Modeling of tertiary and quaternary protein structures by homology

Inauguraldissertation

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

von
Florian Kiefer
aus
Freiburg im Breisgau, Deutschland

Basel, 2012

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Torsten Schwede

Prof. Dr. Manuel Peitsch

Basel, den 13.12.2011

Prof. Martin Spiess

Dekan

List of Publications

1. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. Proteins 2007;69 Suppl 8:129-136.

2. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 Suppl 8:38-56.

3. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 2009;4(1):1-13.

4. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 2009;37(Database issue):D387-392.

5. Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T. The Protein Model Portal. J Struct Funct Genomics 2009;10(1):1-8.

6. Berman HM, Westbrook JD, Gabanyi MJ, Tao W, Shah R, Kouranov A, Schwede T, Arnold K, Kiefer F, Bordoli L, Kopp J, Podvinec M, Adams PD, Carter LG, Minor W, Nair R, La Baer J. The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res 2009;37(Database issue):D365-368.

7. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins: Structure, Function, and Bioinformatics 2011;79(S10):37-58.

Abstract

The structure of a protein is crucial to understand its function. Despite this importance, experimentally solved structures are only available for a small portion of the currently known protein sequences. Comparative or homology modeling is currently the most powerful method used in order to predict the structure from sequence by the use of homologous template structures. Models, hence, need to be accurate regarding their three-dimensional coordinates and must represent the biological active state of the target protein in order to be useful for scientists.

Four goals are pursued in this work in this area of research. Firstly, we increase the coverage of homology modeling by introducing a method which is able to identify and align evolutionary distant template structures. The resulting template search and selection procedure is hierarchical. Closely related template structures are identified accurately and efficiently by standard tools.

A computationally more complex method is invoked in order to identify evolutionary more distant template structures with high precision and accuracy. Integrated into an automated modeling pipeline, the developed method is competitive compared to other protein structure prediction methods.

Secondly, the automated modeling pipeline is applied to a large set of protein sequences to increase the structural coverage of sequence space. The resulting models and associated annotation data are stored in a relational database and can be accessed online in order to allow scientists to query for their protein of interest. Efforts are made to update a selected set of sequences regularly by shortening the update process without losing accuracy. It is found that the structural coverage of seven proteomes is increased considerably by this large scale modeling approach.

Thirdly, the modeling of quaternary structure is addressed. Significant room for improvement in the field of quaternary structure prediction is found when assessing the current state-of-the-art methods in a double blind prediction experiment. Novel similarity measures are therefore developed to distinguish proteins with different quaternary structure. We further create a template library built of structures in their previously defined most likely oligomeric state, to extent the concept of homology modeling towards the prediction of oligomeric protein structures. In order to select template structures which share the same quaternary structure with the target structure, a variety of evolutionary and structural features are investigated. It is shown, that using a combination of these features for the first time predicts the quaternary structure with high accuracy.

Finally, the performances of methods which predict non-folded (intrinsically disordered) protein segments are assessed. Current issues are addressed in a field of very active research as more and more proteins are found to be hubs in interaction networks with considerable disordered portions in their tertiary structure. In general it is found that such methods perform well, even within the limits of the test set.

Contents

1 Introduction

1.1 The importance of protein structures

Proteins are essential components of the cell and are involved into metabolism, signaling cascades, nutrient transports and provide structural stability (e.g. by forming large filaments). They are crucial for function and maintaining the complex cellular machinery and they also ensure the survival and replication of the cell. Understanding the function role of proteins is important to study the mechanism of diseases, thus, understanding the functional aspects of proteins are of great interest for the scientific community.

The function of a protein is ultimately defined by its three-dimensional structure. For example, enzymes achieve their function often by binding substrates. Thereby, the specificity of the substrate is accomplished by the spatial arrangement of amino acids at the protein surface. The shape and the composition of the surface play also an important role for the interaction with other proteins. Overall, a tight relation between function and structure can be observed. As a consequence, the insight into the spatial arrangement of amino acids is an important prerequisite to determine the functional mechanisms of proteins.

Anfinsen pioneering work has shown that the three-dimensional structure is a direct consequence of the amino acid sequence.[8] Based on its observation that proteins refold into the same structure again after removal of a denaturant, he suggested that the native structure of a protein must be the thermodynamically most favorable. However, Levinthal made the argument that the numbers of conformations which are required to find the energetically most favorable structure are far too high to be sampled randomly[9] (known as ”Levinthals paradox”[10]). As a consequence, he suggested that folding happens along “pathways”, which restrict the number of “visited” structures considerably. In a modern view, the pathways can be interpreted as folding funnel whereas parallel routes exist to the bottom of the funnel, the global minimum of free energy of the structure.[11] It is generally assumed that protein folding starts with the forming of local secondary structure elements. During the folding process they are packed closely together to build the tertiary structure of the protein. The spatial arrangement of secondary structure elements are defined as “folds” and it has been thought that only a limited number of such protein folds exist.[12] Driving forces of folding are the formation of stabilizing interactions (i.e. hydrogen bonds, salt bridges and disulfide bonds) and the hydrophobic effect.

Protein sequences can be altered by evolutionary events, such as mutations or deletions and insertions caused by changes of the underlying DNA sequence. In order to maintain the function of a protein, the protein structure must be, to a certain extent, robust against changes in its sequence. Indeed, it has been shown, that the structure is more conserved than the sequence between proteins which share a common ancestor.[13]

1.2 Experimental methods to determine protein structures

In order to investigate the functional mechanisms of proteins, such as the binding of ligands in an active site, protein structures need to be solved with atomic resolution. X-Ray crystallography and Nuclear Magnetic Resonance (NMR) are the two most widely used techniques to solve macromolecular structures experimentally.

X-Ray crystallography

Solving of a protein structure using X-Ray crystallography involves several consecutive steps. The most difficult procedure is the growth of an adequate crystal. Crystallization success is dependent on many factors and requires a high level of expertise. In the final crystal the proteins are arranged in a symmetrical order. Once a sufficiently large crystal has been obtained, the crystal is placed in an intense X-ray beam of a single wavelength. The beam is dispersed by the electrons in the protein and interfering X-Ray waves can be recorded on a screen behind the crystal. The intensity of the reflections is related to the amplitude of the dispersed beam and can be used, in combination with the phases, to calculate the electron density map of the protein. The phases cannot be determined by the experiment itself and have to be estimated using techniques like isomorphous or molecular replacement. Once the electron density map has been built, the protein structure is fitted using standard geometries for bond length and angle. The resolution of a structure (in Angstrom) denotes the distance at which two points can be distinguished in the electron density map. Another qualitative descriptor is the R-Factor, which is calculated by comparing a recomputed diffraction map (derived from the fitted protein structures) with the observed diffraction map in the experiment. An R-Factor of less than the resolution divided by ten characterizes a reliable protein structure.

NMR

Nuclear magnetic resonance (NMR) is used to determine the structure of macromolecules in solution. NMR is a spectroscopic technique and bases on the change of the magnetic spin of nuclei if the protein is irradiated by a short pulse of radiation. The resonance of the nuclei caused by the pulse depends on the direct atomic environment. Based on this effect, couplings between pairs of structurally closed atoms to generate constrains in form of distances and angles. To derive the correct coordinates of a protein structure, these constrains need to be satisfied, however, if not enough constrains were observed or contradict each other, the result is an ensemble of structures rather than one finite solution. Nevertheless, one of the strengths of NMR is that biological relevant changes in the structural conformation can be observed, caused for example by ligand binding. Hence, NMR can be used to examine the dynamics of a protein in solution.

1.3 Resources for protein structures

Experimentally solved structures are deposited in Protein Data Bank (“PDB”)[14]. The PDB was established in the early seventies to make the small but growing number of solved protein structures available to the scientific community. The structure of a protein is deposited with its spatial atomic x,y,z coordinates in a text file. Details about the performed experiment are specified in additional sections of the file. Entries in the PDB database can be identified by a four letter code.

In the last decades, the PDB has become a central place for the deposition of macromolecular structures. This includes also nucleic acids such as DNA and RNA structures. The current release of the PDB (November 2011) consists of 77’000 structures for around 43’000 proteins and the number of structures has been exponentially grown in the last years. The vast majority of structures have been solved by X-Ray techniques followed by NMR.

As of today, the Protein Data Bank is the central place to start with in case of investigating macromolecular structures. Many journals require a deposition in the PDB if structural aspects are discussed on an unpublished protein structure.

Many other databases are derived from the PDB such as CATH[15] and SCOP[16], which classify protein structures in families based on their structural similarity.

1.4 The sequence – structure gap

The discussed techniques for the determination of protein structures are time consuming and not always applicable. Thus, only for a small set of currently known protein sequences structural information can be provided. As can be seen in Figure 1, the number of structures deposited in the PDB is greatly exceeded by the number of curated UniProt protein sequences ("Swiss-Prot")[17] and even more if considering all protein sequences directly derived from known DNA sequences ("TrEMBL")[18]. As of today (November 2011), the UniProtKB (Swiss-Prot + TrEMBL) database[19], consists of 18.7 million protein sequences compared to only 77 000 protein structures deposited in the PDB, leading to an enormous difference between known protein sequences and experimentally determined protein structures.

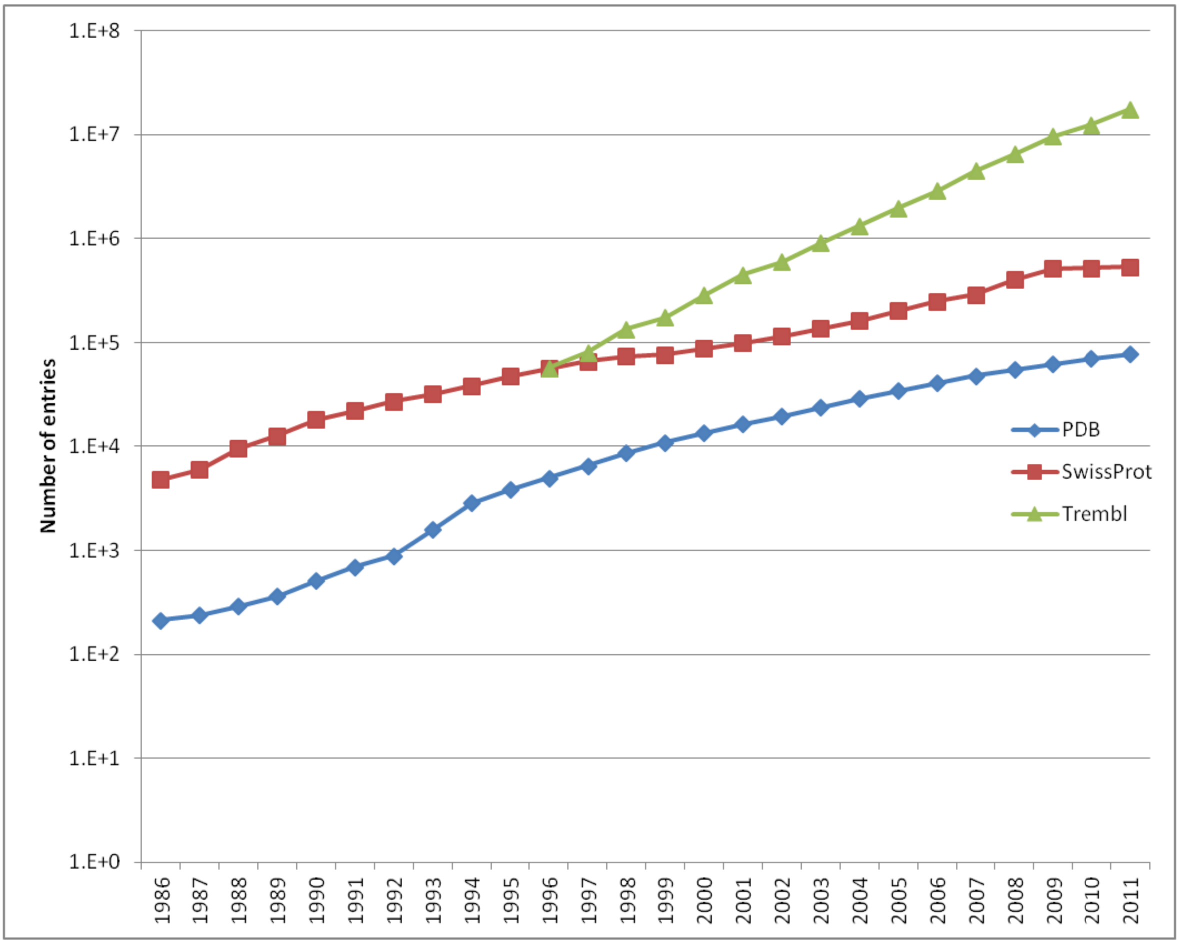


Figure 1 Comparison between sequence and structure databases growth. Swiss-Prot[17] is a set of curated proteins, TrEMBL[18] is the translated DNA-database EMBL. The number of protein sequences is of several magnitudes higher than the number of known protein structures (PDB[20]).

1.5 Modeling of protein structures

As discussed in the previous chapter, the vast majority of proteins do not have experimental information about their structure. It is expected that the number of protein sequences is further growing in an exponential scale, caused by the availability of high-throughput sequence methods, which allow fast and cheap sequencing of complete genomes. The widening gap between protein sequences and structures has caused, over the last decades, the development of a variety of computational approaches in order to "calculate" a protein structures from sequence.

"De novo" or "*ab initio"* methods are based on physical principles and try to imitate the folding process. Such methods have to sample a large number of conformations and require very accurate energy functions to identify structures in the global minima of free energy. To decrease the number of conformations, which needs to be visited, some methods use information of known structures to guide the sampling process.[21-23] Despite the improved strategies for sampling, it remains difficult to distinguish if a protein is in its native state or trapped in a local minimum. Such methods are computationally very demanding and can only be used for small systems.

Homology ("comparative") modeling techniques base on homologue (i.e. share a common ancestor) proteins which serve as structural "templates". It has been demonstrated by Chothia and Lesk in the mid-eighties that the evolutionary distance is directly linked to structural deviation between two proteins.[13] The evolutionary distance between two proteins can be estimated by the number of identical residues after aligning their sequences in an optimal way. Comparative modeling relies on the availability and identification of suitable template structures. By the continuous growth of structures in PDB and resulting increasing availability of template structures, comparative modeling becomes more and more attractive.

However, structural similarity does not require necessarily sequence similarity. It has been observed that the environment around a residue is more conserved than the residue itself. Hence, contact preferences can be derived for a particular type of amino acid.[24,25] "Threading" uses this type of information in order to calculate the fitness of a target sequence in a given environment. Either the environment of the original template structure is used ("frozen approximation") or the environment is replaced by target sequence during the threading process ("defrosted approximation"). To align the sequences optimally, dynamic programming is used; based on an energy function which scores how well the sequence fits in its environment.

Threading is mainly used to identify the correct fold if no homologue template structures can be identified for a given sequence.

Many methods use a combination of the three described principles, for example, by sampling fragments of homologue template structures to explore new conformations[21-23]. However, if homologue template structures can be identified, comparative modeling will be the first choice.

The range of biological questions, which can be answered by protein models, is wide and depends on the quality of the model. The quality is mainly evaluated based on the accuracy of the spatial coordinates compared to real structures. However, the accuracy of a model needs to be estimated, because in a real life scenario the native structure is not known. Often, the accuracy of models is roughly estimated by the fraction of identical residues between the target and template sequence. Additionally, several approaches have been developed in recent years to estimate the quality of a predicted model (for details please see paragraph 2.1.3 below).

Typical applications of protein models are shown in Figure 2. If the sequence identity between template and target is sufficiently high (>50%), models can be used to investigate catalytic mechanism or to design and improve ligands (Figure 2A). Models above this threshold do show only little deviation to the crystallized structure, often caused by incorrect sidechains, small distortions in the arrangement of secondary structure elements, and misplaced loops on the surface of the protein. Models in the medium accuracy region (30%-50%) are useful for example for molecular replacement in order to obtain the phases for the experimental determination of the target structure using X-Ray Crystallography or for site directed mutagenesis (Figure 2B,C). In such models, the overall structural error increases in form of distortion of the core, loop modeling errors and sporadic alignment errors.[26] Models in the low accuracy region often do not exceed more than 30% sequence identity. Errors in such models are often caused by alignment errors. However, even such low accuracy models can be useful in order to investigate the fold of the protein and derive principle functional relationships (Figure 2D,E). It has been also shown that low accuracy models can be used in combination with data from electron microscopy or other experimental data to model large macromolecular complexes.[27] By the combination with experimental data, protein structure prediction widened its range of application considerably.

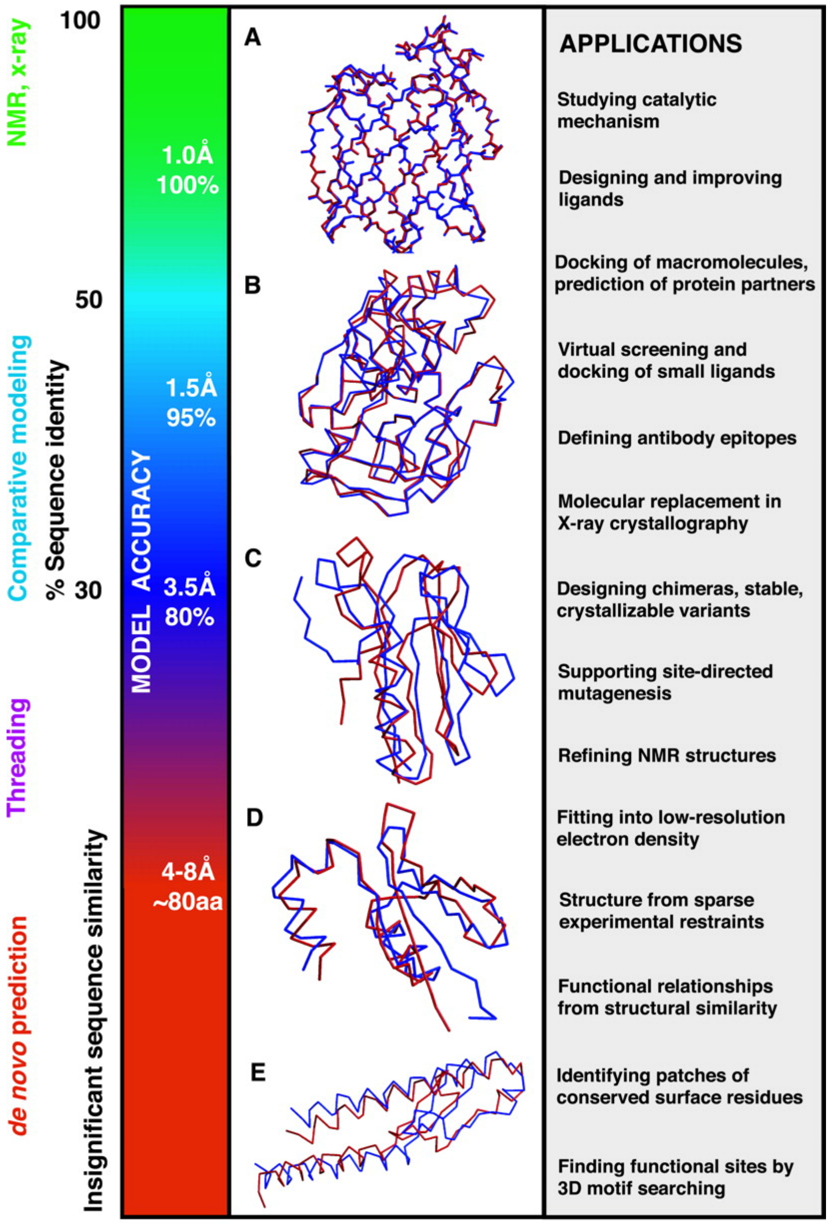

NMR, x-ray
Comparative modeling
Threading
de novo prediction
100
50
30
% Sequence identity
Insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4-8Å
~80aa
A
B
C
D
E
APPLICATIONS
Studying catalytic mechanism
Designing and improving ligands
Docking of macromolecules, prediction of protein partners
Virtual screening and docking of small ligands
Defining antibody epitopes
Molecular replacement in X-ray crystallography
Designing chimeras, stable, crystallizable variants
Supporting site-directed mutagenesis
Refining NMR structures
Fitting into low-resolution electron density
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching

Figure 2 Typical applications of protein models depending on their evolutionary distance to the target (a-e). Figure taken from Baker[28].

1.6 Assessing the accuracy of protein modeling procedures

The accuracy of models is usually noted as the structural deviation from the true protein structure, which is determined by experiment. Two metrics are well established in order to reflect the structural deviance between two protein structures: the Root Mean Squared Deviation ("RMSD") and the Global Distance Test (GDT-TS)[29]:

RMSD reflects the structural divergence between two structures (*a*, *b*) on a common set of residues with *n* atoms and can be defined as:

$$rmsd(a,b) = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(a_{ix}-b_{ix})^2+(a_{iy}-b_{iy})^2+(a_{iz}-b_{iz})^2}$$

The RMSD is expressed in angstrom and calculated after superposition of the two structures. However, the RMSD similarity measure is not optimal for the comparison between a model and its native structure. One reason is the disproportionate impact of large structural deviation even if they occur for example at the termini in one of the structure. In contrast, the GDT-TS score calculates the fraction of residues which can be superposed under a certain threshold and thus reflects more the agreement of the model to the reference structure:

$$GDT-TS = \frac{(GDT-1+GDT-2+GDT-4+GDT-8)}{4}$$

GDT-1, GDT-2, GDT-4, GDT-8 reflect the fraction of model Cα-atoms which are less distant then 1, 2, 4 and 8 Å after optimal superposition with the native structure. However, the selection of cutoffs is somewhat ambiguous. In order to increase the sensitivity of the global distance test, cutoffs of 0.5, 1, 2 and 4 Å have been proposed (GDT-HA).[2]

To estimate the performance of a protein structure modeling method an appropriate benchmark is required to assess the accuracy of the models compared to their reference structure. However, the available benchmark sets can vary in size and difficulty; therefore a comparison between modeling methods become difficult for the user of homology modeling services. To overcome this problem, the accuracy of protein structure prediction methods is evaluated biannually in the CASP ("Critical Assessment of techniques for protein structure prediction") experiment[30]. The CASP installments are organized as a double blind experiments were predictors do not have access to the native structure throughout the modeling period. The native structures are kept on hold and get released to the PDB if the prediction season has finished. On the other hand the assessors of the predictions do not know the real names of the predictors and are not biased by knowing details about the applied methods. This ensures a fair evaluation of the predictions which are based on a predefined set of target structures.

CASP is organized in different modeling categories. The main categories cover the modeling of the three dimensional structure of proteins and consists since 2006 of the following sub-groups:

1. TBM – ‘Template based modeling’

   Homologue template structures can be identified in order to model the structure of the target.
2. FM – ‘Template Free modeling’

   No suitable template structures can be identified for this set of targets.

The prediction of the native protein structures also includes the correct prediction of the quaternary structure and a reliable estimation of the modeling error.

There exist two types of predictors:

- “Server” predictors are asked to model the protein structures in a fully automated fashion and without manual intervention. “Server” groups receive the sequences of the target proteins via email and have to respond within 2 days.
- “Human” predictors can choose the most suitable strategy according to the expertise and knowledge, for example by extracting relevant information from literature. “Human” predictors have a prediction time slot of 4 weeks. Additionally, they can use models submitted by the “server” groups, to either verify their own predictions or use them as input.

The CASP experiment is not limited to the prediction of three dimensional coordinates. The following categories in the context of protein structure prediction are additionally evaluated:

- Prediction of disordered segments
- Prediction of residue-residue contacts
- Prediction of functional binding sites
- The assessment of models regarding their reliability (This includes also the estimation of the modeling errors in a residue wise fashion)

If the prediction slot for a particular target has closed, the native structure is accessible for the assessors. Predictions are assessed according to their accuracy by applying established assessment criteria.

The goal of CASP is twofold. Firstly, the most successful methods are identified and ranked according to different criteria like the overall structural accuracy or the ability of modeling correct side chains. Secondly, the assessment can highlight strengths and weaknesses of the methods, thereby suggesting further areas for future improvements.

All methods are assessed on the same set of targets (according to the category, they attended) using appropriate scores which are selected by the assessors. This guarantees a fair comparison of the results.

Users of protein structure prediction methods can have questions with different biological backgrounds. Modeling a binding site so that the model can be used for docking studies requires a different focus, than modeling of proteins which have only few or no homologue template structures. Thus, a detailed assessment of modeling methods within CASP can help to identify methods which fit best to a specific biological question.

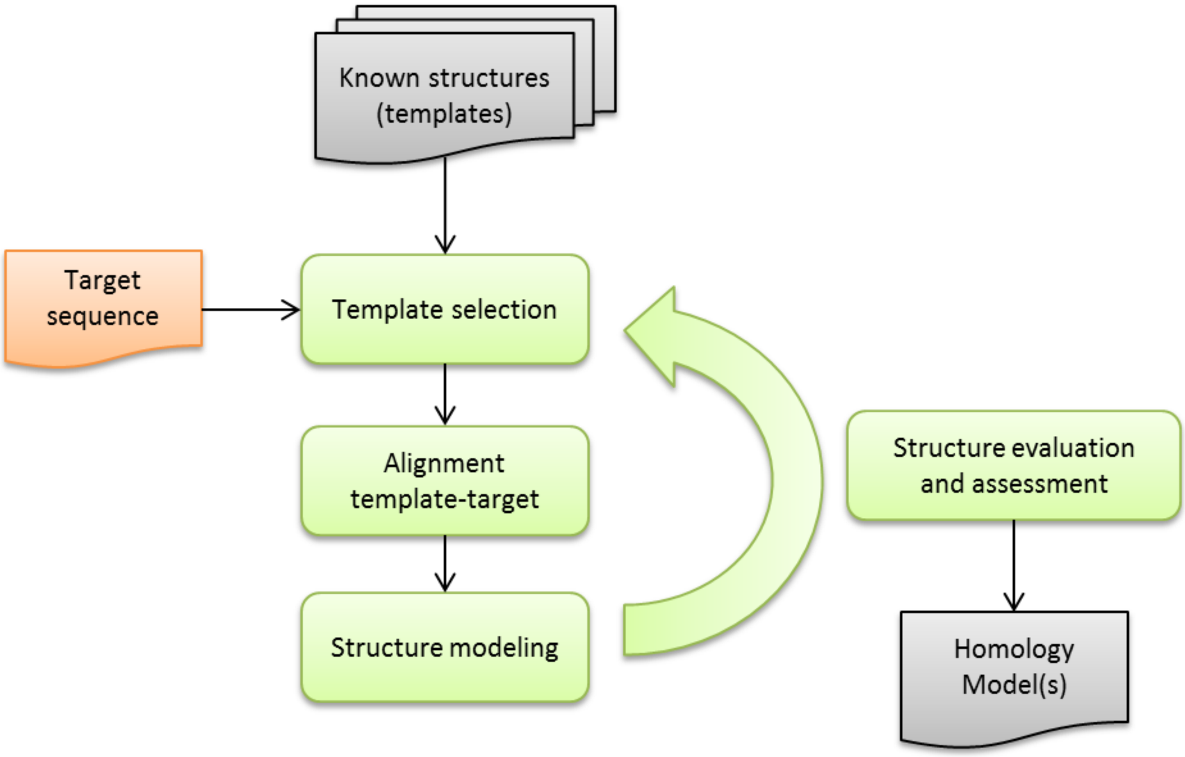
The results of the experiment are discussed during a meeting which is held after the prediction season. Assessors as well as the most successful predictors present and discuss their work and pinpoint the achievements and failures of the applied modeling techniques.

2 Modeling of tertiary protein structures

2.1 The homology modeling approach

Homology or comparative modeling is currently known as the most accurate method to generate protein models.[30] As can be seen in Figure 3, the modeling procedure can be divided into four consecutive steps:

1. The identification and selection of homologue structures ("templates")
2. The alignment between the template and the target sequence
3. The calculation of the model based on alignment information and the identified template structures including the prediction of the regions without alignment information (i.e. loops) and the refinement of the protein model
4. Estimation of the accuracy of the resulting model(s).



Known structures (templates)
Target sequence
Template selection
Alignment template-target
Structure modeling
Structure evaluation and assessment
Homology Model(s)

Figure 3 The four main steps in homology modeling (green boxes).

2.1.1 Template identification and alignment with the target sequence

The first two steps in creating a protein model are the identification of suitable template structures and the generation of an alignment between template and target sequences.

Identification of homologue sequences involves querying a database with experimentally determined structures (“template library”).

For the identification of closely related template structures, sequence-sequence comparison methods like FASTA[31] and BLAST[32] were developed. BLAST stands for “Basic Local Alignment Tool” and became a standard method for the identification and alignment of protein and nucleotide sequences. However, more sophisticated methods are needed for the identification of lower levels of evolutionary relatedness (<40% sequence identity). Based on the assumption that structural and functional important residues are conserved in the family of the protein, position specific scoring matrices (PSSMs) were developed. A PSSM (“Profile”) consist of the probabilities that a particular residue type appear in the column of a multiple sequence alignment which consists of homologue protein sequences. A widely known method using PSSMs is PSI-BLAST (Position Specific Iterative –BLAST)[32]. A typical procedure for template identification with PSI-BLAST is the (iterative) construction of a profile for the target sequence using evolutionary related protein sequences followed by “profile to sequence” search of protein sequences contained in a template library.

Further, profile-profile based methods were developed in order to increase the sensitivity of sequence based fold recognition.[33] Thus, profiles are generated for the target and the template sequences in order identify homologue template structures. More recently, profiles have been replaced by Hidden Markov Models and Generalized Profiles[34-36], which also allow position dependent gap penalties. In addition, structural information may be incorporated into the profile to increase alignment accuracy.[37-39] By using this approach, HHsearch[40] was the first method which allowed the alignment of two HMMs. HMM-profiles are built using culled versions of NCBIs nr database ; redundant sequences are thereby excluded to guarantee high quality of the underlying multiple sequence alignment.

2.1.2 Model building

Protein models are generated based on the structural information given by the template structures and the alignment between templates and target sequences. When applying the rigid body assembly approach, conserved structural parts of the template structure are copied to the model.[41] The generated model is then subjected to refinement methods to account for violations in the stereochemistry and the geometry of the model. Another approach relies on the derivation of spatial constraints (e.g. distances and angles) from one ore multiple template structures.[42] To calculate a model, the violations of the spatial constraints must be minimized. Therefore optimization strategies like the conjugate gradient method[43] are applied.

A variety of refinement strategies were developed in order to optimize regions without sufficient alignment information. This includes loop modeling procedures (see review of Fiser[44]) and the correct placement of sidechains by using rotamer libraries[45].

In summary, the precision of rigid body assembly and restrained based modeling is comparable. Other factors like the identification of suitable template structures and the correct alignment of the target and template sequence play a more important role for the final model accuracy.

2.1.3 Structural evaluation and assessment

It is crucial for the usability of a model to estimate its accuracy. Error in the stereochemistry of models can be detected using tools such as PROCHECK[46] or WHATCHECK[47]. A second type of scoring functions which try to identify structural errors in models are physics based energy functions like VERIFY3D[48] and GROMOS[49] or knowledge based potentials like ANOLEA[50] or QMEAN[51]. The latter are often used to identify the most accurate model amongst a set of alternative models based on either different template structures or provided by different modeling routines. However, the accuracy of a prediction can also vary within a model. Regions which are functional important are known to be more conserved in evolution than for example residues between secondary structure elements which are exposed to the solvent. In such regions (“loops”) structural deviation can often be observed between a model and its native structure. Hence, it is important to identify such regions by assigning an error estimate on a per-residue level.

2.1.4 Automated modeling procedures

Overall, the success of comparative modeling relies on many factors including the availability of suitable homologous structures, the correct alignment between target and template sequences, and the functional divergence between the target protein and the template. Dependent on the given situation, different strategies needs to be applied and the results must be carefully evaluated. This requires a sufficient level of expertise in structural biology and the use of highly specialized programs which are often computationally intensive and thus require adequate hardware settings.

In order to make comparative modeling available for a larger community of biomedical researcher, automate modeling procedure were established. Today, there are a large number of such services accessible over the worldwide web. Biologist can choose an appropriate method according to their needs and expertise. In addition most of the server which participate at the CASP evaluation can be accessed online.[52]

Because an automated pipeline does not know a priori about the difficulty of modeling a particular protein, automated modeling requires “internal” expertise in order calculate accurate models. This includes the fast and accurate modeling of proteins with closely related homologous template structures as well as the identification and correct alignment of template structures with large evolutionary distance to the target protein. A final selection step needs to be applied in order to provide accurate and biological relevant models.

2.1.5 The SWISS-MODEL system

Over 15 years ago, The SWISS-MODEL server was developed in order to make comparative modeling available to a large community. Since then SWISS-MODEL has been constantly developed and is one of the widely used modeling server.[53-55]

The SWISS-MODEL workspace comprises a variety of computational tools which allows predicting structural models for the protein of interest and the analysis of their expected quality.[56] As of today 2000 requests are processed daily by the SWISS-MODEL workspace. According to their expertise, the users can choose between three different modeling approaches.

Automated mode

The automated approach was designed in order to provide an easy to use interface which requires only little user intervention. The user has to specify only the sequence of the target protein or as its Uniprot accession code in order to start the modeling process. Automated modeling involves identification and selection of suitable template structure, calculation of the model including the estimation of the expected quality.

Alignment mode

For more distantly related target and template proteins, multiple sequence alignments can help increasing the quality of the alignment between template and target protein sequences. The alignment mode provides an interface where user can upload a curated multiple sequence alignment of sequences of the template, target and closely related family members.

Project mode

Difficult modeling projects require a more detailed investigation of the alignment between target and template sequences. Visual inspection and manual modification of the alignment often increases the accuracy of the resulting model.[57] The “Project mode” allows the submission of a project file, which contains the template structure and the sequence alignment between target and template sequences. Project files can be generated, modified and displayed by the visualization software DeepView[54]. Project files are also part of the output of the "automated mode" and "alignment mode", thus, a generated model can be further refined and iteratively resubmitted to the "project mode".

2.2 Definition of the Problem

As stated in the introduction of this chapter, one of the most important steps to generate accurate models is the identification of suitable template structures and the correct alignment of their sequences to the target sequence. However, in many cases only templates with low sequence similarity can be identified. Such remote homologue template structures provide often useful information for the protein of interest but also require precise and accurate alignment tools.

In the original version of SWISS-MODEL, BLAST solely was used for template identification and alignment. It has been shown, that BLAST often creates errors in the sequence alignment below 40% sequence identity or is unable to detect remote homologue structures.

Thus, we developed a protocol to improve the sensitivity (i.e. the identification of remote homologue template structures) and the quality of the models of the automated SWISS-MODEL pipeline based on template structures with high as well as low sequence identity by introducing a profile-profile alignment approach.

2.3 Improvement of the SWISS-MODEL homology modeling pipeline

Comparative modeling relies on the identification of protein structures which are homologue to the target protein. Thus, it is essential to apply methods which are sensitive and accurate in respect to the identification of suitable template structures. This is even more important for procedures without any manual intervention, because such applications do not know a priori the difficulty to model a particular protein. Hence, an automated template search routine must be designed in order to find closely related template structures as well as evolutionary distant templates.

The identification of closely related template structure is straightforward, because the sequence alignment is unambiguous. As a rule of thumb, sequence alignments generated with automated procedures can be considered reliable if more than 40% of the residues are identical.[58] BLAST[32] is known as accurate and fast tool for the identification of such closely related template structures and is widely used. In addition, the BLAST package is still under development and updates are released at regular intervals. As a result we have chosen BLAST to identify closely related template structures. However, below 30% sequence identity alignment errors increase rapidly when using sequence-sequence alignment techniques.[28] To overcome this limitation, several methods were developed to increase the specificity of BLAST towards more distant related template structures (see paragraph 2.1.1.).

To identify successful modeling methods, the results of the biannual protein modeling benchmark experiment CASP can be used. Within the “Template based Modeling” (TBM) category it can be expected that top ranked methods are more successful in the detection and alignment of templates for a given target sequence than others.

Hence, we examined the results of the CASP7 sever assessment category[52] to identify accurate template search methods. “Server” groups are asked to process the submission fully automated and have to respond within 48 hours. These guidelines fit best to the needs of the SWISS-MODEL server pipeline, because they reflect real modeling situations where long waiting times are undesirable.

Figure 4 show the assessment of the “server” participants within the TBM-category of the CASP7 installment.[52] Three different evaluation scores were applied (HB, AL0, GDT-HA), where two of them focus more on the global accuracy of the submitted models (AL0, GDT-HA) and one focus on the accuracy of the hydrogen bond network within the model (HB). As shown in Figure 4 (see Battey et al[52] for details of the assessment), two groups (I-TASSER and HHpred) are considerably more accurate compared to the other participants. The top ranked group 25 (“I-TASSER”) is developed by the group of Zhang[23]. I-TASSER uses a threading procedure to identify possible template structures for the target sequence. However, this method is computational expensive (~10 h per query,[52]) and embedded as part of an iterative modeling procedure. The modeling routine of I-TASSER uses fragments of high scoring template structures for the assembly of the model and hence does not represent the classical single template modeling schema. The second top ranked method is HHpred[59], which is based on the identification and alignment of template structures using profiles based on Hidden Markov Models (“HHsearch”)[40], followed by a modeling protocol based on the comparative modeling software MODELLER[60]. The accuracy of the HHpred server relies on its sensitivity to identify evolutionary distant template structures and the correct calculation of the alignment between the template and target sequence. HHpred had an average response time of ~10 minutes in CASP7 and can be considered as a fast and reliable modeling server. The template search routine (“HHsearch”) is freely accessible and can be installed as stand-alone program. To run the program, firstly a HMM-profile for the target sequence needs to be built. Secondly, the target profile is queried against a template library of HMM-profiles. The quality of the alignment is further increased by a realign procedure which uses the Maximum Accuracy algorithm. Based on the result of our CASP7 assessment of server predictions and the performance and availability of the tool we decided to use HHsearch as template identification tool for more distantly related structures.

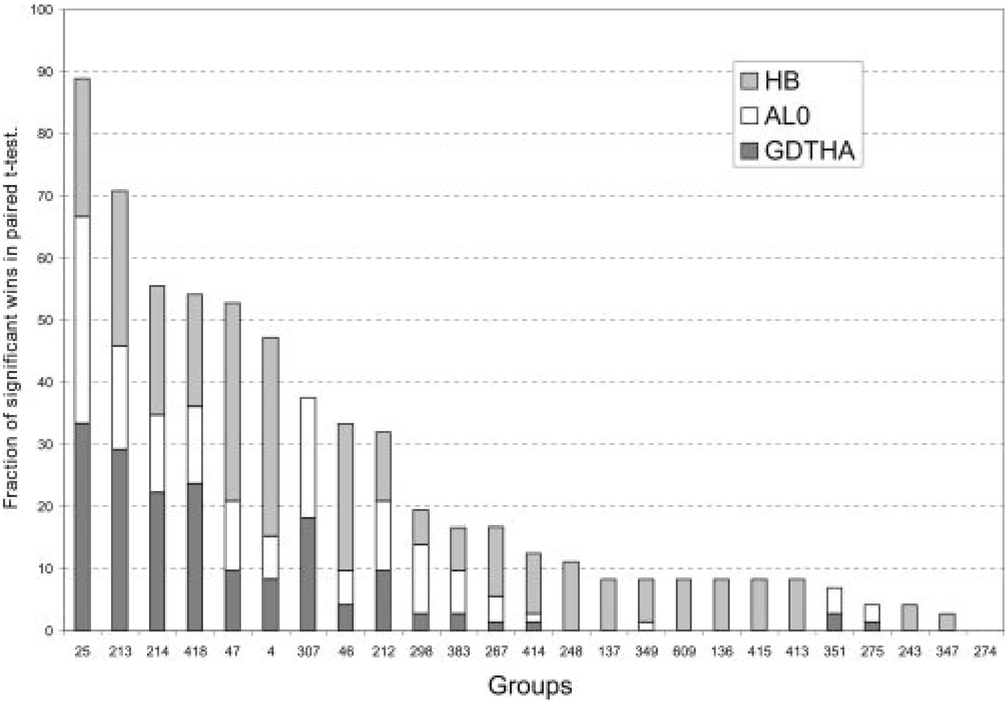


Figure 4 Performance of CASP7 server groups. Two groups 25 and 213 (“I-TASSER” and “HHpred”) outperform clearly the other methods in their number of significant wins on common predicted target structures. Figure taken from Battey et al.[52]

Hierarchical combination of template search methods

To combine the speediness of BLAST and its accuracy to detect closely related template structures with the ability of HHsearch to identify and align evolutionary distant template structures, we deployed a hierarchical template search protocol. (See Figure 5 for a schematic representation).

Firstly, BLAST is launched to search for closely related template structures within our template library. We use very conservative thresholds, to ensure high alignment accuracy. BLAST hits are only retained if more than 60% of the residues are conserved within the sequence alignment and the E-value does not exceed 0.0001.

Secondly, HHsearch is started if either (1) no suitable template structures were found by BLAST (2) if the target sequence was not fully covered by BLAST hits. For the latter criteria we used a threshold of 25 residues, which reflects roughly the size of a small domain. If an additional HHsearch run is required, a profile-HMM of the target sequence is built. This involves several rounds of PSI-BLAST against culled versions of the NCBI-nr database. The target HMM-profile is then queried against the templates HMM-profile library which is culled so that two sequences in template library do not share more than 70% sequence identity. Templates are retained according to the recommended cutoffs (P-value > 50) by the authors of the programs. Finally, the list of template structures is subjected to the template selection procedure described below (paragraph 2.3.1).

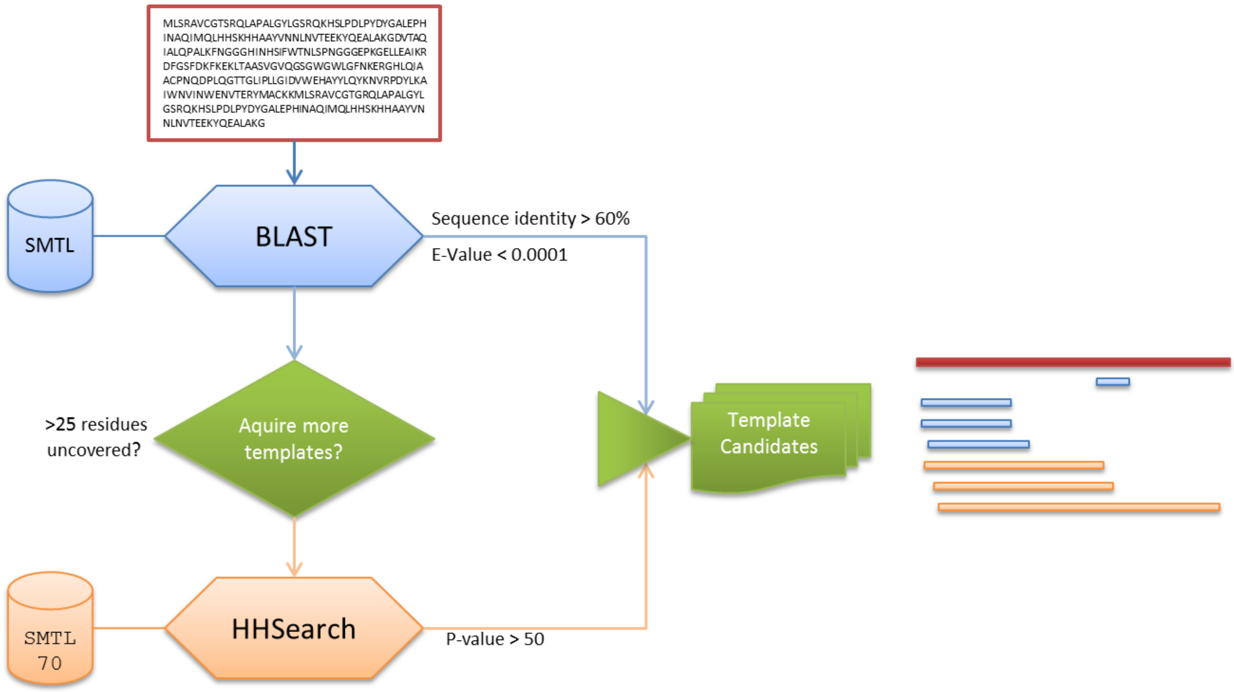

MLSRAVCGTSRQLAPALGYLGSRQKHSLPDLPYDYGALEPH
INAQIMQLHHSKHHAAYVNNLNVTEEKYQEALAKGDVTAQ
IALQPALKFNGGGHINHSIFWTNLSPNGGGEPKGELLEAIKR
DFGSFDKFKEKLTAASVGVQGSGWGWLGFNKERGHLQIA
ACPNQDPLQGTTGLIPLLGIDVWEHAYYLQYKNVRPDYLKA
IWNVINWENVTERYMACKKMLSRAVCGTGRQLAPALGYL
GSRQKHSLPDLPYDYGALEPHINAQIMQLHHSKHHAAYVN
NLNVTEEKYQEALAKG
SMTL
BLAST
Sequence identity > 60%
E-Value < 0.0001
>25 residues
uncovered?
Aquire more
templates?
Template
Candidates
SMTL
70
HHSearch
P-value > 50

Figure 5 Schematic workflow of the hierarchical template selection used in the workspace. Submitted target sequences are subjected to BLAST. If necessary an additional HHsearch query is performed to identify more distant related template structures. Identified template structures are merged and subjected to the template selection procedure (paragraph 2.3.1).

Secondly, HHsearch is started if either (1) no suitable template structures were found by BLAST (2) if the target sequence was not fully covered by BLAST hits. For the latter criteria we used a threshold of 25 residues, which reflects roughly the size of a small domain. If an additional HHsearch run is required, a profile-HMM of the target sequence is built. This involves several rounds of PSI-BLAST against culled versions of the NCBI-nr database. The target HMM-profile is then queried against the templates HMM-profile library which is culled so that two sequences in template library do not share more than 70% sequence identity. Templates are retained

according to the recommended cutoffs (P-value > 50) by the authors of the programs. Finally, the list of template structures is subjected to the template selection procedure described below (paragraph 2.3.1).

