ASU generated by the f th symmetry operator the f th ASU.

Suppose we want to find the voxel corresponding to a given point in 3D space. We first convert the Cartesian coordinate to a fractional coordinate by

$$f = U^{-1} * (p - o)$$

where U^{-1} is the inverse of the 3x3 matrix whose columns are formed by the basis vectors of the unit cell and o is the origin of the crystal. We then make sure that the components of f fall into first unit cell by removing the integral part of f and only keeping the fractional part. Multiplying by the number of voxels along each axes of the unit cell then gives the integer pixels into the 3D data array of the whole unit cell.

We now have to figure out the symmetry operator that transform the point f to the symmetry equivalent in the 0th ASU.

This is essentially the inverse operator that brings us from the zero-th ASU to the ASU containing f . Since the the symmetry operators of a space group are a closed set, the inverse is already part of the symmetry operators in the space group and we *only* have

to loop over all the symmetry operators in the list of symmetry operators to find the right one. To quickly find out whether one of the pixels is part of the 0th asymmetric unit, we keep a byte array for book-keeping. This array has the same size as the ASU map. For each voxel, we store the symmetry operator number that transforms the voxel coordinate to the zeroth ASU. For the 0th ASU itself, this is the identity operator that always has number 0. We apply the symmetry operators of the space group one after the other until we end up at a voxel that contains a zero in the bookkeeping array. What did we win? Random access still scales linearly with the number of symmetry operators. However, random access to pixel values is rarely used, usually one is interested in a continuous region of the crystal.

SYMMETRY-EXTENDING A CONTINUOUS REGION OF THE CRYSTAL | For the first point of the region, we have to perform the $O(n)$ lookup as described above. This will give us both the symmetry number as well as the index into the data array. We now would like to move to the next voxel in X direction in global coordinates. Using again the example of P 1 21 1 space group, if we are currently in the 0th ASU, this translates to a movement along X in the data array, and in the 1th ASU a movement along -X in the data array. This movement is expressed as an offset and the new *index* is calculated by adding the offset to the current index. If the bookkeeping array contains a zero at the new position, we are still in the same ASU and we are done, if not, we are changing from one ASU to the other. The new symmetry operator is then the product of the current symmetry operator and the symmetry operator stored in the bookkeeping array.