This hierarchical template search protocol has several advantages. The closely related template structures with unambiguous alignment can be identified quickly with very high accuracy. Because building a HMM-profile for the each query sequence is time-consuming and computational demanding, the usage of BLAST as a first template search tool decreases the computational load without a loss of alignment quality. Further, it has been shown that BLAST performs more accurately in identifying closely related template structures than methods which rely on profile information.[61] The applied procedure ensures that information about close sequence relationships is not dispersed by the subsequent profile based search strategies. By the hierarchical combination of both search approaches we merge the speediness and accuracy of BLAST for closely related template structures, with the ability of HHsearch to identify distant related template structures and align them correctly. In summary the current procedure increases the sensitivity of the previous SWISS-MODEL automated pipeline which was based on BLAST solely.

2.3.1 Template selection procedure

Template selection is an important step in modeling especially when more than one template is identified for the same target sequence. In this case, a decision has to be taken about which template structure(s) will be subjected to the modeling routine. This task becomes difficult if the available template structures cover different regions of the target sequence. This is a common scenario for multidomain proteins which appear frequently in eukaryotic systems.[62] In addition, users of an automated modeling routine may want to answer different biological questions with the help of comparative models. For example, one user is interested in the active site of a catalytic domain whereas another needs information about the relative orientation between two domains. The first user requires an accurate model of the binding site: however, the second user will be more interested in a model which covers both domains. In addition, computing all possible models for a given protein based on all available template structures in not necessarily the best solution. In fact, users which are non-experts in the field of comparative modeling are often confused if several models for the same target region are computed and presented without any biological information attached.

To meet the expectations of users with different biological questions, we deployed a method which selects the evolutionary closest templates on one hand but also tries to achieve the best coverage of the target sequence.

Proteins consist often of more than one functional domain, especially in Eukaryotic organisms.[63] As a consequence, a more sophisticated template selection protocol is required to guarantee the best template for a particular region of the target sequence. We have therefore developed a template selection approach which uses different types of sequence features to select the best template for a given region of the target sequence. It has been shown that the evolutionary distance (i.e. the sequence identity between target and template sequences) is a good indicator for the expected quality of a model.[13,28,58,64]

We therefore use the sequence identity as a first criterion to rank all detected template structures. If more than one template can be identified with a given sequence identity, we chose the template based on the E-value reported by the template search methods. The E-value implicitly combines the sequence identity between target and template sequences with the length of the alignment. In general, lower E-values are assigned to longer hits. As a consequence, a template which covers more residues of the target sequence will be preferred over another template with the same sequence identity. If the sequence identity and E-value of the templates are non-distinguishable, the experimental resolution of the template structure will be taken into account, favoring a X-ray diffraction derived structure with the highest resolution. The target-template alignment is then submitted to the SWISS-MODEL algorithm ProModII[54] in order to calculate the three-dimensional coordinates of the model. Afterwards, the model is refined by using the GROMOS force field[49]. If the modeling of the select template structure fails due for example difficulties in the loop reconstruction process, the next template which is identified using the described criteria is submitted to the modeling process.

New templates are added recursively if they either increase the coverage of the target sequence or elongate a model by at least 25 residues. Modeling is terminated if all selected templates have been analyzed.

2.3.2 Accuracy of the SWISS-MODEL Pipeline

The accuracy of the automated SWISS-MODEL pipeline is evaluated within the CAMEO ("Continuous Automated Model EvaluatiOn") project. CAMEO (www.cameo3d.org) continuously benchmarks the accuracy of automated protein modeling methods. CAMEO submits the sequences of protein structures which will become public in the next official PDB release to the participating prediction servers. The timeframe for model prediction is 48 hours; the assessment of the models takes places if the PDB structure is released. Hence, the model quality evaluation occurs as blind experiment, where the predictors do not have access to the structure during the prediction period. Currently there are three automated modeling server registered to CAMEO, and data has been accumulated for a period of 16 weeks:

1. "server0" : SWISS-MODEL[56] "
2. "server1" : ModWeb[65]
3. "server2": HHpred[59]

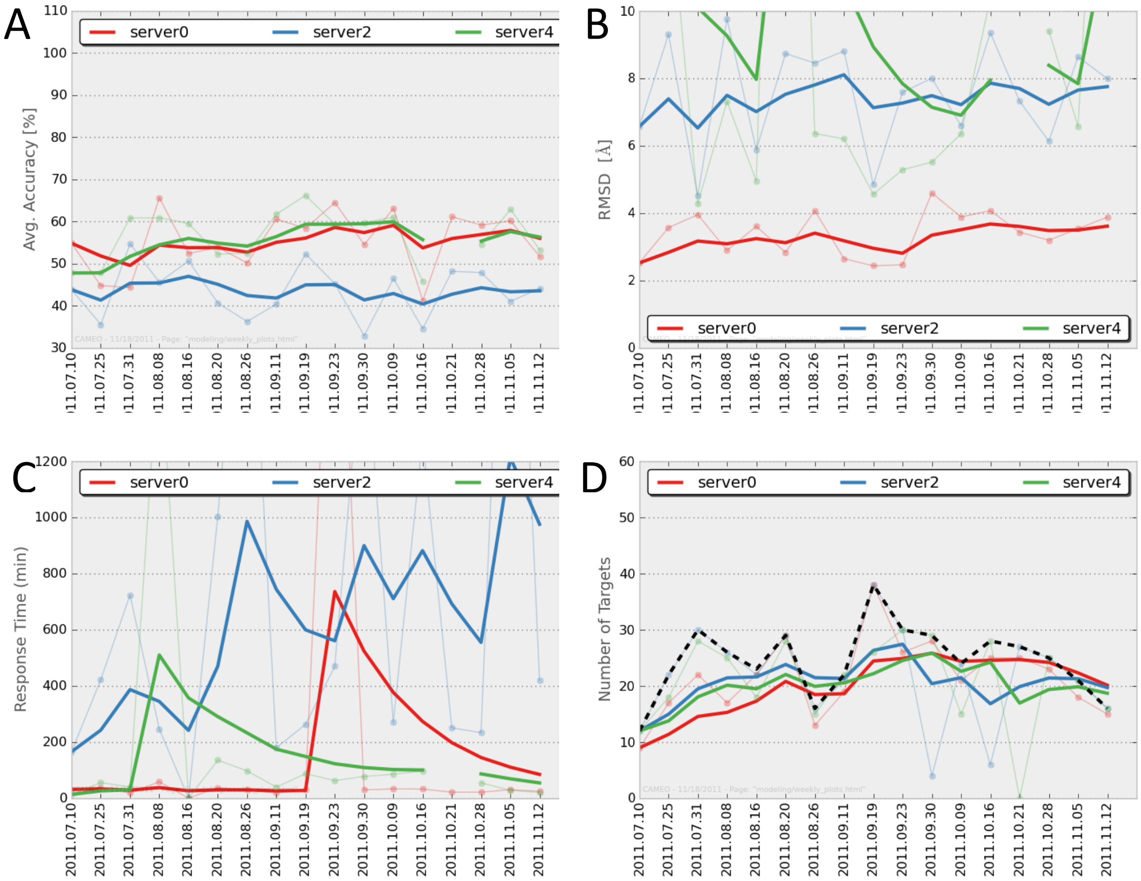


Figure 6 Performance of the SWISS-MODEL pipeline compared to HHpred and ModWeb. The panels show the performance in terms of average accuracy, RMSD, response time and the number of target for which at least one model was produced.

Figure 6A shows the average accuracy for each week and server. The average accuracy combines coverage of the target sequence with the structural accuracy of the model and is comparable to the GDT-HA score. It can be observed that SWISS-MODEL performs comparably well to HHpred and better than ModWeb. If using RMSD as similarity measurement, SWISS-MODEL outperforms ModWeb as well as HHpred. The analysis of average accuracy and RMSD indicates that SWISS-

MODEL builds shorter but more accurate models than HHsearch, which however outperforms SWISS-MODEL in terms of model coverage. In addition, CAMEO analyzes the modeling time and the number of targets for which at least one model is returned by the target. SWISS-MODEL performs comparable to HHpred regarding the computation time. Both methods had one outlier indicating a high load on their server in that period. Regarding the number of submitted models all three servers perform similarly.

In summary, the results show that the fully automated SWISS-MODEL pipeline server performs well compared to other standard modeling servers in the field in terms of both, the accuracy and responsiveness. Nevertheless, the benchmarking period is restricted to 17 weeks and 418 targets, which limits the significance of the evaluation results. More detailed results are expected with the assessment of coming PDB released target sequences.

2.3.3 Discussion

2.3.3.1 Template identification

In the SWISS-MODEL expert system we apply a hierarchical template search to handle the different levels of difficulty for identifying and aligning target to templates sequences. In comparison, ModWeb uses PSI-BLAST for the identification of template structures whereas HHpred is based on HHsearch as template search tool. PSI-BLAST as well as HHsearch are known for their strength in identifying remotely related templates. However a recent study has shown that simple sequence-sequence alignment tools are often superior to tools which use evolutionary information if the template is a close homologue of the target protein. As a result, we apply BLAST for the identification of closely related templates. The higher accuracy in terms of RMSD is likely an effect of BLAST. BLAST builds typically rather short alignments. If analyzing the results for RMSD in combination with the “average accuracy”-score, which accounts also for coverage, it seems that SWISS-MODEL predicts shorter models with high accuracy. In contrast, HHpred focus more on the prediction of models with high coverage. In addition, methods which require the generation of a profile are usually computational intensive and significantly extend the overall modeling time. By using BLAST as first template identification tool, we shorten the computation time without losing sensitivity and alignment accuracy. However, the computationally efficiency of modeling routines itself is hard to estimate using the data from CAMEO, because the response time also includes the overall load of server, hardware archicture etc.

Currently, a culled version (max. 70% sequence identity between two sequences) of the PDB is used to compile the HMM-template library, mainly because of performance reasons and the fact that clusters of proteins which share more than 70% sequence identity are highly similar regarding their structure. It can be expected, that by using all structures, the accuracy could be improved, when applying an appropriate template selection.

2.3.3.2 Template selection

The selection of the templates which are submitted to the modeling routine is mainly based on the evolutionary distance as quality criteria combined with the optimal coverage of the target sequence.

Additionally, it is has been shown in recent CASP editions that the use of quality estimation methods can help to distinguish near-native from non-native protein structure models[66]. Many methods model the target sequence based on all template structures at hand and use model quality estimation methods (MQE) for the selection of the final model. As a consequence the incorporation of such methods into the template selection process should increase the accuracy in identifying suitable template structures.

Finally, it has to be noted that the purpose of protein models submitted to benchmark experiments like CASP may differ from that of a model used by biomedical researchers in order to guide their experiments. The evaluation of the template based modeling category within CASP requires the submission of one model which ideally covers the complete target sequence. As such the predictors have to find an optimal trade-off between coverage and quality of the submitted models. The limitation of submitting only one model make sense within the CASP experiment in order to force groups to develop methods which model accurate and complete models, however, it is may be less relevant for a biomedical researcher. For the latter, a shorter but more precise model would be preferred for investigating for instance an active site, whereas a longer model could provide information about the relative domain orientation, likely with a lower accuracy. As a consequence, the “quality” of the model depends on the biological application and can be hardly expressed in numbers. Because SWISS-MODEL was designed to provide models for non-experts, we have chosen the approach which selects models according to biological applications rather than maximizing the accuracy for one single model.

2.4 Implementation

The described automated homology modeling routine is implemented as a modular PERL framework and integrated in the SWISS-MODEL workspace (“automated mode”) and the SWISS-MODEL Repository. In the “automated mode” of the SWISS-MODEL Workspace the user have to specify the target sequence or its UniProt accession code in order to obtain the protein structure models. A detailed description of the various modeling modes and the general use of the SWISS-MODEL workspaces is presented in a protocol which was recently published in Nature protocols[3]. Currently the SWISS-MODEL workspace is one of the mostly used homology modeling server in the biomedical community, with about 2000 requests for the automated SWISS-MODEL pipeline per day. The applications of the automated modeling pipeline within the SWISS-MODEL Repository are discussed in the next chapter.

2.5 The SWISS-MODEL Repository and associated resources

The following chapter was published as journal paper.[4]

My contributions were the follow:

- Development of the automate modelling pipeline
- Application of the automated modelling pipeline to a large set of protein sequences
- Development and design of a relational database
- Development of an incremental update procedure

SWISS-MODEL Repository (“http://swissmodel.expasy.org/repository/”) is a database of three-dimensional protein structure models generated by the SWISS-MODEL homology-modelling pipeline. The aim of the SWISS-MODEL Repository is to provide access to an up-to-date collection of annotated three-dimensional protein models generated by automated homology modelling for all sequences in Swiss-Prot and for relevant models organisms. Regular updates ensure that models are based on the current state of sequence and structure databases, including new template structures and building models for new target sequences, as well as accounting for improvements in the underlying modelling pipeline. As of September 2008, the database contains 3.4 million entries for 2.7 million different protein sequences from the UniProt database. SWISS-MODEL Repository allows the users to assess the quality of the models on the database, search for alternative template structures, and to build models interactively via SWISS-MODEL Workspace (http://swissmodel.expasy.org/workspace/). Annotation of models with functional information and cross-linking with other databases such as the Protein Model Portal module (http://www.proteinmodelportal.org) of the PSI Structural Genomics Knowledge Base facilitates the navigation between protein sequence and structure resources.

Introduction

Three dimensional protein structures are crucial for understanding protein function at a molecular level. In recent years, tremendous progress in experimental techniques for large scale protein structure determination by X-ray crystallography and NMR has been achieved. Structural genomics efforts have contributed significantly to the elucidation of novel protein structures[67], and to the development of technologies, which have increased the speed and success rate at which structures can be determined and lowered the cost of the experiments[68,69]. However, the number of known protein sequences grows at an ever higher rate as large scale sequencing projects, such as the Global Ocean Sampling expedition, are producing sequence data at an unprecedented rate[70]. Consequently, the last release of the UniProt[19] protein knowledgebase (version 14.0) contained more than 6.5 millions sequences, which is about 100 times the number protein structures currently deposited in Protein Data Bank[20] (~ 53’000, September 2008) . For the foreseeable future, stable and reliable computational approaches for protein structure modelling will therefore be required to derive structural information for the majority of proteins, and a broad variety of *in silico* methods for protein structure prediction has been developed in recent years.

Homology (or comparative) modelling techniques have been shown to provide the most accurate models in such cases, where experimental structures related to the protein of interest were available. Although the number of protein sequence families rises at a rate that is linear or almost linear with the addition of new sequences[70], the number of distinct protein folds in nature is limited[12,67] and the growth in the complexity of protein families appears as a result of the combination of domains. Complete structural coverage of whole proteomes (on the level of individual soluble domain structures) by combining experimental and comparative modelling techniques appears therefore as a realistic goal, and is already been pursued e.g. by the Joint Center for Structural Genomics for the small model organism *Thermotoga maritima* (JCSG)[71,72]. Assessment of the accuracy of methods for protein structure prediction, e.g. during the bi-annual CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments[2,73] or the automated EVA project[64], has demonstrated that comparative protein structure modelling is currently the most accurate technique for prediction of the 3D-structure of proteins. During the CASP7 experiment, it became apparent that the best fully automated modelling methods have improved to a level where they challenge most human predictors in producing the most accurate models[23,40,52]. Nowadays, comparative protein structure models are often sufficiently accurate to be employed for a wide spectrum of biomedical applications, such as structure based drug design[74-78], functional characterisation of diverse members of a protein family[79], or rational protein engineering for e.g. the humanization of therapeutic antibodies, or the study function properties of proteins[80-84].

Here, we describe the SWISS-MODEL Repository, a database of annotated protein structure models generated by the SWISS-MODEL Pipeline, and a set of associated web based services that facilitate protein structure modelling and assessment. We emphasize the improvements of the SWISS-MODEL Repository which have been implemented since our last report[85]. These include a new pipeline for template selection, the integration with interactive tools in the SWISS-MODEL Workspace, the programmatic access via DAS (Distributed Annotation System)[86], the implementation of a reference frame for protein sequences based on md5 cryptographic hashes, and the integration with the Protein Model Portal module (http://www.proteinmodelportal.org) of the PSI Structural Genomics Knowledge Base[87,88].

Repository Contents, Access and Interface

Homology Modelling

The SWISS-MODEL Repository contains models that are calculated using a fully automated homology modelling pipeline. Homology modelling typically consists of the following steps: Selection of a suitable template, alignment of target sequence and template structure, model building, energy minimization and / or refinement, and model quality assessment. This requires a set of specialized software tools as well as up to date sequence and structure databases. The SWISS-MODEL pipeline (version 8.9) integrates these steps into a fully automated workflow by combining the required programs in a PERL based framework.

Since template search and selection is a crucial step for successful model building, we have implemented a hierarchical template search and selection protocol, which is sufficiently fast to be used for automated large scale modelling, sensitive in detecting low homology targets, and accurate to correctly identify close target structures. In the first step, segments of the target sequence sharing close similarity to known protein structures are identified using a conservative BLAST[32] search with restrictive parameters (E-value cut-off : 10^{-5}, 60% minimum sequence identity to sequences of the SWISS-MODEL Template Library SMTL[56]). This ensures that information about close sequence relationships is not dispersed by the subsequent profile based search strategies[61]. If regions of the target sequence remain uncovered, in the second step a search for suitable templates is performed against a library of Hidden Markov Models for SMTL using HHSearch[40]. Templates resulting from both steps are ranked according to E-Value, sequence identity, resolution and quality of the template structures. From this ranked list, the best templates are progressively selected to maximize the length of the modelled region of the protein. New templates are added if they significantly increases the coverage of the target sequence (spanning at least 25 consecutive residues), or new information is gained (e.g. templates spanning several domains help to infer relative domain orientation). For each selected target-template alignment, 3-dimensional models are calculated using ProModII[54] and energy minimized using the Gromos force field[49]. The quality of the resulting model is assessed using the ANOLEA mean force potential[50].

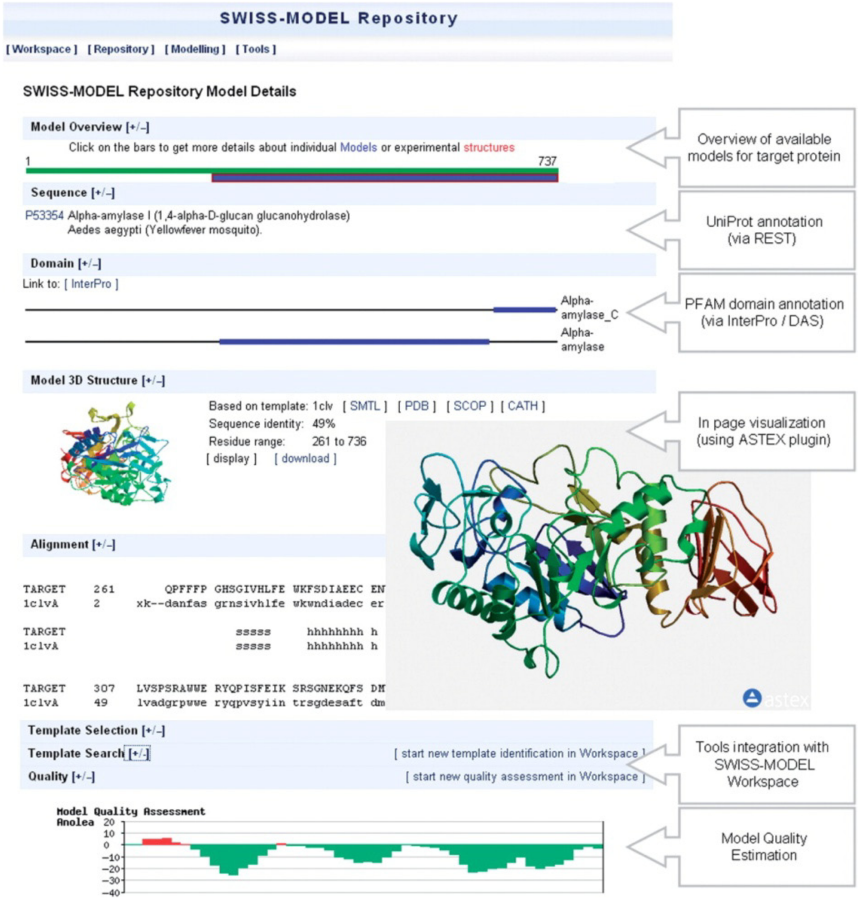
Depending on the size of the protein and the evolutionary distance to the template, model building can be relatively time-consuming. Therefore, comprehensive databases of pre-computed models[85,89,90] have been developed in order to be able to cross-link in real-time model information with other biological data resources, such as sequence databases or genome browsers.

Model Database

The SWISS-MODEL Repository is a relational database of models generated by the automated SWISS-MODEL pipeline based on protein sequences from the UniProt database[19]. Within the database, model target sequences are uniquely identified by their md5 cryptographic hash of the full length raw amino acid sequence. This mechanism allows reducing the redundancy in protein sequence databases entries, and facilitates cross-referencing with databases using different accession code systems. Mapping between UniProt and various database accession code systems to our md5 based reference system is derived from the iProClass data base[91]. Regular updates are performed for all protein sequences in the SwissProt database [17], as well as complete proteomes of several model organisms (*Homo sapiens, Mus Musulus, Rattus norvegicus, Drosophila melanogaster, Arabidopsis thaliana, Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Caenorhabditis elegans , Hepacivirus*). Regular incremental updates are performed to include new target sequences from the UniProt database and to reflect new template structure information becoming available, whereas full updates are required to account for major improvements in the underlying modelling algorithms. The current SWISSMODEL-Repository release contains 3,45 million models for 2,72 million unique sequences, built on 26,185 different template structures (34,540 chains), covering 48.8% of the entries from UniProt (14.0.), and more specifically 65.4% of the unique sequences of Swiss-Prot (56.0.), the manually annotated section of the UniProt knowledgebase. The size of the models ranges from 25 up to 2059 residues (e.g. fatty acid synthase beta subunit from *Thermomyces lanuginosus*) with an average model length of 221 residues.

Graphical User Web Interface

The web interface at http://swissmodel.expasy.org/repository/ provides the main entry point to the SWISS-MODEL Repository. Models for specific proteins can be queried using different database accession codes (e.g. UniProt AC and ID, GenBank, IPI , Refseq) or directly with the protein amino acid sequence (or fragments thereof, e.g. for a specific domain). For a given target protein, a graphical overview illustrating the segments for which models (or experimental structures) are available is shown (Figure 7). Functional and domain annotation for the target protein is retrieved dynamically in real time using web service protocols to ensure that the annotation information is up-to-date. UniProt annotation of the target protein is retrieved via REST queries (http://www.uniprot.org). Structural domains in the target protein are annotated by PFAM domain classification[92], which is retrieved dynamically by querying the InterPro[93] database using the DAS protocol[86]. The md5 based reference frame for target proteins allows to update the database accession mappings in between modelling release cycles. This ensures that cross references with functional annotation resources such as InterPro correspond to proteins of identical primary sequence, thereby avoiding commonly observed problems with incorrect cross-references as a result of instable accession codes or asynchronous updates of different data resources. Finally, for each model, a summary page provides information on the modelling process (template selection and alignment), model quality assessment by ANOLEA[50] and Gromos[49], and in page visualization of the structure using the Astex Viewer [94] plugin.



SWISS-MODEL Repository

[Workspace] [Repository] [Modelling] [Tools]

SWISS-MODEL Repository Model Details

Model Overview [+/–]

Click on the bars to get more details about individual Models or experimental structures

Sequence [+/–]

P53354 Alpha-amylase I (1,4-alpha-D-glucan glucanohydrolase)
Aedes aegypti (Yellowfever mosquito).

Domain [+/–]

Link to: [InterPro]

Model 3D Structure [+/–]

Based on template: 1clv [SMTL] [PDB] [SCOP] [CATH]
Sequence identity: 49%
Residue range: 261 to 736
[display] [download]

Alignment [+/–]

```
TARGET     261        QPFFFP GHSGIVHLFE WKFSDIAEEC EN
1clvA      2      xk--danfas grnsivhlfe wkwndiadec er

TARGET                        sssss      hhhhhhhh h
1clvA                         sssss      hhhhhhhh h


TARGET     307    LVSPSRAWWE RYQPISFEIK SRSGNEKQFS DM
1clvA      49     lvadgrpwwe ryqpvsyiin trsgdesaft dm
```

Template Selection [+/–]

Template Search [+/-] [start new template identification in Workspace]

Quality [+/–] [start new quality assessment in Workspace]

Overview of available models for target protein

UniProt annotation (via REST)

PFAM domain annotation (via InterPro / DAS)

In page visualization (using ASTEX plugin)

Tools integration with SWISS-MODEL Workspace

Model Quality Estimation

Figure 7 Typical view of a SWISS-MODEL Repository entry. For the UniProt entry P53354, the α-amylase I (EC 3.2.1.1; 1,4-α-D-glucan glucanohydrolase) from Aedes aegypti (Yellowfever mosquito), a model covering the active amylase domain is shown, including information on the template structure used for model building, the target–template sequence alignment, and quality assessment of the model. Functional annotation such as PFAM domain structure and UniProt annotation of the protein sequence is retrieved dynamically. Links to SWISS-MODEL Workspace enable the user to run additional model quality assessment tools on the model, or search the template library for alternative template structures.

Integration with SWISS-MODEL Workspace

The SWISS-MODEL Repository is a large-scale database of pre-computed three-dimensional models. Often however, one may be interested in performing additional analyses either on the models themselves, or on the underlying protein target sequence. We have therefore implemented a tight link between the entries of the SWISS-MODEL Repository and the corresponding modules in the SWISS-MODEL Workspace, which provides an interactive web based, personalized working environment [54,56,95]. Besides the functionality for building protein models it provides various modules to assess protein structures and models. The estimation of the quality of a protein model is an important step to assess its usefulness for specific applications. In particular, models based on template structures sharing low sequence identity require careful evaluation. Therefore, entries from the Repository can be directly submitted to the Workspace for quality assessment using different global and local quality scores such as DFire[96], ProQRes[97] or QMEAN[51].

The default output format for models in the Repository is DeepView project files[54], which provides the possibility to manually adjust errors identified in the underlying alignment and resubmit the request to Workspace for modelling. Since new protein structures are deposited daily in the PDB, databases of pre-computed models will exhibit a certain delay in incorporating new templates, depending the respective update cycles. The Repository links therefore directly to the corresponding template search module in Workspace, which allows to interactively running searches for newly released templates. The direct cross-linking between Repository and Workspace allows combining the advantages of the database of pre-computed models with the flexibility of an interactive modelling system.

Interoperability

Programmatic Access

One of the major challenges of computational biology today is the integration of large amounts of diverse data in heterogeneous formats. Very often, data exchange within one domain, e.g. sequence-based data resources, is relatively straightforward, but seamless exchange between resources serving different data types, such as genome browsers and protein structure databases, is less common due to the lack of common and accepted standards. The "Distributed Annotation System" (DAS)[86] is a light-weight mechanism for webservice-based annotation exchange. The DAS concept relies on a XML specification which defines the communication between server and client. Queries can be executed by sending a specific http-request, to the DAS server. The result of the DAS-Server request is a human readable and easy-to-parse XML-document following the Biodas specifications (http://www.biodas.org).

The DAS-Server of the SWISSMODEL-Repository is based on the DAS/1 standard and can be queried by primary UniProt accession codes or md5-hashs of the corresponding sequences. Individual models for a query sequence (“SEGMENT”) are annotated as “FEATURE”, with information about the start and stop position in the target sequence, template-sequence identity, and the URL to the corresponding SWISS-MODEL Repository entry. The DAS service allows SWISS-MODEL Repository be cross-linked with other resources using the same standards, e.g. genome browsers. The SWISS-MODEL Repository DAS service is accessible at http://swissmodel.expasy.org/repository/das/xxxxx.

PMP – The Protein Model Portal

One of the major bottlenecks in the use of protein models is that, unlike for experimental structures, modeling resources are heterogeneous and distributed over numerous servers. However, it is often beneficial for the user to directly compare the results of different modeling methods for the same protein. We have therefore developed the Protein Model Portal as a component of the PSI structural genomics knowledge base[87,88]. This resource provides access to all structures in the PDB, functional annotations, homology models, structural genomics protein target tracking information, available protocols, and the potential to obtain DNA materials for many of the targets. The Protein Model Portal currently provides access to several million pre-built models from four PSI centers, ModBase[90], and SWISS-MODEL Repository[85,89].

Future Directions

SWISS-MODEL Repository will be updated regularly to reflect the growth of the sequence and structure databases. Future releases of SWISS-MODEL Repository will include models of oligomeric assemblies, as well as models including essential cofactors, metal ions, and structural ligands. Structural clustering of the Swiss Model Template Library will also allow us to routinely include ensembles of models for such proteins, which undergo extensive domain movements.

Citation

Users of SWISS-MODEL Repository are requested to cite this article in their publications.

Acknowledgements

We are grateful to Rainer Pöhlmann, [BC]^2 & Biozentrum University of Basel, for professional systems support, Pascal Benckert for fruitful discussions on model quality assessment, and Jürgen Kopp for pioneering work on earlier versions of SWISS-MODEL Repository. We thank James Battey for critically reading the manuscript. We are indebted to Dr. Michael Podvinec for his enthusiastic support and excellent coordination of the Scrum process for the SWISS-MODEL team. We grateful to Eric Jain for the swift implementation of md5 based REST queries on the UniProt server, and Wendy Tao, John Westbrook, and Helen Berman (RCSB) for the great collaboration on the PSI SGKB Protein Model Portal. Computational resources for SWISS-MODEL Repository are provided by [BC]^2 Basel Computational Biology Center (http://www.bc2.ch) and Vital-IT (http://www.vital-it.ch). The PSI SGKB Protein Model Portal was supported by the National Institutes of Health NIH as a sub-grant with Fox Chase Cancer Center grant 3 P20 GM076222-02S1, and as a sub-grant with Rutgers University, under Prime Agreement Award

Number: 3U54GM074958-04S2. SWISS-MODEL Workspace and Repository have been supported by the Swiss Institute of Bioinformatics (SIB).

2.6 Current status of the SWISS-MODEL Repository

This section describes the current status of the SWISS-Model Repository and highlights recent improvements. In addition a more detailed view about the technical implementation is given.

The current release of the SWISS-MODEL Repository consists of 3,2 million model entries for 2,3 million unique sequences of the UniProt database (release 2011_09, September 21, 2011). In the last update cycle (2011-10-06, October 10, 2011) 122,348 unique sequences were considered for revision (for details of the update procedure please see paragraph 2.6.1).

One of the major advantages of a protein structure model database is the possibility to include annotation of other data resources, thus providing biological relevant information in one view. Cross-references were established with two other important protein resources.

Cross-references to STRING

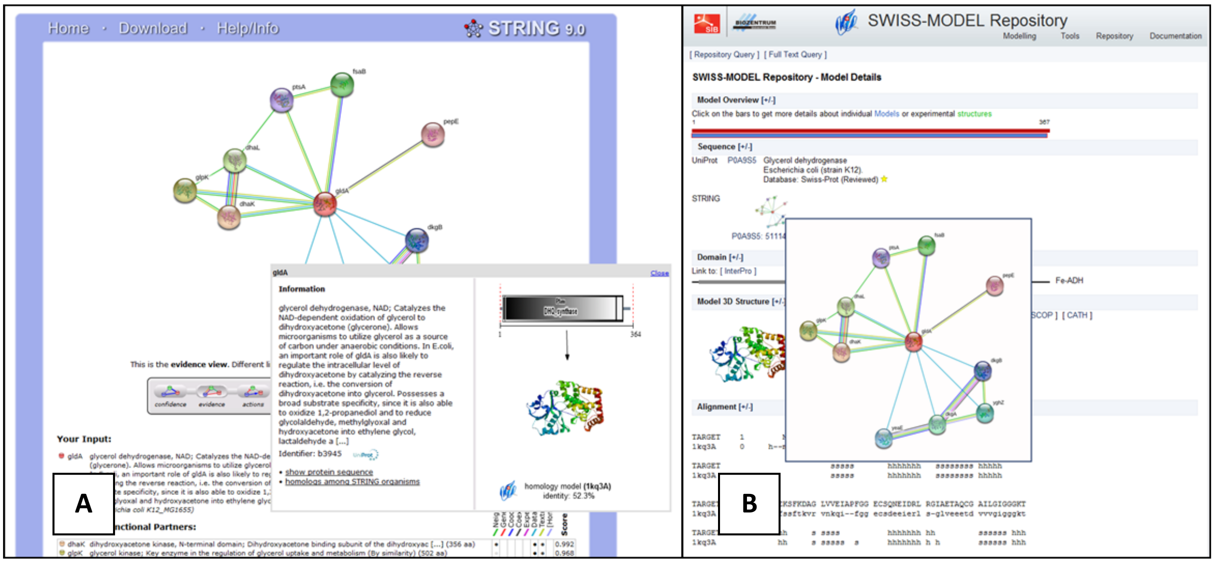


Figure 8 Cross reference to the STRING database[98]. A) If a homology model for a protein described in STRING exists, the according structure is displayed. B) If for a particular homology model a STRING protein entry exists, a crosslink to STRING including the interaction network is displayed.

Cross references were established between the SWISS-MODEL Repository and STRING[98], a database of known and predicted protein interactions. As such, STRING represents a valuable resource for user to investigate if a particular protein is involved into protein interaction networks. Screenshots of the STRING website and the SWISS-MODEL Repository website are shown in Figure 8. Within STRING, proteins for which a homology model exists are linked to the SWISS-MODEL Repository via a mouse over menu (Figure 8, panel A). Likewise, the SWISS-MODEL Repository displays a thumbnail of the protein interaction network if the protein was analyzed by STRING (panel B). Currently there are 587’855 proteins linked to STRING (v 9.0) and regular updates on both sites are performed.

Cross-references to UniProt-KB

Cross-links were also established from the UniProt database[19] to the SWISS-MODEL repository. If a particular UniProt sequence has been modeled by the SWISS-MODEL Repository pipeline, the target sequence is annotated in the Cross-reference section of the UniProt entry.

2.6.1 Update procedure

A protein model which was built by comparative modeling relies on the template situation at the time it is calculated. Hence, a protein model is per definition outdated if new template structures are released. As a consequence, it is essential for the accuracy of the models that regular updates are carried out in order to examine if new template structures become available.

Because the number of sequences in protein sequence databases is rapidly increasing, an update of the complete SWISS-MODEL Repository based on the latest release of the UniProt database (~17million sequences in Nov. 2011), would be computationally very intensive and could not be performed regularly using the SWISS-MODEL modeling pipeline. We therefore decided to limit the number of sequences for regular updates and selected seven proteomes which are of interest to the scientific community. They are frequently refereed as “model organisms” to examine important biological questions. The sequences of complete proteomes are available for download on the UniProtKB website (“http://www.uniprot.org”).

Table 1 shows seven proteomes updated regulary in the SWISS-MODEL Repository and their actual number of sequences (Uniprot release 2011_08). In total 185 206 UniProt entries are revised in a four weeks update cycle or if new UniProt sequences are released.

Two different situations need to be addressed during an update cycle:

1. New target sequences were released (e.g. by UniProt).
2. New (potential) template structures were added to the template library.

The first scenario can be approached by modeling the new target sequences using the standard modeling pipeline. The latter scenario, which will affect all proteins during the update cycle, needs more efforts, because it is not known a priori if a recently published template is homologous to a previously modeled target sequence. Thus, it needs to be proven if any of the template structures provide new structural information for the target sequence. However, this step is computational very intensive. We have therefore modified the automated SWISS-MODEL pipeline in order to decrease the overall CPU time:

Scientific name	**Taxonomy id**	**Number of Sequences**
Homo Sapiens	9606	56,582
Mus Musculus	10090	44,837
Escherichia coli (strain K12)	833333	4,304
Saccharomyces cerevisiae (Bakers yeast)	559292	6,627
Arabidopsis thaliana	3702	32,689
Caenorhabditis elegans	6239	22,630
Drosophila melanogaster	7227	17,537
Total		185,206

Table 1 Proteomes in the current SWISS-MODEL Repository update cycles. The number of sequences was calculated based on UniProt release 2011_08.

BLAST-Batch mode

All sequences in the current update are subjected to BLAST in a "batch" mode. Therefore, groups of 100 target sequences are consecutively queried against the current template library. Thereby the BLAST database is kept in memory. This procedure decreases the load onto the file system by avoiding I/O intensive operations for loading the BLAST database into the memory for each single sequence.

Incremental template search using HHsearch

Profile generation is a computational intensive approach, because several rounds of PSI-BLAST are required. To avoid profile generation for target sequences which were already modeled in previous updates, all HMM- profiles were stored and reused for further updates.

If the HMM-profile of the query and the HMM-profile of the templates are identical to the last update, a new alignment step is not required. We therefore queried the target profile against a template library which consists only of template HMM-profiles which have been released since the last update. In order to select the template structures for modeling, a list was built containing the templates from BLAST, the incremental HHSearch run and the template structures identified from the last update cycle.

Modeling

The list of identified templates is then compared with the list of successful modeled templates in the previous update cycle. If the current selection of templates is different from the previously selected, the current set of templates is subjected to ProModII; if no new templates were selected, the coordination files from the previous update were used.

Figure 9 shows a schematic representation of the SWISS-MODEL repository update procedure as described in the previous paragraphs. CPU intensive operations such as BLAST, HHsearch and modeling with ProModII are calculated in parallel on our computational powerful inhouse cluster.

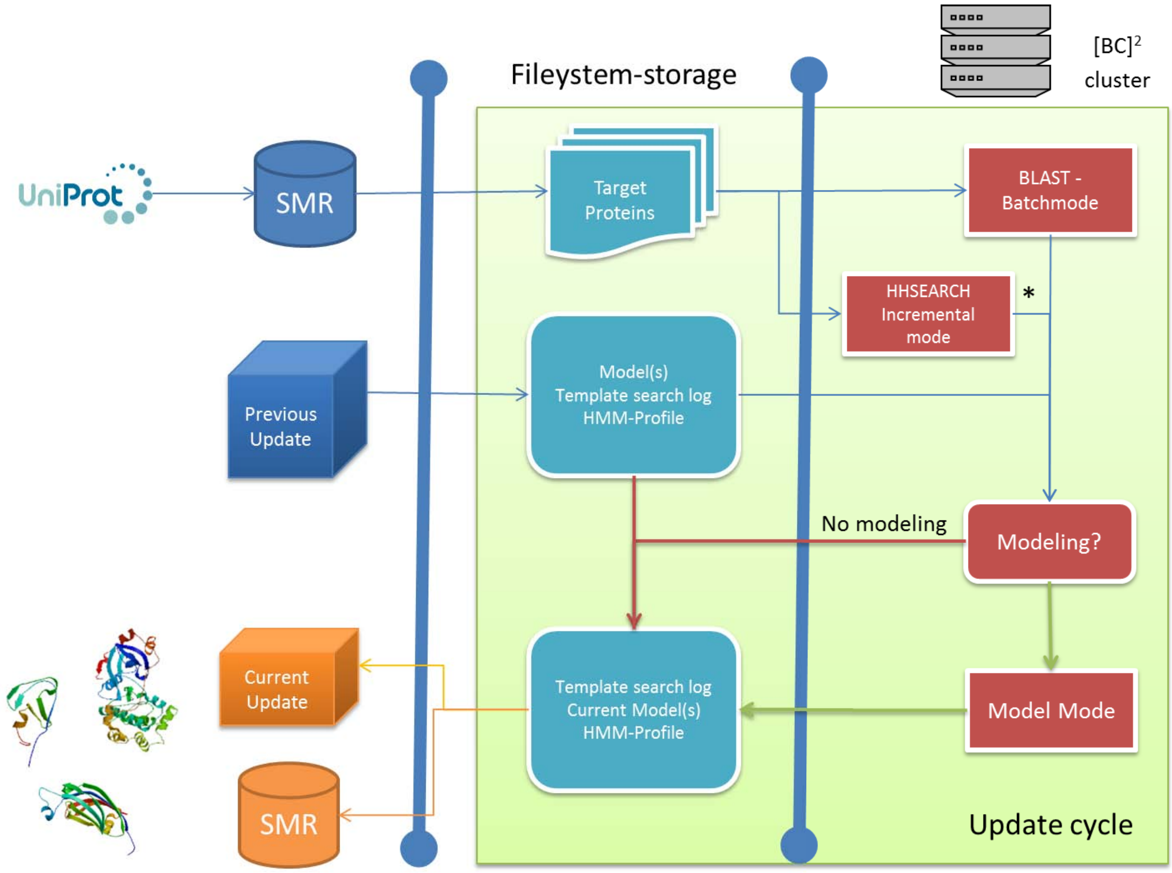


Figure 9 The SWISS-MODEL Repository update schema. The target sequences are fetched from the UniProt database and subjected to BLAST in Batch mode. Based on information of the previous update an incremental HHSearch template library is built and queried with the cached HMM-profile of the target sequence. If the selected template structures are different from the previous set of models, the templates are modeled by ProModII. The relational database is updated and the coordination files as well as the profiles and template reports are deposited for further updates.

2.6.2 Performance

In order to evaluate the performance of the update procedure the average CPU time per protein has been determined. Thus, we compared the CPU times of the automated pipeline of the SWISS-MODEL Workspace server to the procedure applied in a regular update cycle of the SWISS-MODEL Repository. In both cases the difference between “Start” and “Finish” time in the corresponding status file was calculated for all sequences with less than 1500 residues. In case of the SWISS-MODEL workspace (“workspace pipeline”), submissions for a period of 2 weeks were considered (~10,700 modeling jobs), for the SWISS-MODEL Repository (“repository update pipeline”) all new sequences of the last update cycle (2011-10-06, October 10, 2011) were considered (~115 000 target sequences).

The increase in CPU time ranges from about 3.4 minutes for sequences up to 50 residues to 4 hours for sequences with 1500 residues (Figure 10, upper panel). For proteins of size of 300-350 residues the standard “workspace pipeline” is a factor of 2.5 slower than the incremental “repository update pipeline”. With increasing sequence length the gap between the “workspace pipeline” and the “repository update pipeline” is steadily increasing. Considering protein models with 1000 residues, the standard “workspace pipeline” requires 5 times more CPU time.

To calculate the absolute CPU time difference we used the distribution of sequence lengths in the latest SWISS-MODEL update (Release 28) (see Figure 10, lower panel) and multiply them with the average times calculated for the standard “workspace pipeline” and the pipeline implemented in the Repository update procedure. If one would use the standard ”workspace pipeline” for updating the SWISS-MODEL Repository ~83 000 CPU hours in total would be necessary. This shows the tremendous computational effort which would be required. In contrast the “repository update pipeline” is able to update all proteins in about 800 CPU hours.

In summary, the comparison shows that incremental update procedure decreases the overall CPU time considerably. Despite the exclusion of the initial profile generation for HHsearch, it has been shown that regular updates for a selected set of proteins can be performed in reasonable time. Currently, the time limiting step in the SWISS-MODEL homology pipeline is the profile generation by HHsearch for the target sequences. Profile-generation involves several rounds of PSI-BLAST against protein sequence databases to search for homologous proteins. Such sequence databases grow exponentially, as an example, TrEMBL doubled its size in the last two years and holds now about 17.8 million sequences. This implies also higher demands for the computational systems (e.g. RAM memory), especially if parallel computing is performed. As a consequence the generation of profiles for newly added proteomes will considerably influence the overall computation time.

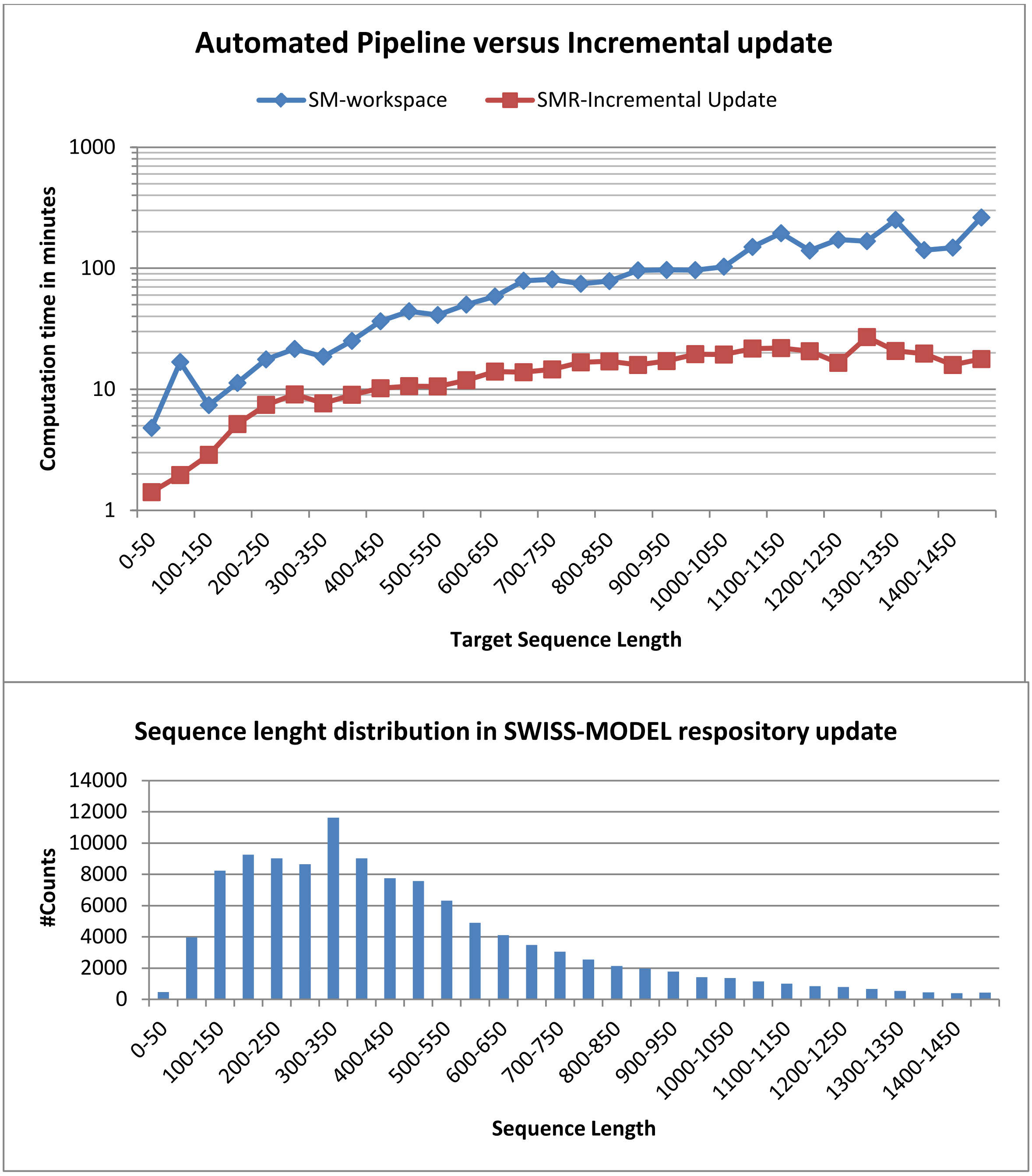


Figure 10 Performance of the incremental update procedure. (Upper panel) CPU times for a protein of a given length. A considerable acceleration can be observed if using the incremental update procedure. Please note that the BLAST-Batch process and the Profile building were not taken into account. (Lower Panel) Distributions of sequence length in the last regularly update cycle.

2.6.3 MySQL database-schema

The usefulness of calculated protein models relies on the visualization and accessibility of the data. The SWISS-MODEL Repository can be accessed via website or programmatically by the DAS protocol[86]. In order to enhance the functionality of the resource, different database identifier can be used to query the SWISS-MODEL repository. Various sequence databases often rely on the same set of sequences but use different accession code systems for the identification of sequences. It is therefore useful to provide an interface which can be queried with accession codes of various databases or the sequence itself. Therefore the accession code systems are mapped to the UniProt accession code. This information is provided by the IProClass database[91]. However, the mapping is time-critical and requires a fast mapping between accession code and md5 hashes of the target sequences to display the model quickly after entering the accession code on the website.

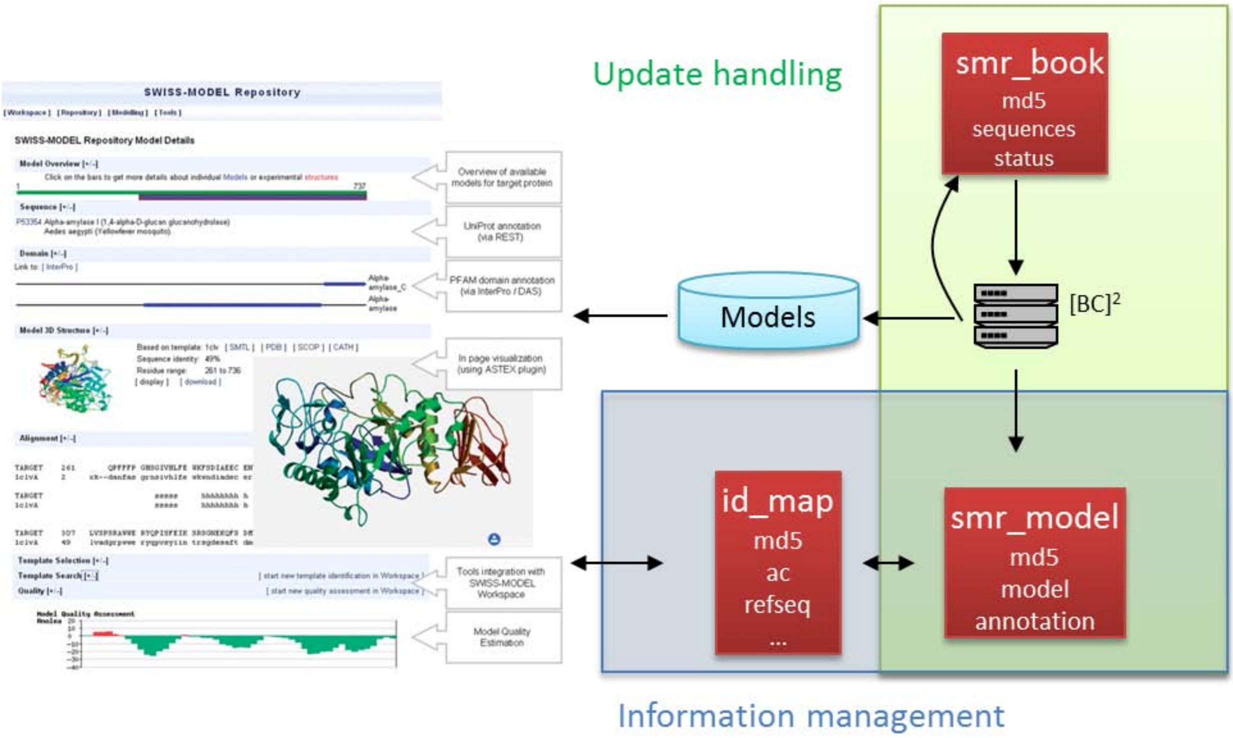

SWISS-MODEL Repository
[Workspace] [Repository] [Modelling] [Tools]
SWISS-MODEL Repository Model Details
Model Overview [+/-]
Click on the bars to get more details about individual Models or experimental structures
Sequence [+/-]
P53354 Alpha-amylase I (1,4-alpha-D-glucan glucanohydrolase) Aedes aegypti (Yellowfever mosquito)
Domain [+/-]
Link to: [InterPro]
Alpha-amylase_C
Alpha-amylase
Model 3D Structure [+/-]
Based on template: 1clv [SMTL] [PDB] [SCOP] [CATH]
Sequence identity: 49%
Residue range: 261 to 736
[display] [download]
Alignment [+/-]
Template Selection [+/-]
Template Search [+/-]
Quality [+/-]
[start new template identification in Workspace]
[start new quality assessment in Workspace]
Model Quality Assessment
Anolea
Overview of available models for target protein
UniProt annotation (via REST)
PFAM domain annotation (via InterPro / DAS)
In page visualization (using ASTEX plugin)
Tools integration with SWISS-MODEL Workspace
Model Quality Estimation
Update handling
smr_book
md5
sequences
status
[BC]²
Models
id_map
md5
ac
refseq
...
smr_model
md5
model
annotation
Information management

Figure 11 Database schema of the SWISS-MODEL Repository. The smr_model table stores relevant meta information of the protein models. Database identifiers were mapped to md5 hashes to identify the correct model entries. smr_book stores the status of performed updates.

To achieve fast response times, we implemented a MySQL database to store relevant information about the protein models, the mapping between the different accession code systems and the md5 of the target sequence and information about the update process itself.

Figure 11 shows a schematic representation of the database design. The database schema can be divided into two functional parts:

Information management

- smr_model: Holds all relevant annotation to display a particular model. This includes information about the used template structure, the sequence identity between template and target sequences, coverage of the target, model and revision dates. This information is primarily used for visualization on the SWISS-MODEL repository page
- uniprot: Consists of a representation of the current UniProt Knowledge database. This includes the UniProt accession code, the length of the sequence, taxonomy information and the md5 hash of the corresponding sequence.
- Mapping tables translate common used database identifier to UniProt accession code. These are derived using the PIR database[91]. The corresponding model(s) in the smr_table can be identified using the UniProt accession code-md5 relation. The mapping procedure was developed within the Protein Model Portal project[5].

Update handling

- smr_book: The content of this table helps to organize the regularly updates of the SWISS-MODEL Repository. It stores the sequence, md5, status, uniprot and pdb-release of the last update cycle. If a regularly update is performed, sequences are fetched from this table.

2.6.4 Statistics and structural coverage of proteomes

The current SWISS-MODEL Repository (release of October 10, 2011, based on UniProt release 2011_09, September 21, 2011) consist of totally 3’223’059 models for 2’293’270 distinct UniProt sequences. The total number of modeled sequences is composed by the initial modeling of the complete UniProtKB (14.0) in 2008 and the seven proteomes which have been regularly updated (see Table 1). To estimate the structural coverage of these proteomes, we calculated the structural coverage of all residues in a particular proteome. The *E.coli* proteome display the highest structural coverage; more than 45% of the residues in the proteome can be modeled with sequence identity greater than 30% (Figure 12). The structural coverage for eukaryotic systems is much lower ranging from 18% to 21% for *C.elegans* and human, respectively. However, a study estimated that about 18% of all residues in the human proteome are disordered and hence do not show a regular protein structure[99] . Similar percentages are reported for other eukaryotic systems and thus lowering the fraction of residues without any structural information.

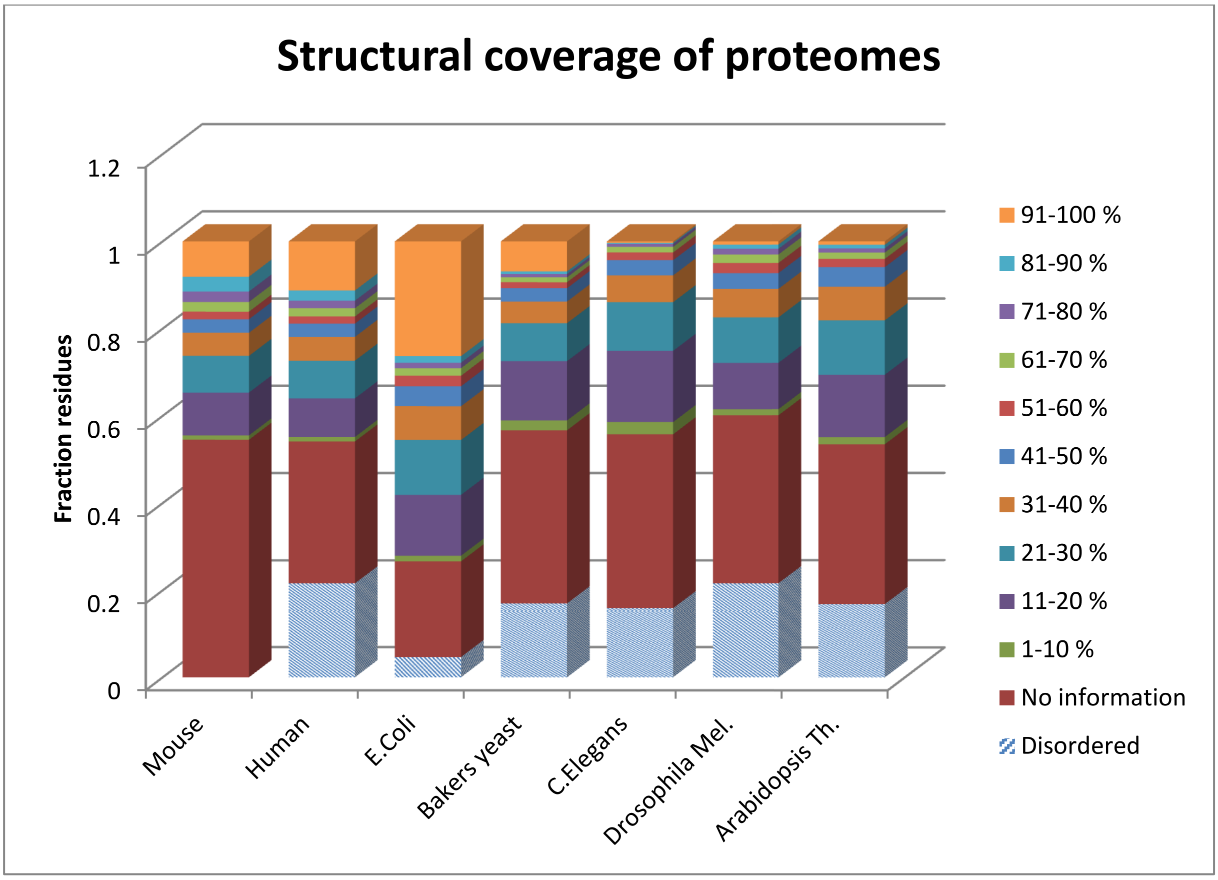


Figure 12 Structural coverage of residues in various proteomes. In addition the predicted fraction of disordered residues is shown [99]. Disorder predictions for the mouse genome were not found but can be assumed to be comparable to other closely related organisms.

3 Modeling of quaternary protein structures

3.1 Introduction

3.1.1 Function of oligomeric proteins

Proteins often accomplish their function either by interacting with other proteins or by forming macromolecular complex by self-assembly. Protein complexes are involved into any type of cellular processes, thus contributing considerably to the vital ity and survival of the cell.

Types of oligomeric proteins

In the enzyme database BRENDA[100] more than 75% of the enzymes are annotated as complexes. Two types of protein complexes can be distinguished. Homo-oligomers are formed by the assembly of self-interacting units of the same protein. In contrast, hetero-oligomers are formed by the assembly of subunits which are different in sequence and structure. It is estimated that between 50%-70% of oligomers are homo-oligomers[101,102] and the majority of the complexes annotated in BRENDA are homo-oligomers[100].

Furhter, oligomers can be divided into permanent or transient complexes.[103] Permanent complexes are usually very stable and disassembly of the complex leads to unfolded monomers. In contrast, transient complexes may exist also as stable monomers and assemble temporarily based on physical and chemical interaction with other proteins. Transient complexes are proposed to control signaling cascades and pathways in vivo and thus are very important for the functioning of the cell.

Reasons for oligomeric assembly

Despite the abundance of oligomers in the cell only little is known about the mechanism of oligomerization and their general benefits. Goodsell and Olson[101] proposed advantages for larger proteins in general and the formation of oligomeric proteins:

1. Building large complexes out of small subunits reduces errors due to protein translation. Because the error rate of translation increases with increasing gene length, smaller genes are less likely to contain translational errors. In addition, error prone subunits can be detected and exchanged in the systems with identical subunits (homo-oligomers). As a result, this leads to a more error tolerant system.

2. The genetic space required for coding the information of one single subunit of a protein complex is smaller than for monomeric proteins of the same size. Examples are capsids of viruses, which are typically, build by only a small number of distinct subunits. As a consequence, the required genetic space is only little compared to the space needed for coding larger monomers. On the other side, the discovery of large amounts of non-coding DNA in higher organisms[104] argues against this argument as general driving force for oligomerization.

3. The relative orientation of subunits can have regulatory functions, e.g. by relative changes in the conformation of the subunits. One example is the allosteric regulation of hemoglobin.

Finally, Marianayagam[105] proposed advantages for enzyme regulation and activation: for example, the close location of multiple active sites in oligomeric structures places enzyme activity under the multifaceted regulation of oligomerization. Other benefits can be identified for signal transduction in pathways and the regulation and construction of large structural units in the cell (e.g. actin filaments). Another example: the assembly of complexes can be triggered by the protein concentration in the cell. If the function is coupled to the oligomeric state (e.g. if the binding pocket consists of residues from more than one chain), the assembly can be seen as a censoring system which reacts on the cellular environment.

Symmetry in oligomers

In general, oligomers can be divided into two groups of symmetry: open and closed symmetry. Open symmetry can be observed in large complexes responsible for the cellular stability (e.g. Actin and Tubulin). In theory, complexes having an open symmetry can assemble infinitely in space if no limiting factors are present. In contrast, proteins adopting closed symmetry are finite in space and are mainly built by cyclic or dihedral symmetry. In addition, a small fraction of protein complexes have cubic symmetry. The following description of symmetry in protein complexes was adopted from Goodsell[101].

Cyclic symmetries consist of one single axis of rotational symmetry. C1 consists of one subunit (monomers), whereas C2 denotes a dimeric protein and so forth. In general higher cyclic symmetries (>2) require face-to-back interfaces, with at least two different types of interfaces on the surface.

Dihedral symmetries consist of an additional perpendicular axis of two-fold symmetry. The lowest dihedral state is D2, which consists of two C2-Dimers. Oligomeric complexes having dihedral symmetry can consist of several types of interfaces. This implies advantages, e.g. for allosteric control.

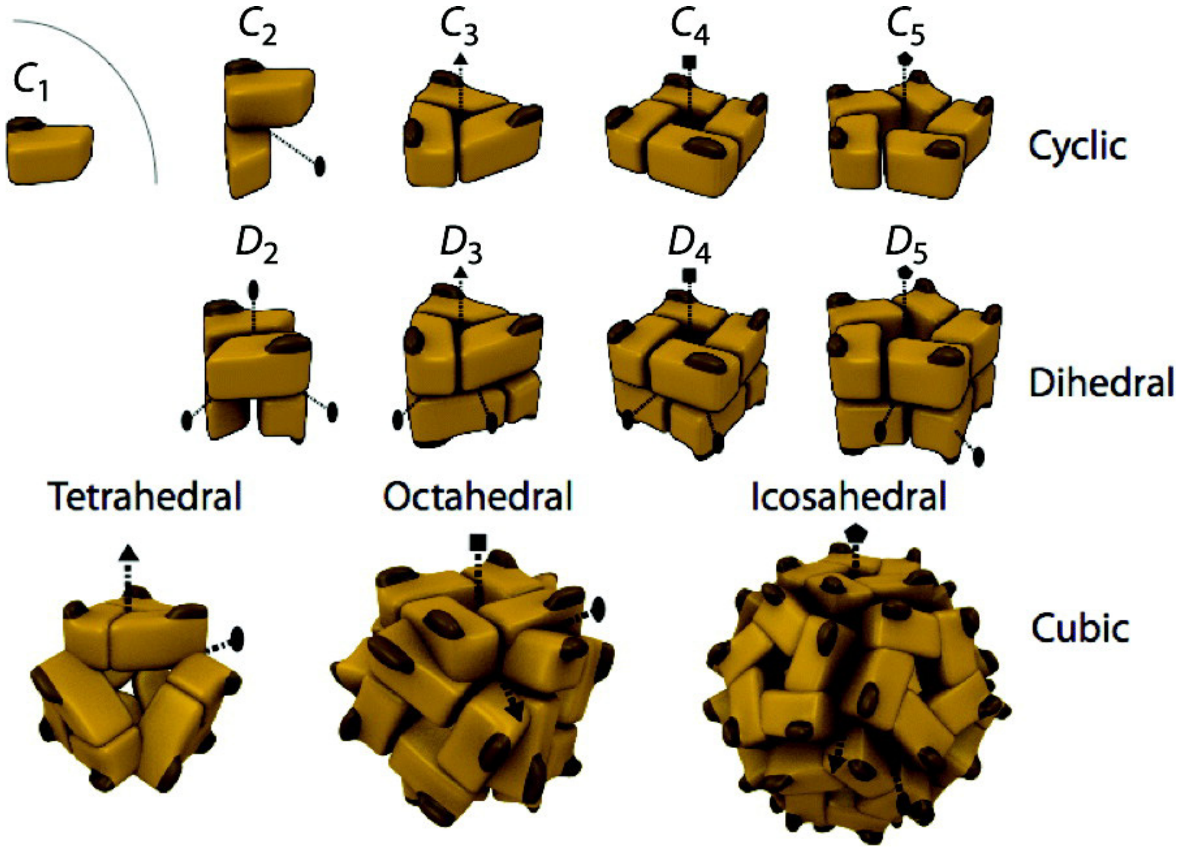


Figure 13 Different types of oligomeric symmetries. Courtesy of E.Levy (Cambridge, UK).

Protein complexes having cubic groups are usually involved into storage and transport. They contain threefold symmetries with another nonperpendicular rotational axis. It has been proposed very early by Watson and Crick[106], that icosahedral symmetries play an important role in the formation of virus capsids.

3.1.2 Stability of Interfaces

The disassembly of protein complexes can be described using the standard Gibbs free energy:

$$\Delta G = \Delta H - T\Delta S$$

The Gibbs free energy consists of two terms, which describe the enthalpic and entropic change in the system. A spontaneous chemical reaction requires in total a gain in the overall entropy (second law of thermodynamics). In this respect, the increase in enthalpy (i.e. the energy which is transferred from the system to the surrounding) must be higher than the loss of entropy (e.g. the restriction of the translational or rotational freedom). Krissinel concretize the concept of Gibbs free energy in order to describe the assembly or disassembly of macromolecular complexes.[107]

$$\Delta G^0_{diss} = -\Delta G_{int} - T\Delta S$$

The dissociation energy can be written as ΔG^0_{diss} ; a negative delta G indicates the disassembly of a complex whereas a positive value denotes a stable complex.

ΔG_{int} reflects the binding enthalpy between the subunits. As such it contributes positively to the interface stability. In contrast, the absolute temperature (T), and the change towards lower entropy in the system (ΔS) drive the assembly to less stable states.

The main forces for interface stability, as proposed by Chothia and Janin[108], are the hydrophobic interactions between non-polar residues. Presenting hydrophobic residues at the protein surface lowers the overall entropy in the system and is considered to be energetically unfavored. Thus, the burial of hydrophobic patches contributes positively to interface stability.

Other contributors to interface stability are contact interactions like hydrogen bonds, salt bridges and disulfide bonds. Out of these three types of interaction, hydrogen bonds appear to be the most frequent (6-8 hydrogen bonds per 1000Å^2)[109] and are likely the most important contributor to interface stability. Salt bridges appear to be less frequent (~1 salt bridges per 1000Å^2)[109], and have about the same energetic contribution to the interface stability like hydrogen bonds (~0.6-1.5 kcal/mol). Disulfide bonds occur even less frequent but contribute 2-8 kcal/mol due to their covalent binding character. Hence, the main effectors for interface stability are the burial of hydrophobic atoms and non-covalent interactions between subunits.

Interface destabilizing effects are caused by a decrease in entropy which reduces the dissociation energy and as a consequence leads to less stable complexes. The calculation of the absolute entropic contribution is not yet solved, but can be estimated by the summation of translation, rotational, vibrational and symmetrical entropy.[107] Entropic contributions are rather mass and size dependent and are, in contrast to enthalpy contributions, only marginally influenced by a specific residue distribution of the interfaces.

Janin[110] compared the composition of groups of atoms in the interface compared to the surface. He observed that non-polar atom groups are observed more frequently in interfaces than in surfaces. The amino acid composition of interfaces has been investigated in detail in many publications[103,111-114]. The conclusion is that aromatic and aliphatic residues are more frequent in the interfaces, i.e. they occur on average twice as often in the interface than in the whole surface. Conversely, charged residues (with the exception of Arginine) are less frequent in the interface by the same order of magnitude.

3.1.3 Evolution of oligomeric complexes

Proteins with similar sequences are likely to also share their quaternary structure. In an earlier study, Alloy[115] found that the interaction between complexes is conserved for structures sharing more than 30% sequence identity. Later, Levy[116] has shown that in the range between 30%-40% sequence identity, 30% of the proteins described in the PiQsi[117] resource have a different quaternary structure and below 30% sequence identity half of the homologues changes their quaternary structure. Another study reported that the probability to find a pair of proteins with similar quaternary structure is given for the majority of cases if the identity between the sequences of the two proteins is greater than 50%.[118]

Levy[116] also investigated the occurrence of cyclic and dihedral symmetries during evolution. They found that whenever there is a chance to choose between cyclic or dihedral symmetries, an 11-fold preference for the latter is observed. This can be explained by the interface geometry, whereas for cyclic symmetries two interfaces were involved into complex assemblies (face-to-back), dihedral symmetries often consist of interfaces which are face-to-face (or back-to-back). See Figure 14 for a schematic representation.

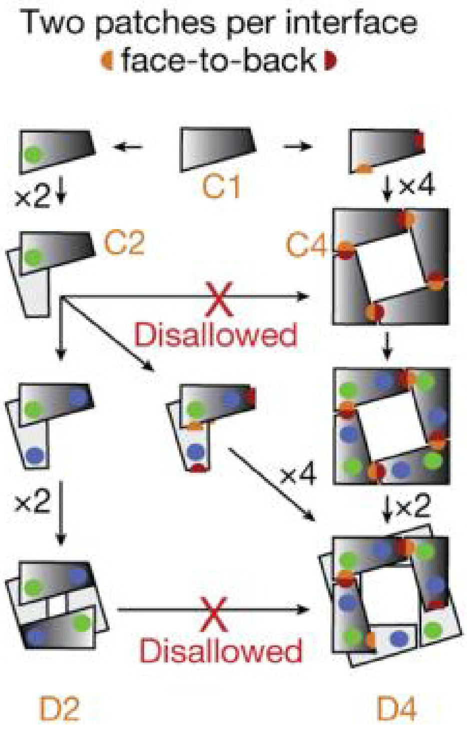


Figure 14 Evolutionary paths of dihedral and cyclic oligomeric assemblies. Assembly of dihedral complexes can take several pathes, whereas cyclic symmetries can evolve only by one pathway. Figure taken from Levy et al[116].

Assuming the same number of subunits, interfaces formed by dihedral symmetries are more likely to occur by random mutations than complexes with cyclic symmetries[119] [120]. Dihedral systems can evolve in multiple steps (C1-C2-D4) whereas cyclic system must evolve in one step (C1-C4), (See Figure 14,[116]). As a consequence complexes of dihedral symmetries evolve often through their cyclic intermediates.[116,121]

Amino acid substitutions may affect the stability of the protein complex interfaces and thereby lead to the assembly or disassembly of protein complexes. The exchange of surface residues by hydrophobic and large protruding residues may lead the formation of oligomers.[122] The design of oligomeric proteins based on single mutations supports this hypothesis[123]. One example is the mutation of a single surface residue to a nonpoloar residue in order to promote the assembly of a symmetric tetramer based on a dimer. Single point mutations can be disruptive for the function of the protein and the cause of several diseases. One example is the disease fructose intolerance caused by a mutation in the interface of the enzyme 1,6 biphosphate aldolase A. The consequence of this mutation is a decreased stability of the tetrameric protein, associated with a decreased activity of the enzyme.[124]

Another explanation for the modulation of oligomeric protein during evolution is the insertion and deletion of small fragments in the interfaces. Hashimoto[125] has shown that about one quarter of the insertions and deletions in homologue proteins is located in the interface and has an impact onto the stability of the complex.

Similar quaternary structure between two proteins implies to some extent evolutional pressure to the interface residues. Several studies have used the Shannon entropy to calculate the flexibility in evolution for a single interface residue on a multiple sequence alignment which consists of homologue sequences.[126-128] It has been shown that interface residues are more conserved than surface residues. Elcock & McCammon[126] uses the ratio between interface and surface conservation to distinguish between biological interfaces and crystal contacts, because the latter behave more like surface residues in respect to their evolutionary conservation.

3.1.4 Comparison of oligomeric complexes

The comparison of quaternary structures was the target of many studies.

For example Levy et al[102] use a graph-based approach to explore differences in topology between several types of oligomeric assemblies. Subunits are considered as nodes, whereas interfaces between subunits are defined as edges. In order to compare two complexes based on their topology, a modified version of the A* algorithm[129] is used. This method does not engage in a deeper evaluation of the interface geometry.

To describe differences in the relative orientation between subunits, several concepts were developed in recent years. Aloy et al[115] developed iRMSD (interaction RMSD) to calculate the similarity between two binary complexes. The RMSD is calculated by firstly superpose A on its equivalent chain in the other complex (A'). The same procedure is applied to chain B in order to get a superposition on B'. After superposition, 14 coordinates are used to calculate the final root mean square deviation: the center of mass of the subunit plus six additional points which were calculated by adding or subtracting 5Å to the x,y and z coordinates of the center of mass. Similar approaches involve only the interface residues for the superposition, or limit the superposition only to one chain, when calculating the RMSD[130].

For the evaluation of predicted protein complexes within the CAPRI experiment[131], the fraction of correctly predicted contacts is computed. This is defined as the number of correct residue-residue contacts in the predictions divided by all contacts in the native complex.[130] A more fine grained score was developed by Xu et al[132] for the comparison of homodimeric complexes. Q score calculates the weighted mean of differences in distances between equivalent residue pairs. Equivalent residue pairs are identified by a sequence alignment or structural superposition of the subunits. The weighting function decreases the influence of distant residue-residue contacts to the overall score.

3.1.5 Modeling of quaternary structures

Protein structure elucidation at atomic level is needed to determine the function of a protein in detail, e.g. to characterize the interface area between two subunits or identifying residues which are in contact with the ligand. However, there are only about 75 000 experimentally solved structures deposited in the Protein Database[20] compared to the number of known protein sequences, which is exponentially increasing: i.e. there are more than 17 million sequences in the UniProt Knowledge Database(release 2011_09, September 2011). One reason for this huge gap between the number of protein sequences and known protein structures is the development of high throughput sequencing methods in recent years. Comparative or homology modeling can help narrowing this gap. These methods are known to be the most accurate to calculate a protein structure based on evolutionary related protein (template) structures [2],[73]. Because of the large impact and abundance of protein complexes in nature, several studies has investigated evolutionary conservation of protein complexes

As described in the paragraph "Evolution of oligomeric complexes", several studies have shown that the quaternary structure between two evolutionarily closely related proteins can be assumed to be similar. In recent years, several methods were published which use this relation to model the quaternary structure of a structural unknown target proteins based on the quaternary structure of related template structures. The M-Tasser approach by Chen [133] was developed for the prediction of homodimers. It first builds a monomeric model using the protein prediction method TASSER[22] and then superposes the model onto the members of the template library in order to generate a dimeric complex. The generated dimers are then subjected to a refinement method, which improves the overall interface geometry. Weerayuth[134] developed Protinfo PPC, a webserver which queries its template library using PSI-BLAST[32] and SSEARCH[135] to identify template structures which match the submitted query sequences. The complexes in the template library are based on biological unit files provided by the protein database. Models are built based on a multiple sequence alignment which is built between the target sequence and homologues template sequences. Additional homologue sequences from the UniProt sequences are added to increase the overall alignment quality. BISC[136] calculates models based on experimentally verified protein-protein interactions from functional genomics databases. The sequences of the interacting domains are used to query a template library which is based on PISA[107] (see next section for details) and the identified template structures are processed using MODELLER[60].

Further, interface residues are predicted by sequence alone. For example, ISIS[137] uses different input features to train a neural network.

Another approach to calculate the structure of protein complexes are protein-protein docking procedures. Most methods rely on an exploration step to determine the initial configuration of the involved subunits. Later, near-native solutions are identified using scoring functions which incorporates energy, geometric complementarity, propensities, and other terms which distinct interfaces from surfaces. Promising candidates are subjected to further refinement steps. In addition methods like HADDOCK[138] or RosettaDock[139] allow the predictor to incorporate additional data like NMR data, sequence conservation or mutation data to restrain the conformational search.

The accuracy of docking routines is assessed in the CAPRI experiment (Critical Assessment of Predicted Interactions)[131], which is organized in a similar way as CASP[140]. Predictors are asked to submit their predictions for protein complexes which are structurally characterized but not yet public available. The accuracy of the submitted models is then evaluated by assessors, which do not know the real names of the predictors. Practically, docking techniques are often used for exploring transient interaction between structures which are already known.

3.1.6 Annotation of quaternary structures

Despite recent advances of protein structure determination by X-Ray crystallography the true biological active state of a protein often remains unknown. The correct quaternary state must hence be estimated by analyzing the given conformations in the crystal cell if no additional information from direct solution experiments is available. In addition to its native state, the protein in a crystal often undergoes interactions caused by the dense packing. Hence, it is important to distinguish true biological interactions from interactions caused by crystallization (so called "crystal contacts"). This effect cannot be observed for structures solved by NMR. Nevertheless only few NMR structures address oligomeric proteins or protein-protein interactions.

It has been observed that the extent of the buried surface area upon complex formation is a very important descriptor to distinguish crystal contacts from biological interfaces[141]. Contacts caused by crystal packing bury in many cases less surface area than biological interfaces. Since this is not always the case, methods were developed to incorporate other properties of physiological interfaces like evolutionary conservation, electrostatic potentials, hydrophobicity, shape complementary and aminoacid composition. Machine learning techniques like neural networks[142], random Forests[143], Support Vector Machines[144] or Bayes classification[145] were frequently used to combine such attributes. Other approaches try to verify quaternary structure annotation by literature review[117,146], or consider the crystal packing of homologue proteins[147] to decide if a particular interface is biological relevant.

However, many of these studies rather describe theoretical concepts and do not apply their methods to structural databases in a regular and up-to-date fashion. Annotation methods which apply their prediction method in a consequent way to new released structures are the Biounit annotation within the PDB entry itself and PISA (Protein Interfaces, Surfaces and Assemblies; [148]). PQS was disabled in summer 2010 and replaced by PISA as the main quaternary structure annotation system supported by the European Bioinformatics Institute.

PISA

PISA[148] estimates the thermodynamical stability of protein complexes by calculating a pseudo dissociation energy. This includes an enthalpic (interface stabilizing) and an entropic (interface destabilizing) term. Interface stabilizing contributions are hydrogen bonds, salt bridges, disulfide bridges and buried hydrophobic area between interacting subunits. In contrast, the entropic term consists of translational, rotational and symmetrical entropy. Complexes which have positive dissociation energy are considered to be stable. The final score is trained and optimized on a set of protein complexes with manually characterized oligomeric states and reaches an average accuracy of 83%.[148] If more than one complex can be built by applying the symmetry operators to the chains in the asymmetric unit, the assemblies are ranked according to their oligomeric state (larger assemblies supersede smaller assemblies). Thus, one configuration is always ranked first. Predictions for all structures solved by X-ray in the current release of the PDB can be accessed and downloaded via the EBI website (www.ebi.uk.co). PISA replaced PQS as the default tool for the automated annotation of biological assemblies in the PDB.

PDB

The quaternary structure annotation of a deposited protein structure can be found in the REMARK300/350 sections of the header. The REMARK300 section contains information of experiments which were used to determine the correct oligomeric state or other information about quaternary structure assignments. This section is free text and can only be hardly used for automated procedures. Nevertheless it contains often important information (e.g. if a particular complex can be supported by other experiments). The subsequent part ("REMARK350") consists of information on how to build the biological unit(s) using the chain(s) in the asymmetric unit file. According to the PDB, all likely quaternary structures, which can be built by applying the symmetry operators of the crystal cell are computed and reported ("SOFTWARE DETERMINED QUATERNARY STRUCTURE"). The applied software is PQS (for earlier deposited entries) or PISA. If such an assembly is considered to be biological relevant by the authors, a corresponding REMARK can be given ("AUTHOR DETERMINED BIOLOGICAL UNIT"). Additionally, the matrices to build the biological unit (i.e. translation and rotation matrices for all chains in the asymmetric unit) are denoted.

An example for the REMARK300/350 section is given in Figure 15. The authors of a dimeric histidine triad protein from *Sinorhizobium meliloti 1021* (PDB-ID: 3nrd) give additional information about their preference for a particular quaternary structure.

```
REMARK 300
REMARK 300 BIOMOLECULE: 1, 2, 3, 4
REMARK 300 SEE REMARK 350 FOR THE AUTHOR PROVIDED AND/OR PROGRAM
REMARK 300 GENERATED ASSEMBLY INFORMATION FOR THE STRUCTURE IN
REMARK 300 THIS ENTRY. THE REMARK MAY ALSO PROVIDE INFORMATION ON
REMARK 300 BURIED SURFACE AREA.
REMARK 300 REMARK: CRYSTAL PACKING AND ANALYTICAL SIZE EXCLUSION
REMARK 300 CHROMATOGRAPHY ANALYSES SUPPORT THE ASSIGNMENT OF A DIMER AS A
REMARK 300 SIGNIFICANT OLIGOMERIZATION STATE IN SOLUTION.
REMARK 350
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW.  BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN.
REMARK 350
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: DIMERIC
REMARK 350 SOFTWARE DETERMINED QUATERNARY STRUCTURE: DIMERIC
REMARK 350 SOFTWARE USED: PISA
REMARK 350 TOTAL BURIED SURFACE AREA: 5160 ANGSTROM**2
REMARK 350 SURFACE AREA OF THE COMPLEX: 11880 ANGSTROM**2
REMARK 350 CHANGE IN SOLVENT FREE ENERGY: -124.0 KCAL/MOL
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 350
REMARK 350 BIOMOLECULE: 2
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: DIMERIC
REMARK 350 SOFTWARE DETERMINED QUATERNARY STRUCTURE: DIMERIC
REMARK 350 SOFTWARE USED: PISA
REMARK 350 TOTAL BURIED SURFACE AREA: 3820 ANGSTROM**2
REMARK 350 SURFACE AREA OF THE COMPLEX: 12060 ANGSTROM**2
REMARK 350 CHANGE IN SOLVENT FREE ENERGY: -99.0 KCAL/MOL
REMARK 350 APPLY THE FOLLOWING TO CHAINS: C, D
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 350
REMARK 350 BIOMOLECULE: 3
REMARK 350 SOFTWARE DETERMINED QUATERNARY STRUCTURE: TETRAMERIC
REMARK 350 SOFTWARE USED: PISA
REMARK 350 TOTAL BURIED SURFACE AREA: 10420 ANGSTROM**2
REMARK 350 SURFACE AREA OF THE COMPLEX: 22500 ANGSTROM**2
REMARK 350 CHANGE IN SOLVENT FREE ENERGY: -238.0 KCAL/MOL
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C, D
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 350
REMARK 350 BIOMOLECULE: 4
REMARK 350 SOFTWARE DETERMINED QUATERNARY STRUCTURE: OCTAMERIC
REMARK 350 SOFTWARE USED: PISA
REMARK 350 TOTAL BURIED SURFACE AREA: 26040 ANGSTROM**2
REMARK 350 SURFACE AREA OF THE COMPLEX: 39790 ANGSTROM**2
REMARK 350 CHANGE IN SOLVENT FREE ENERGY: -581.0 KCAL/MOL
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C, D
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT1   2  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   2  0.000000 -1.000000  0.000000        0.00000
REMARK 350   BIOMT3   2  0.000000  0.000000 -1.000000       85.24100
```

Figure 15 REMARK300/350 section of 3nrd, a histidine triad protein. Additional information about the performed experiments is given in the REMARK300 section. All relevant PISA predictions are annotated in the REMARK350, however the author supports only the dimer hypothesis.

In the REMARK350 section four different assemblies were specified. All assemblies were calculated using the PISA algorithm. The oligomeric state was predicted by PISA to be dimeric ("BIOMOLECULE 1,2"), tetrameric ("BIOMOLECULE 3") or octameric ("BIOMOLECULE 4"). Based on their own analysis (see REMARK300 section) the author decided to assign dimeric to be the most likely quaternary state.

This example highlights several issues. Firstly, more than one quaternary structure can be assigned for the chains present in the asymmetric unit. Thus, a unique annotation for a given set of chains cannot be assumed. Secondly, the annotation given by the author can support one or many of these assemblies, none of them or can suggest other assemblies which were not predicted by the software.

Finally, if the authors do not have evidence for one of the proposed assemblies, annotation by the authors is absent.

This results in a mixture of sources of quaternary structure annotation. Different types of software can be used to calculate assembly structures based on the asymmetric unit. Author annotation can then support one of these hypotheses.

PiQsi

Another resource of quaternary structure information is the manually curated PiQsi database[117]. PiQsi relies on the analysis of literature and closely related homologue structures. The structural sources for the annotation are biological unit files provided by the PDB, annotated either by the authors or by the automated approaches like PQS and PISA. The human annotator can choose between the following states: "NO", "PROBNOT", "PROBYES" and "YES" depending on his own investigation and interpretation of the data:

1. "NO" indicates that the quaternary structure as annotated by the PDB is correct.
2. "PROBNOT" indicates that the quaternary structure as annotated by the PDB is "likely" correct.
3. "YES" indicates an erroneous assignment by the PDB.
4. "PROBYES indicates a likely erroneous assignment by the PDB.

The curator's assignment of states depends on his investigation and interpretation of the data. Additionally, the annotator can create his own annotation. However, the deposition of the corrected coordination file is not possible. Unfortunately, PiQsi annotation is only available for around 15 000 structures, which is less than one fifth of the currently available structures in the PDB and currently, no new annotations are added.

3.2 Definition of the problem

Modeling of proteins in their correct quaternary structure results in a more detailed view of their biological function. If the concept of comparative or homology modeling is applied, the quaternary structure must be deduced from homologue template structure. The following questions are addressed:

Firstly, the accuracy of state-of-the-art template based oligomer modeling methods needs to be evaluated in order to identify the weakness and strengths of already existing methods.

Secondly, similarity scores need to be developed to elucidate the similarity in quaternary structure between two homologues proteins. It is thereby important to incorporate the oligomeric state as well as the geometrical accuracy between the interfaces. A score which classifies the quaternary structure between two proteins as similar or dissimilar is essential in order to develop methods for the prediction of quaternary structures.

Thirdly, it has to be analyzed which descriptors are required to identify template structures, which share their quaternary structure with the target protein. The quaternary structure of template structures needs to be estimated using automated annotation tools.

3.3 Material and Methods

Assessment of quaternary structures in CASP9

Target preparation

For all CASP9 TBM targets, we determined the most probable biological active quaternary structure in the following way: For the definition of the oligomeric assembly state, we relied primarily on the assignment by the authors (“REMARK 350”). For targets solved by NMR, having no “REMARK 350” section, the oligomeric state was defined by their assembly of chains in the PDB entry. Targets without or with ambiguous assignments by authors were inspected manually taking into account PISA annotation[148] and the “REMARK 300” section. Targets with ambiguous assignment of the oligomeric state which could not be resolved satisfactorily by visual inspection were excluded from the evaluation. For two targets, the structure was not yet deposited in the PDB. Table SI provides the oligomeric state assignment for all targets used in this assessment.

Coordinate sets representing the biological units were downloaded from the PDB protein database or PISA respectively using the PDB code for the targets reported on the CASP9 target website. Residues in the experimental structure of the oligomeric assembly were mapped to the CASP target sequence chain-by-chain, and only amino acid residues corresponding to the CASP target sequence were included.

Oligomeric Predictions

Predictions were considered as oligomer predictions if a model consisted of multiple chains, and the oligomeric state of a prediction was interpreted as the number of chains found in the corresponding coordinate file submitted to the prediction center. Groups with at least one oligomeric model submission were included in the evaluation. Groups “55 MUFOLD-MD”, “117 3-D JIGSAW_V4-5” and “333 DELCLAB” submitted models with inconsistent chain naming and were therefore excluded from this evaluation. Human groups were evaluated using the targets labeled as “human/server”, Server groups on all targets. Group “353 SAMUDRALA” (registered as “human”) submitted in total only one oligomeric prediction (T0516) which is classified as “server” and was therefore not included in the assessment.

Numerical Oligomeric State Assessment

We calculated the fraction of correctly predicted oligomeric states (dimer, trimer, tetramer, etc.) by normalizing with the maximum number of oligomeric structures, either in the target or in the prediction set, in order to account for over-prediction of oligomeric states:

$$Acc_{Oli} = \frac{number\ of\ correctly\ predicted\ multimeric\ targets}{\max(number\ of\ multimeric\ targets, number\ of\ multimeric\ predictions)}$$

In order to assess the quality of the structure of the predicted complex, we evaluated how accurately the interfaces of the oligomeric complexes were modeled. This analysis accounts for the correct number of interfaces as well as the correct orientation by calculating a “Contact Agreement Score” S$_{agree}$ which reflects the fraction of correctly modeled interface contacts in the complex:

$$S_{agree} = \frac{\sum_{i,j} f(x_{ij}, y_{ij})}{\sum_{i,j} g\,(x_{ij}, y_{ij})}$$

$$f(x_{ij}, y_{ij}) = \begin{cases} 1 - \dfrac{|x_{ij} - y_{ij}|}{\max(x_{ij}, y_{ij})}, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

$$g(x_{ij}, y_{ij}) = \begin{cases} 1, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

With x_{ij}, and y_{ij} being the total number of contacts between residues *i* and *j* in the target sequence. Two residues were considered to be part of the protein-protein contact interface, if they were located in different chains and their corresponding Cβ atoms (Cα for GLY) were less than 12 Å apart. S_{agree} ranges from 0 and 1, with $S_{agree} = 1$ indicating that all contacts in the target complex are present in the model. $S_{agree} = 0$ indicates, that none of the contacts in the target complex are present in the model. Note that S_{agree} is undefined if either one of the two compared structures is monomeric, and was set to zero in this case.

Naïve oligomeric assembly predictors

To be able to estimate the difficulty of the different targets, two naïve predictors were included in the assessment: group "998 NaïveSeqId", and group "999 NaïveCoverage". HHSearch[40] was used to identify homologue template structures in the PDB, and all hits showing sequence identity with the target of less than 15% or coverage less than 15% were discarded. Group "998 NaïveSeqId" sorted possible templates first by highest sequence identity in the target-template alignment and second by coverage of the target sequence by the alignment, giving precedence to the first criterion, and selected the highest ranked template. Conversely, group "999 NaïveCoverage" selected the model, where its target-template alignment reached the highest coverage to the target sequence, and within the same coverage had the highest sequence identity, in this order of importance.

For the selected template, the first oligomeric assembly assigned by PISA[148] was used to build oligomeric pseudo-models. Models were built by copying the backbone atoms and the Cβ atoms (except for Glycine) of the aligned regions in the sequence alignment

Development of a score to derive similarity between homologous quaternary structures

Similar to the procedure described by Xu[132], an interface was defined if two chains have at least 10 residue contacts and at least one atomic contact. A residue contact was defined as two Cβ-atoms with a distance less than 12Å. An atomic contact was defined if two non-hydrogen atoms were not more than 5Å apart.

All accessible surface areas were calculated using NACCESS[149] with its default settings.

Analysis of the QscoreOligomer in dependence to the similarity of the quaternary structure was performed on the same dataset as described in the next section.

Error bars were generated by calculating 100 intervals using 70% of the targets in the set.

Development of a method for template based modeling of oligomeric protein structures

General definitions

Definition of interfaces

Accessible surface areas were calculated using NACCESS[149] with its default settings. Residues were labeled as “surface” if at least 5% of its relative surface area can be accessed. A residue was labeled as “interface” if at least 0.1Å^2 were buried. A residue was considered as “core” residue if at least one atom was fully buried. The total buried surface area per chain was defined as follows:

$$BSA_{Chain} = \sum_{i \leq chains} ASA_i - ASA_{Complex}$$

We defined an interface to be biological relevant if at least 500 Å^2 were buried.

Classification procedure

Random forests[150] were used with its default configuration (500 trees, number of variables used at each split: 5, “random Forest” library[151] within R) for all classification procedures. Unlike otherwise stated, two thirds of the targets in the dataset were used for training and one third for testing. In order to calculate the sensitivity and specificity we used the probabilities if being correct as emitted from the random forest.

Classification accuracy

Three different score were applied in order to calculate the classification accuracy:

- $Specificity = \frac{TN}{TN+FP}$
- $Sensitivity = \frac{TP}{TP+FN}$

With TP, TN, FP, FN, being true positives, true negative, false positives and false negative respectively.

- AUC ("area under curve")

The raw class probabilities of the classifier were used as input to visualize ROC curves and to calculate the area under curve (ROCR package[152] in R).

Oligomeric Template Library

Benchmark sets

The literature test set was compiled by using the dataset published by Gorin and Bordner[143]. Entries which had distinct SEQRES sequences were removed from the set. The PiQsi test set was compiled by downloading the complete list of all annotated protein structures in PiQsi[117]. All proteins which had differing SEQRES entries within their PDB header were removed, the remaining set of consisted of 10 865 sequences and was submitted to the PISCES server[153] and culled on 70% sequence identity level (max R-factor:0.3, maximum resolution: 3.0Å).

Quaternary Structure annotation

The PISA annotation was derived from the corresponding XML file downloaded from the EBI website (http://www.ebi.ac.uk/msd-srv/prot_int/pi_download.html). The top ranked annotation was used to determine the oligomeric state by counting the number of macromolecular chains in the XML file. If no XML file was found, the structure was flagged as "No annotation".

PDB files were downloaded from the official RCSB repository. The PDB author annotation was extracted from the "REMARK350" section (given as "AUTHOR DETERMINED BIOLOGICAL UNIT").If the authors did not assign an oligomeric state to the structure or one chain was annotated by the authors in more than one "biological unit", the structure was flagged as "No annotation". The oligomeric state was derived by the counting the number of chains annotated in the "biological unit" multiplied by the number of translation/rotation matrices.

Annotations by PiQsi were derived using the field ("No. sub (corrected)") in the annotation file provided by PiQsi[117].

Dataset

Target proteins

The dataset of target proteins was compiled using quaternary structure annotations of PISA[107] and PiQsi[117].

Only those entries which are labeled “NO” in their PiQsi error state (i.e. the annotation in the “REMARK 350” section of the PDB is correct) and the oligomeric state is confirmed by PISA (i.e. using the “first” annotation in the corresponding assembly XML file) were used. If a complex consists of more than one chain in the asymmetric unit, the entry was only kept if all chains had the same SEQRES sequence (=Homo-oligomer). The remaining 5328 sequences were submitted to the PISCES server [153] and culled on a 30% sequence identity level, using only entries with a maximum resolution of 3.0Å and R-factor of maximum 0.3.

Template identification

To detect homologue template structures a profile-profile search of the target sequence against the oligomer template library using the HHSearch Package[40] was performed. Only templates with a reliable sequence alignment, all hits having a P-value greater than 50, a sequence identity greater than 15% and a minimum of 20% coverage of the target sequence were retained. If multiple instances of the same PDB target were found (e.g. a dimer appears more than once in the asymmetric unit), the one with the highest coverage to the target sequence was used. Only monomeric or homo-oligomeric template structures were considered (see definitions above).

Model selection

Models were calculated by the following procedure:

1. Conserved Residues were copied using the coordinates of all heavy-atoms in the template structure
2. Non-conserved residues were modeled by copying only the main chain atoms. In an additional step sidechain-atoms were modeled using SCRWL4[154] with its default settings.

The relative ratio between surface and interface area was defined as follows:

$$ASA_{ratio} = \ln \frac{ASA_{Chain}}{BSA_{Chain}}$$

The largest chain in the complex was selected to calculate $\text{ASA}_{\text{ratio}}$. $\text{Asa}_{\text{Chain}}$ denotes the accessible surface area where as $\text{BSA}_{\text{Chain}}$ reflects the buried surface area of that chain.

Grouping of templates with similar quaternary structure

To detect groups of similar quaternary structures agglomerative hierarchical clustering was applied.

Firstly, all templates were considered as single groups. Two groups were merged if their average QscoreOligomer was greater than 0.5. The average QscoreOligomer between two groups was defined as follow:

$$QscoreOligomer_{G1G2} = \frac{1}{n_{G2}n_{G1}}\sum\nolimits_{x\in G1}\sum\nolimits_{y\in G2} QscoreOligomer(x,y)$$

G1,G2 stand for group 1, group 2 respectively and x,y for template 1 in group1, and template 2 in group 2.

For sequence clustering CD-HIT [155] was used in its default settings.

Evolutionary conservation of interfaces

To calculate the multiple sequence alignments (MSA), the SEQRES sequence of the target protein was searched against the UniRef100[156] using three iterations of PSI-BLAST[32]. A sequence was only added to the MSA if it covered at least 80% of the target sequence and had a maximum e-value of 0.001. The template alignment sequence was mapped on to the complete target sequence. Thus, deletions in the target sequence were removed.

The Shannon entropy[157] S_{Column} of each alignment column was calculated using either all 20 standard residues types or 6 groups of physiological similar residue types (as defined by Guharoy [127]):

$$S_{column} = \sum_i p_i \ln p_i$$

The average conservation was calculated by weighting the column entropies with the corresponding accessible surface area, buried surface area for surfaces $\langle S_{surface}\rangle$ and interfaces $\langle S_{interface}\rangle$ respectively:

$$\langle S_{surface}\rangle = \frac{\sum_i asa_i S_i}{\sum_i asa_i}$$

$$\langle S_{interface}\rangle = \frac{\sum_i bsa_i S_i}{\sum_i bsa_i}$$

The log-ratio of entropy between surface and interface residues was finally defined as followed:

$$S_{conservation} = \ln(\frac{1+\langle S_{interface}\rangle}{1+\langle S_{surface}\rangle})$$

$S_{conservation}$ is defined as the log-ratio between the average interface and surface entropy. The addition to one prevents undefined values if $\langle S_{interface} \rangle$ or $\langle S_{surface} \rangle$ are zero.

To calculate $S_{conservation}$ on different multiple sequence alignments, all sequences were sorted according to decreasing evolutionary distances. We then decreased successively (in steps of 5% sequence identity) the level of sequence identity which is included into the multiple sequence alignment. To distinguish models with correct quaternary structure from models with incorrect quaternary structure, a random Forest was trained using the 20 $S_{Conservation}$ as input variables. The binary output variable was set 0 if the quaternary structure between template and target was different (QscoreOligomer <0.5) and to 1 if the quaternary structure of the template was similar to the template. To ensure that coverage is not the main reason for being labeled incorrect, we excluded all templates structures if the sequence alignment covers not more than 80% of the target sequence. Additional we excluded all template structures burying less than 500Å^2 (BSA_{Chain}).

QMEAN

For the analysis of the discriminative power of mean force potentials the normalized version QMEAN4[51] was used. Structural deviation between model and target structure were calculated by the program TMscore[158].

3.4 Results & Discussion

3.4.1 Assessment of oligomeric models in CASP9

The section "Assessment of oligomeric models in CASP9" was published recently as part of the official CASP9 TBM assessment.[7]

Evaluation of oligomer modeling

Many proteins form stable higher order quaternary structures in the form of complexes or oligomeric assemblies. In fact, proteins which form exclusively monomeric structures are a minority, while the majority of proteins in a cell is involved in complexes and assemblies in some form[159]. The prediction targets in CASP do not form an exception, and a large fraction of them is forming stable oligomeric assemblies in their native state: 53 prediction targets in the assessment were homo-oligomeric complexes, while only 43 were monomers[1]. Frequently, it is only in the context of this assembly that we can understand their function, e.g. when active sites are located in the interface between subunits, and the protomeric structure is often not sufficient to study functional aspects in an accurate and complete fashion. It is also often observed that the isolated subunits do not represent self-sufficient globular structures without the interactions with their neighboring subunits. In these cases, predicting (and also assessing) targets as monomeric structures does not make much sense.

The task of predicting protein structures in CASP includes prediction of the quaternary structure. According to the CASP format definition, quaternary structure predictions should be submitted in the same frame of reference, with the first chain labeled as A and subsequent chains following the latin alphabet, e.g. a tetramer's chains should be labeled as A, B, C, D. Here, we evaluated how well predictors identified the correct oligomeric state of the target, and how accurately the predictions resembled the structure of the complexes. To be able to estimate the difficulty of the different targets, two naïve predictors were included in the assessment: group "998 NaïveSeqId", and group "999 NaïveCoverage". These naïve predictors assume that the oligomeric state of the target is the same as the one of the closest template which can be identified by standard sequence search methods. In case of group "998 NaïveSeqId", the template with the highest sequence identity to the target was selected, in case of group "999 NaïveCoverage", the one covering the largest fraction of the target sequence.

The oligomeric state of CASP9 targets ranges from "Monomeric" to "Tetrameric", with a majority of multimeric targets. For multimeric targets the most abundant state is "Dimer" (41 Targets), followed by "Tetramer" (9) and "Trimer" (3). Hetero complexes were not observed. The "human/server" category consists of 20 monomers, 16 dimers, 3 trimers and 2 tetramers.

We included all groups in the evaluation which submitted at least one model as oligomer. The majority of targets was predicted as monomeric (83% "human/server", 78% "server" category), followed by dimeric predictions (13% "human/server", 18% "server"). This indicates that many of the groups only submitted a fraction of their predictions as oligomers. Only for a small number of groups does the oligomeric state distribution of the predictions resemble the one of the experimental target structures (See Figure 16).

[1] Some targets, for which the oligomeric state assignment by the experimentalist was ambiguous, were excluded from this part of the assessment.

Successful modeling of oligomeric complexes relies on the identification of the correct oligomeric state of the unknown target protein, and then on constructing a realistic model for the quaternary structure. The fraction of correctly identified oligomer states over the number of submitted models (Acc_{Rel} column in Table 2) ranges between 79% ("282 Taylor") to 36% ("20 dokhlab") for "human" and 66% ("452 Seok-server") to 45% ("102 Bilab-ENABLE") for "server groups", respectively. However, these numbers are dominated by the assignment of monomeric structures. In the context of CASP it is not possible to determine if a monomeric submission was an explicit choice for monomer or just a "non-oligomeric" prediction. (If a predictor would submit only monomeric models, an accuracy of 47% and 53% would be achieved for "human/server" and all targets, respectively.) Also, submitting oligomeric predictions only for a small set of targets ("cherry picking") increases the chance to achieve good accuracy in this measure.

We therefore calculated the fraction of correctly predicted states for oligomeric targets only, normalizing by the maximum number of oligomeric structures, either in the target or in the prediction set (See Materials and Methods for details). Two human groups ("458 Bilab-solo", "242 Seok") and the two naïve predictors predicted more than 50% of their oligomeric targets correctly. Remarkably, group "458 Bilab-solo" was able to characterize more than three quarters (76%) of the oligomeric states in the "human/server" targets correctly. In the "server" category, the naïve predictor based on sequence identity, classified most accurately 55% of the targets (See Table 2).

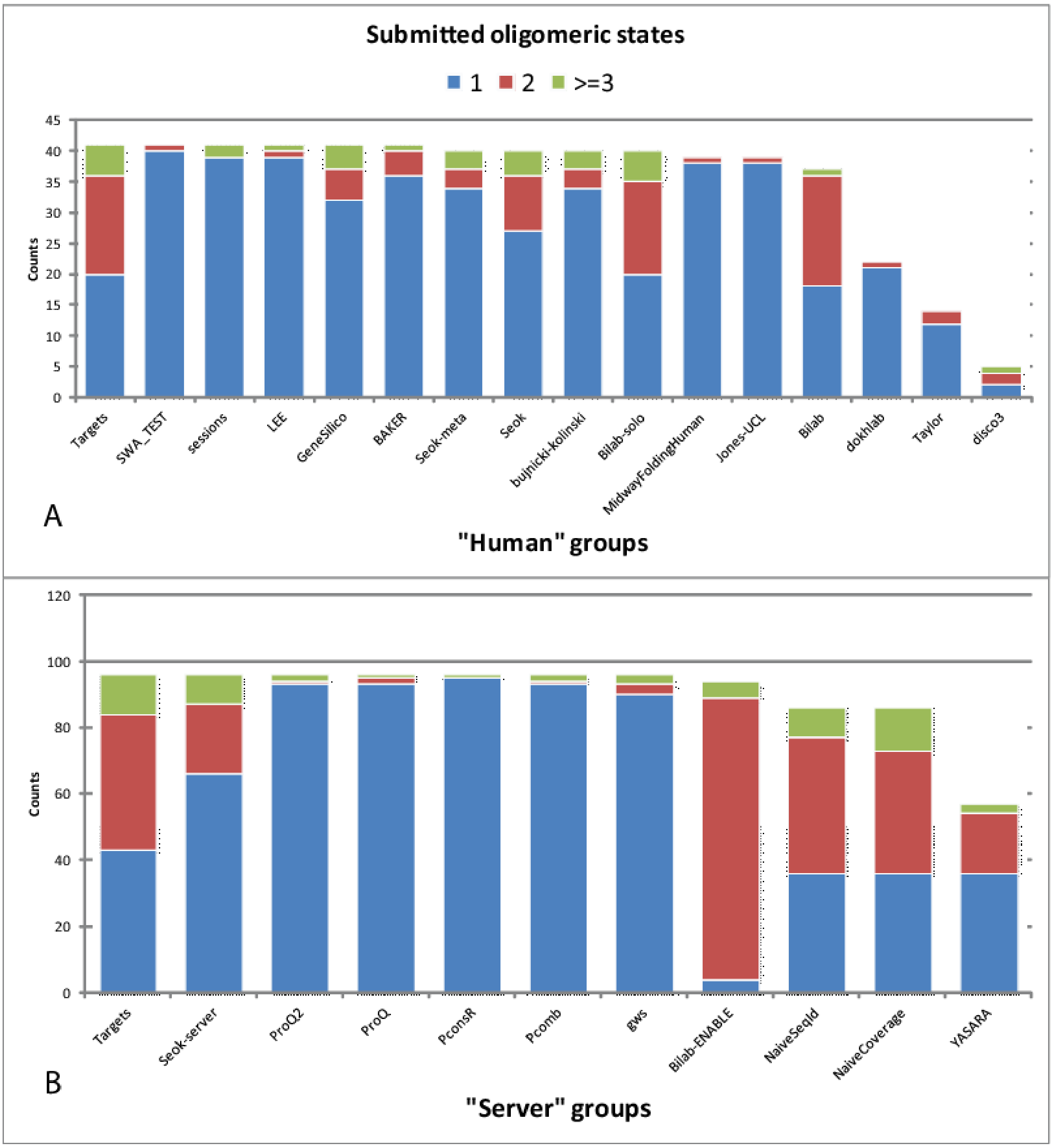


Figure 16 Distribution of oligomeric states amongst the submitted models. Panel A shows the fraction of different oligomeric states in the set of predictions made by "human" groups for the targets in the "human/server" category. Panel B displays the same data for the set of the submissions made by "server" groups for all evaluated targets. Both panels show on the left the actual distribution of the oligomeric states in the experimental target structures for comparison (see text).

In our visual inspection of the predictions, we observed a substantial fraction of physically impossible models with parts of the structures overlapping, severe steric clashes, or isolated chains in space - lacking a protein-protein interface. We calculated the fraction of models which contained more than 10 backbone-backbone clashes or were lacking inter-chain contacts (all C-β –C-β distances > 12Å). In general, server groups tended to build more unrealistic models than human groups. Among the groups having at least 5 oligomeric predictions, the server group “102 Bilab-ENABLE” had with 36% the highest fraction of unrealistic models (SeeTable 2).

Table 2 Summary table of oligomeric predictions

Category	Group	Group#	∑oligomer predictions	Total	$\mathrm{Acc}_{\mathrm{Rel}}$	$\mathrm{Acc}_{\mathrm{Oli}}$	% unrealistic	$\sum S_{\mathrm{Agree}}$
Human	Seok-meta	16	6	40	57.5%	23.8%	0.0%	1.2
Human	dokhlab	20	1	22	36.4%	0.0%	0.0%	0.0
Human	Jones-UCL	104	1	39	53.8%	4.8%	0.0%	0.0
Human	LEE	114	2	41	51.2%	4.8%	20.0%	0.6
Human	GeneSilico	147	9	41	61.0%	28.6%	11.1%	2.3
Human	BAKER	172	5	41	61.0%	23.8%	0.0%	1.8
Human	Seok	242	13	40	75.0%	57.1%	7.7%	3.1
Human	sessions	278	2	41	51.2%	4.8%	0.0%	0.4
Human	Taylor	282	2	14	78.6%	9.5%	0.0%	1.3
Human	SWA_TEST	297	1	41	51.2%	4.8%	0.0%	0.8
Human	bujnicki-kolinski	299	6	40	62.5%	23.8%	0.0%	2.2
Human	Bilab	423	19	37	51.4%	42.9%	15.8%	1.3
Human	disco3	439	3	5	60.0%	9.5%	0.0%	1.0
Human	Bilab-solo	458	20	40	77.5%	76.2%	0.0%	4.9
Human	MidwayFoldingHuman	477	1	39	48.7%	4.8%	0.0%	0.6
Server	ProQ2	1	3	96	46.9%	3.8%	33.3%	0.9
Server	Bilab-ENABLE	102	90	94	44.7%	45.6%	35.6%	7.6
Server	PconsR	173	1	96	44.8%	0.0%	100.0%	0.1
Server	YASARA	228	21	57	63.2%	28.3%	9.5%	5.1
Server	Gws	236	6	96	50.0%	9.4%	33.3%	1.9
Server	Pcomb	273	3	96	46.9%	3.8%	33.3%	1.0
Server	ProQ	296	3	96	46.9%	3.8%	0.0%	0.8
Server	Seok-server	452	30	96	65.6%	43.4%	6.7%	6.4
Server	NaiveSeqId	998	50	86	60.5%	54.7%	20.0%	8.1
Server	NaiveCoverage	999	50	86	57.0%	50.9%	8.0%	9.8

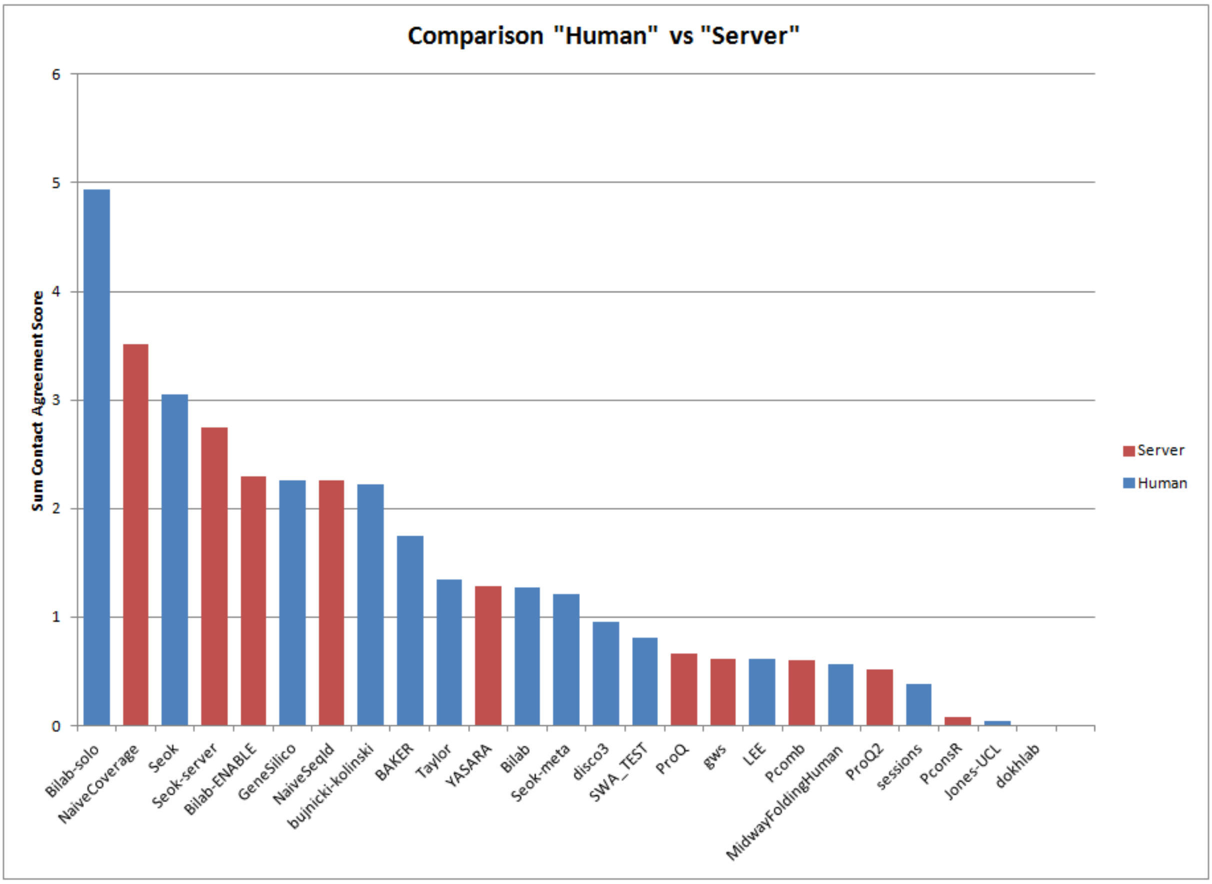


Figure 17 Interface agreement scores S(Agree) summed over all targets in the “human/Server” category. “458 Bilab-solo” clearly outperforms all other groups. In general, predictions by “human” predictors (blue) are not more accurate than “server” groups (red), and only “458” outperforms a naïve control predictor in this analysis.

In order to characterize the accuracy of the predicted oligomeric structures, we calculated a “Contact Agreement Score” S_{agree} which reflects the fraction of correctly modeled interface contacts in the complex and thereby accounts for the correct number of interfaces as well as their correct orientation (see Materials and Methods). S_{agree}–score ranges from 1 for a perfectly predicted complex to 0 for a completely incorrect one. To estimate the overall performance for each group, we summed up the S_{agree}–score for each target. This procedure rewarded successful modeling of the complex but did not penalize unsuccessful attempts, which also accounts for the fact that is was often not clear if a “single chain” submission should be considered as explicit choice for a monomer or no attempt was made to predict the oligomeric state. When considering the “human/server” subset of targets, group “458 Bilab-solo” submitted overall the most accurate predictions (see Figure 10,$\sum$ $S_{agree} = 4.9$). By evaluating the server groups on all targets, the naïve predictor which uses the template with best coverage was the most accurate (“999 NaïveCoverage”), followed by the second naïve predictor (“998 NaïveSeqId”). To investigate if human predictors were able to submit more accurate models than servers, we compared server and human groups on the subset of “human/server” targets. Among the best five groups there were two human predictors, two server predictors and the naïve predictor “999 NaïveCoverage”. The human group “458 Bilab-Solo” outperformed all other groups, but a general trend that human groups predict more accurately could not be observed. For most targets (except T0584, T0517, T0598), “458 Bilab-solo” predictions achieved an accuracy similar to the top predicting groups and thus performed on a constant level (Figure 18).

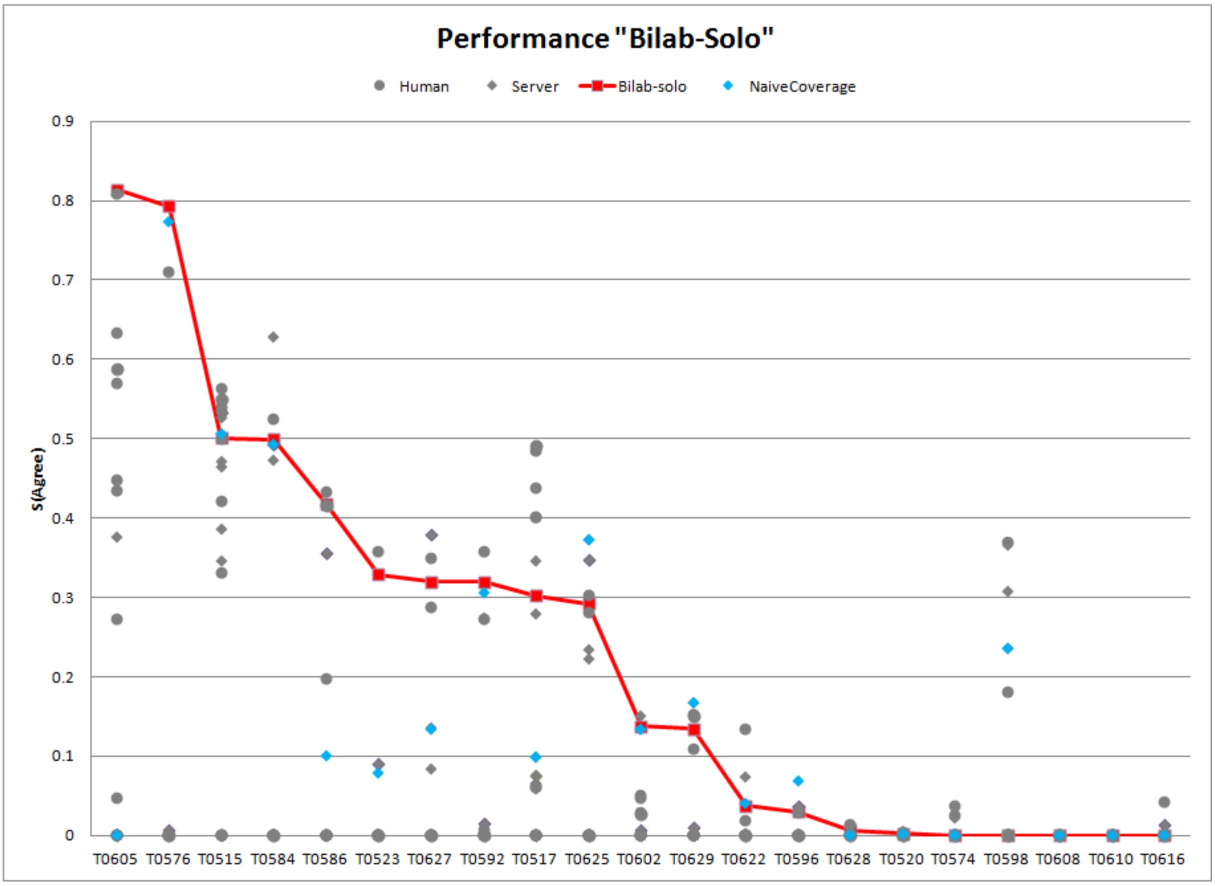


Figure 18 Interface agreement scores S(Agree) summed over all targets in the “human/Server” category. “458 Bilab-solo” clearly outperforms all other groups. In general, predictions by “human” predictors (blue) are not more accurate than “server” groups (red), and only “458” outperforms a naïve control predictor in this analysis.

Despite the fact that the quaternary structure is often essential for the understanding of the biological function of a protein and more than half of all CASP9 targets are oligomers, only a small fraction of the participating groups in CASP9 submitted oligomeric models. The overall performance of the participating server groups compared to the two naïve predictors was rather poor. As shown in Figure 16B most of the sever groups submitted mainly monomeric submissions, except group “102 Bilab-ENABLE”, who submitted mainly oligomeric models. Unfortunately, a significant part of these models contained clashes or have non-interacting subunits. In general, the accuracy of the server oligomer predictions appeared rather low. The high rate of unrealistic models reveals that the complexity of oligomeric modeling is currently not handled properly by automated approaches. Obviously, there is a great opportunity for significant improvement in modeling of quaternary structures of proteins in future CASPs.

Example of an oligomer target

Ignoring the quaternary structure of a protein can lead to models which cannot explain important physiological properties, or even to structures with a disrupted functional site. One example is T0576 (PDB 3NA2), a functionally uncharacterized protein from *Leptospirillum rubarum*. This target shows in its oligomeric form an extensive interaction network at the interface between the monomers. If only the monomeric structure is considered, one beta sheet remains exposed, extending into the solvent in a situation that is obviously not energetically favorable. In the dimeric form, sheet pairing causes the exposed hydrophobic residues to become buried (Figure 19). A homologue structure of this target (a putative heme-binding protein from *Anabena Variabilis* - PDB: 3FM2), shows a very similar configuration of the binding site, and includes two zinc ions located in the clefts formed by the two chains. This finding supports the hypothesis that the dimeric form of target T0576 is a requirement for its biological function.

The group that submitted the best prediction for this target, “458 BILAB-solo” (S_{agree}=0.79), was successful in modeling the interface region, including the sheet pairing. Almost all inter-chain contacts present in the interface region of the target have been correctly modeled in the prediction. For biological applications, the usefulness of a correct oligomeric model like the one discussed here, which shows the protein in what is likely to be its functional state, clearly exceeds the one of any monomeric model.

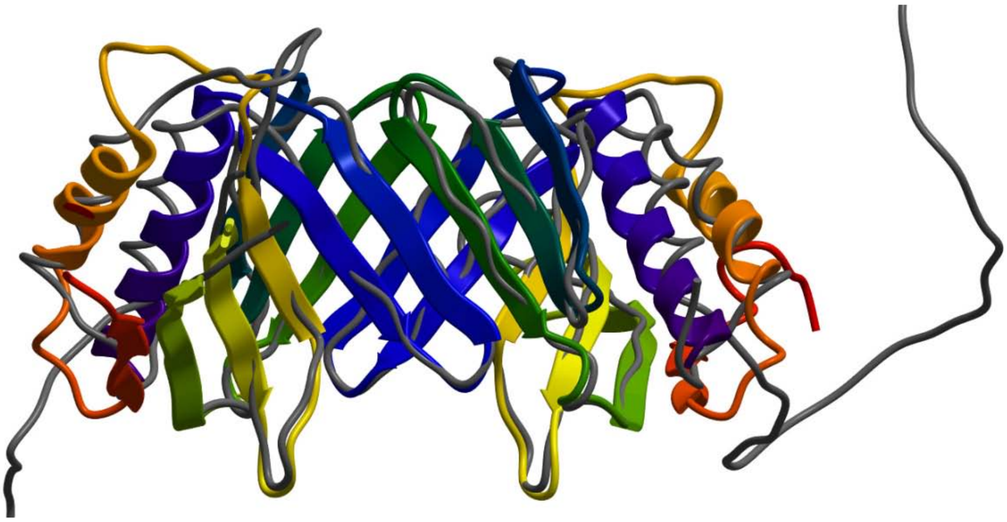


Figure 19 T0576 – example for a dimeric prediction target in CASP9 (group "458 Bilab-solo"). An isolated monomeric subunit would not form a self-contained structure, and leave several of the β-sheet interactions unsatisfied.

3.4.1.1 Discussion

Despite the fact that the quaternary structure is often essential for the understanding of the biological function of a protein and more than half of all CASP9 target appeared to be oligomers, only a small fraction of the participating groups in CASP9 submitted oligomeric models.

Among the groups which submitted at least one oligomeric model, the group of M. Kakuta, with its predictors" Bilab", "Bilab-solo" and "Bilab-ENABLE", performed best. In particular, the human predictor "Bilab-solo" was best in terms of oligomer classification and modeling accuracy (S_{Agree}). Excluding our naive predictors, the server "Bilab-ENABLE" was the best server, even if it had a very high number of unrealistic models. The dominance of the method "Bilab-solo" can may be explained by the extensive use of manual inspection as stated in the submitted CASP-abstract. It seems that manual investigation of the template situation for the target of interest increase the overall accuracy considerably.

The performance of the participating server groups compared to our naïve predictors is rather poor. Most of the sever groups submitted mainly monomeric submissions, except group "Bilab-ENABLE", who submitted mainly oligomeric models. Unfortunately, a significant part of their models contained clashes or have non-interacting subunits. In general, the accuracy of the participating server groups is rather low. The high rate of unrealistic models reveals that the complexity of oligomeric modeling can yet not be handled properly by automated approaches.

In contrast to the overall server performance, the naïve predictor “NaiveCoverage” has shown a remarkable performance in oligomer classification and modeling performance. This indicates that a reasonable classification rate and modeling accuracy can be achieved with standard tools in the field.

Although different scores were used for evaluation, the discrepancy between the observed accuracy in tertiary structure modeling and quaternary structure modeling is large. The accuracy of the top performing modeling servers (i.e HHpred[59] or I-Tasser[23]) has been slowly saturated over the last CASP experiments, however as described in the introduction, the biological usefulness of single chain models is limited. In contrast, the accuracies of oligomeric modeling methods observed in this experiment are rather low.

One may explain the weak performance by the large evolutionary distance between template and the target and the resulting difficulty to identify the correct oligomeric state. It can be hypothesized that predictors submitted their models monomeric, if evidence for oligomeric complexes was not given. Currently the assessors cannot distinguish between explicitly modeled monomers and “single chain” models, which were submitted because of lacking evidence for a particular oligomeric state.

Further, none of the top ranking “server” groups released their methods for the public audience. As a consequence, the results of this CASP assessment have only limiting implications for real life modeling.

In summary, the field of oligomeric modeling is only slowly emerging in terms of accuracy of the models and availability of the methods and provides the opportunity for scientific achievement. As a consequence, the organizer of coming CASP experiments should account more for the biological completeness of models. This includes, besides the correct prediction of tertiary structure, the prediction of quaternary structure, biological relevant ligand and cofactors and implies an adequate error estimation of such models.

3.4.2 Development of a score to derive similarity between homologues quaternary structures

A requirement for the comparison of quaternary structures is a metric which reflects the structural divergence between two protein complexes. Naturally, the difference between oligomeric states of proteins can be described by their change in the number of subunits. However, this does not allow a comprehensive analysis of the differences between protein complexes: the relative orientation of subunits can be different between protein complexes having the same oligomeric state. Conversely, oligomeric assemblies with different number of subunits can have structurally equivalent interfaces. For example, the structure of a D2-Tetramer can be generated by applying relevant symmetry transformations to the structure of a C2-Dimer. When comparing the D2-Tetramer with the corresponding C2-dimer, the same interface type can be identified, even if the oligomeric states of the complexes are different. The assembly of dihedral oligomers from cyclic intermediates is a common path in evolution and appears in nature 11 times more frequently than expected.[116] When the concept of comparative modeling is applied to quaternary structures, one can often identify templates which are, with respect to quaternary structure, evolutionary intermediates of the target protein. To study the mechanisms of assembly or disassembly of homologue oligomeric structures, a target function which evaluates both the differences in the number of subunits and the correct interface geometry is needed.

Several scores were developed in the past, some, like the already mentioned iRMSD[115], are based on superposition and focus on the differences in translation and rotation between subunits. These generally do not account explicitly for differences in the oligomeric state, mainly because such scores focus on the assessment of protein-protein docking methods, where the correct evaluation of the relative orientation of subunits is most important.

Other scores are superposition-free, like the Q score[132] and intent to describe small differences in the geometry of interfaces. However, the number of subunits is not considered explicitly.

In this chapter we introduce a new score for the comparison of homo-oligomers, which accounts for differences in the number of subunits as well as the overall interface accuracy.

Contact agreement function

The focus in this study is on homo-oligomeric structures, which consists of multiple copies of the same subunit. As a consequence a particular residue *i* in the protein sequence can be observed in each subunit, and a contact x_{ij} between residue i and j appear at least once in a dimer and twice in tetramer etc. (see Figure 20 for a schematic representation).

A relationship between two residues can be called a contact when the spatial distance between the two corresponding residues in different subunits is below a certain threshold. We used a 12Å threshold between the Cβ-atoms (Cα for Glycins) to define contacting residues. This threshold was introduced by Xu[132] to calculate the Q score (see subchapter "weighting function" for more details).

The difference in in the total number of contacts between two equivalent residue pairs can be defined as follow:

$$f(x_{ij}, y_{ij}) = \begin{cases} 1 - \dfrac{|x_{ij} - y_{ij}|}{\max(x_{ij}, y_{ij})}, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

where x_{ij} and y_{ij} reflect the total number of contacts between residue *i* and *j* in structure A and B, respectively. $f(x_{ij}, y_{ij})$ =0 indicates that none of the contacts between residue i and residue j in structure A can be observed in structure B (left example, Figure x). $f(x_{ij}, y_{ij})$ =0.5 indicates that half of the contacts between residue i and residue j in structure A are present in structure B (right example, Figure x). $f(x_{ij}, y_{ij})$ =1 indicates that the number of contacts between residue i and j in structure 1 (x_{ij}) and structure 2 (y_{ij}) is identical. As a result, the function $f(x_{ij}, y_{ij})$ reflects the contact agreement of corresponding residues in two complexes, which can differ either by the absolute number of contacts (i.e. different number of subunits) or by the orientation of their interfaces.

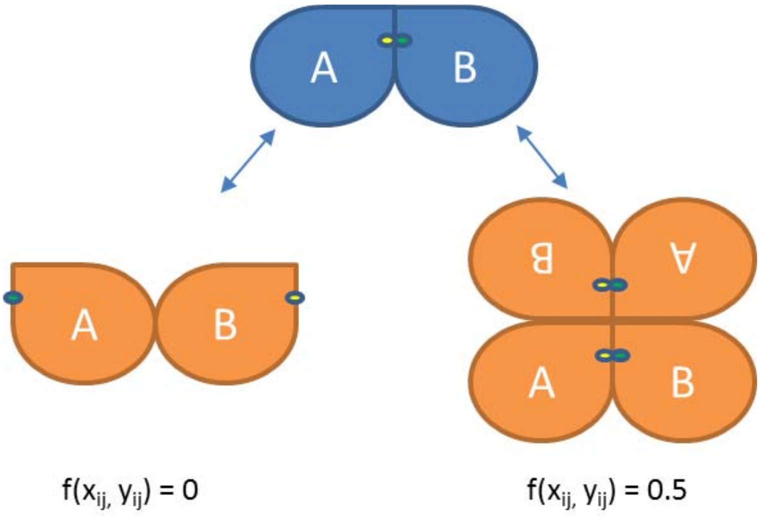


Figure 20 A contact between residue i (yellow circle) and j (green circle) appear at least once in a dimeric complex and at least twice in a tetrameric homo-oligomer. The fraction of common contacts results in a score of 0 for the left comparison and a score of 0.5 for the right comparison if considering one residue pair.

Weighting function

It is worth examining the effect of long-range contacts on the score obtained by the f function. As shown in Figure 21 the number of long distance contacts for a particular residue is higher than the number of short distance ones. As a consequence the long-range contacts contribute on average more to the overall score than short-range. Furthermore, once again as shown in Figure 21, small changes in the position of the reference residue result in a very different contact pattern; mainly because contacts close to the threshold of 12Å easily cross into the non-contact zone, and vice-versa. For example, in Figure 3, structure 2 loses three contacts and gains 2 new ones compared to Structure 1.

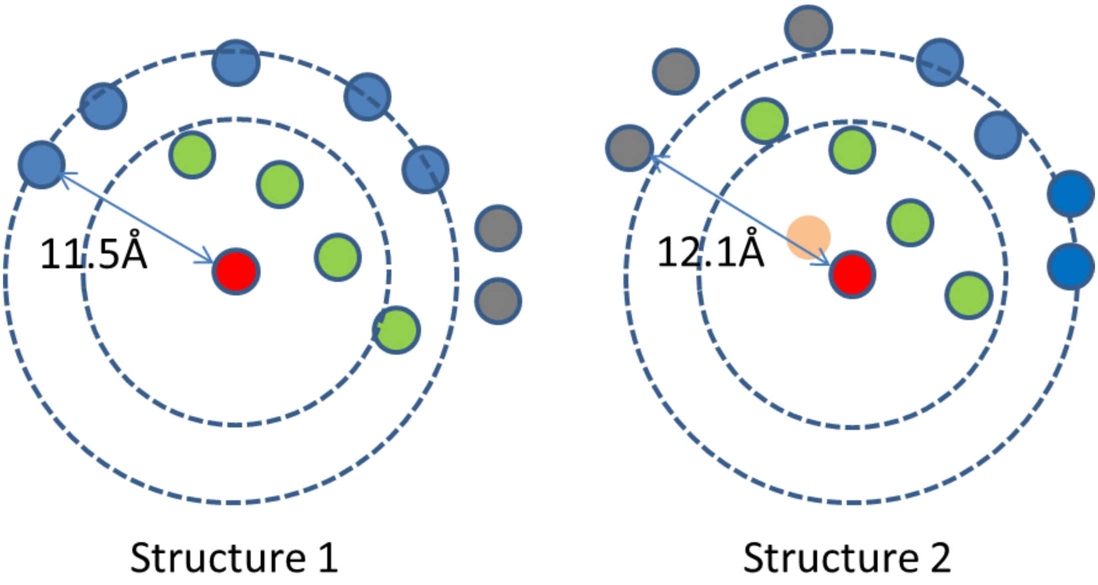


Figure 21 Impact of the absolute distance between two residues. Distant contacts (blue circles) were more frequent and contribute more to the score than close contacts (green circles). Further, small variations in the position of residue i (red circle) lead to a different contact pattern (3 contacts were lost, two new contacts were formed).

To account for the overrepresentation of distant contacts, we introduce a weighting function $w(d_{ij})$, conceptually similar to the one introduced together with the Q score scoring function[132]. The weighting function has been derived by Xu by evaluating a non-redundant set of dimeric protein structures and defines the probability to observe an atomic contact on given residue-residue distance (Cβ-Cβ). As shown in Figure 22 the probability is equal to 1 if a residue-residue contact is closer than 5Å and decreases monotonically with increasing distance to 0 for residue-residue distances larger than 12Å. The probability function was fitted to a Gaussian distribution with the following parameters [132]:

$$w(d_{ij}) = f(x) = \begin{cases} 1\,, & d_{ij} < 5 \\ e^{-2(\frac{d_{ij}-5}{4.28})^2}, & d_{ij} \geq 5 \end{cases}$$

With $d_{ij} = \min\,(\overline{d_{ij}^1}, \overline{d_{ij}^2})$. $\overline{d_{ij}^1}, \overline{d_{ij}^2}$ denotes the average distance between residue i and j in structure 1, structure 2 respectively.

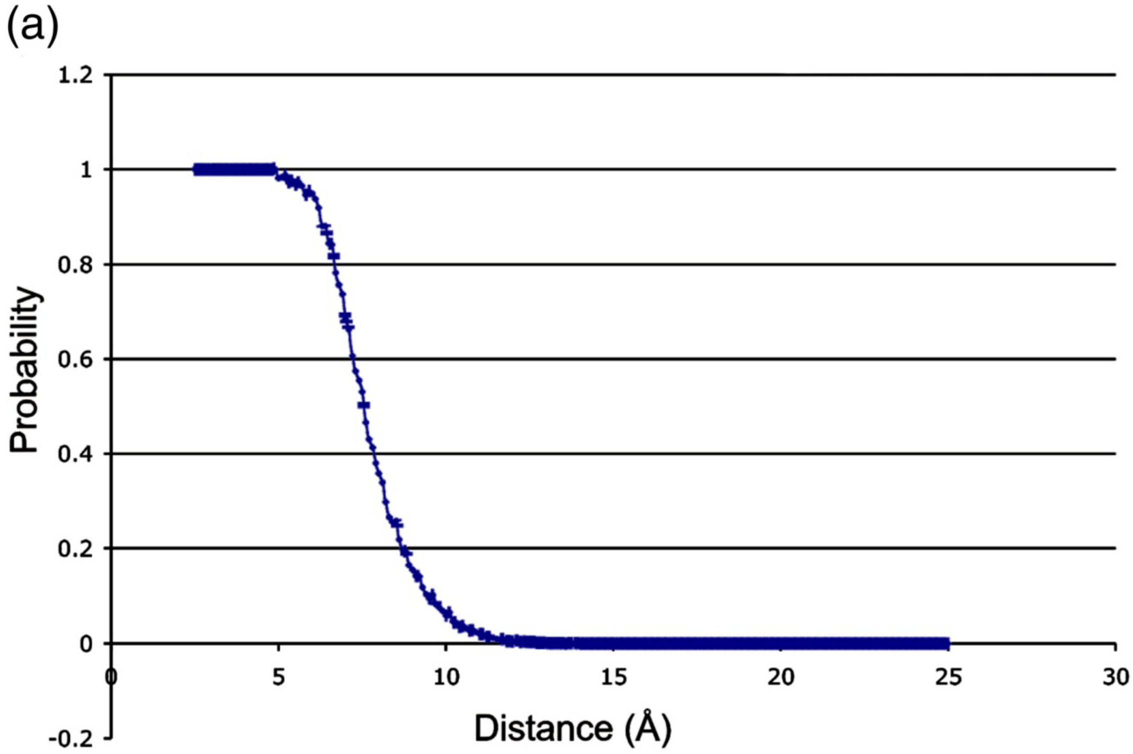


Figure 22 Probability to observe and atomic contact under different residue-residue distances. The probability function was derived by Xu[132] using a set of non-redundant set of dimeric PDB structures. Figure taken from Xu [132].

By combining the weighting function with the contact agreement function f described previously, the final QscoreOligomer metric can be defined as the weighted mean of the contact agreement scores for all equivalent residues in structures A and B:

$$QscoreOligomer = \frac{\sum_{i,j} w(d_{ij})\, f(x_{ij}, y_{ij})}{\sum_{i,j} w(d_{ij})}$$

QscoreOligomor ranges from 0 to 1. A value of 1 means that all contacts in structure A are present in structure B, with the correct number of subunits and identical interface geometry.

When considering monomeric structures, QscoreOligomer obviously cannot be defined according to the formulas described previously, as no intersubunit contacts can be observed. We then define the QscoreOligomer between two monomeric proteins to have a value of 1, and one between a monomeric protein and an oligomer to have a value of 0.

Separation of similar and dissimilar quaternary structures

QscoreOligomer reflects the similarity between two proteins regarding their quaternary structure. However, numbers are not very intuitive to answer questions about the qualitative similarity between two structures. In terms of tertiary structure similarity, the question if two proteins share a common fold requires a binary answer. In terms of quaternary structure similarity, we evaluated on which level of QscoreOligomer, two proteins can be considered as "similar" regarding their quaternary structures.

We applied the following test to evaluate if the quaternary structure between two structures is 'similar' and calculated QscoreOligomer-scores for a non-redundant set of target-template pairs. To establish a QscoreOligomer cutoff which defines similarity of oligomeric complexes, the range of possible QscoreOligomer values was partitioned into bins, and the fraction of similar quaternary structures for a given QscoreOligomer bin was calculated. We define the quaternary structure between two complexes as similar if:

1. The number of subunits in both structures is identical
2. A biological relevant interface (BSA_{Chain} >500$Å^2$), between two chains in structure 1 has at least one structurally equivalent in structure 2. To evaluate structural equivalence we used the Q score as developed by Xu. Xu proposed a cutoff of 0.1[132] and of 0.2[147] to distinguish interfaces which are similar in orientation and overall geometry.

As shown in Figure 23, more than 80% of the target-model pairs were correctly classified as similar for QscoreOligomer > 0.5 and Q score >0.1 (Figure 23 "blue line"). If applying the more stringent interface similarity cutoff for Q score (0.2), the fraction of similar quaternary structures reaches 80% for QscoreOligomer > 0.55. We investigated some model-target protein pairs in the "grey zone" range from QscoreOligomer of 0.50 to 0.55. The number of subunits can be expected to be identical in this QscoreOligomer range and the main difference seems to come from coverage (i.e. missing residues) in one of the two complexes. Since the orientation and the number of subunits were in most of the cases identical, a cutoff value of 0.5 QscoreOligomer was in the end chosen to differentiate similar and dissimilar quaternary structures.

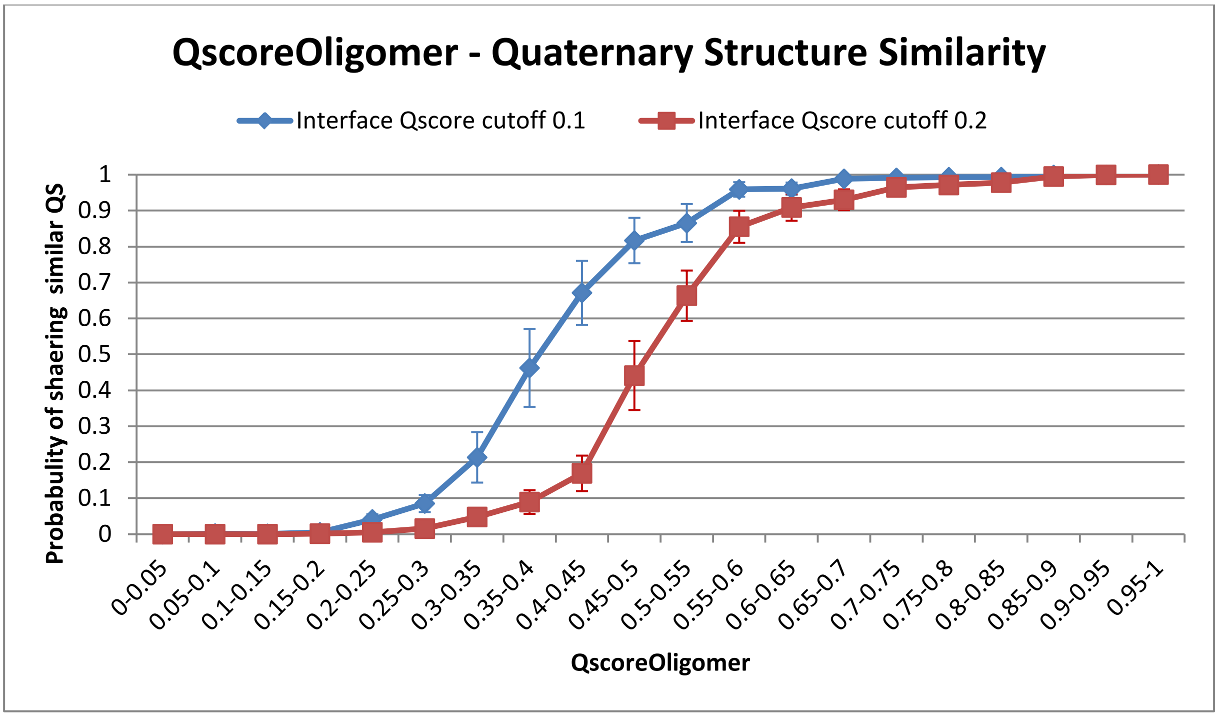


Figure 23 QscoreOligomer - Quaternary structure Similarity. Two complexes were considered similar if all interfaces BSA_{Chain} >500Å^2 have a Q score > 0.1 (0.2). A cutoff of 0.5 was chosen to distinguish between similar and dissimilar quaternary structures. Error bars were calculated by leave-out 30% of the targets in the testset.

Discussion

The developed QscoreOligomer introduces a new scoring function for the comparison of the quaternary structure of oligomeric proteins. Conceptually, the QscoreOligomer score is very similar to the Q scoring function developed by Xu. However, the original term describing differences in distances is replaced by a contact agreement score. As a result, QscoreOligomer can be applied on the direct comparison of complexes which are different in their number of subunits, and is not limited to the analysis of binary interaction as the Q score is. We also show that a QscoreOligomer cutoff differentiating pairs of complexes with “similar” quaternary structure from “dissimilar” ones can be defined.

QscoreOligomer score is symmetrical, which means the result does not depend on the order of the two structures being compared, and does not require choosing one of them as reference.

However, QscoreOligomer depends on the identification of equivalent residue contacts in the two structures, and this requires a sequence alignment between the two proteins. This limits the use of QscoreOligomer to homologue protein structures which can be aligned properly. This also makes QscoreOligomer strongly dependent on the coverage of the interfaces. A possible solution for this problem could be the structural superposition of the subunits of the two structures. Equivalent residues could then assigned depending on spatial proximity (e.g. < 3.5Å).

An extension of QscoreOligomer towards the comparison of hetero-oligomers is in general possible but would require a clear matching of different subunit sequences between the two protein complexes. One could then in principle work with a hypothetical sequence resulting from the concatenation of the sequences of single subunits.

In the following paragraph we used QscoreOligomer to compare the quaternary structure of templates with their corresponding target structure.

3.4.3 Development of a method for template based modeling of oligomeric protein structures

3.4.3.1 Oligomeric Template Library

Modeling of protein complexes based on comparative modeling techniques relies on the accuracy and completeness of the used library of template structures. This requires regular updates, by adding recently published experimentally solved structures on one hand, but most importantly it requires the correct annotation of the template structures regarding their biological relevant quaternary structure. Besides the correct quaternary structure, which is required for the analysis described in this thesis, correct template annotation regarding the presence and position of ligands and cofactors in the complex are also necessary, even if they are not necessary for the assembly of the complex.

An accurate oligomeric template library consists of template structures in their biologically relevant form. To select an appropriate annotation method for all PDB entries we benchmarked the assignment of the PDB versus the heuristic classification algorithm of PISA. As described earlier, the PDB annotation consists of multiple sources. Since the author assignment also includes human expertise we focused on this type.

Though, to ensure good performance of a public available modeling service, a template library which represents the current state of biological relevant protein structure space is needed. Continuous and regularly quaternary structure annotation for protein structures is only provided by a limited number of services (i.e. PISA and PQS).

To estimate the classification error of the PDB and PISA annotations, a dataset from literature[160] with manually verified oligomeric structures was used. The resulting dataset consisted of 390 protein structures. Among them 138 Structures were annotated monomeric (See Table 3).

Oligomeric State	**Testset Literature**	**Testset PiQsi**
1	138	853
2	133	671
3	36	87
4	51	265
5	2	2
6	19	83
7	0	3
8	5	22
10	2	8
11	0	1
12	4	23
13	0	1
14	0	2
16	0	1
24	0	12
60	0	1
Total	**390**	**2035**

Table 3 Composition of the investigated testsets. The literature testset was based on a study from Bordner and Gorin[143]. The PiQsi testset was compiled by using sequences which have an evolutionary distance of maximal 70% sequence identity to each other sequence in the testset. Only entries which share the same sequence were retained.

As described in the introduction of this chapter, a specific biological unit within the PDB REMARK350 section can be either annotated by the author, based on results from software, or any combination of them. Since we wanted to benchmark explicitly the ability of the authors to annotate a structure correctly only those annotations were investigated. If author assignment was missing or ambiguous (i.e. one chain appears in at least two authors' assigned complexes), no correct annotation was assumed. For the "PISA" annotation, the top ranked assembly calculated by the PISA server was used. Additionally we included also PiQSi annotations for those entries which were found in the database (43% of the structures in the testset). (See Material and Methods for more details)

It was assumed that PISA annotates biological active state most correctly or at least as accurate as the notation of the PDB[107,132 148] . As seen in Table 4, the comparison between PISA and PDB reveals a better performance of the annotation given by the PDB authors (87.9% versus 82.6%). Both methods tend to predict higher oligomeric states. For 2.8% of the targets, PDB annotation provides no information (1.3% for PISA). PiQsi reaches an overall accuracy of 97.6%.

	Correct	**Lower predicted**	**Higher predicted**	**No annotation**
Literature				
PDB	87.9%	2.8%	6.4%	2.8%
PISA	82.6%	6.2%	10.0%	1.3%
SMOTL	91.5%	2.8%	5.4%	0.3%
PiQsi	97.6%	1.8%	0.6%	-
PiQsi				
PDB	82.1%	5.3%	10.7%	1.9%
PISA	81.1%	8.0%	10.0%	1.0%
SMOTL	83.3%	5.3%	11.2%	0.1%

Table 4 Observed accuracies using the predictions of the annotation methods of the PDB, PISA and the schema of the SWISS-MODEL oligomeric template library. A prediction was considered to be correct, if the oligomeric state was annotated correctly. Results were reported for the literature and PiQsi testset. The results for PiQsi were normalized against the total number of annotations given by PiQsi.

Since PiQSi seems to have a very high classification rate, we used the PiQsi database as an additional benchmark set. We removed all Hetero complexes (entries with distinct SEQRES sequences) and clustered the remaining sequences in a way that two sequences from the testset do not share more than 70% sequence identity. This resulted in a set of 2035 protein structures (see Table 3). The difference in the accuracy of predicted results between the two methods is not as clear as for the literature testset (82.1% accuracy for PDB, 81.1% for the PISA method, Table 4); in this case only a small improvement can be observed. Further, as shown in Table 4, for this testset 32 (1.9% of the total number) structures did not have an author assignment; two of them were assigned ambiguously. PISA did not provide annotation for 20 (1.0%) entries. The overall analysis of the classification accuracy based on two different testset reveals that the annotation given by the authors in the PDB is superior to the prediction accuracy of PISA.

We visually inspected predicted assemblies and the following observations have been made. Firstly, ligand and cofactors were removed by PISA if they do not contribute to the interaction energy. This can be neglected, when investigating only the quaternary structure of a protein. However, the biological usefulness of such structures is dramatically increased by ligand and co-factor information. Thus, the annotations of the PISA webserver were only of limited use for the purpose at hand. In contrast, the biological unit files of the PDB contained all relevant ligand and cofactors, even if the prediction was accomplished by PISA. Secondly, NMR structures were obviously not supported by PISA, since they do not have crystal contacts. The quaternary structure must be calculated from the NMR experiment itself. Because such structures can provide structural information when other information is not available, they have to be included into a complete and up-to-date template library. Thirdly, author annotation is often missing (i.e. if the authors of the PDB entry do not have any information of the correct quaternary state) or ambiguous(i.e. if more than one biological active state exists or cannot be distinguished, a particular chain appears in more than one biological unit file). Fourthly, PISA renames it chains when applying symmetry operators to a particular set of chains, while PDB splits chains with the same name by introducing a MODEL tag. In both cases the chain of a complexes has to be renamed so that a chain letter appear only once in a complex. Fifthly, PDB and PISA consider small peptides as macromolecular chains. For example, the HIV-Protease (PDB-ID: “1hiv”) is annotated as hetero-trimer, the structure consists of a peptide which is bound between the two protease chains (see Figure 24).

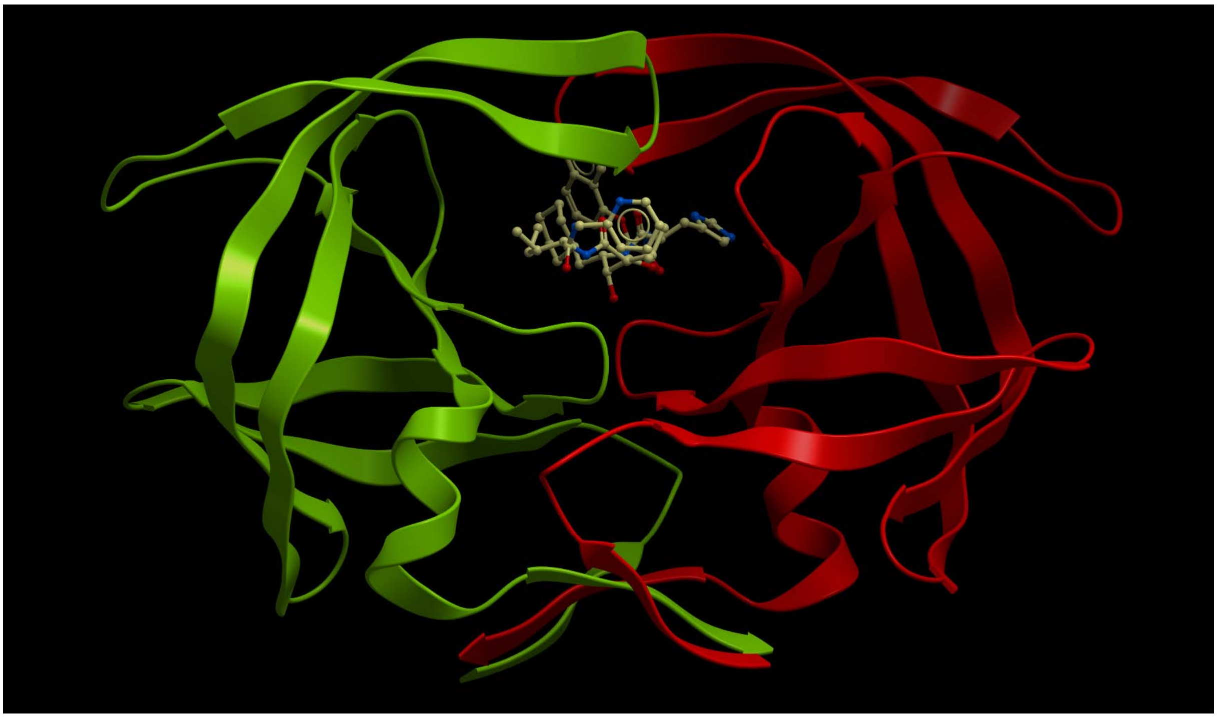


Figure 24 Structure of HIV-Protease (PDB-ID: “1hiv”). The complex consists of one peptide (shown in ball and stick) which binds in the interface of two macromolecular chains. Contradictory to PISA and PDB, this structure is labeled as dimeric Homooligomer in the SWISS-MODEL oligomer template library

Because the annotations of the authors superseded the predictions of the automated PISA server, an oligomeric template library using the PDB annotation system was compiled. Only if an author statement was missing or ambiguous, the annotation of PISA was taken into account.

Figure 25 shows the decision tree to build the oligomeric template library. As can be seen only structures classified as proteins (“prot”) or proteins complexed with nucleotides (“protein-nuc”) were considered. In addition, we included only structures solved by diffraction or NMR techniques.

More precise, the REMARK350 section was parsed and only biological units annotated by the authors were used as oligomeric templates. If a PDB files missed the REMARK350 section and the structure was solved by NMR, the chains as found in the PDB entry were considered as the correct quaternary structure. We excluded all other cases without REMARK350 section from the template library. If the author state was missing or ambiguous (i.e. if a chain appears in more than one author annotated biological unit), the top ranked PISA annotation was used instead. Since PISA often does not annotate ligands correctly, we checked if one of the biological unit files of the PDB entry corresponded to the assembly suggested. The identified biological unit file was then used to create the corresponding template structure; otherwise the complex was built using the transformation matrices as stated in the PISA XML file.

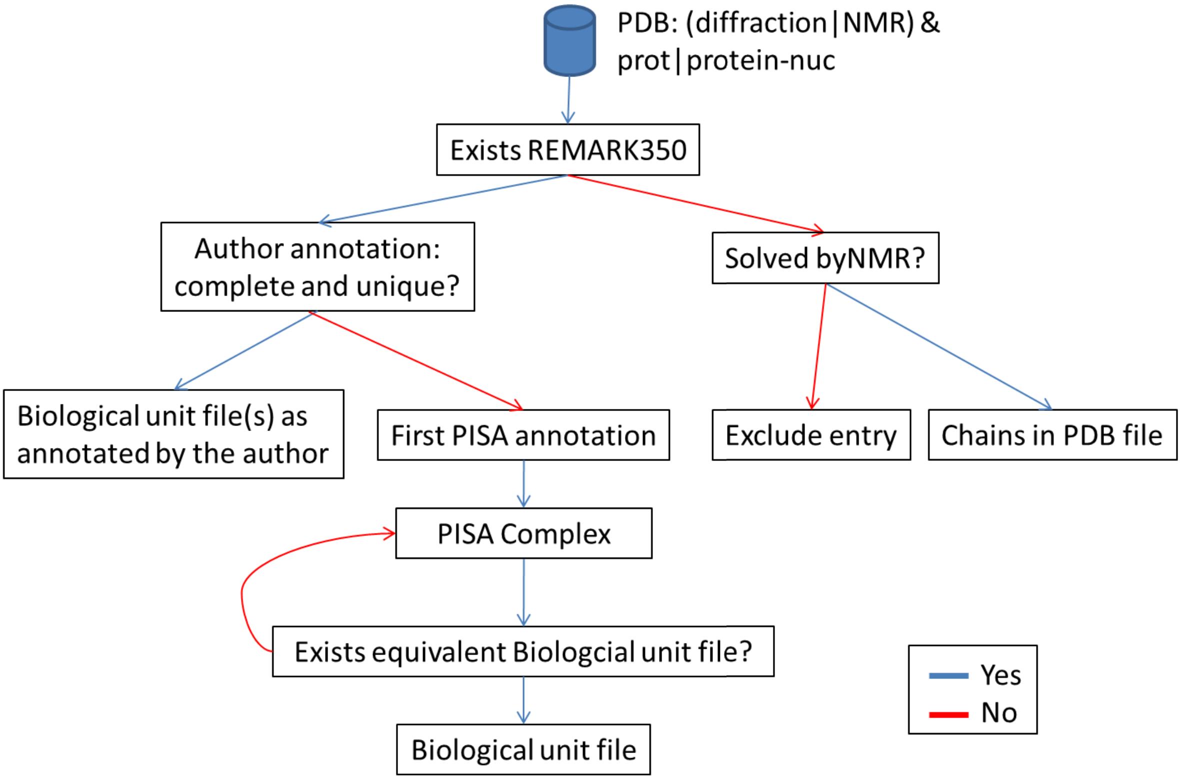


Figure 25 Decision schema of the SWISS-MODEL oligomer template library (SMOTL). Since PISA predictions often lack essential ligands and cofactors we used PDB anntotation where possible. X-Ray structures without a REMARK350 were excluded and manually checked. NMR structures lacking a REMKAR350 section were annotated as they appear in the PDB file.

The determination of the oligomeric state was based on the number of macromolecular chains (> 10 residues) in the template complex. Thus, the oligomeric state of 1hiv changes from Hetero-trimer to Homo-dimer. To model the complex correctly, only the two protease chains have to be identified.

All ligands/cofactors and modified residues which do not have a SEQRES entry were merged into one chain ('z' chain), as well as all peptide chains with less than 10 residues. We then rename all chains for a particular PDB entry so that a particular letter is unique for a PDB identifier. The current PDB format provides only one column for chain naming; this implies a maximum number of 62 chains in the assembly. Thus, we excluded structures with more than 60 chains (e.g virus capsides).

To ensure the template library of quaternary structures being up-to-date, a fully automated update procedure was generated. The up-date script downloads, parses and processes all recently released PDB entries as described above. Relevant information about the generation of an entry of the template library of quaternary structures is given in the header of file (see **Figure 26**). This includes the original chains of the PDB entry, the used annotation (PDB/PISA), the oligomeric state (number of protomers) and the type (i.e Homo/hetero oligomer).

```
REMARK    SwissModel Oligomer Template File Version. 0.1  9-Oct-2007
REMARK       EXPDB  File  3nrv  based  on  PDB  3nrv  and  Biological  unit  file
3nrv.pdb1.gz
REMARK    Generated at BIOZENTRUM http://www.swissmodel.unibas.ch
REMARK    DeepView:    http://www.expasy.org/spdbv/
REMARK    SWISS MODEL: http://swissmodel.expasy.org/
REMARK    ----------------------------------------------------------
REMARK    This is the first draft of the oligomeric template library
REMARK    ----------------------------------------------------------
REMARK    EXPDB_Oli ORI_CHAINS  AC
REMARK    EXPDB_Oli ASSIGNMENT  PDB:AUTHOR
REMARK    EXPDB_Oli SOURCEFILE  3nrv.pdb1.gz
REMARK    EXPDB_Oli PROTOMERS   2
REMARK    EXPDB_Oli STATE       HOMO
REMARK    ----------------------------------------------------------
SEQCOR  A QKINIDRHATAQINMLANKLMLKYTQKFGIGMTEWRIISVLSSASDCSVQ
SEQCOR  A KISDILGLDKAAVSRTVKKLEEKKYIEVYAINLTEMGQELYEVASDFAIE
SEQCOR  A REKQLLEEFEEAEKDQLFILLKKLRNKVDQM
SEQCOR  B INIDRHATAQINMLANKLMLKSSTAYTQKFGIGMTEWRIISVLSSASDCS
SEQCOR  B VQKISDILGLDKAAVSRTVKKLEEKKYIEVNGHSEDKRTYAINLTEMGQE
SEQCOR  B LYEVASDFAIEREKQLLEEFEEAEKDQLFILLKKLRNKVDQM
REMARK  A CHAIN A TO A
REMARK  A NA
REMARK  B CHAIN C TO B
REMARK  B NA
```

Figure 26 Header of structure 3nrv_1.pdb in the SWISS-MODEL template library.

We used the two previously described benchmark sets to estimate the accuracy of this procedure compared to PDB and PISA alone. As shown in Table 4 the classification accuracy is higher compared to the PDB author annotation and the automated PISA method. This indicates that the strategy to use the top ranked PISA annotation is a valuable approach if no or ambiguous PDB-Author annotation is available. Annotations made by PISA lack often important ligand/cofactor and were, as a consequence not suitable for the systematic deployment in template pipeline.

3.4.3.2 Categorization of template structures in terms of quaternary structure similarity to the target

In order to apply the concept of homology modeling for the prediction of oligomeric complexes, suitable template structures need to be identified. This implies the identification of templates which have the same quaternary structure than the target protein. In the following chapter we discuss a function which classifies template structures according to quaternary structure similarity.

3.4.3.2.1 Dataset of oligomeric protein structures

In literature one can find several test sets which were used to assess the performance of methods detecting biological relevant interfaces among all inter-protein contacts occurring in a crystal.[114,143,161] To ensure that the structures in the benchmark set are correctly annotated, an extensive literature review process is necessary. Typically this limits the number of protein structures in a given set considerably.

To overcome the problem of non-correctly annotated proteins we used the annotation of the PiQsi database[117]. We used only those PDB entries for the set, which are flagged as “correct” regarding their PDB annotation. Because the annotation within the PDB is not always unambiguous (see section “Oligomeric Template Library” for details), we used only those entries confirmed by PISA[107].

Another problem is the evolutionary distance amongst all proteins in the dataset. Because an overrepresentation of closely related proteins introduces a bias in the results of the analysis (which cannot be easily generalized to the overall protein structure landscape), the remaining set of sequences was culled to reduce bias towards large protein families with similar quaternary structure.

The resulting dataset consisted of 571 target structures. The number of subunits varied between 1 (monomeric) and 24 with dimeric proteins as the most abundant oligomeric type. Even numbers of subunits were overrepresented (except monomers).

We are aware that the ratio between the homo-oligomeric and monomeric protein structures most likely does not represent the true distribution of oligomeric state in nature (Monomeric proteins are expected to appear more frequent). However, because the focus is set on the modeling of oligomeric structures, we used monomeric target proteins as negative control.

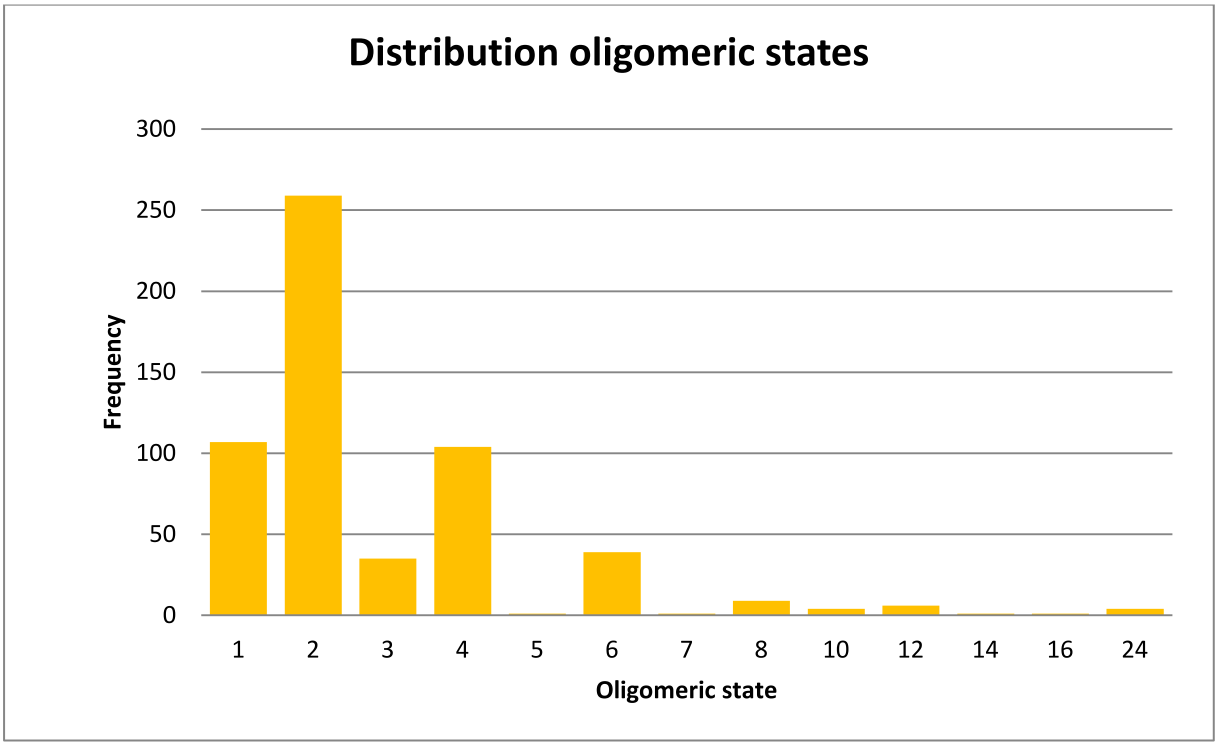


Figure 27 Composition of the dataset. The most abundant state is dimeric, followed by monomers and tetramers

Template selection

The way how template structures are identified and aligned is crucial for the success for comparative modeling. The last CASP installment has shown that methods which detect and align templates using profile-profile comparison were most accurate[7], notably also for identifying and aligning distant related protein sequences. We therefore decided to use HHsearch[40] to identify homologue templates.

Modeling

The correct and accurate modeling of quaternary structures includes modeling of insertions and deletions and the refinement of the final model under a symmetrical perspective. However, this can be considered as a non-trivial task and is beyond the scope of this work. In addition, the accuracy of the final model would be influenced by the modeling routine itself, for example if loops would be wrongly predicted and thus falsify the analysis. As a consequence, we used only a rudimentary approach for the modeling of the three-dimensional coordinates. Coordinates were taken from the template structure and if the residues were not identical in the sequence alignment subjected to SCWRL4[154], a standard sidechain modeling method. Regions without alignment information were not modeled. This procedure guarantees that the performance of the used modeling engine does not influence the evaluation. It has to been noted that for some parts of the analysis the model was used, whereas in others the template structure itself.

Models with a low coverage of the target sequence may lack significant parts of the interface and do not show any biological relevance. An analysis has shown that such models do not have any residue-residue contacts or consisted of very unrealistic interfaces. In this study we excluded all models which do not cover at least 20% of the target sequence. However, there are likely unrealistic models having higher coverage. Therefore, an additional filtering procedure has been put into the place to exclude such models.

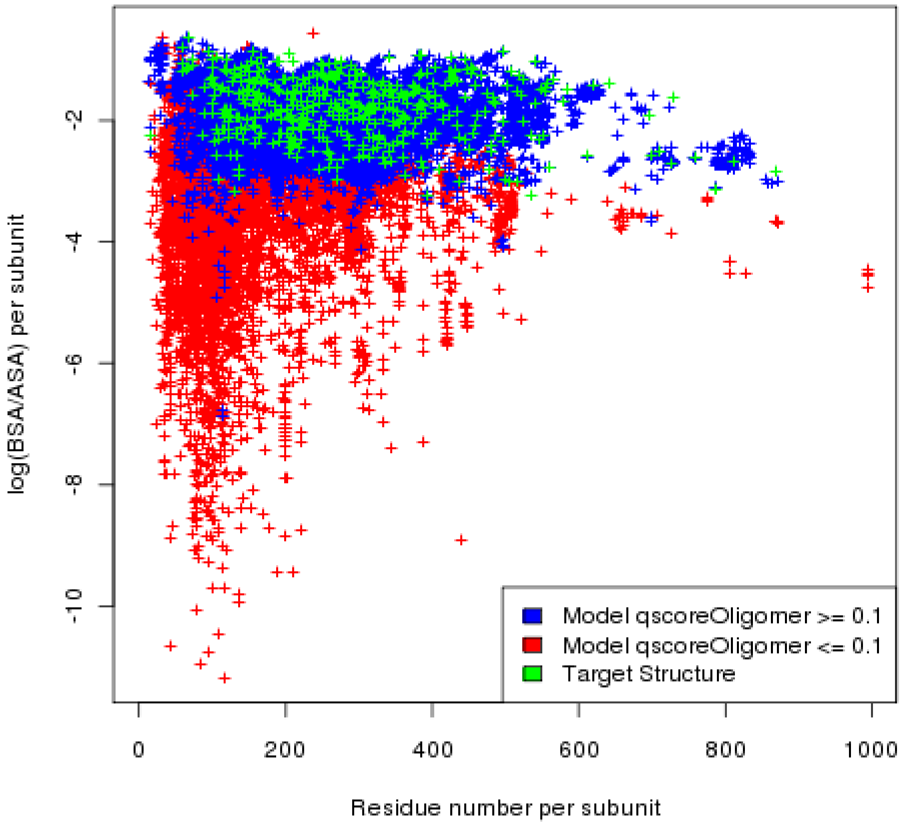


Figure 28 The log ratio between surface and interface area (ASA_{ratio}). Models with a QscoreOligomer of at least 0.1 are colored in blue and perform similar as native structures. Models below a QscoreOligomer of 0.1 are colored in red. Considerably negative ratios were observed for the latter. Visual inspection supports the non-native character of such models.

A rigid interface area cutoff, as proposed previously is not applicable, because small proteins form often small interfaces.[161] We therefore calculated the log-odd ratio between interface and surface area (ASA_{ratio} , see Material and Methods for details) and evaluated to what extend unrealistic structures can be distinguished from near-natives ones. As can been seen in Figure 28 the ASA_{ratio} of native homo-oligomeric proteins varies between -0.5 and -3, indicating a coherence between interface and surface size (i.e. small surfaces correlate with small interfaces) . The ASA_{ratio} of all models was calculated and split into two groups based on a QscoreOligomer cutoff of 0.1. Thus, the models were divided into the groups "realistic" (QscoreOligomer >=0.1) and "unrealistic/incorrect" (QscoreOligomer <0.1). The ASA_{ratio} – values have been analyzed according to the groups they belonged to. Hence, the "unrealistic/incorrect" group of models spans a very wide range of ASA_{ratio} (from 0 less than -10). Contradictory, the models having at least a QscoreOligomer of 0.1 have shown values in the same range as the target structures. Visual inspection of models below a threshold of -4 supports the hypothesis that such models do not represent realistic biological complexes. Such models were often incomplete and cover only a small part of the template structure. Thus, we decided to exclude all models ASA $_{ratio}$ < -4. Further, all models with at least 2 subunits, but without any buried residue were excluded. The remaining set of models and their corresponding template structures were used as dataset.

Grouping of templates with similar quaternary structure

We analyzed the sequence identity distribution of all target proteins in the set. For 88% of all targets, a template with more than 80% of the sequence identity was available. To a certain extent, this is an artifact by using the PDB as source for the target proteins and does not represent a typical model situation, whereas template structure with lower sequence identity would be expected. In addition, such closely related templates bias, for example, the mean or the width of particular cluster considerably. As a result, we excluded all template structures sharing more than 80% sequence identity to the target.

To analyze common characteristics of templates with similar quaternary structure we grouped all templates of a specific target according their quaternary structure. In order to do that, we used the QscoreOligomer as a distance metric and applied hierarchical clustering to generate groups of "similar" quaternary structure (see Material and Methods for details).

Based on this initial set of template groups we applied three additional filters which resulted finally into four template sets:

1. **All** – All models/template structures after processing as described in the previous sections
2. **Culled** – To decrease the impact of templates which are very similar in sequence and quaternary structure (e.g. identical structures which were deposited in the PDB with different ligand), all sequence in a group were culled at a 90% sequence identity level
3. **Culled/NoMonomers** – Monomers have taken a special role in the dataset, because they are always clustered into one group. To avoid biases in the evaluation, we removed all monomeric templates in this set
4. **Culled/NoSingletons** – All groups which consist of only one template were excluded.

The four template sets were used in the section "Characteristics of groups with similar quaternary structure"

3.4.3.2.2 Quaternary structure similarity between template and target

The template sets contained 68875 homologue template structures for 571 target proteins (107 monomers, 464 multimers). The number of templates for each target ranged from 2 (PDB-ID: "1b25", "1sg3") to 849 (Cyclin-dependent kinase, PDB-ID: "1gz8") (mean =123 templates). All template structures were grouped according to their quaternary structure similarity (see Material and Methods). Merging templates within a cluster at a sequence identity level of 90% decreases the average number of templates to 38 (min: 1 – max: 207). This indicated a high level of redundancy in the template set. Out of 571 target proteins, 32 did not have any homologue templates with similar quaternary structure; another 54 shared their quaternary structure only with templates having more 80% sequence identity. In total 15% of the targets did not have similar templates below a sequence identity of 80%. In order to determine the reason for this remarkably high fraction, we visually investigated some of the targets from the set. 63% of the targets lacked template structures at an evolutionary distance where similar quaternary structures can be expected (above 40% sequence identity). Other reasons for non-conservation were domain-swapping events (e.g Focal Adhesion Kinase, PDB-ID: "1k04") or pH dependent assembly (Ricin A chain, PDB-ID: "1uq5"). One could argue against the decision of using such targets to investigate the evolution of oligomeric structures. On the other hand, their inclusion makes the test set mirror more closely a real life modeling situation, where efforts have to be taken not to assign an incorrect quaternary state to the target. With increasing evolutionary distance, the chance to observe a template with similar quaternary structure is constantly but slowly decreasing (Figure 29). For more than 50% of the targets, a correct (i.e. similar in quaternary structure to the target) structural template sharing less than 30% sequence identity to the target could be identified. Analyzing these highly conserved templates has shown deviation in tertiary structure, like shifts in the sequence of helices, loops with different geometry, but conservation of the overall interface geometry.

Summarizing the given data reflect the large variability between protein families regarding their conservation of the quaternary structure. It has been shown that similar quaternary structure can be identified for a considerable fraction of targets. However, proposing a general cutoff is not useful in this context. As a result, the classification of template structures regarding their quaternary structure similarity requires additional criteria.

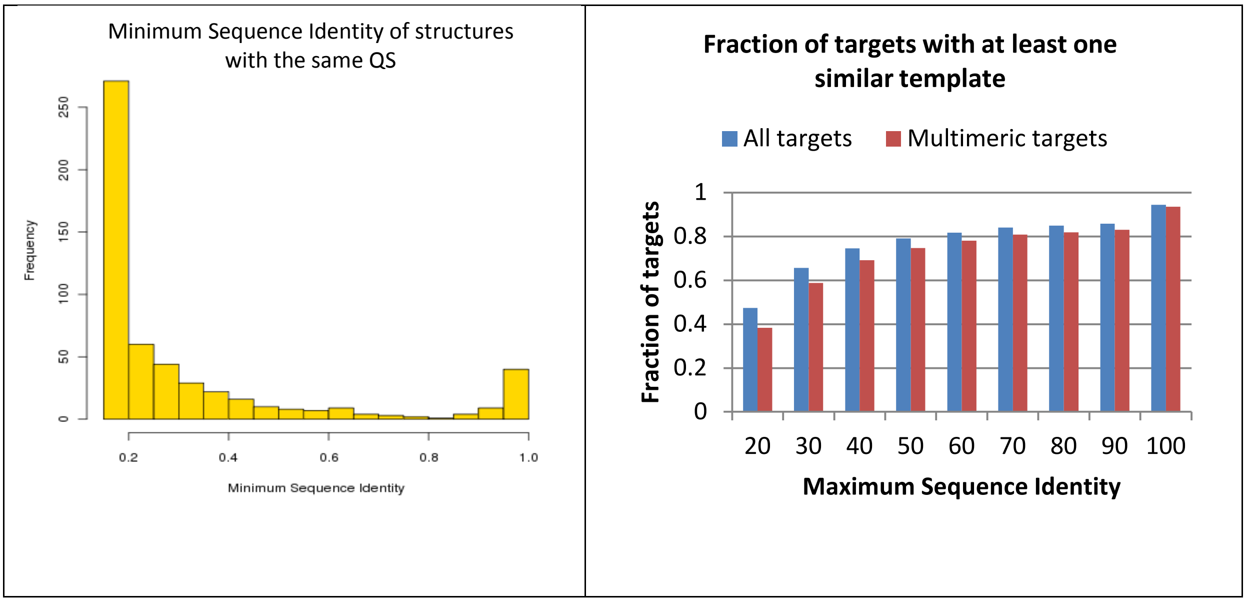


Figure 29 Lowest sequence identity of template structures sharing the same quaternary structure as the target. 15% of the target sharing their quaternary structure only with templates which are less distant than 80% sequence identity. However, for more than half the targets templates with similar QS can be identified below 30% sequence identity.

We calculated the probability to observe a correct template structure at a given evolutionary distance in order to the estimate how accurate the conservation of quaternary structure can be explained by the sequence identity alone.

As shown in Figure 29, the chance to find a correct template structure is slowly decreasing to a sequence identity of 40%. Below 40% sequence identity the chance to identify a template structure with similar quaternary structure is considerably decreasing. This is in agreement with previous studies.[115,116] However the probability to be correct for closely related template structures (>40% sequence identity) is rather low, from 0.6 up to 0.8 (if sequence identity >90%).

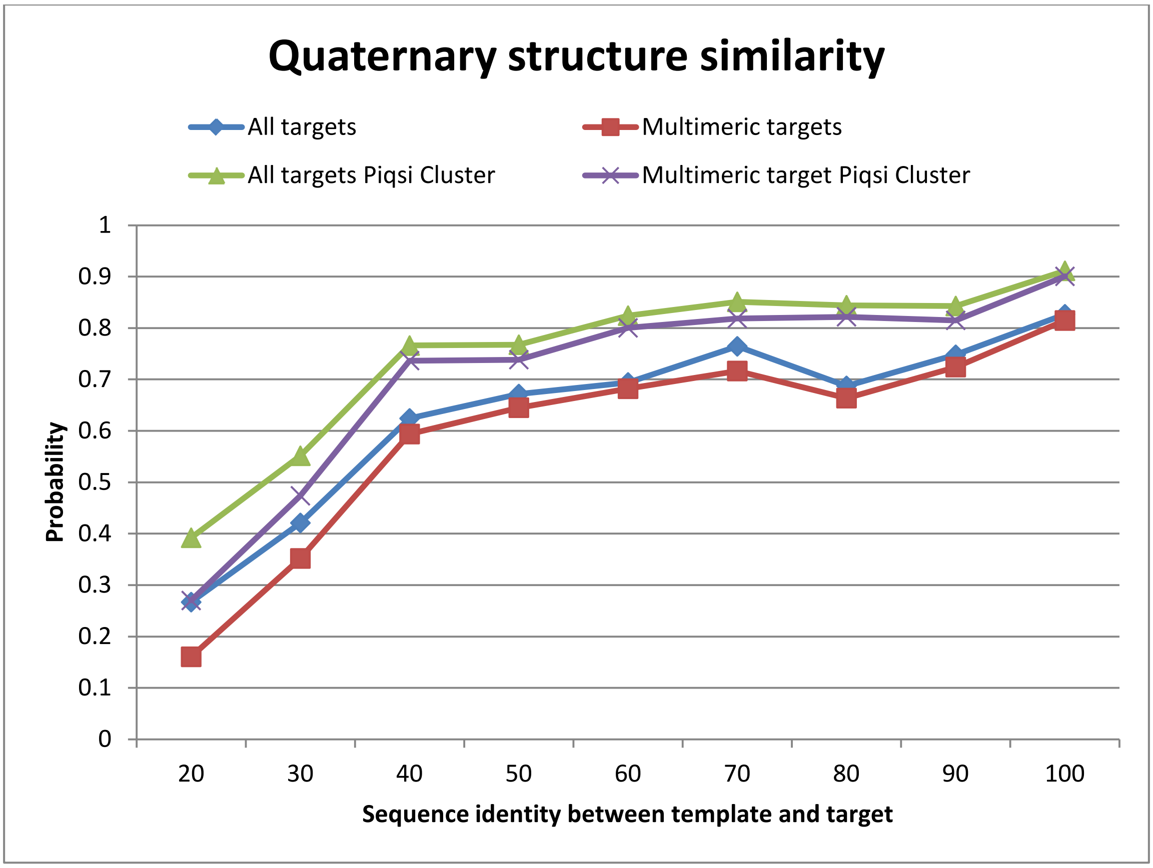


Figure 30 Probability of finding a template with similar quaternary structure on a given evolutionary distance. A) all targets including their template structures B) Only multimeric targets C) All targets with PiQsi approved clusters D) Only Multimeric targets with approved PiQsi clusters.

To ensure that the success of identifying the correct quaternary structure is not biased by identifying monomeric templates for monomeric targets, we investigated if the removal of monomeric targets (19% of all targets) decreases the level of conservation. A significant decrease cannot be observed (see Figure 30, “red line”), indicating that the error rate for monomeric targets is only slightly decreased compared to multimeric targets.

The overall results show, firstly, that for the majority of targets remote template structures with similar quaternary structure can be identified. Secondly, the range of evolutionary distance between target and template structures which share the same quaternary structures is wide and flexible. Choosing an absolute sequence identity cutoff is thus not useful in that context, because it neglects the variety in conservation between protein families. Thirdly, selecting templates with similar quaternary structure needs consideration of additional attributes and cannot be distinguished by evolutionary distance alone.

3.4.3.2.3 Verified quaternary structure annotation

To investigate to what extent wrongly assigned quaternary structures (e.g. Complexes consisting of interfaces made by crystal contacts) decrease the probability to find a correct template structure, all groups without at least one template structure having a PiQsi error state “NO” or “PROBNOT” were excluded.

As shown in Figure 30, the average accuracy on a certain sequence identity level increased by approximately 8-10% for templates which share more than 30% sequence identity with the target. This indicates the importance of correctly annotated template structures. Unfortunately, manual annotation of quaternary structure is only available for a small part of currently deposited proteins in the PDB.

3.4.3.2.4 Characteristics of groups with similar quaternary structure

The selection of a single template structure based on sequence identity alone involves risk. Since the change in quaternary structure (e.g. from dimeric to tetrameric) appear in a rather direct way, the selection of a template structure based on sequence identity alone is somewhat random, especially if template structures with different quaternary structures exist (a likely scenario for templates sharing low sequence similarity) at a particular sequence identity level. One example is the dimeric aldehyd dehydrogenase (PDB-ID:”1ad3”) which appear in three different oligomeric states (dimeric, tetrameric, hexameric) between 20%-30% sequence identity. We therefore asked if the analysis and comparison of the quaternary structure of all templates can be used to identify the correct quaternary structure for the target.

Xu[132] have shown that crystal interfaces can be distinguished from biological interfaces by the analysis of interface clusters with similar geometry. Large clusters, which contain interfaces of X-Ray structures having different crystal parameters have been identified to be biological relevant.[132] Based on these results, we decided to group the identified template structures for a specific target according to their quaternary structure. The study of Xu[132] is based on dimeric proteins and uses an interfacewise score as distance measure. Since we are interested in identifying the correct quaternary structure, our recently developed QscoreOligomer was used as a distance criterion (see “Dataset” for details) to group templates with similar quaternary structure.

The total number of groups for all 571 targets is 5865; on average a target consists of approximately 10 groups, with one group as minimum and 40 groups as maximum (PDB-ID:

“1jbe”). The number of group does not necessarily represent the number of possible oligomeric states, because it includes singletons that represent different alignments among similar structures or structures with an incorrectly assigned quaternary state (e.g. crystal contacts). In total 2991 singleton groups were detected. 72 targets have only one template with similar quaternary structure and thus representing singletons.

Here we tested, if the analysis of groups of similar quaternary structures provides additional information for the correct classification of template structures. Four different set of templates were compiled as described in Material and Methods.

3.4.3.2.5 Average distance of group members to the target

It has been shown that groups of similar binding modes are on average more conserved in evolution than non-conserved binding modes (binding modes reflect the difference in orientation between two subunits[162]).[118] This is in agreement with the observation that the quaternary structure is more likely to be similar if the evolutionary distance is decreasing.[115,116]

We therefore calculated for all four template sets the probability to have similar quaternary structure to the target depending on the average evolutionary distance of the whole group to the target.

As can be seen in Figure 31, three out of four template sets have shown very similar performances. The probability is constantly increasing to its maximum of 0.3 at a level of 50-60% sequence identity for “All”, “Culled”, “Culled/NoMonomers”. If considering the set “Culled/NoSingletons” at its maximum, the probability is doubled compared to the remaining three sets. This indicates a large impact of the group size to the classification accuracy. Interestingly, the probability decreases if the average sequence distance of the group is higher than 60% sequence identity. The number of templates for a particular target is decreasing significantly if the sequence identity between template and target ranges between 50% and 80%. Groups with an average sequence identity in this region are representing groups with only a little number of members. In addition, more than 50% of the targets share their quaternary structure with template structures of less than 30% sequence identity. Since the number of templates is increasing with increasing evolutionary distance, distantly related template structures dominate the average distance to the target sequence.

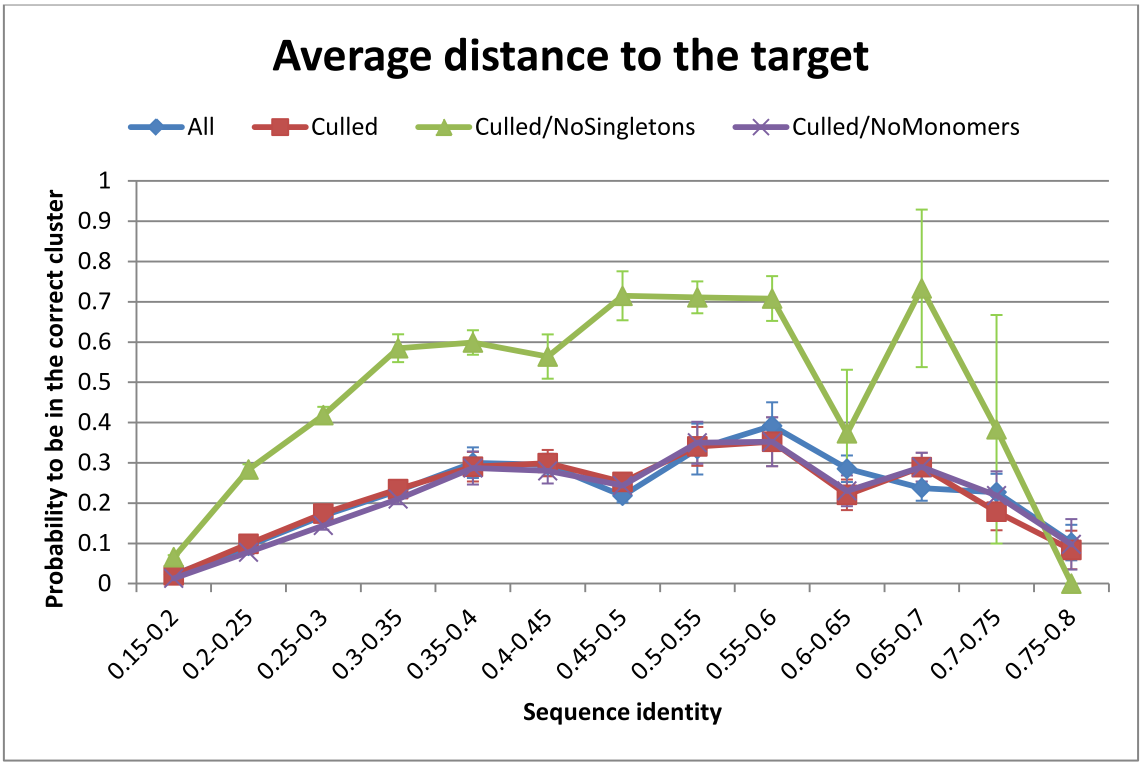


Figure 31 The probability to observe a correct group based on the average distance to the target sequence. The probability to identify the group reaches it maximum between 50-60% sequence identity. Groups with more than one template structure are twice as likely to be correct at that level of conservation.

3.4.3.2.6 Consensus among template structures

Encouraged by this dependency of the overall probability on the group size, we analyzed to what extend the size of the group can be used for classification.

Thus, we calculated the probability of observing a group with the same secondary structure as the target depending on the number of templates in the group. Only three template sets were analyzed, because the only difference between the set "Culled" and "Culled/NoSingletons" was the deletion of groups with only one template in the latter. As expected from the analysis shown in Figure 31, the probability to observe the same quaternary structure as the target was very low for groups made of only one template structure. With increasing number of template structures in a group, the probability to observe a correct quaternary structure increased. The lower probabilities seen for the dataset containing redundant sequences ("All") showed a large bias introduced by the overrepresentation of some protein families in the template set.

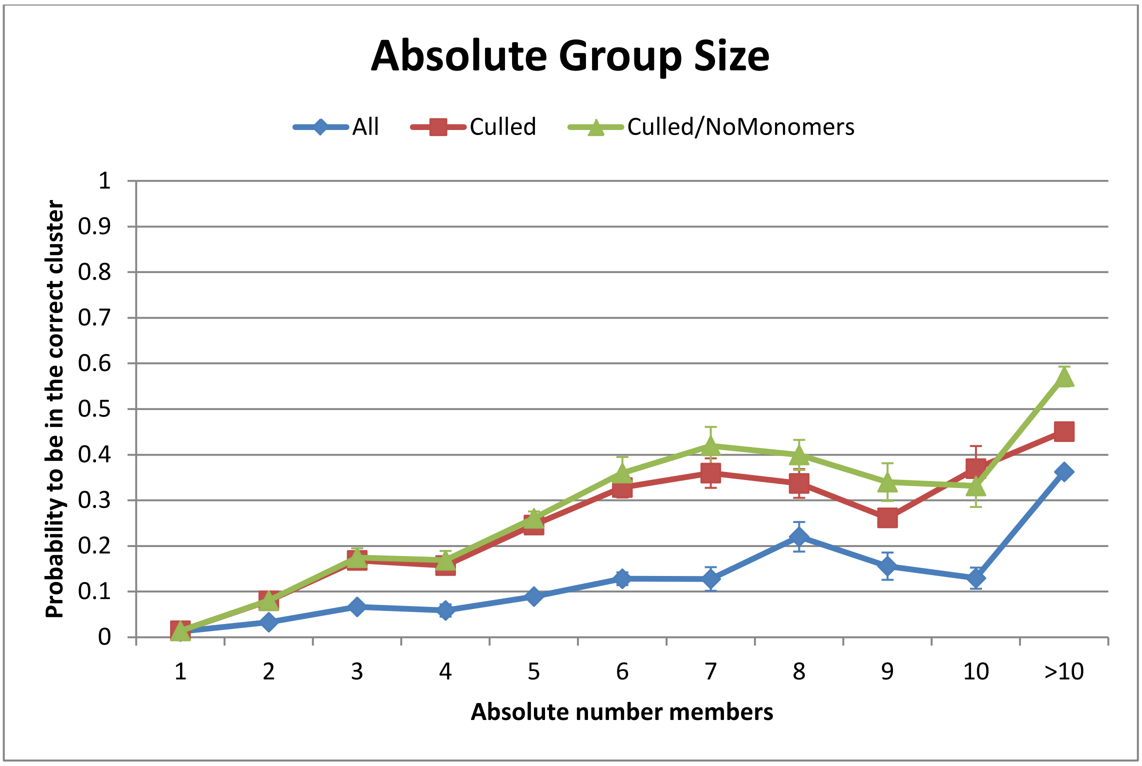


Figure 32 Absolute number of members in the group. The probability that a particular group is correct increases with the number of members. Reducing the redundancy within a group increases the probability to be correct considerably.

The number of templates varied from target to target (as described in the section “Quaternary structure similarity between template and target”) and the absolute number of templates in a particular group revealed not to be a reliable descriptor if only few templates were present in the group. Hence, we took as reference the fraction of templates in each group (i.e. the number of templates in the group normalized by the absolute number of templates for that target).

As shown in Figure 33, the probability is steadily increasing with increasing relative group size. In general two observations can be made. First, the non-redundant templates set with monomers removed have shown the best performance. Groups consisting of monomeric templates alone were on average larger than groups with multimeric template structures, independently from the quaternary structure of the target. Secondly, all four sets showed a lower probability to be identify the correct quaternary structure if the relative group size became very large. This effect was more apparent in targets which did not have any template group with the correct quaternary structure, and a very low total number of templates.

In general the analysis revealed a strong dependency of the conservation of quaternary structure within the family on the group size. The probability to observe a group with similar quaternary structure to the target can be clearly linked to the relative group size.

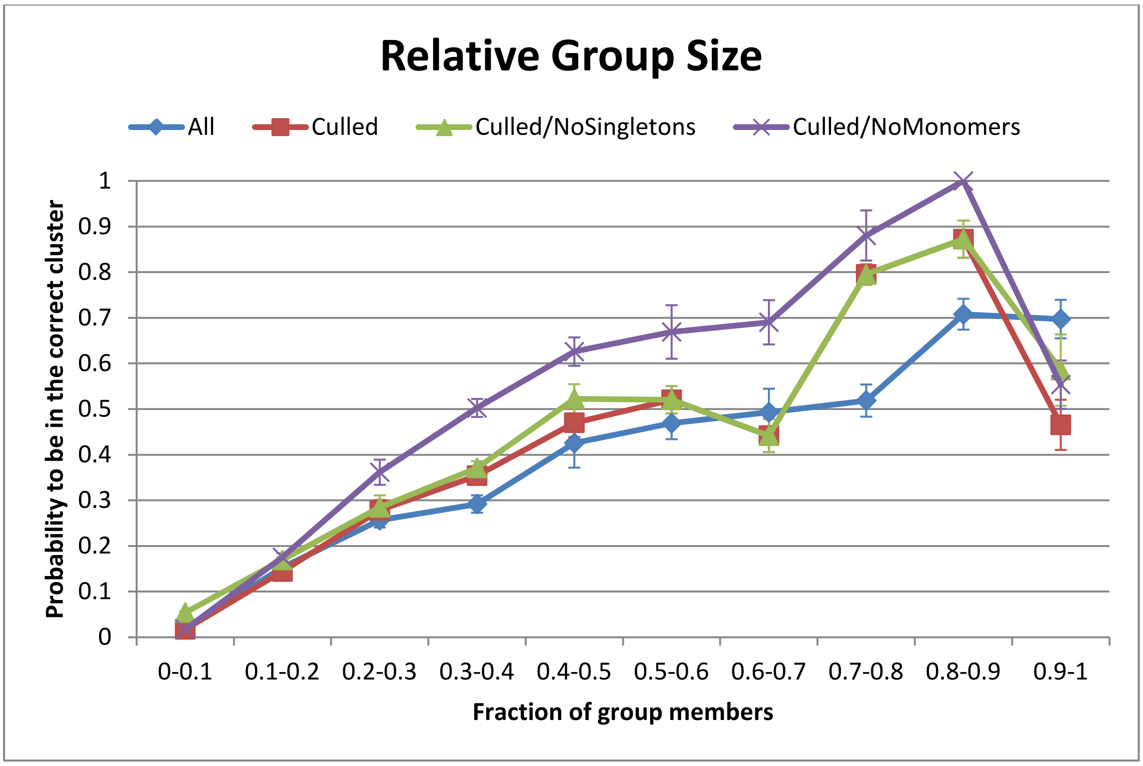


Figure 33 Relative group size. Instead of the absolute number of templates within a group, the relative group size was considered. The probability increased compared to the absolute number of members.

3.4.3.2.7 Evolutionary range within groups of similar quaternary structure

We hypothesized, that groups consisting of a wide range of evolutionary distances to the target (=large width), were more likely to be correct than groups of templates which share similar evolutionary distances to the target. Since the evolutionary distances to the target within a group is not necessarily normal distributed, we used four scores to define the evolutionary range within a group.

- The maximal width of the group characterized by the distance between the maximum and minimum sequence identity.
- The distance between 90% and 10% quantiles of the sequence identities within the group.
- The distance between the average evolutionary distance and minimal evolutionary distance.
- The standard deviation within the group.

The four calculated scores are shown in Figure 34. Independent of the used criteria, a high correlation between the width of the group and the probability to be correct can be observed. Most reliable were the maximum evolutionary width and the distance from the average of the group to the minimal evolutionary distance as they did not show bins with unexpected low accuracy. The standard deviation has shown a similar trend, but most of the groups were located in the first bin. Groups which had an evolutionary width of 15% within the group were correct in more than half of the cases. For all four scores, the deletion of monomeric groups increased the probability on given width.

Discussion of analyzing the properties of grouped quaternary structures

Group characteristics were identified as good descriptor for the classification of templates regarding to their quaternary structure similarity to the target.

The probability of sharing similar quaternary structure as the target structure increased when considering the size and the evolutionary range within the groups. The average distance has shown a rather weak performance mainly because a considerable amount of targets share their quaternary state also with very remotely relate template structures. Thus, the average mean is pulled towards larger evolutionary distances.

In general, the range of evolutionary distances and the size of groups reflect the overall conservation of the quaternary structure in the protein family of the target very well. The dependence of the probability on size and width of a particular group indicates that for most of the targets the quaternary structure is conserved within the family and can be identified using these two attributes.

However, the analysis of such group characteristics requires a sufficient number of template structures. This cannot be always guaranteed; in many cases a classification between correct and incorrect quaternary structure is required if only one template structure is available.

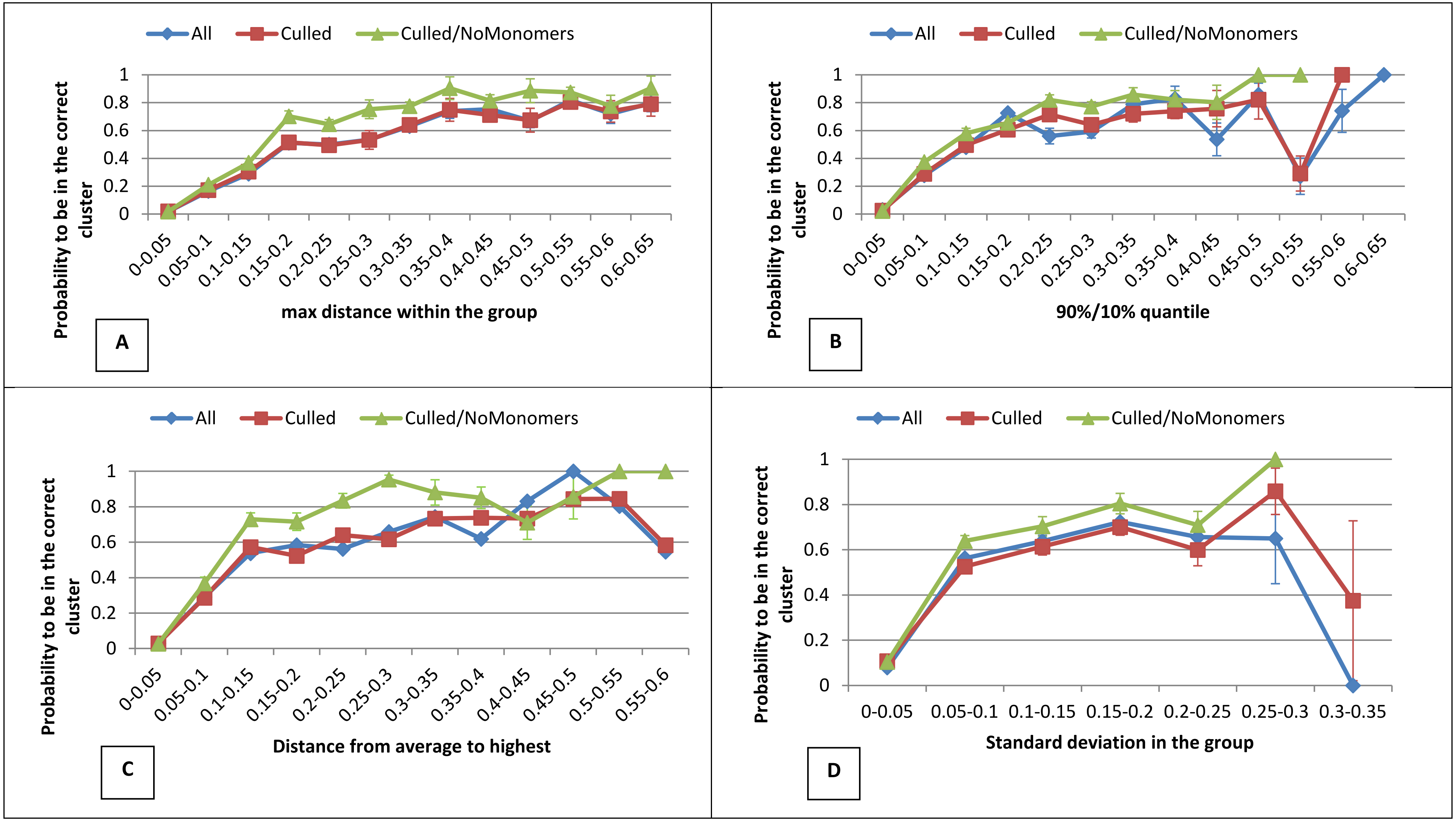


Figure 34 Four different distances were used to describe the evolutionary range within groups. A) Maximal distance width B) 90%/10% quantiles C) Distance from the average of the group to the member with the highest sequence identity D) The standard deviation within a group. For all four criteria, the probability to observe a group with similar quaternary structure is increasing with increasing distances.

3.4.3.2.8 Conservation of interface residues in evolution

If an interface is relevant for the function of a protein, it requires some form of robustness against evolutionary events.[163] Interface residues must retain their interface stabilizing character in order to contribute to the complex stability yielding to a high conservation. For example, the mutation of a large hydrophobic residue to a small hydrophilic residue is not favored, because the burial of hydrophobic atoms has been proposed as a main contributor to interface stability.[122] Surface residues are known to be more tolerant to mutations because a direct coupling to functionality is in general not given (an exception are surface residues which are involved into functional task such as active sites or DNA/RNA binding). As a consequence, the average surface conservation is considered to be lower than the average interface conservation. An example for evolutionary conservation within an interface is given in Figure 35. A dimeric superoxide dismutase was queried against the UniRef50[156] in order to build a multiple sequence alignment. Residues in the interfaces (marked with a red bar) shown indeed a higher level of conservation, as indicated by the blue background color in the corresponding rows. In contrast, surface residues (not marked) are not conserved.

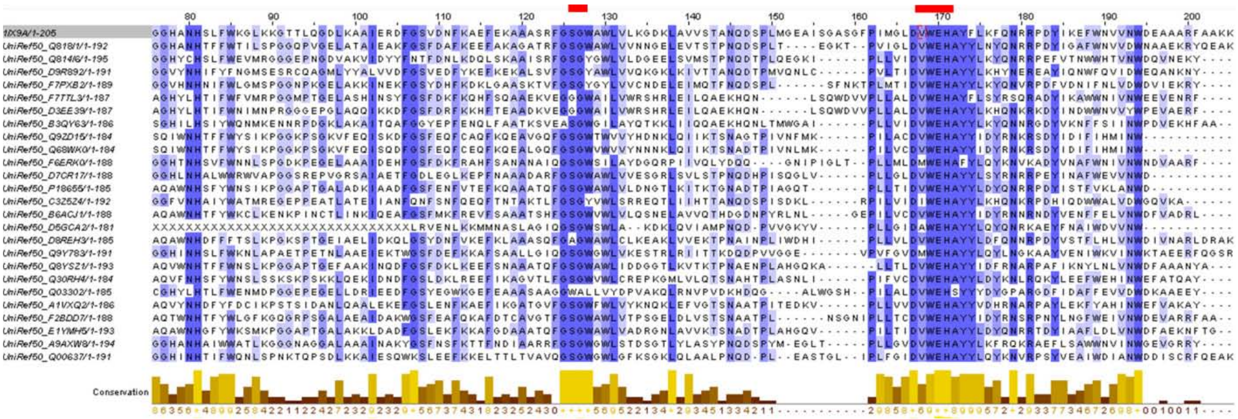


Figure 35 Multiple sequence alignment (MSA) of manganese superoxide dismutase ("1ix9"). Residues which were involved into interface assembly (indicated by red bars) were more conserved than surface residues. For clarity not all sequences of the MSA are shown.

Several studies have used the ratio between interface and surface conservation to characterize biological interfaces and distinguish them from non-biological ones (e.g. crystal contacts).[126,164] It has been shown that interfaces caused by crystal contacts are more similar to surfaces than to biological interfaces regarding their conservation in evolution.

We examined if this principle can be used in order to classify template structures according to their quaternary structure similarity to the target protein. We hypothesized that the target interface is more conserved than the target surface. Conversely, the interface conservation is similar to the surface conservation if the interface disappeared during evolution (i.e. by mutations). In the next chapter we will formulate a score, which describes the ratio between interface and surface conservation.

The Shannon entropy as measure for conservation

A widely used metric, which reflects the conservation of a particular residue during evolution, is the Shannon entropy[157]. The Shannon entropy, named after its inventor Claude E. Shannon, quantifies the expected value of information contained in a message. It can be written as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

Basically, the Shannon entropy can be seen as measure of uncertainty of a system with discrete states ($x_i: i = 1, \dots, n$). If all states occur with equal probability, H(X) reaches its maximum, whereas if only one state occurs H(X) = 0. Equivalently, the Shannon entropy is low if a residue is conserved and high if many mutations have occurred during evolution. The per residue entropy (S_{Column}) is calculated using the amino acid frequencies derived of the particular column in a multiple sequence alignment. In order to tolerate mutations which does not change the essential nature of the side chain(e.g. Val to Ile), it has been proposed to group residues by their physiochemical character.[126] Hence, two sets (20 standard amino acids; 7 groups suggested by Guharoy[127]) of residues were used in order to calculate the column wise entropy S_{Column}.

It has been proposed that not all interface residues equally contribute to the interface stability.[165] The burial of large hydrophobic residues upon assembly has been identified as a major contributor to interface stability[165] and the mutation of them often leads to interface disruption[122]. In order to account for the varying contribution of residues, based on their buried surface area, it was proposed to implement the average entropy as weighted mean[126] (see Material and Methods for details). As a result, interface residues were weighted by their buried surface area, surface residues by their accessible surface area in order to calculate the average interface entropy and average surface entropy respectively. Thus, the ratio between interface and surface conservation defines if a particular side is more conserved than the surface.

In order to calculate the ratio between interface and surface conservation for the target protein, we used the accessible surface area/buried surface areas from the template structure and mapped them to the corresponding target residues based on alignment information. According to this procedure, we formulated $S_{\text{Conservation}}$ as the log ratio between interface and surface entropy (see “Material and Methods” for details). Thus, negative values indicate that, on averge, the interface is more conserved than the surface.

$S_{Conservation}$ relies on the number of sequence in the multiple sequence alignment and the evolutionary distance between them. However, the alignment depth (i.e. the evolutionary distance between the target and the sequence with the highest sequence identity) is assumed to play an important role for the absolute value of $S_{Conservation}$. An alignment, built of very closely related sequences, does not contain much evolutionary information because of missing evolutionary events. As a consequence, the average entropy of the surface and the interface are expected to be low, yielding $S_{Conservation}$ scores of zero. A similar ratio was expected if all sequences are included. In this scenario, mutations frequently occurred in the interface and the surface, yielding to a balanced entropy ratio.

Introducing the concept of evolutionary fingerprints

We investigated how $S_{Conservation}$ changed for an increasing alignment depth and therefore we successively added sequences with a decreasing sequence identity to the multiple sequence alignment.

As an example, we calculated the $S_{Conservation}$ values of a dimeric d-lactate dehydrogenase (PDB-ID: “2dld”) as shown in Figure 36 (A, “green line”). In the beginning, the multiple sequence alignment only consisted of sequences which are closely related to the target sequence. With increasing alignment depth the evolutionary distance increased. The resulting $S_{Conservation}$-scores were analyzed according to their absolute values. As a result, the interface was only slightly more conserved than the surface, indicated by an almost flat line. Similar results were observed for the remaining targets in the set (data not shown).

In order to reason these findings, we investigated the surface of the dehydrogenase regarding its residue wise conservation. Therefore, we used the multiple sequence alignment depth when $S_{Conservation}$ reaches its minima (55% sequence identity). In panel Figure 36 (B), the molecule is shown in its dimeric form and the surface of one subunit is colored by applying a gradient from blue (low S_{Column}) to red (high S_{Column}) according to the residue wise level of conservation. The interface, built by the catalytic domain of 2dld, buried about 8600Å^2. Analyzing the entropy in the interface reveals high conservation (low S_{Column}) for most of the interface residues (Figure 36 C). In contrast, the surface is less conserved than the interface indicated by red patches. However, conserved residues have been identified around the active site of the protein. As shown in Figure 36 (D), residues, which were involved into the binding of NADH, show a similar level of conservation as the interface residues. As a result, surface residues which were involved in other functional tasks, like ligand binding or nucleotide binding, increase the overall conservation of the surface, thus lowering the discriminative power.

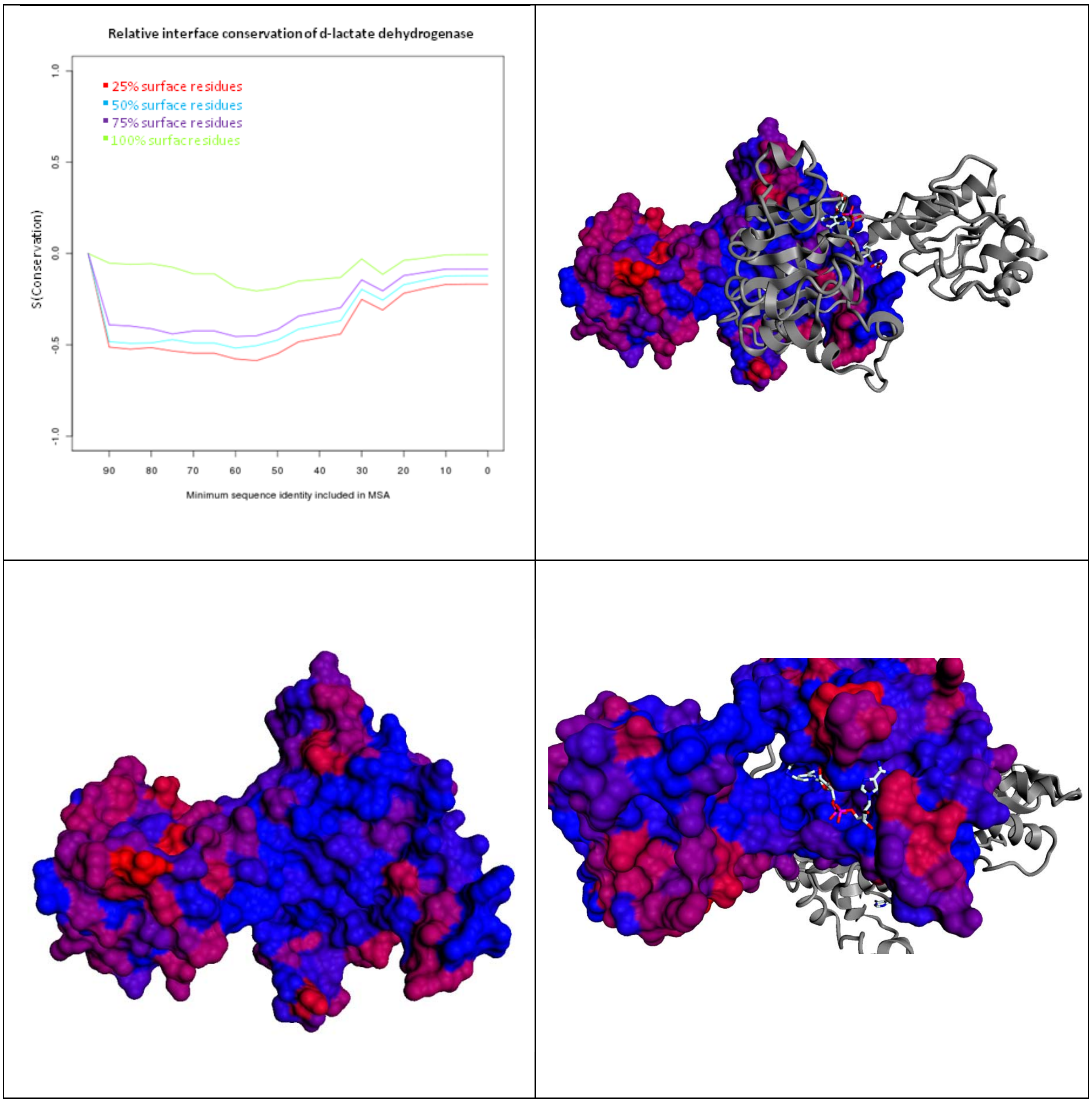


Figure 36 Analysis of the interface conservation for the d-lactate dehydrogenase (PDB-ID: "2dld"). The surface is colored according to the residue specific conservation (low entropy(blue); high entropy(red)) A) Using all surface residues results in a flat $S_{Conservation}$. Excluding the most conserved surface residues decreases $S_{Conservation}$ considerably B) The structure of 2dld, one chain is colored according to the entropy of each residue on a max alignment depth of 55%. C) A subunit of 2dld. The interface region has clearly shown lower entropy than the rest of the surface. D) The active site of 2dld is evolutionary constrained and increased the averaged surface entropy.

To exclude functional important surface residues other than interface residues, we only retained surface residues which were among the 75%/50%/25% of the residues reflecting the highest values for $S_{Column.}$ As illustrated in Figure 36 (A), the majority of $S_{Conservation}$-values were considerably lower with a decreasing cut-off of residues with high S_{Column} We recalculated the $S_{Conservation}$ scores for the remaing targets using in total 4 sets of surface residues.

As a result, three consecutive domains of $S_{Conservation}$ have been identified:

1. Decreasing $S_{Conservation}$-values:
   Sequences added to the MSA had a higher mutation rate at the surface than in the interface, which led negative $S_{Conservation}$ values.
2. Constant $S_{Conservation}$-values:
   Sequences added to the MSA had a stable mutation rate between surface and interface.
3. Increasing $S_{Conservation}$-values:
   By adding more distant related sequences, $S_{Conservation}$ increased. With increasing evolutionary distance the conservation of the quaternary structure was less likely, especially if the sequence identity was very low (<25% sequence identity). In addition, the number of sequences added at each bin increased exponentially. In combination with low sequence identities this decreased the chance to include sequences which shared the same quaternary structure.

The described characteristics can be observed in almost all targets in the testset. Differences were observed in the length of the section and the absolute values of $S_{Conservation}$.

The level of conservation of the interface compared to the surface calculated on different alignment depths can be seen as evolutionary snapshots. The shape and form of such curves follow a certain pattern and are therefore “evolutionary fingerprints”, as they represent the conservation of interfaces during evolution.

Analyzing the evolutionary fingerprints of superoxide dismutase

We evaluated in what sense the evolutionary fingerprint was different from what we would expect from native structures. We therefore mapped the surface/interface classification from the template to the target sequence and calculated an evolutionary fingerprint for each target/template (see Material and Methods for details) investigating how well one can distinguish correct from incorrect templates.

Figure 37 shows in the lower part the subunit of a dimeric manganese superoxide dismutase mutant from E.Coli (PDB-ID “1ix9”). Superoxide Dismutases (SOD) degrade superoxide anion radicals, which were toxic to biological systems [166] and appear mainly homodimeric or homotetrameric *in vivo*.[167]

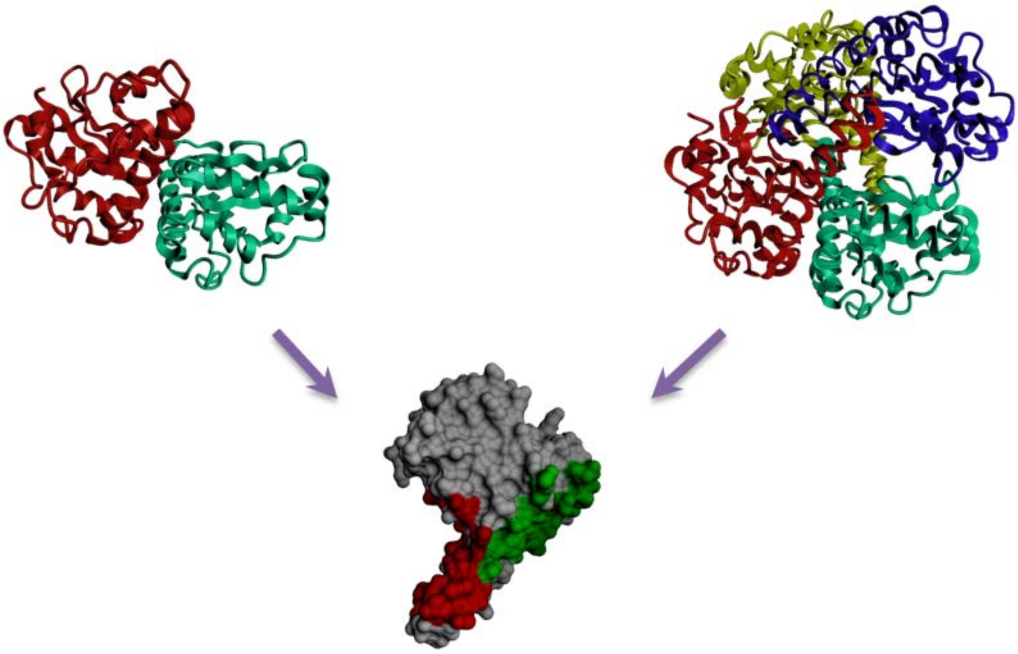


Figure 37 Template structures were identified at a sequence identity level of 40% in dimeric and tetrameric form (upper structures) for the manganese superoxide dismutase (SOD). The interface and surface residues were mapped onto one subunit of the SOD. The dimeric interface of the template is colored green. The tetrameric interface consists of two patches, the first is similar to the dimeric form (because of dihedral symmetry), the second patch is colored red.

The template structures, which were identified at an evolutionary distance around 40% sequence identity, were either dimeric or tetrameric (See Figure 37 for a schematic representation). We mapped the surface and interface residues of each template onto the target sequence and calculated the evolutionary fingerprint for each mapping separately. As shown in Figure 38, two diverse evolutionary fingerprints patterns can be observed. Template structures in tetrameric form show only slightly increased conservation compared to the surface. Evaluating the interface residues of the dimeric form results in an increased level of conservation compared to their surface.

This supports our hypothesis that interfaces occurring in the template but not in the target perform more “surface” like according to their conservation in evolution.

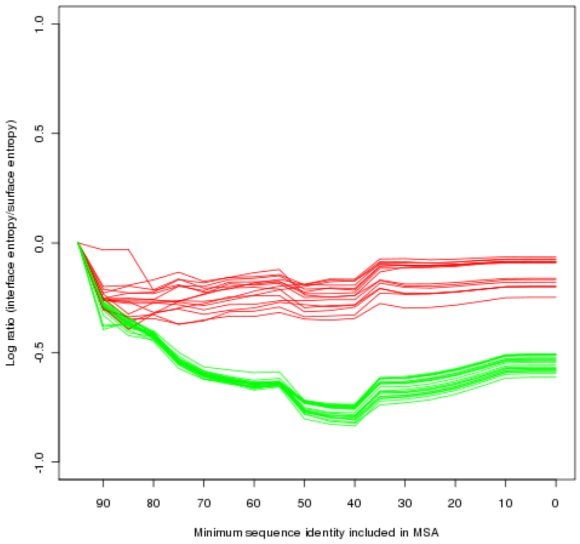


Figure 38 The evolutionary fingerprints of dimeric and tetrameric template structures for SOD, after mapping their interface and surface definitions to the target sequence. The correct ("dimeric") and incorrect ("tetrameric") quaternary structure is colored red and green, respectively. The set with 75% of the fewest conserved surface residues was used.

Distinguish conserved interfaces from incorrect interface using evolutionary information

Firstly, we investigated if the minimal $S_{conservation}$ can be distinguished between native and non-native quaternary structures. Therefore we identified the minimal $S_{Conservation}$ -values per template and target. The resulting distributions are shown in Figure 39. On average, $S_{Conservation}$ for native quaternary structures (-0.45) is lower than for non-native quaternary structures (-0.38). The difference was statistically significant on a 95% confidence level ("students t-test", $p < 2.2e-16$). However visual inspection of the distribution shows large overlapping areas between both distributions. As a result, defining an absolute cutoff to distinguish both classes was not useful in this context.

In order to evaluate if a classifier was able to learn the difference in shape and orientation between correctly assigned and incorrectly assigned quaternary structures, we used the machine learning method random forest[151]. All four surface residue sets were used in order to test if random forests perform considerably better on one set.

Figure 40 shows the ROC curves based on the classification of the random forests. The area under curve (AUC) was 0.65, 0.67, 0.67, and 0.66 for 100%, 75%, 50%, 25% of the fewest conserved surface residues, respectively. Only a minor decrease in AUC was observed if all surface residues were used. The fraction of true positives over all "correct" labeled fingerprints (Sensitivity) was rather low for all four sets. Again the sets with excluded surfaces residues perform slightly better than if all surface residues were used. However, all four classifier labeled correct quaternary structure assignments frequently as “incorrect”, thus under predicting “correct” fingerprints.

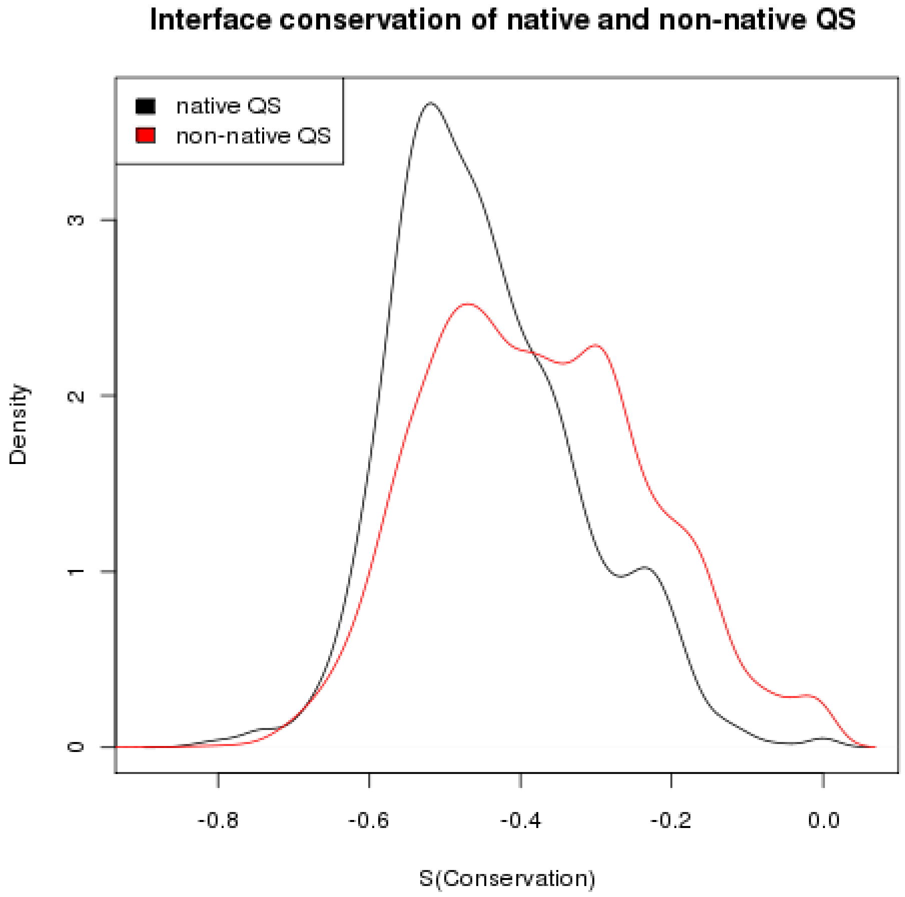
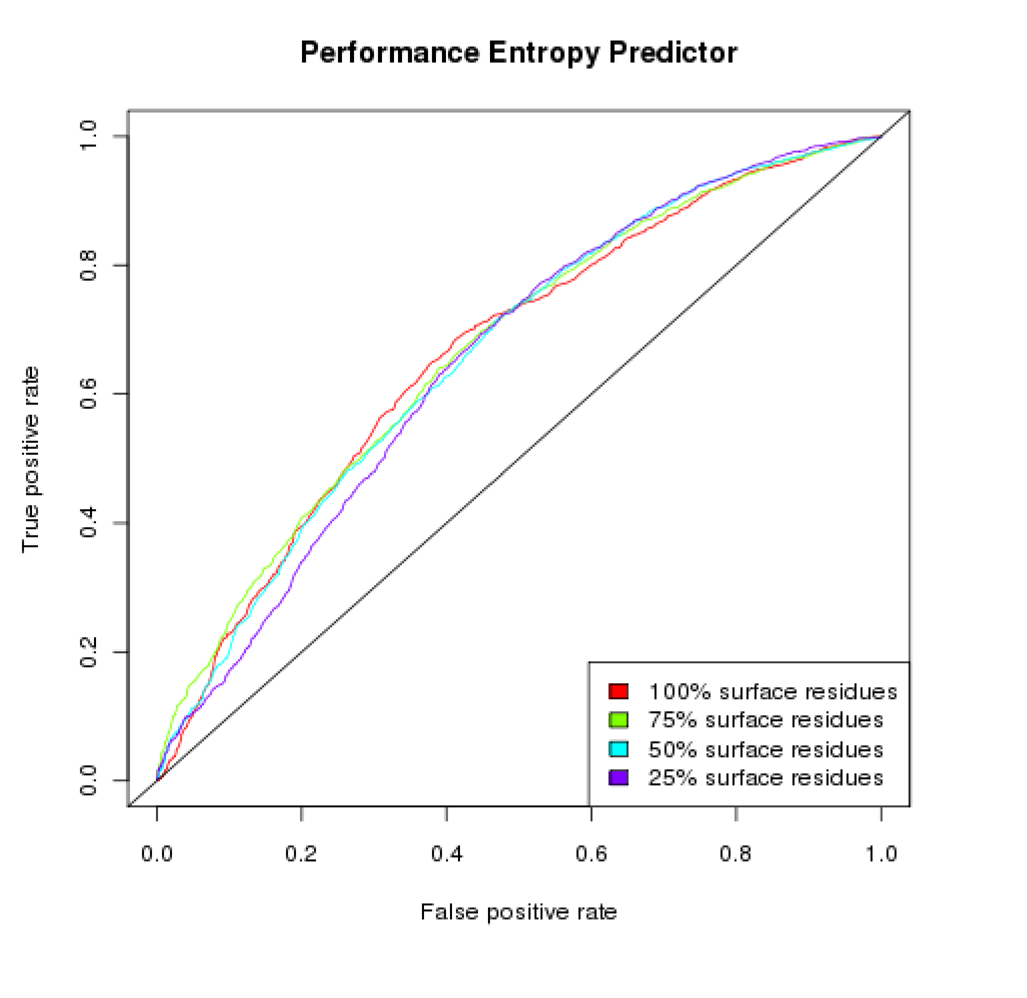


Figure 39 Distribution of the correct (black line) and incorrect assigned quaternary structures. The residue set with 75% of the surface residues was used.

The reason for this is twofold. Firstly, an interface was defined as buried area of one subunit, even if the interactions origin from different subunits. Due to symmetrical reasons, a tetramer can be seen as a dimer of dimer[116], leading to at least two distinct chainwise interfaces. When assigning surface and interface residues from a tetrameric template structure to a dimeric target structure, a significant part of the interface residues show interface typical conservation pattern despite dissimilar quaternary structures.

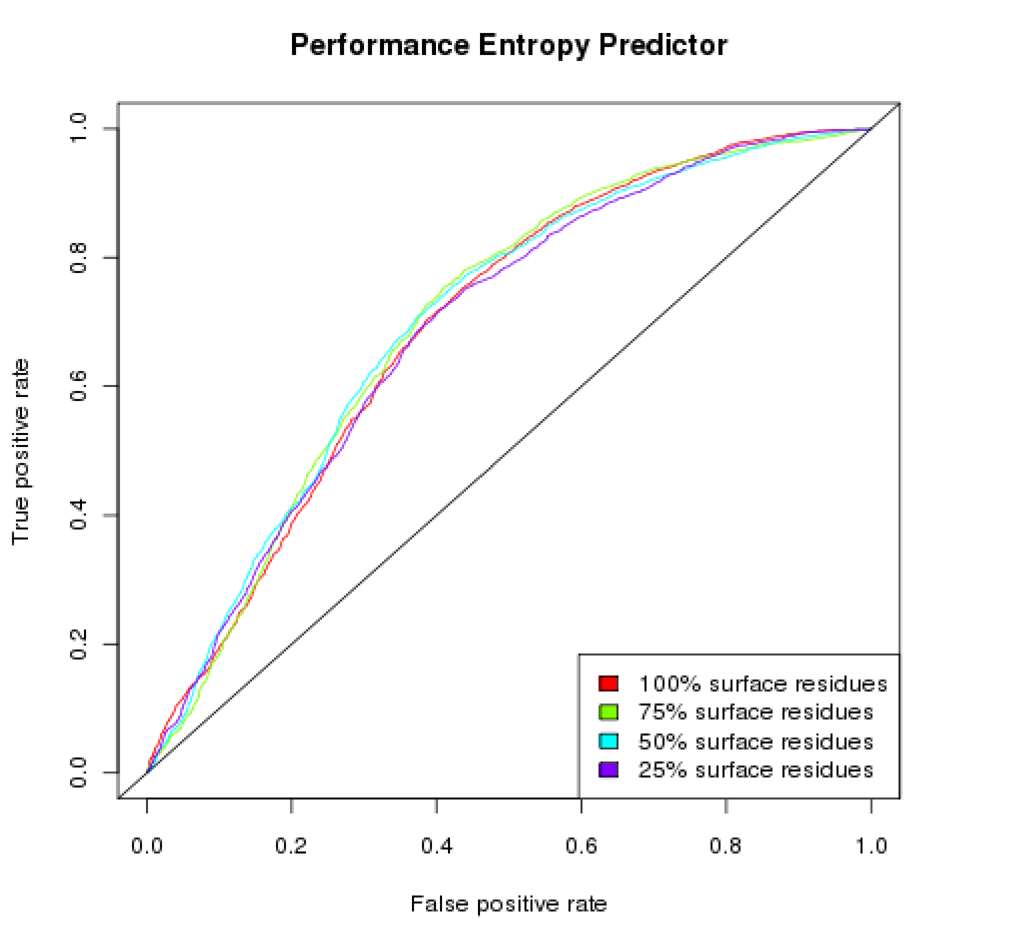
This leads to an incoherent signal, since some parts of the interface were more conserved compared to the surface whereas as other interface residues were not. Indeed if reviewing evolutionary fingerprints of all targets after assigning interface/surface residues of the template, a high number of targets have shown slightly lower $S_{Conservation}$ scores for dissimilar quaternary structures than for similar ones. However the change in shape and orientation between targets is much larger and thus the evolutionary fingerprints can be not separated by the classifier.



	Specificity	Sensitivity	AUC
100% surface residues	0.79±0.03	0.37±0.04	0.65±0.02
75% surface residues	0.83±0.03	0.37±0.05	0.67±0.01
50% surface residues	0.83±0.03	0.37±0.04	0.67±0.02
25% surface residues	0.84±0.03	0.35±0.05	0.66±0.02

Figure 40 Performance of the targetwise interface entropy calculation. The achieved area under the curve is improved compared to the complexwise .

To increase discrimination power, we calculated the evolutionary fingerprints of interfaces in a chainwise manner. The resulting ROC curves (Figure 41) reaches higher classification accuracy compared to the complexwise interface definition. Similar AUCs were observed for all four classifiers and were increased on average by 0.08 units. If considering only chain wise interfaces the sensitivity decreases to 0.63 for the set using 75% and 25% surface residues set and 0.59 for the set of 100% and 50% surface residues.



	Specificity	Sensitivity	AUC
100% surface residues	0.68±0.03	0.60±0.04	0.69±0.01
75% surface residues	0.64±0.04	0.64±0.05	0.70±0.02
50% surface residues	0.62±0.03	0.66±0.06	0.70±0.02
25% surface residues	0.59±0.04	0.66±0.06	0.68±0.01

Figure 41 Performance of the chainwise interface entropy calculation. The achieved area under the curve is improved compared to the complexwise.

The performance of identifying truly similar interfaces (sensitivity) is roughly doubled and being highest among the classifier using 100% and 25% of the fewest conserved surface residues.

The classification per chainwise interface requires additional efforts to estimate a probability of the whole complex.

To decide if a particular template structure is a "quaternary homologue" of the target, the probabilities of each chainwise interface classification has to be converted in order to calculate the likelihood of the total complex. This became important if the oligomeric state is larger than 2, which implies at least two interfaces. For example, a tetramer is often built by dimerization of an already assembled dimer (hexamers by dimerization of trimers etc.). This assembly path is observed more frequently than expected[116] and it is likely that larger interfaces were built earlier in evolution than smaller interfaces[118]. To decide if a proposed quaternary state is homologue to the target structure, a similar assembling procedure based on the predicted probabilities has to be performed. To calculate an overall probability a graphwise consideration of the problem is likely to be useful. Similarity between target and template regarding their quaternary state can be assumed if the given complex can be assembled by interfaces predicted to be similar.

The analysis has shown a better classification performance if excluding conserved surface residues. However the argumentation that conserved surface residues are always involved in functions which were not in related to complex assembly is precarious. If testing a dimeric template structure on a target protein which is tetrameric (D2-symmetry of the template dimer), the surface would consist of conserved residues, the ones which are in the tetrameric interface but not in the dimeric. Comparing the surface and interface entropy would result in a more balanced conservation ratio indicating that assigning a dimeric state is probable wrong. If excluding such wrongly assigned surface residue, the discrimination power decreases.

In summary, the applied approach of using evolutionary fingerprints to distinguish native from non-native interfaces works well, at least if binary interactions were used. The concept relies mainly on the hypothesis that assembly of disassembly of interface occur at a certain evolutionary distance. The assumption that assembly or disassembly of quaternary structures is a pure function of evolutionary distance was simplified. As discussed in the introduction single point mutation can lead to oligomerization[123] on one hand or disrupting of the complex on the other. Therefore the weighted mean was likely not sensitive enough to cover mutations of only a small number of residues.

Another source of errors are functional residues which are, for example, part of an interface in one complex, but also involved important for the active sites. The same residue would be still conserved in a monomeric version of the protein, even if the oligomeric state has changed.

Further work is required to characterize the function of conservation in the interface compared to the surface and the implication to oligomerization.

3.4.3.2.9 Analysis of mean force potentials

Mean force potentials were designed to discriminate between native and non-native protein models. We investigated to what extent mean force potentials can be used to discriminate models with correct quaternary structure from models with incorrect quaternary structure. Therefore, we calculated the QMEAN4norm score[168] for all models to assess the overall quality of the model. A QMEAN4norm of one indicates high agreement with experimentally solved structures, whereas a QMEAN4norm score of zero reflects non-native properties of the investigated protein structure.

Firstly, we examined if mean force potentials were able to distinguish between models with correct quaternary structure and incorrect quaternary structure. Therefore, the models were splitted according to their similarity to the target's quaternary structure. Figure 42 shows the quartiles of QMEAN4-norm scores of correctly and incorrectly models. On average, models with incorrect quaternary structure got lower QMEAN4norm-scores assigned than models with the correct quaternary structure. The difference between their medians was significant at the 5% level which is indicated by the non-overlapping notches in the boxplot.

The analyzed sets of models included also models which show large structural deviations to the target structure and hence got penalized by a low QMEAN4norm-score independent if the quaternary structure was correct or not. To exclude this effect, the structural agreement between target and model on the level of tertiary structure was calculated using the global distance test (GDT_TS)[29]. Hence, models with large structural deviations were excluded (GDT_TS <0.6). The resulting medians of QMEAN4-norm score were increased by 0.06 for correct models and 0.08 for incorrect models compared to the set which consisted of all models. The difference between the medians of the correct and incorrect set was statistically significant.

Finally, we investigated if QMEAN4-norm was able to distinguish between correct and incorrect quaternary structure if the models was built on an evolutionary distant template structures. Thus, all models based on templates sharing more than 30% sequence identity were excluded.

The QMEAN4-norm levels for both classes were decreased compared to the previous described testsets, but can be still distinguished on a significant level.

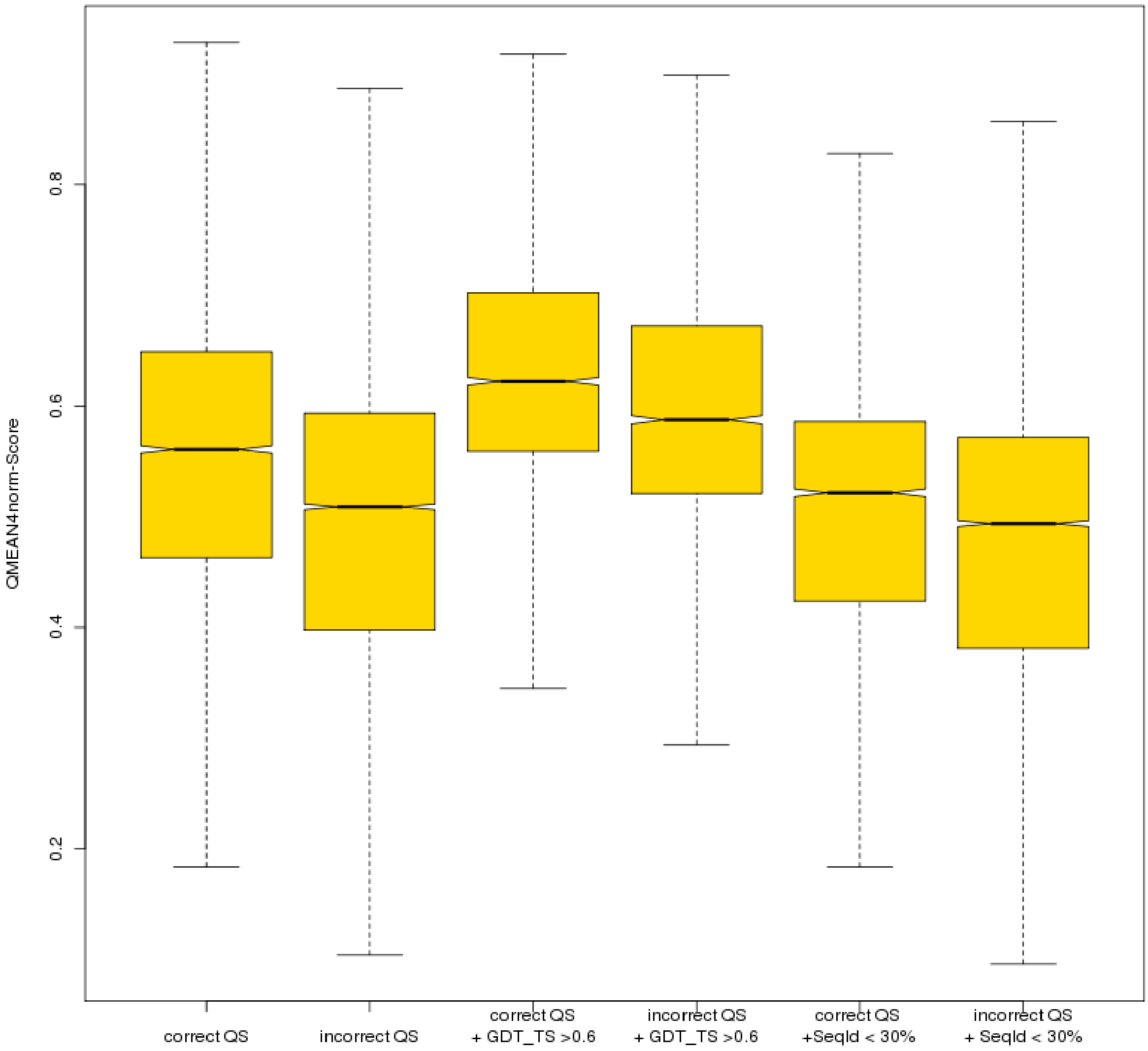


Figure 42 Differences in QMEAN score for different sets of target structures. Different set of models were used A) all models in the dataset B) All models where a subunit has a better GDT_TS score than 0.6 C) All models with sequence identity between model and template sequence is lower than 30%. In all three sets QMEAN was able to discriminate between models with the correct and incorrect quaternary structure.

In order to evaluate if the difference in QMEAN4norm-score was a general feature for all targets, we repeated the analysis on target level and asked if correct models received on average higher scores than incorrect models. As shown in Figure 43, 473 out of 571 targets had templates with correct and incorrect models. For 74.6% of the targets, models with correct quaternary structure reached a better QMEAN energy than incorrect models. For 120 targets the QMEAN energy was not able to discriminate between incorrect and correct template structures. We manually investigated some of the outliers having very low energy for their correct models. As a result, low energies were frequently assigned if the sequence identity between template and target is very low (< 20% sequence identity, although the quaternary structure was very similar) or a model consisted of large unaligned regions.

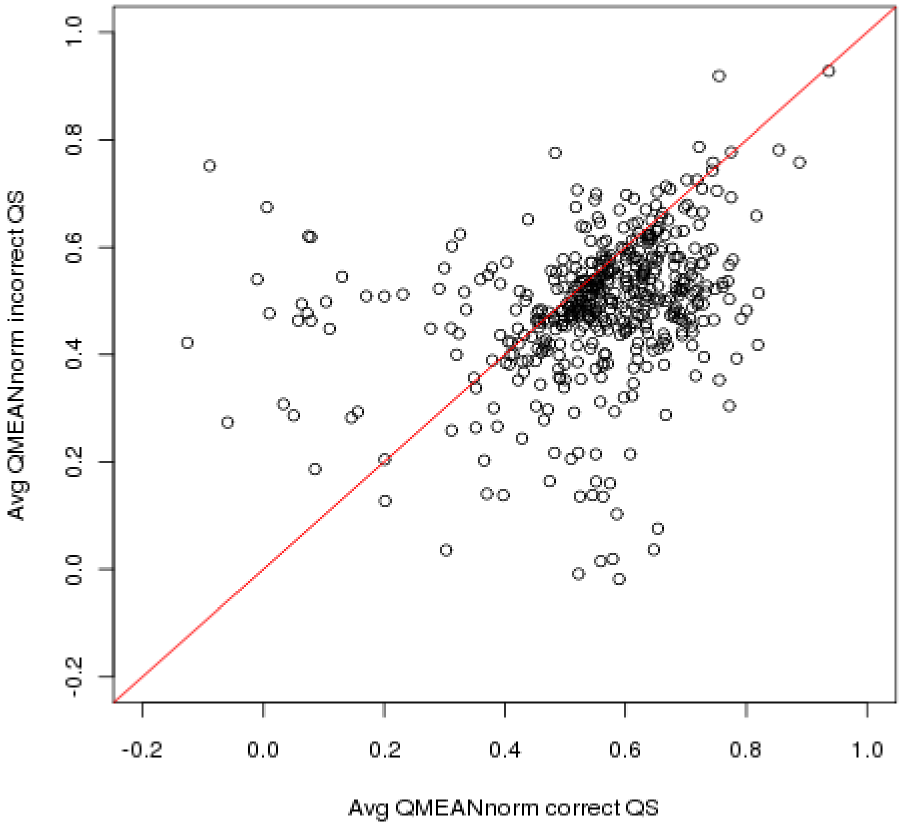


Figure 43 Averaged difference in QMEAN4norm between correct and incorrect groups. The average energy was calculated using the median, to prevent that outlier dominate the Energy. For a majority of target structures QMEAN4norm assigns higher scores for groups with the correct quaternary states.

In general, it can be concluded that the QMEANnorm was able to distinguish between models with correct and incorrect quaternary structure. However, models with incorrect quaternary structure but correct tertiary structure were superior to models with correct quaternary structure and unknown tertiary structure accuracy. The per target analysis of correct and incorrect models have shown that the majority of models were classified to be more accurately than incorrect models. It can be said that the quality as estimated by QMEANnorm-score depends much on other factors for example the accuracy of the tertiary structure. To estimate the interface accuracy directly a score which was trained on the physiochemical properties of native interfaces is required. Hence, energy functions used in Docking methods (e.g. RosettaDock[139]) can be used as a basis for further developments.

3.4.3.2.10 Final Classification of template structures

To classify if the quaternary structure of a template is similar to the target, we used random Forests. Random Forests consist of a large number of decision trees, built by random selection of an ensemble of input variables.[150] They are known to be very accurate and insensitive for overfitting effects. We used random Forest in order to train a classifier which predicts if the template shares its quaternary structure ("correct") with the target or not ("incorrect").

To assess the binary classification performance of the predictors we calculated the sensitivity and specificity. To evaluate the reliability of the given probabilities, receiver operating characteristics (ROC) curve were calculated. ROC curves reflect the true positive rate and the false positive rate under varying thresholds. The area under curve (AUC) can be interpreted as the ability of the classifier to rank positive hits higher than negative hits.[169]

In order to estimate the robustness of the classifier we applied bootstrapping to calculate the sensitivity, specificity and AUC after randomly labeling two thirds of the targets as training set and one third as test set (see Material and Methods for details).

The following input variables were used to train the classifier:

1. Sequence related characteristics
    a. Sequence identity in target-template sequence alignment
    b. Coverage of the target sequence
2. Group attributes (parameters were similar for all templates one group):
    a. PiQsi state of group
    b. Average evolutionary distance to the target sequence
    c. Absolute group size
    d. Relative group size
    e. Min-Max width of the group
    f. Distance between the 90%/10% quantiles
    g. Distance from the average sequence identity to the highest sequence identity in the group
3. Conservation in the protein interface

    a. Evolutionary fingerprint after mapping the interface/surface definition to the target

4. Model related characteristics
    a. QMEANnorm -score
    b. ASA_{ratio}

Firstly, we evaluated the performance of a predictor using sequence related characteristics. This includes the sequence identity and coverage between the target and template sequence. As shown in Table 5, predicting the reliability of the template structure only with sequence features is rather weak. Despite a good performance in identifying true negatives (i.e. templates which do not share their quaternary structure with the target), the sensitivity of 0.38 indicates limited capability to identify true positives (i.e. templates with similar quaternary structures to the target).

	Specificity	**Sensitivity**	**AUC**
Sequence	0.79±0.02	0.38±0.03	0.65±0.01
Sequence+Conservation	0.90±0.02	0.36±0.03	0.69±0.03
Sequence+Conservation+Model	0.90±0.01	0.40±0.03	0.72±0.03
Sequence+Conservation+Model+Group	0.89±0.02	0.75±0.07	0.92±0.02

Table 5 Performance of different characteristics. The specificity, sensitivity and area under curve were used as evaluation criteria. The use of group information increases the performance considerably.

By adding the information about the conservation of the interface in evolution to the classifier, the specificity was increased to a value of 0.9. This indicates that considering the evolutionary fingerprints improved the identification of incorrect templates. Compared to the classifier which used solely sequence features, the sensitivity remains on a similar level; AUC was slightly increased. Further, we trained a classifier using sequence features, conservation in the interface and model related features (i.e. the QMEANnorm score and the ASA_{ratio} of the model). As can be seen in Table 5, the accuracy was only slightly increased compared to the previous classifiers.

Finally, we added group specific characteristics of templates, which share similar quaternary structure among themselves. The overall performance increased significantly for two out of three scores. Whereas the specificity remains on a similar level, sensitivity and the AUC outperformed the previous classifier clearly. The sensitivity has been increased to 0.75, which indicates a fairly good performance in the identification of correct templates. In order to visualize the differences in the AUC among the different classifiers, we plotted the corresponding receiver operating curves; the AUC can be directly identified as area under the

ROC curve. Because the reported values in Table 5 are averaged over 10 independent training and testing iterations, an arbitrarily classifier was chosen for visualization. As shown in Figure 44, the ROC curve of the classifier based on all input features (colored in red) clearly outperforms the remaining three classifiers in terms of AUC. This indicates the group characteristics contribute significantly to the classification accuracy.

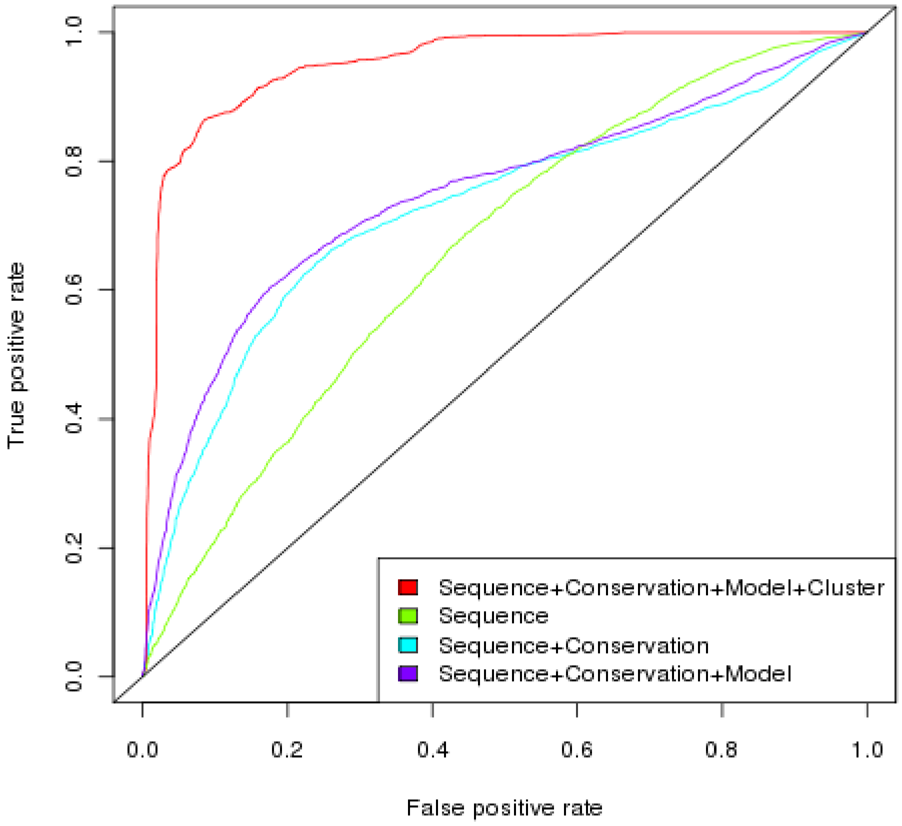


Figure 44 ROC curves for different classifiers. The classifier which uses all features clearly outperformed the remaining set of classifier.

To estimate if it is more difficult to classify templates of oligomeric target structures or of monomeric target structures, we evaluated the performance separately using a classifier with all input features (Sequences+Conservation+Model+Group).

Monomeric templates were correctly identified for the vast majority of the monomeric target structures (Sensitivity: 0.96), however the specificity decreases compared to the complete testset by 0.13. For oligomeric targets, the specificity is similar to the overall testset but lower compared the monomeric set. In contrast the sensitivity has improved to 0.81 (compared to 0.75 for the complete testset and 0.62 for the monomeric testset). The area under curve is similar for both sets and slightly better than for the complete set. As a result, the classifier is more successful on classifying template structures for oligomeric targets than for monomeric targets.

	Specificity	Sensitivity	AUC
Monomeric targets	0.96±0.04	0.62±0.14	0.95±0.02
Oligomeric targets	0.89±0.03	0.81±0.08	0.94±0.02

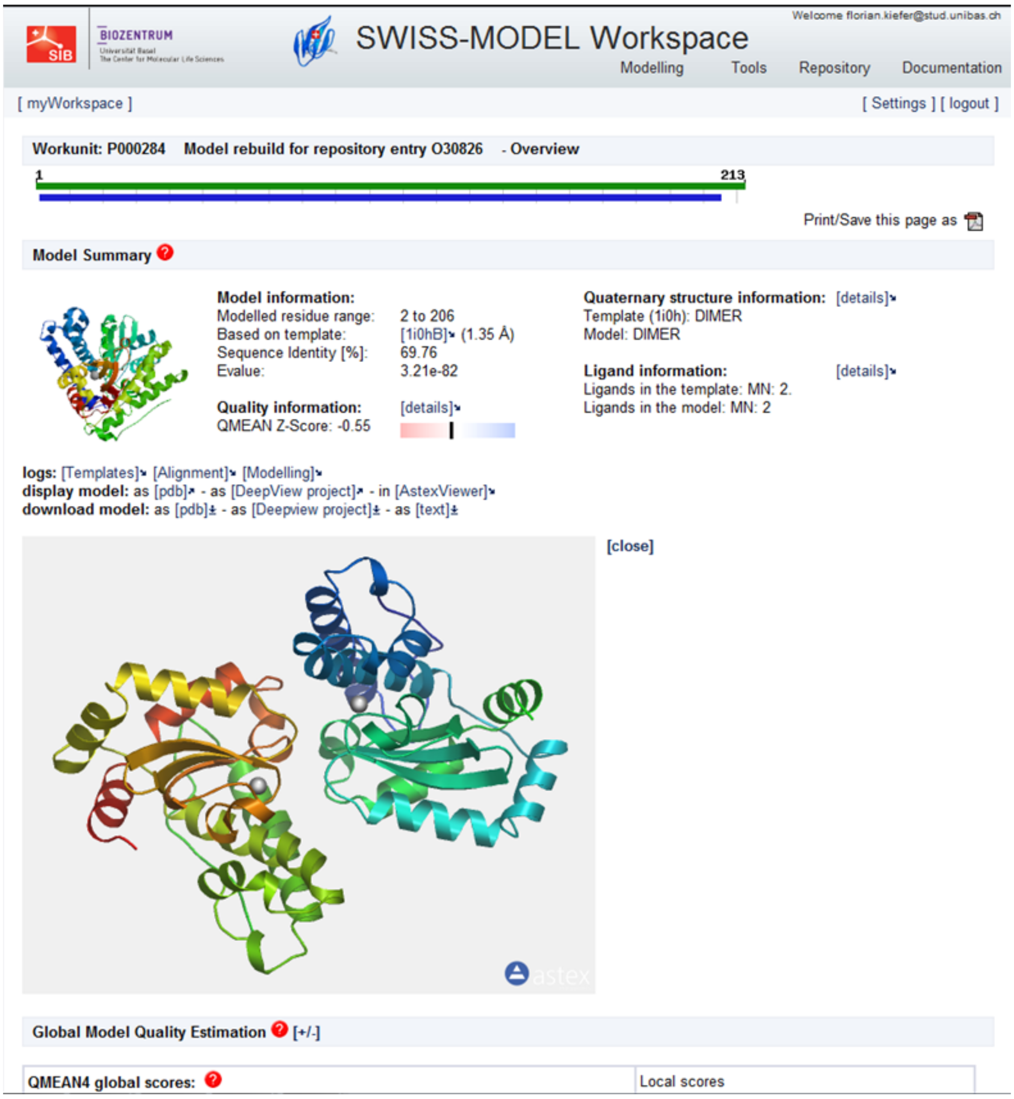
Table 6 Comparison of prediction accuracy between oligomeric and monomeric targets. All input features were used to train the random Forest. Predicting monomeric targets reveal in an higher sensitivity but lower specificity compared to oligomeric targets.

In summary, a classifier based on input variables deduced from sequence, conservation and model based parameters reaches a remarkable classification accuracy. When considering the balance between sensitivity and specificity, the classifier tends to be rather conservative in labeling template structures “correct” which is indicated by a low sensitivity. Based on the given results it became evident that the classifier has problems to classify correctly monomeric templates for monomeric targets. A likely explanation is the different performance of monomeric templates in terms of clustering. An analysis of the input parameters which contributed most to prediction success have shown that attributes of the groups are most responsible for the classification success. Monomeric templates are in that sense different because they always become merged in the same cluster. Based on these observations, it is likely that the accuracy can be improved if training and testing a classifier for the prediction of monomeric targets only.

A rigorous classification of templates regarding their quaternary structure similarity to the target is not known to the authors. Modeling of quaternary structures is currently mainly assessed by two blind tests. Predictions are submitted in the context of CAPRI installment involves often the assembly of complexes by docking techniques, such as RosettaDock[170] or HARDDOCK[138]. However, docking procedures can be compared with ab initio and thus is out of the scope of this method. Another category within CAPRI is the ranking of complexes and the identification of the nearest native model. It cannot be expected that our classification is accurate in this type of benchmark, because our method requires a reasonable amount of templates. The template based modeling category within the CASP installments provides a benchmark sets which fits well the criteria for an external benchmark set. As shown in the last CASP assessment most predictors do not submit any oligomeric model, and those which did, performed worse (except one) than a naïve predictor[7]. We will assess the accuracy of our predictors by using the CASP testset and try to identify bottlenecks which need to be addressed in the future.

3.5 Current Implementation within the SWISS-MODEL homology pipeline

A rudimentary oligomer modeling protocol was integrated in the SWISS-MODEL homology pipeline (see Chapter 2). Based on our findings of the conservation of the quaternary structure related to the evolutionary distance (Figure 30), we decided to classify the templates quaternary structure as correct, if the sequence identity exceeds 60%. On this level the probability to observe a correct template was almost 70%. Currently only homo-oligomeric complexes are supported.



Workunit: P000284   Model rebuild for repository entry O30826   - Overview

Model Summary

Model information:

Modelled residue range:	2 to 206
Based on template:	[1i0hB] (1.35 Å)
Sequence Identity [%]:	69.76
Evalue:	3.21e-82

Quality information: [details]
QMEAN Z-Score: -0.55

Quaternary structure information: [details]
Template (1i0h): DIMER
Model: DIMER

Ligand information: [details]
Ligands in the template: MN: 2.
Ligands in the model: MN: 2

logs: [Templates] [Alignment] [Modelling]
display model: as [pdb] - as [DeepView project] - in [AstexViewer]
download model: as [pdb] - as [Deepview project] - as [text]

Global Model Quality Estimation

QMEAN4 global scores:

Figure 45 Quaternary structure prediction for a Mn superoxide dismutase of Haemophilus ducreyi. The model was predicted as dimer and consists of two Manganese ions.

Figure 45 shows a model of a superoxide dismutase from *Haemophilus ducreyi*. The model was built on 1i0h, a manganese superoxide dismutase of *E. Coli*., sharing around 70% sequence identity. It has to be noted that the model contains two manganese ions and thus represent a protein in its biological active form. Based on the discussed results, an improved oligomeric modeling protocol will be soon implemented.

3.6 Outlook

3.6.1 Amino acid composition

A well-known characteristic of interfaces is the different amino acid composition in the interface compared to the surface. Several studies have shown that large hydrophobic residues are more abundant in the interface than in surface.[103,113,171,172] In contrast, polar residues like lysine or aspartic acid appear more frequent in the surface due to their hydrophilic character. A preliminary analysis confirmed this also for our dataset. Figure 46 (upper panel) shows the amino acid composition of all targets in the dataset. As a result, hydrophobic residues appear more frequent in interfaces than in the surface (as indicated by a positive log odd ratio). However, it is less likely to observe large hydrophobic residues in general (Figure 46, lower panel).

In general this could be used as a criterion to distinguish if an interface remains its typical "composition" during evolution. In principle one could calculate the difference between the residue composition of the template and the residue composition in the hypothetical target interface. However, such a procedure would neglect the fact, that for example large hydrophobic residues could be mutated to residues which are involved into the formation of salt bridges and thus remaining their interface stabilization character.

A likely more successful approach would be the combination of evolutionary related sequence. The evolutionary fingerprint introduced in the section "Conservation of interface residues during evolution" could be adapted to evaluate how the amino acid composition changes by evolutionary events. Hypothesizing, that a proposed interface which is conserved in evolution would remain interface typical residues. Nevertheless, it was shown that the mutation of single residues can lead to interface disruption[123] in such cases the interface composition of the remaining interfaces remain identical.

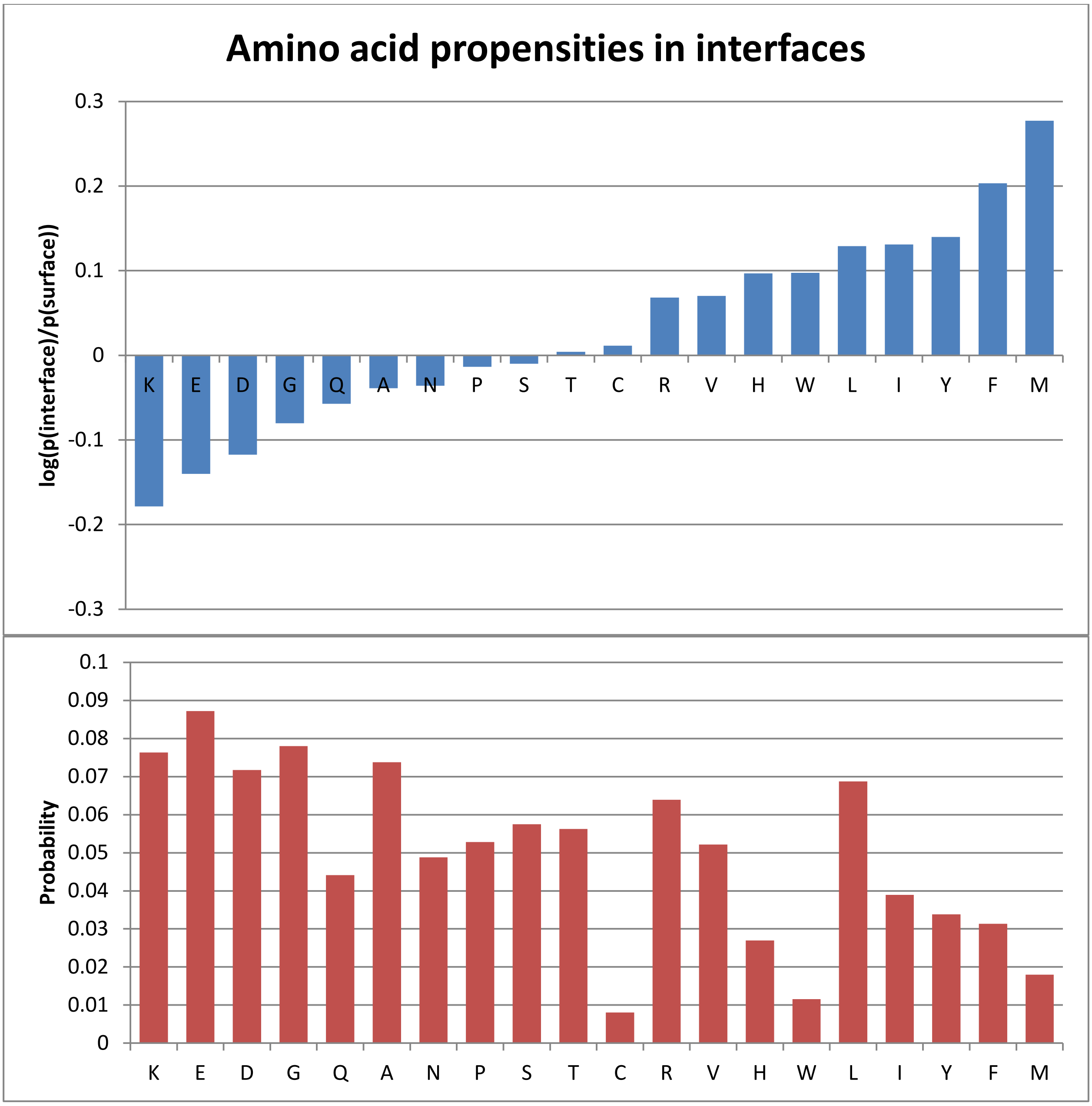


Figure 46 Amino acid preferences for all target interfaces in the dataset. Hydrophobic residues are overrepresented in interfaces compared to surfaces (upper panel), although their occurrence is in general lower (except Leucine). Only surface/interface residues were used to calculate the statistics

3.6.2 Coevolution of residues

Another aspect which was used to predict interaction sites, are coevolving residues. Interface residues which are essential for interface stability are known to be more conserved in evolution than surface residues (see introduction “Evolution of oligomeric complexes”). As a consequence the mutation of interfacial residues requires also adaption of the interacting residues to avoid electrostatically non-favored situations. For example, Burger[173] uses a Bayesian approach to identify pairs of residues which had similar mutation pattern and thus are more likely to be close in space in than pairs of residues which uncorrelated mutations.

This is a valuable approach for the identification of interaction sites between two proteins. However, homo-oligomeric complexes are not well suited for this type of analysis because of their symmetrical arrangement of interacting resides. Residue *i* can coevolve with residue *j* because of “internal” reasons, e.g. they are close in space within one subunits, or they can coevolve because residue *i* interacts with residue *j* in the other subunit. In addition, it becomes obvious that a residue which is conserved in one interface is also conserved in the other interface. This can be explained by the symmetrical arrangement of most oligomeric assemblies.

3.6.3 Hetero-Oligomer

When predicting homo-oligomeric assemblies, it comes naturally to the question of prediction of hetero-oligomeric complexes.

The prediction of hetero-oligomeric complexes was targeted by many researchers. The field can be divided into two parts. The prediction of obligate hetero-oligomeric complexes can be compared to the prediction of homo-oligomeric assemblies as described in this work. As an extension to the described comparative modeling protocol, all sequences which are involved into the target protein must be known beforehand. For example the prediction of the hetero-dimeric complex hemoglobin, requires the sequences of the α-chain and the β-chain. Both sequences need to be identified in a template structure to apply comparative modeling and to avoid protein-protein docking.

The described techniques for the evaluation of the similarity between template and target interface can be in principles also used for the evaluation of hetero-oligomeric complexes. It is assumed that the clustering approach as well as the conservation in interfaces gives similar results as for homo-oligomeric complexes. For prediction of hetero-oligomeric complexes, the identification of coevolving residues is likely beneficial.

Protein-Protein Interaction (PPI) plays an important role in the cell since they are often involved into regulation of signaling pathways. The application of large scale experiments like yeast to hybrid methods, results in a large amount of characterized protein-protein interactions. Such data consists often of binary interactions between two proteins. In general, interfaces of transient complexes are not showing such extensive interface characteristics as described for obligate interfaces and, thus, are difficult to predict by such features.

4 Assessment of disorder predictions in CASP7

The widely accepted sequence-to -structure-to-function has been established for many years in the scientific community and has been derived from structural genomics.[174] However, in the previous decades functional active but intrinsically flexible proteins has been observed. For a long time such proteins have been seen as rather as an exception then the rule. However, it has been shown that the functional impact of such proteins is wide. As a consequence, a new class of proteins can be defined, where the function is not directly linked to a well-defined three-dimensional structure and determined by the "unstructured" character itself. Such proteins are usually highly dynamic and non-uniform and are therefore often call "intrinsically disordered".[175] Disordered proteins can be broadly classified into two types: (1) Proteins which are disordered throughout their complete length ("natively unfolded proteins") and (2) proteins which consist of long disordered stretches (>30-40 residues) but are structural well-defined otherwise.

Since the sequence of a protein determines the three-dimensional structure, it can be assumed that the sequence also determines the disorder of non-folding protein structures. It has been shown that sequences of disordered protein are depleted by hydrophobic residues such as C, W, Y, F, I, V, and L and enriched in M, K, R, S, Q, P, and E.[176] It makes it easy to understand, that a reduced level of hydrophobic residues and a higher level of polar residues hinder the folding process.[177]

Several studies applied prediction methods to various genomes in order to predict the percentage of disordered proteins in a certain proteome.[99,178] It has been proposed that about 25 to 30% of eukaryotic proteins are mostly disordered[178] and that more of the half of them have at least long regions of disorder[179]. In addition, it has been shown that more than 70% of the signaling proteins have long disordered regions.[180] In contrast, bacteria and archaea were predicted to have much lower rates of long disordered regions in their genomes, ranging from 16-45% and 26-51%, respectively.[178,179] The increased level of disorder in eukaryotic systems if very likely related to increased cellular signaling.[181]

More than 30 different types of functions which are linked to disorder have been identified. Most of them are connected to cell cycle control, and transcriptional and translational control and indicate the large functional importance of disordered structures for the cell.[182]

In the following, a published manuscript is included:

“Assessment of disorder predictions in CASP7”

My contributions were the follows:

- Implementing a framework for the statistical analysis of the submitted predictions
- Calculation of the scores

Abstract

Intrinsically unstructured regions in proteins have been associated with numerous important biological cellular functions. As measuring native disorder experimentally is technically challenging, computational methods for prediction of disordered regions in a protein have gained much interest in recent years. As part of the seventh Critical Assessment of Techniques for Protein Structure Prediction (CASP7), we have assessed 19 methods for disorder prediction based on 96 target proteins. Prediction accuracy was assessed using detailed numerical comparison between the predicted disorder and the experimental structures. On average, methods participating in CASP7 have improved in accuracy in comparison to the previous assessment in CASP6. Overall, however, no improvement over the best methods in CASP6 was observed in CASP7. Significant differences between different prediction methods were identified with regard to their sensitivity and specificity in correctly predicting ordered and disordered residues based on a protein target sequence, which is of relevance for practical applications of these computational tools.

Introduction

Intrinsic disorder in proteins, i.e. the presence of unstructured regions in functional proteins, has been a focus of much attention recently, as it has been shown to be implicated in important biological roles, such as translation and transcriptional regulation, cell signaling and molecular recognition in general. Several studies report indeed examples of disordered proteins implicated in important cellular processes, undergoing transitions to more structured states upon binding to their target ligand, DNA, or other proteins (for review see references [177,183,184]).

In recent years much effort has been invested in the experimental characterization of native disorder in proteins as well as in the development of predictive methods to gain more insights into the functional and biological role of natively unfolded proteins.[185-187] For instance, whole genomes have been scanned *in silico* for the presence of disordered regions in order to examine the frequency of unstructured regions in different organisms and to provide hints of the different biological role they might be involved.[99,182] New biological functions linked to native disorder are emerging, such as self-assembly of multiprotein complexes or involvement in RNA and protein chaperones.[188,189] Therefore computer aided methods for detecting disordered regions are becoming a valuable tool for the functional annotation of proteomes and the design of laboratory experiments aimed at identifying interaction or regulatory sites.

The presence of unstructured regions in proteins is also known to complicate high-throughput structural determination, as they can hinder the crystallization of proteins or interfere with NMR experiments. To overcome these problems, computational approaches have been used to screen for such elements, thus complementing the use of programs to detect low complexity regions in protein sequences .[190,191]

Since the first disorder prediction method was developed a decade ago,[192] an increasing number of groups have been developing methods to predict the occurrence of native disorder in proteins. Starting with CASP5 in 2002, the accuracy of disorder prediction methods has been assessed as part of the experiment.[193,194] In this paper we present the detailed numerical evaluation of the predictions submitted by 19 groups participating in CASP7 in the category of disorder prediction. Predictions were compared with experimental structures for 96 target proteins, 85 of which solved by X-ray and 11 by NMR experiments.[195] The number of structures solved by NMR available for the assessment has increased since 2002, but still the dataset is mainly characterized by protein structures solved by crystallography, which are generally known to contain only relatively short unstructured regions. For this reason the result of the present assessment may only be partially indicative for the accuracy of the participating methods in predicting longer regions of disorder.

Methods

Data processing and definition of disordered residues in CASP7 targets

The assessment of the disorder category of the CASP7 experiment consisted of the evaluation of 1694 predictions for 96 protein targets from 11 expert and 8 server groups (see Table I). The majority of the groups submitted predictions for more than 80% of the targets, one group made predictions for only 10 targets. The format for the submitted predictions corresponds to the format of previous CASP experiments.[193,194] For each residue of the target a binary classification for order or disorder should be assigned (O/D) together with a measure of the probability (P) for the residue of being disordered, a real number between 0 and 1.

For the assessment of disorder prediction in CASP7, residues in 96 target structures were classified as "ordered" or "disordered" respectively. Residues in targets solved by X-ray structures were classified as disordered if no coordinates for the crystallized residues were present. For targets solved by NMR, those residues whose conformation was not sufficiently defined by NMR restraints, i.e. exhibit high variability within the ensemble or were annotated as disordered in REMARK 465 by the experimentalists, were considered as disordered. At the time of our assessment, the target sequences given to the predictors were compared with the structures of the targets deposited in the Protein Data Bank (PDB)[14] or submitted to the organizers of the experiment. In case of discrepancies, the sequence deposited in the PDB (or by the organizers) was considered. In addition, for 5 out of 96 targets no information about the sequence of the expressed protein used for the experiment (SEQRES) was reported in the submitted structure file. For these targets, the N- and C- termini of the assessed target sequence were defined by the first and last residue for which coordinates were present. Thus, 19,816 of the initial 19,891 residues were used for the assessment.

The fraction of disordered residues of the 96 targets ranges from 0% (targets T0283, T0286, T0290, T0297, T0303, T0308, T0319, T0329, T0340, T045, T0346, T0367, T0371, T0374, T0375 and T0386) up to 60% for target T0352, which is 117 resides long and contains 66 disordered residues. Figure 1 shows the distribution of the length of disordered regions in CASP7. The dataset is characterized by a relatively high number of short disordered regions (smaller than 10 residues) whereas few long disordered regions are present. Target T0351 contains the longest disordered stretch which is 47 residues long. Overall, the number of disordered residues in the 96 targets is 1189, representing 6% of the total number of residues.

For the analysis of the accuracy of the predictions, we have chosen to evaluate only disordered segments in the experimental target structures with more than 3 contiguous disordered residues. Shorter regions are more likely to represent experimental artifacts than intrinsic disorder, and have therefore been omitted in the present work.

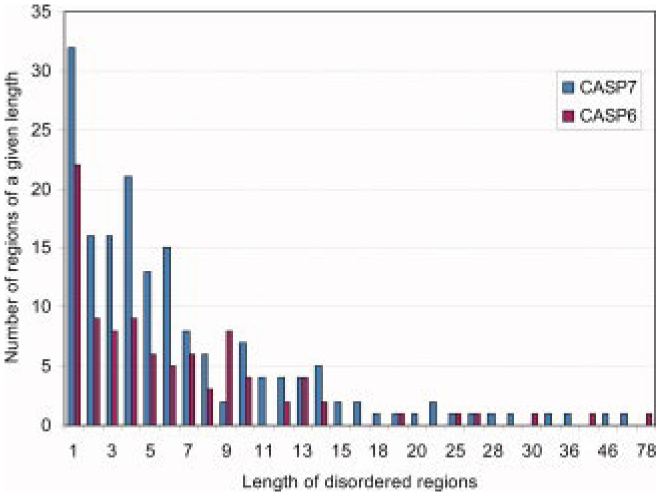


Figure 47. Length distribution of disordered regions in CASP6 and CASP7 targets. Bars in the graph correspond to the number of regions of a given length specified on the x-axis of the plot. Both dataset are biased towards a higher number of of relatively short disordered regions.

Evaluation Criteria

The predictions of the different participating groups were assessed on a per-residue level. Based on the P values assigned to each prediction, the results were compared using receiver operating characteristic (ROC) curves, a widely used method to assess the performance of a classifier system and its dependence upon its discrimination threshold. This method has been applied by several investigators in the field and in the previous CASP experiment. [99,194] For each threshold value of P (the probability of being disordered) the fraction of true positive predictions was plotted versus the fraction of false positives, whereby residues with a P value equal to or greater than the threshold value were considered disordered. The performance criterion used for our analysis is the area under the curve (AUC), which was computed using the trapezoid rule.[152] It has been demonstrated that there is a clear relationship between this quantity and the Wilcoxon (or Mann-Whitney) statistics.[196] The value for the AUC ranges from 0.5 to 1, in the case of a random classifier and perfect predictor respectively. In contrast to CASP6, the CASP7 predictor groups assigned sufficiently distinct P values to their yes/no predictions, allowing comparison of the results with smooth ROC curves (Figure 2, Table II). The only exception is group 284 who made use of only P values of 1 and 0.

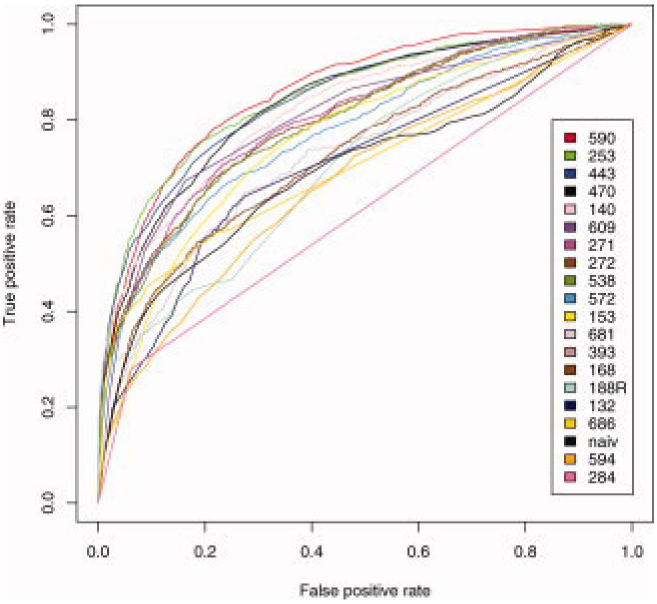


Figure 48. ROC curves of disorder predictions submitted by 19 groups and generated by the naïve predictor. The area under the curve (AUC) was used as a measure for the accuracy of the individual methods. The majority of the groups used continuous *P*-values associated with their disorder predictions. Group 284 used only two values 0 and 1 for *P*. For the naive predictor *P*-values were assigned as*P* = 1/(1 + separation from terminus).

In addition, the different group's predictions were evaluated by how well their binary classification correctly identifies the negatives cases (Specificity) or the positive cases (Sensitivity):

$$Sensitivity = S_{sens} = \frac{TP}{TP + FN} = \frac{TP}{N_{disorder}}$$

$$Specificity = S_{spec} = \frac{TN}{TN + FP} = \frac{TN}{N_{order}}$$

where TP is the number of true positives (correctly identified disordered residues), FP false positives (predicted as disordered, but experimentally ordered), TN true negatives (correctly identified ordered residues), and FN false negatives (predicted as ordered but experimentally disordered), respectively. A well performing prediction method would have both high sensitivity and specificity. These measures can therefore be combined into a single one as the product or the average, which has been used by some investigators as a measure of the overall accuracy (ACC):[197]

$$\sqrt{S_{product}} = \sqrt{S_{sens} \times S_{spec}} = \sqrt{\frac{TP \times TN}{N_{disorder} \times N_{order}}}$$

$$ACC = \frac{S_{sens} + S_{spec}}{2}$$

As already indicated in the previous two CASP experiments, a simple Q2 measure, similar to the Q3 measure for the evaluation of secondary structure prediction algorithms, is not appropriate due to the unbalanced rate of ordered versus disordered residues in the dataset. By simply predicting all the residues as ordered, this would yield a Q2 of on average 90%. A weighted score, rewarding a correctly predicted disordered residue more than an ordered one, overcomes this problem. Such measure adopted for the assessment of the present experiment is the weighted score Sw introduced by Dunbrack and coworkers in CASP6:[194]

$$S_w = \frac{S}{S_{\max}} = \frac{W_{disorder}TP - W_{order}FP + W_{order}TN - W_{disorder}FN}{W_{disorder}N_{disorder} + W_{order}N_{order}}$$

The S_w score ranges from -1 to +1 and predicting all the residues in the targets to be ordered would results in a score equal to 0. $W_{disorder}$ and W_{order} are adjustable weights which in the present work were set to the rates of ordered and disordered residues respectively, i.e. $W_{disorder}$ = 94.53 *and* W_{order} = 5.47 for groups predicting all targets. As different groups may have predicted different subsets of targets, the specific values for $W_{disorder}$ and W_{order} for each group vary slightly (Table II). In general, the frequency of disordered residues at the N- and C-termini of proteins is higher than average. For this reason we decided to include in our assessment several naïve predictors, assigning between 1 and 10 residues at the N- and C-termini of the target sequences as disordered.

To test the statistical significance of the assessment we used a bootstrapping procedure: 80% of randomly chosen target structures were assessed repeatedly 1000 times to derive standard error of each binary score reported in Table II. For ROC curves, the standard errors were estimated according to Hanley and McNeil.[196] Finally, the statistical significance of the difference in performance between the various groups was tested by comparing the ROC AUC measures using two different non-parametric tests for comparing ROC curves from correlated samples.[198,199] This statistical analysis was performed using the statistical packages MedCalc version 9.2.1.0 (MedCalc Software, Mariakerke, Belgium) and Accumetric version 1.1 (Accumetric Corporation, Montreal, Quebec).

Results

Evaluation of prediction accuracy

Figure 2 shows the ROC curves for the 19 groups that participated in the experiment, and additionally for the different naïve predictors. As noted before only the results of group 284 are not assessed well by ROC curves since they used only two discrimination cut-offs. The values of the area under the curve together with the results of the different binary predictions are summarized in Table II and plotted in Figure 3 ranked by the ROC AUC. According to the accuracy measured by ROC curves, groups 590, 253, 443 and 470 are the best performing groups and differ significantly from the others as evidenced in the statistical analysis test reported in Table III. Group 590 has the highest sensitivity and Sw score of all the participating groups. The Sw score rewards correct disorder predictions but penalizes to a lesser extent incorrect disorder prediction and this is reflected in the relatively high (compared to the top groups) FP rate (1-S_{spec}) of this group. The same is true for the ACC, which is highly correlated to the Sw score. On the other hand group 253 has a lower Sw score than 590 - still in the same range of the rest of the top groups - but compensates with a lower FP rate. The results of group 443 lie in between these two examples, being on average less sensitive but more specific than group 590 and less specific but more sensitive than group 253. Group 470 is characterized by a relatively high specificity, comparable to group 253, although it is slightly less sensitive. The results of the other participating groups are either characterized by a FP rate in the range of group 590 or 443, but do not match the same high Sw or Sensitivity scores. Or possess a lower FP rate but are less sensitive than groups 253 and 470.

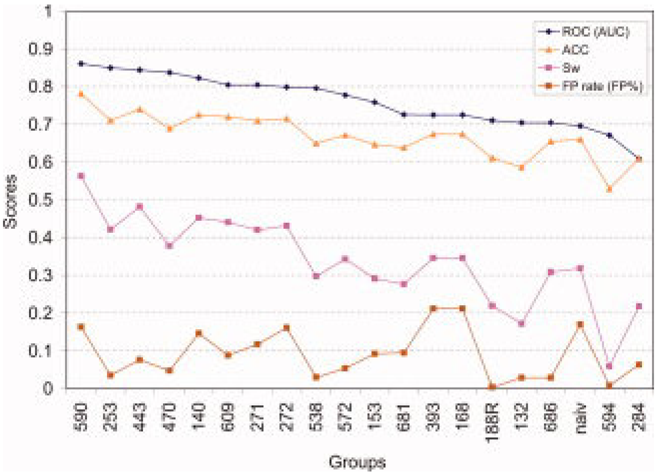


Figure 49 Assessment of disorder prediction by different scores. For each group, the area under the ROC curve (AUC), ACC, S_w scores, and FP rate are shown. In this plot, groups are ordered according to decreasing AUC of their respective ROC curves. The numerical values of each score are reported in Table II.

We were further interested to see if the methods would perform differently on short or long segments of protein disorder and evaluated target regions with disordered stretches longer or shorter than 10 residues separately. This slightly deviates from the definition of long and short regions of other authors.[99,197] However as mentioned earlier, the CASP7 dataset is biased towards relatively short disordered regions and in this way the two target subsets would contain a comparable number of disordered residues: 646 for the “longer” subset and 431 for the “shorter” one. To obtain these two subsets, regions shorter (but longer than 3) or longer than 10 residues were eliminated from the target sequences for the calculation of the different scores. The results are heterogeneous: some groups perform better on "longer" disordered regions (e.g., group 253 and group 470), whereas others appear more accurate on "shorter" regions. Group 590 has comparable accuracy on both long and short disordered regions. The prediction method described in the abstract of group 590 is indeed a specialized predictor for both short (≤ 30) and long (> 30) disordered regions.[197] Group 253 also describes a method specially designed for short (< 40) or long (> 40) disordered regions.[200] However, since only 2 regions longer than 40 residues are present in the CASP7 targets, this dataset is not suitable for the assessment of prediction methods specialized for long disordered regions of more than 40 residues.

Since disordered regions are commonly found at the amino and carboxyl ends of proteins we also compared the results of predicting terminal versus internal disordered regions in proteins. 10 consecutive residues were removed at both N- and C- termini from the target sequences. About 42% of the total number of disordered residues of the CASP7 targets are located within the 10 residues at the terminal regions of the target proteins and the results of Figure 4 unambiguously show that for all the methods assessed it is easier to detect terminal disordered regions than internal ones. This is most likely explained by the bias of the available training data, where terminal disordered regions are overrepresented compared to those in the middle of protein sequences.

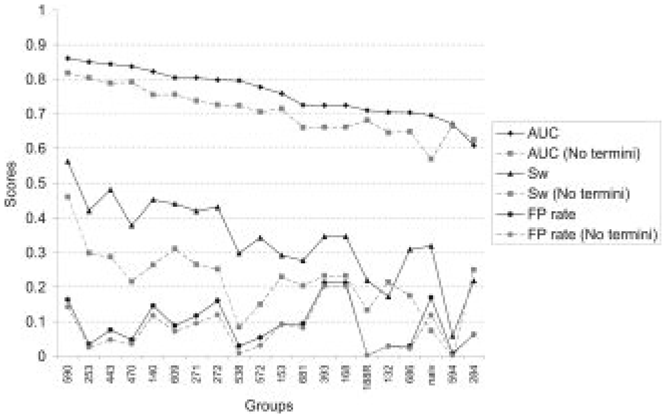


Figure 50 Internal versus terminal disordered regions. Disorder prediction was evaluated for full-length target sequences (black), and sequences with 10 residues removed at both the N- and C-termini (grey). ROC AUC, $S_w$scores, and FP rate indicate that disorder in terminal residues is easier to identify than in internal regions of the protein.

To investigate whether 3-dimensional information from homologous proteins would influence the accuracy of detecting disorder in protein, we separated the CASP7 targets into two sub-sets, “3D-homologous” and “3D-non-homologous”, depending on whether a homologous protein with known three-dimensional structure could be found by PSI-Blast.[32] Based on this definition the “3D-homologous” subset contained 59 targets and the “3D-non-homologous” 37 targets, respectively. The results for the “homologous” and “non-homologous” categories are summarized in Table IV and show that the influence of related proteins with known structure differs from group to group. Methods of e.g. groups 443 or 140 seem to take advantage (directly or indirectly) of three-dimensional information, more than other groups like 590, 253 or 470. In particular predictions of group 443 for the homologous set of proteins are more accurate (based on ROC curve analysis) then the rest of the groups ($p_{(443,140)}$ = 0.002 when comparing groups 443 and 140 on common targets, $p_{(443,590)} < 0.0001$, $p_{(443,470)} < 0.0001$ and $p_{(443,253)} < 0.0001$ respectively). There are some cases where the results of the "3D-non-homolgous" sub-set seem to outperform the prediction obtained in the "3D-homologous" sub-set. However, this is due to an overall rather poor performance of these methods for the "3D-homologous" subset of proteins.

To conclude, the results of the best 6 predictor groups are depicted in form of pie charts in Figure 5. It is evident that in general these methods are more accurate in correctly predicting ordered residues than in predicting disordered ones, due in part to the lower number of disordered residues available for the training of these type of predictors.

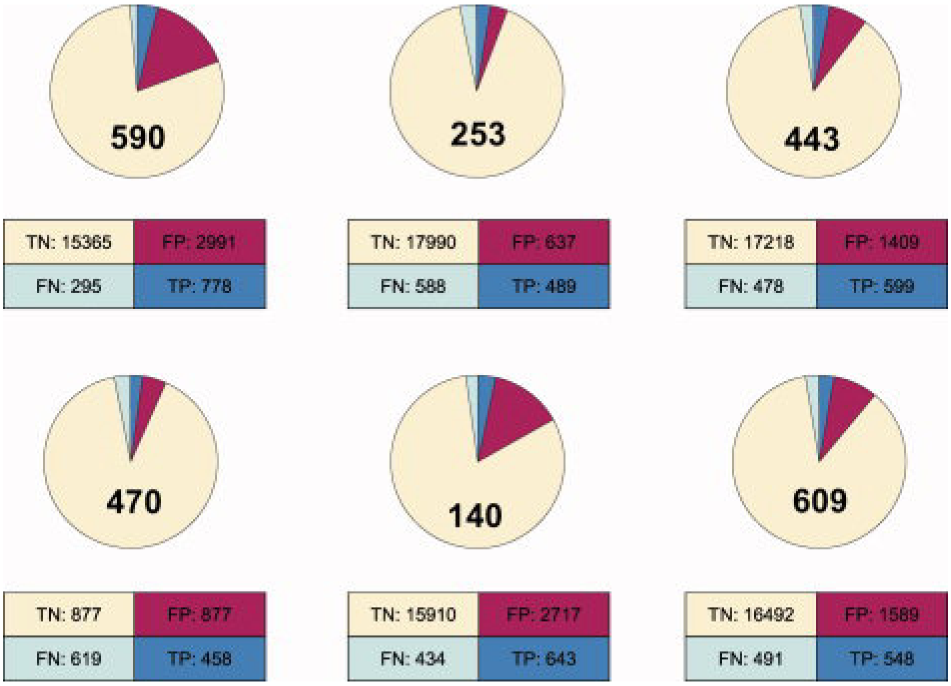


Figure 51 Confusion matrix. Confusion matrix with the TP (blue), FP (red), FN (green), and TN (yellow) predictions of groups 590, 253, 443, 470, 140, and 609 are shown as numerical values and graphical representation (pie chart).

While the assessment so far was based on a per residue basis, we also analyzed the results based on a per target basis. The ROC AUC and the Sw scores of individual targets averaged over the top 6 groups were analyzed. There is indeed a negative correlation between the ROC AUC and the Sw scores and the percent of disordered residues of the targets, indicating that the higher the proportion of disordered residues in a target is, the less accurate the predictions are. From the pie charts is also evident that the false positive rate is in general higher than the true positive rate, although groups like 253 and 470 have a more balanced error rate than others. There is in general a trade-off between sensitivity and specificity and more sensitive methods pay the price of having a 2- to 4- fold higher number of FP than TP predictions.

Comparison with CASP6

The prediction results for the present experiment were compared with the results of CASP6. A comparable number of groups participated in both CASP experiments and although this year 96 targets were available for prediction, compared to the 66 of CASP6, figure 1 shows that the distribution of the length of contiguous disordered regions is similar to the distribution of two years ago. We compared the predictions of CASP6 and CASP7 on same criteria (ROC curves, Sw and FP rate). The results indicate that on average the methods participating in CASP7 perform better than those in CASP6. Additionally we evaluated the results of a subset of groups which participated in both CASP6 and CASP7 (groups 590, 470, 443 and 140). None of the methods of CASP7 outperform the best methods of the past experiment. In conclusion, if we assume the comparison based on the same scores for different data sets to be valid, it appears that several groups have improved the accuracy of their individual methods in comparison to CASP6. However, no improvement over the best methods in CASP6 was observed in CASP7.

Discussion

In our assessment of disorder prediction submitted for the CASP7 target proteins, the accuracy of the predictions by four groups was higher than the remaining methods: group 253 (K. Shimizu, S. Hirose, N. Inoue, S. Kanai and T. Noguchi, National Institute of Advanced Industrial Science and Technology, Japan), group 443 (T. Ishida, and K. Kinoshita, Human Genome Center, University of Tokyo, Tokyo), group 470 (K. Bryson, D.T. Jones, University College London, UK) and 590 (K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker, Z. Obradovic, Temple University, Philadelphia, PA). However, individual methods are characterized by different trade-offs between sensitivity and specificity, either by correctly identifying more disordered residues at the cost of over prediction, or by having a more modest sensitivity but a more balanced error rate. Depending on the type of application, methods with either higher sensitivity or specificity may be more appropriate. While not correctly identifying disordered regions could cause experimental difficulties e.g. for protein expression and crystallization, over-prediction of disordered regions might also lead to misinterpretations with serious consequences, e.g. designing expression constructs for experimental structure solution by X-Ray or NMR which are too short may result in dysfunctional or instable proteins.

In comparison with the previous CASP experiment, it appears that the field has converged, however it seems that with the present type of algorithms it is difficult to improve on the plateau reached by the best groups. An obvious limitation is that the available dataset for training of the algorithms is still suboptimal since it is biased by structures solved by X-ray crystallography. Indeed, crystallographic data may be misleading as regions that are disordered in solution may adopt an ordered conformation upon crystallization. Conversely, the absence of well defined electron density may not necessarily prove the lack of structure of a protein region. Furthermore, X-ray datasets contain notably fewer and shorter disordered regions than would be expected for in vivo native protein disorder. Therefore, other experimental techniques such as NMR may be preferred for the detection of native disorder, and in the future provide a better dataset for both development and assessment of the accuracy of disorder prediction. It has been suggested that new impulse in the development of new algorithms may result from incorporation of other data source than sequences.[177] One attractive idea could be to make use of detectable homology to proteins with known structures. Although this might on average improve the accuracy of the prediction, this approach could introduce additional bias towards the dataset of the currently known domain structures within the PDB, and might not improve predictions for proteins where no structural homology can be detected, i.e. the proteins for which sequence-based prediction methods are most needed.

Acknowledgements

We would like to thank John Moult and Anna Tramontano for helpful discussion, and Andriy Kryshtafovych and Krzysytof Fidelis from the Predictioncenter @ UC Davies for their perfect support in providing the data for the analysis. We are grateful to the previous CASP assessors Roland Dunbrack and Yumi Jin for valuable advice on the assessment and for providing the CASP6 dataset, and Jürgen Kopp and James Battey for their constructive comments on the manuscript. We gratefully acknowledge support of our group by the Swiss Institute of Bioinformatics.

Group	Targets predicted	Residues predicted	P-values	D(cutoff)	O(cutoff)
132	96	19816	cont.	>=0.50	<0.50
140	96	19816	cont.	>=0.50	<0.50
153	95	19723	cont.	>=0.50	<0.50
168	96	19816	cont.	>=0.594843	<=0.553115
188	10	2221	cont.	<0.50	>=0.50
253	96	19816	cont.	>=0.50	<0.50
271	94	19368	cont.	>=0.50	<0.50
272	96	19816	cont.	>=0.50	<0.50
284	95	19494	0.0,1.0	1	0
393	96	19816	cont.	>=0.594843	<=0.553115
443	96	19816	cont.	>=0.50	<0.50
470	96	19816	cont.	>=0.50	<0.50
538	96	19816	cont.	>=0.50	<0.50
572	95	19816	cont.	>=0.50	<0.50
590	95	19541	cont.	>=0.50	<0.50
594	82	16687	cont.*	>=0.85	<=0.85
609	93	19231	cont.**	>0.5	<=0.50
681	75	15180	cont.	>=0.50	<=0.50
686	96	19816	cont.	>0.5	<=0.5

Table I: Overview of disorder prediction data assessed in CASP7. For each participating group, the number of predicted targets and number of predicted residues is shown. Groups registered as prediction servers are underlined. Most groups provided continuous P-values for disorder prediction, except groups 284, providing only binary predictions (0/1), group 594 (*) providing P-values in steps of 0.01, and group 609 (**) in steps of 0, 0.25, 0.33, 0.5, 0.66, 0.75, 1. Group 188 used “inverse” P values, in other words a lower (instead of a higher) P value is associated with greater probability of disorder.

Groups	**Nb targets**	**Nb residues**	**W_{ord}**	**W_{dis}**	**S_{sens}**	**S_{spec}**	**S_w**		**ACC**	**ROC (AUC)**		**FP rate**
590	95	19429	5.52	94.48	0.837	0.725	0.562	± 0.043	0.781	0.860	± 0.007	0.163
253	96	19704	5.47	94.53	0.966	0.454	0.420	± 0.052	0.710	0.850	± 0.007	0.034
443	96	19704	5.47	94.53	0.924	0.556	0.481	± 0.049	0.740	0.844	± 0.008	0.076
470	96	19704	5.47	94.53	0.953	0.425	0.378	± 0.043	0.689	0.837	± 0.008	0.047
140	96	19704	5.47	94.53	0.854	0.597	0.451	± 0.049	0.726	0.822	± 0.008	0.146
609	93	19120	5.43	94.57	0.912	0.527	0.440	± 0.047	0.720	0.804	± 0.008	0.088
271	94	19256	5.57	94.43	0.883	0.536	0.419	± 0.043	0.710	0.804	± 0.008	0.117
272	96	19704	5.47	94.53	0.839	0.591	0.430	± 0.044	0.715	0.798	± 0.008	0.161
538	96	19704	5.47	94.53	0.971	0.327	0.298	± 0.045	0.649	0.796	± 0.008	0.029
572	96	19704	5.47	94.53	0.947	0.396	0.343	± 0.035	0.672	0.777	± 0.008	0.053
153	95	19613	5.49	94.51	0.908	0.383	0.291	± 0.042	0.646	0.758	± 0.009	0.092
681	74	15098	6.25	93.75	0.906	0.371	0.277	± 0.058	0.639	0.726	± 0.010	0.094
393	96	19704	5.47	94.53	0.788	0.558	0.346	± 0.040	0.673	0.724	± 0.009	0.212
168	96	19704	5.47	94.53	0.788	0.558	0.346	± 0.040	0.673	0.724	± 0.009	0.212
132	96	19704	5.47	94.53	0.971	0.201	0.172	± 0.055	0.586	0.704	± 0.009	0.029
686	96	19704	5.47	94.53	0.971	0.338	0.309	± 0.038	0.655	0.704	± 0.010	0.029
594	82	16586	5.87	94.13	0.993	0.066	0.058	± 0.013	0.529	0.671	± 0.100	0.007
naiv10	96	19704	5.47	94.53	0.926	0.367	0.293	± 0.043	0.646	0.646	± 0.009	0.074
naiv9	96	19704	5.47	94.53	0.934	0.341	0.275	± 0.040	0.638	0.638	± 0.009	0.066
naiv8	96	19704	5.47	94.53	0.943	0.313	0.256	± 0.037	0.628	0.628	± 0.009	0.057
naiv7	96	19704	5.47	94.53	0.951	0.285	0.237	± 0.034	0.618	0.618	± 0.009	0.049
284	95	19382	5.53	94.47	0.937	0.280	0.218	± 0.053	0.609	0.609	± 0.009	0.063
naiv6	96	19704	5.47	94.53	0.960	0.254	0.214	± 0.031	0.607	0.607	± 0.009	0.040
naiv5	96	19704	5.47	94.53	0.968	0.217	0.185	± 0.026	0.592	0.592	± 0.009	0.032
naiv4	96	19704	5.47	94.53	0.975	0.178	0.154	± 0.022	0.577	0.577	± 0.009	0.025
naiv3	96	19704	5.47	94.53	0.983	0.134	0.117	± 0.016	0.558	0.558	± 0.009	0.017
naiv2	96	19704	5.47	94.53	0.990	0.089	0.079	± 0.011	0.539	0.539	± 0.009	0.010
naiv1	96	19704	5.47	94.53	0.995	0.045	0.040	± 0.005	0.520	0.520	± 0.009	0.005
188(*)	10	2204	4.08	95.92	0.997	0.222	0.219	n.a.	0.610	0.290	± 0.024	0.003

* Group 188 submitted predictions for only 10 targets and used "inverse" P values, in other words a lower (instead of a higher) P value is associated with greater probability of disorder. Calculating the area under the curve (AUC) for inverted P-values would result in AUC = 0.710.

Table II: Assessment results of disorder prediction. See text for details

	590	253	443	470	140	609	271	272	538	572	153	681	393	168	132	686	594
590	-------	0.072	**0.002**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
253	0.170	-------	0.328	**0.010**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
443	**0.013**	0.382	-------	0.153	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
470	**<0.001**	**0.044**	0.266	-------	**0.004**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
140	**<0.001**	**<0.001**	**<0.001**	**0.020**	-------	**0.027**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
609	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.068	------	0.957	0.613	0.352	**0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
271	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	0.964	------	0.177	0.066	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
272	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.674	0.233	------	0.335	**0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
538	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.445	0.147	0.442	------	**0.002**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
572	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.004**	**<0.001**	**0.007**	**0.021**	-------	0.052	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
153	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.074	------	0.198	**0.001**	**0.001**	**<0.001**	**<0.001**	**<0.001**
681	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.207	------	0.688	0.688	**0.002**	**0.001**	**<0.001**
393	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	0.693	-----	1.000	0.057	**0.048**	**<0.001**
168	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.002**	0.693	1.000	-----	0.057	**0.048**	**<0.001**
132	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.005	0.084	0.084	-----	0.982	0.128
686	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.068	0.068	0.984	------	**0.003**
594	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.146	**0.004**	-----
284	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.049**

Table III. Statistical comparison between the 19 groups based on the ROC area under the curve (AUC). The top right of the table shows the comparison calculated by applying the non parametric test by De Long et al. and in the lower left by the non parametric test by Hanley and McNeil on a common set of targets. Significant differences between the groups (*p*-values of less than 0.05) are highlighted in grey.

	3D-Homologous	**3D Non-Homologous**	
Groups	ROC (AUC)	ROC (AUC)	*p*-value
590	**0.860**	0.838	0.159
253	**0.838**	0.820	0.249
470	**0.840**	0.815	0.110
609	0.796	**0.808**	0.481
443	**0.885**	0.806	< 0.001
140	**0.862**	0.788	< 0.001
572	0.765	**0.781**	0.369
681	0.560	**0.776**	< 0.001
272	**0.796**	0.762	0.046
271	**0.816**	0.759	0.001
538	**0.812**	0.757	0.001
132	0.663	**0.725**	0.001
153	**0.840**	0.704	< 0.001
393	**0.757**	0.704	0.004
168	**0.757**	0.704	0.004
naiv10	**0.649**	0.641	0.677
686	**0.808**	0.627	< 0.001
594	**0.775**	0.619	< 0.001
284	**0.648**	0.580	0.004

Table IV: Comparing 3D-homologous versus 3D-non-homologous target subsets. The predictions of 19 groups for the two subsets of targets are compared using the ROC - AUC. The statistical significance between the ROC curves is assessed by the non parametric test by Hanley and McNeil on non correlated samples. Significant differences between the two subsets (*p*-values of less than 0.05) are highlighted in grey.

5 Conclusion

The number of available protein sequences greatly exceeds the number of available protein structures. Reliable and automated modeling procedures are therefore required to close the gap between the number of experimentally determined structures and of known protein sequences. An essential step in modeling is the identification and alignment of suitable template structures. In order to improve the sensitivity of identifying template structures which are evolutionary distant we used a method which was among the top predictors in the double blind experiment CASP. A hierarchical template selection was developed in order to favor fast and accurate sequence-sequence alignment tools over the computationally demanding HMM-HMM alignment procedures when evolutionary distance between the target and template sequence is small. Additionally, an approach that guarantees an optimal balance between model accuracy and target coverage was developed. The template search and selection routine was integrated into the homology modeling pipeline of SWISS-MODEL (as automated mode). The SWISS-MODEL homology modeling pipeline was then benchmarked against two other widely used modeling servers in a blind fashion. The benchmarking has shown comparable results in terms of model accuracy and response time. When considering the root mean square deviation between the model and target structures, SWISS-MODEL was ranked first amongst the three servers.

The homology modeling pipeline was then applied to a large number of protein sequences deposited in the UniProt-knowledge database, in order to generate structural information. A regular updating procedure was set in place for a selected set of proteomes, which are of interest for the scientific community, in order to improve the quality of the models and the structural coverage. In order to reduce computational time, an incremental update procedure was developed. The database of models can be queried online using common database accession code or the sequence itself.

To evaluate the current status of comparative quaternary structure modeling, we assessed those prediction methods which submitted oligomeric models within the CASP9 installment. The rigorous assessment of oligomeric prediction methods was performed for the first time. Because a systematic evaluation was not carried out in the past, we developed a novel set of scores which reflects the accuracy of oligomeric models. Two conclusions can be drawn. Firstly, only a minority of the predictors submitted oligomeric protein models. Secondly, the accuracy is not as high as it could be; two naïve predictors which rely on standard techniques were able to outperform all participating methods, except one. The results have shown that additional efforts are required in order to push quaternary structure modeling towards higher accuracy.

To develop an adequate modeling method for oligomeric protein structures, we identified in this study central aspects which are essential for the success of oligomeric modeling:

To distinguish between similar and dissimilar quaternary structure, a metric which could be a used to compare the model and the target structure and reflect the difference in terms of number of subunits (i.e. oligomeric state) and accuracy of the interface modeling was determined to be essential. We established *QscoreOligomer* which can be seen as weighted mean of differences in the absolute number of contacts. *QscoreOligomer* weights down the influence of long range contacts and thus is robust against small positional changes of interfaces residues in one structure compared to the other. *QscoreOligomer* can be used for the assessment within further rounds of CASP, other systematic evaluation procedures, or by research groups which wants to improve their methods and thus need a robust score to benchmark their methods.

To carry out homology modeling successfully, the availability of a template library that is updated often and promptly includes newly-released structures is crucial. This is even more important for oligomeric template structures, because the assignment of the biological relevant quaternary structure is not always unambiguous and often contains errors. We benchmarked the accuracy of different methods and approaches that were used to assign a biologically relevant quaternary structure to the majority of structures deposited in the PDB.

We have shown that author assignment is the most valuable approach, at least for homo-oligomers. Not all structures deposited in the PDB were annotated by the authors, nevertheless the ones that have been annotated can be used very successfully as template structures for homology modeling. The annotation given by PISA can be used as a second-line solution. Based on these criteria, we developed a template library is both complete and accurate and thus represents a valuable basis for the correct modeling of quaternary structure.

One important, if not the most important, aspect when modeling the quaternary protein by comparative techniques, is the question whether a particular template structure has the same quaternary structure than the target. For that reason we investigated to what extend and under which conditions the oligomeric state of a protein being modeled can be deduced from template structures.

The analysis of similarity between the quaternary structure of the template and the one of the target has confirmed trends uncovered by previous studies, with the number of pairs with the same quaternary structure increasing linearly for sequence identity higher than 40% and with a sharp drop in structure conservation for lower sequence identity. However, it has been shown that even for closely related template structures it cannot be guaranteed that the quaternary structure is similar to the target. Furthermore, it was found that the evolutionary distance range at which at least one template with similar quaternary structure can be identified varies considerably among the target proteins. For the majority of targets suitable template structures can be found even if the evolutionary distance falls below 30%. Based on these observations, the conclusion can be drawn that the concept of comparative modeling can be applied to the prediction of homo-oligomeric protein structures, but for the selection of suitable templates, using exclusively evolutionary distance is a weak choice.

In order to determine which strategies could be followed to accurately identify templates with the correct quaternary structure amongst many candidates for a given target, we applied clustering techniques and analyzed the characteristics of the resulting clusters. We also took into account manually annotated information about the quaternary structure of the template, evolutionary conservation in the interface and information about the protein complex being modeled. We found out that size and width of template clusters can be used successfully to select relevant templates. In general, a good performance is obtained by combining all cluster attributes using a random forest algorithm. A classifier based on this features was able to reach a high accuracy in identifying templates with correct quaternary structures compared to a classifier with relied on sequence features solely. To our knowledge this was the first attempt to classify template structures according to their quaternary structure similarity to the target. Thus, the discussed approach has the ability to enhance the accuracy of modeling routines considerably.

Proteins which lack a well-defined three-dimensional structure are common among eukaryotic species and involved into many important functions. The prediction accuracy of sequence based methods which predict the occurrence of intrinsically disordered segments was evaluated in a double blind experiment. Four methods have been identified with a higher accuracy than the competitors. Highlighting the strengths and weaknesses of individual methods can help scientists choose a method which fits best their needs.

A possible future enhancement of this work is a more in-depth analysis of the template cluster properties, which will lead to improved accuracy in template selection. This will require the introduction of model quality estimates which are exclusively trained on oligomeric interfaces. Moreover, a more precise model-building procedure will be required in order to generate more realistic models. A first step in the future development of this project, however, will be the release of the developed software to the public, to allow other scientists to understand, validate and expand our analysis.

6 Acknowledgement

Firstly, I would like to thank Prof. Torsten Schwede for giving me the opportunity to perform my PhD studies in his group at the Biozentrum.

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

The Swiss Institute of Bioninformatics is acknowledged for the financial support.

I am indebted to Lorenza Bordoli, Jürgen Haas and Valerio Mariani for critically reading and correcting my thesis and providing constructive comments.

I would like to thank my colleagues in the structural bioinformatics group and the working friendly environment.

And last but certainly not least I would like to thank my family. Saskia for their tireless support and the acceptance of many limitations in the past; Mila for bringing always a smile in my face and my sister and my parents for their help and support.

7 Supplementary Material

Table SI: Oligomeric states of CASP9 targets. The definition of the oligomeric state for the assessment is based primarily on the assignment by the depositor (“REMARK 350”). The data set includes monomers, homo dimers, trimers and tetramers (column “state”). Targets without or with ambiguous assignments by authors were inspected manually taking into account PISA annotation and the “REMARK 300” section (See “Comments” column). Coordinate sets representing the biological units (column “PDB unit”) were downloaded from the PDB protein database based using the PDB code for the targets reported on the CASP9 target website.

Target	PDB-ID	Category	State	PDB unit	Comments
T0515	3mt1	HS	2	1	
T0516	3no6	S	4	1	
T0517	3pnx	HS	3	1	
T0518	3nmb	S	1	1	
T0519	-	-	-	-	Cancelled
T0520	3mr7	HS	2	1	
T0521	3mse	S	2	1	
T0522	3nrd	S	2	1	
T0523	3mqo	HS	2	1	
T0524	3mwx	S	1	1	
T0525	3mqz	S	1	1	
T0526	3nre	HS	1	1	
T0527	3mr0	S	1	1	
T0528	3n0x	S	1	1	
T0529	3mwt	HS	-	-	Excluded due to ambiguous assignment .
T0530	3npp	S	2	1	
T0531	-	-	-	-	FM Target
T0532	3mx3	S	1	1	
T0533	3mwb	S	2	2	
T0534	-	-	-	-	FM Target
T0535	-	-	-	-	Cancelled
T0536	3mxq	S	4	1	
T0537	-	-	-	-	FM Target
T0538	2l09	S	1	1	
T0539	2l0b	S	1	1	
T0540	-	HS	-	-	No PDB entry.
T0541	2l0d	S	1	1	

T0542	3n05	S	2	1	
T0543	2xrg	HS	1	1	
T0544	-	-	-	-	FM Target
T0545	2l3f	S	1	1	
T0546	-	-	-	-	Cancelled
T0547	3nzp	HS	-	-	REMARK 300: DIMER IN SOLUTION AND CRYSTAL, HOWEVER, THE BIOLOGICAL UNIT IS TETRAMER.
T0548	3nng	S	-	-	REMARK 300: AUTHORS STATE THAT THE BIOLOGICAL UNIT IS A DIMER, NOT A TETRAMER. THE DIMER IN THE ASYMMETRIC UNIT MAY NOT BE THE REAL DIMER IN SOLUTION, HOWEVER.
T0549	-	-	-	-	Cancelled
T0550	3ngk	HS	-	-	REMARK 300: ANALYTICAL SIZE EXCLUSION CHROMATOGRAPHY WITH STATIC LIGHT SCATTERING SUPPORTS THE ASSIGNMENT OF A TRIMER AS A SIGNIFICANT OLIGOMERIZATION STATE IN SOLUTION.
T0551	3obh	S	2	1	
T0552	2l3b	S	1	1	
T0553	-	-	-	-	FM Target
T0554	-	-	-	-	Cancelled
T0555	-	-	-	-	FM Target
T0556	-	-	-	-	Cancelled
T0557	2kyy	S	1	1	
T0558	3no2	HS	1	1	
T0559	2l01	S	2	1	
T0560	2l02	S	2	1	
T0561	-	-	-	-	FM Target
T0562	2kzx	HS	1	1	
T0563	3on7	S	4	1	
T0564	2l0c	HS	1	1	
T0565	3npf	S	2	1	Complex assigned by PISA was used.
T0566	3n72	HS	1	3	Hypothetic dimer interface is classified as only rarely stable. The largest monomeric chain was used.
T0567	3n70	S	1	1	Dimer interfaces are mainly stabilized by SO4 (buffer). The largest monomeric chain was used.
T0568	3n6y	HS	1	1	
T0569	2kyw	HS	1	-	NMR
T0570	3no3	S	1	1	
T0571	3n91	HS	1	1	
T0572	2kxy	S	1	-	NMR

T0573	3oox	S	1	1	
T0574	3nrf	HS	4	1	
T0575	3nrg	S	1	1	
T0576	3na2	HS	2	1	Author suggests two conformations. The dimer consistent with PISA assignment was used.
T0577	-	-	-	-	Cancelled
T0578	-	-	-	-	FM Target
T0579	2ky9	HS	1	-	NMR
T0580	3nbm	HS	1	1	Monomeric
T0581	-	-	-	-	FM Target
T0582	3o14	HS	1	1	
T0583	-	-	-	-	Cancelled
T0584	3nf2	HS	2	1	
T0585	3ne8	S	2	1	
T0586	3neu	HS	2	1	
T0587	-	-	-	-	Cancelled
T0588	3nfv	HS	1	1	
T0589	3net	S	2	1	
T0590	2kzw	HS	1	1	
T0591	3nra	S	2	1	
T0592	3nhv	HS	3	1	
T0593	3ngw	S	1	1	
T0594	3ni8	HS	1	1	
T0595	-	-	-	-	Cancelled
T0596	3ni7	HS	2	1	
T0597	3nie	S	1	1	
T0598	3njc	HS	2	1	
T0599	3os6	S	2	2	
T0600	3nja	S	2	2	REMARK 300: EXPERIMENTALLY UNKNOWN. THE CHAINS A AND B, C AND D MAY FORM DIMERS RESPECTIVELY.
T0601	3qtd	S	2	1	
T0602	3nkz	HS	2	2	REMARK 300: EXPERIMENTALLY UNKNOWN. THE CHAINS A AND B, C AND D MAY FORM DIMERS RESPECTIVELY.
T0603	3nkd	S	2	1	
T0604	3nlc	HS	1	1	
T0605	3nmd	HS	2	2	
T0606	3noh	HS	1	1	
T0607	3pfe	S	2	1	
T0608	3nyy	HS	2	1	
T0609	3os7	S	1	1	
T0610	3ot2	HS	2	1	

T0611	3nnr	S	2	1	
T0612	3o0l	S	2	1	
T0613	3obi	S	4	1	
T0614	-	-	-	-	No PDB file
T0615	3nqw	S	2	1	
T0616	3nrt	HS	2	2	
T0617	3nrv	S	2	1	
T0618	-	-	-	-	FM Target
T0619	3nrw	HS	1	1	
T0620	3nr8	S	1	1	
T0621	-	-	-	-	FM Target
T0622	3nkl	HS	2	1	
T0623	3nkh	S	2	1	
T0624	-	-	-	-	FM Target
T0625	3oru	HS	2	1	
T0626	3o1l	S	2	1	
T0627	3oql	HS	4	1	
T0628	3nuw	HS	2	5	PISA suggested Dimer with 11 H-bonds and 7 salt bridges in the interface; classified as stable
T0629	2xgf	HS	3	1	
T0630	2kyt	HS	1	1	
T0631	-	-	-	-	Cancelled
T0632	3nwz	S	4	1	Authors assigned different states, but the tetramer is confirmed by PISA as most stable complex. The observed coenzyme A is interacts with both sides of the tetrameric interface.
T0633	-	-	-	-	Cancelled
T0634	3n53	S	2	3	
T0635	3n1u	S	4	1	
T0636	3p1t	S	2	1	
T0637	-	-	-	-	FM Target
T0638	3nxh	S	1	1	
T0639	-	-	-	-	FM Target
T0640	3nyw	S	4	1	
T0641	3nyi	S	1	2	REMARK 300: EXPERIMENTALLY UNKNOWN. IT IS LIKELY MONOMERIC.
T0642	-	-	-	-	Cancelled
T0643	3nzl	HS	1	1	

8 References

1. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. Proteins 2007;69 Suppl 8:129-136.
2. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 Suppl 8:38-56.
3. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 2009;4(1):1-13.
4. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 2009;37(Database issue):D387-392.
5. Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T. The Protein Model Portal. J Struct Funct Genomics 2009;10(1):1-8.
6. Berman HM, Westbrook JD, Gabanyi MJ, Tao W, Shah R, Kouranov A, Schwede T, Arnold K, Kiefer F, Bordoli L, Kopp J, Podvinec M, Adams PD, Carter LG, Minor W, Nair R, La Baer J. The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res 2009;37(Database issue):D365-368.
7. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins: Structure, Function, and Bioinformatics 2011;79(S10):37-58.
8. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181(96):223-230.
9. Levinthal C. Are there pathways for protein folding? Journal of Medical Physics 1968;65(1):44-45.
10. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proceedings of the National Academy of Sciences 1992;89(1):20.
11. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nature structural biology 1997;4(1):10-19.
12. Chothia C. Proteins. One thousand families for the molecular biologist. Nature 1992;357(6379):543-544.
13. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5(4):823-826.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
15. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.
16. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536-540.
17. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods Mol Biol 2007;406:89-112.
18. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28(1):45-48.
19. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33(Database issue):D154-159.
20. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2007;35(Database issue):D301-303.

21. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.

22. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61 Suppl 7:91-98.

23. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69 Suppl 8:108-117.

24. Ponder JW, Richards FM. Internal packing and protein structural classes. Cold Spring Harb Symp Quant Biol 1987;52:421-428.

25. Bowie JU, Clarke ND, Pabo CO, Sauer RT. Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 1990;7(3):257-264.

26. Sanchez R, Sali A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A 1998;95(23):13597-13602.

27. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Rout MP, Sali A. Determining the architectures of macromolecular assemblies. Nature 2007;450(7170):683-694.

28. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294(5540):93-96.

29. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370-3374.

30. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15(3):285-289.

31. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227(4693):1435-1441.

32. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-3402.

33. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9(2):232-241.

34. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14(10):846-856.

35. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-763.

36. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3(3):265-274.

37. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299(2):499-520.

38. Hargbo J, Elofsson A. Hidden Markov models that use predicted secondary structures for fold recognition. Proteins 1999;36(1):68-76.

39. Kawabata T, Nishikawa K. Protein structure comparison using the markov transition model of evolution. Proteins 2000;41(1):108-122.

40. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.

41. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326(6111):347-352.

42. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234(3):779-815.

43. Hestenes M, Stiefel E. Methods of Conjugate Gradients for Solving Linear Systems. Journal of Research of the National Bureau of Standards 1952;49(6):409-436.

44. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9(9):1753-1773.

45. Dunbrack RL, Jr. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12(4):431-440.

46. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: A program to check the stereochemical quality of protein structures J Appl Cryst 1993;26:283-291.

47. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381(6580):272.

48. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356(6364):83-85.

49. van Gunsteren WF, Billeter SR, Eising A, Hünenberger PH, Krüger P, Mark AE, Scott WRP, Tironi IG. Biomolecular Simulations: The GROMOS96 Manual and User Guide. Zürich: VdF Hochschulverlag ETHZ; 1996.

50. Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 1998;277(5):1141-1152.

51. Benkert P, Tosatto SC, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008;71(1):261-277.

52. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins 2007;69 Suppl 8:68-82.

53. Guex N, Diemand A, Peitsch MC. Protein modelling for all. Trends Biochem Sci 1999;24(9):364-367.

54. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18(15):2714-2723.

55. Schwede T, Diemand A, Guex N, Peitsch MC. Protein structure computing in the genomic era. Res Microbiol 2000;151(2):107-112.

56. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 2006;22(2):195-201.

57. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2001;Suppl 5:39-46.

58. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12(2):85-94.

59. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33(Web Server issue):W244-248.

60. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci 2007;Chapter 2:Unit 2 9.

61. Sadowski MI, Jones DT. Benchmarking template selection and model quality assessment for high-resolution comparative modeling. Proteins 2007;69(3):476-485.

62. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem 1995;64:287-314.

63. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001;310(2):311-325.

64. Koh IY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B. EVA: Evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31(13):3311-3315.

65. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A. Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 2003;31(13):3375-3380.

66. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins 2007;69 Suppl 8:175-183.

67. Levitt M. Growth of novel protein structural data. Proc Natl Acad Sci U S A 2007;104(9):3183-3188.

68. Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, Godzik A. The challenge of protein structure determination--lessons from structural genomics. Protein Sci 2007;16(11):2472-2482.

69. Manjasetty BA, Turnbull AP, Panjikar S, Bussow K, Chance MR. Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. Proteomics 2008;8(4):612-625.

70. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 2007;5(3):e16.

71. McCleverty CJ, Columbus L, Kreusch A, Lesley SA. Structure and ligand binding of the soluble domain of a Thermotoga maritima membrane protein of unknown function TM1634. Protein Sci 2008;17(5):869-877.

72. Xu Q, Kozbial P, McMullan D, Krishna SS, Brittain SM, Ficarro SB, DiDonato M, Miller MD, Abdubek P, Axelrod HL, Chiu HJ, Clayton T, Duan L, Elsliger MA, Feuerhelm J, Grzechnik SK, Hale J, Han GW, Jaroszewski L, Klock HE, Morse AT, Nigoghossian E, Paulsen J, Reyes R, Rife CL, van den Bedem H, White A, Hodgson KO, Wooley J, Deacon AM, Godzik A, Lesley SA, Wilson IA. Crystal structure of an ADP-ribosylated protein with a cytidine deaminase-like fold, but unknown function (TM1506), from Thermotoga maritima at 2.70 A resolution. Proteins 2008;71(3):1546-1552.

73. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. Proteins 2007;69 Suppl 8:194-207.

74. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. Drug Discov Today 2004;9(15):659-669.

75. Tan ES, Groban ES, Jacobson MP, Scanlan TS. Toward deciphering the code to aminergic G protein-coupled receptor drug design. Chem Biol 2008;15(4):343-353.

76. Thorsteinsdottir HB, Schwede T, Zoete V, Meuwly M. How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-I protease inhibitor binding. Proteins 2006;65(2):407-423.

77. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J Med Chem 2003;46(13):2656-2662.

78. Oshiro C, Bradley EK, Eksterowicz J, Evensen E, Lamb ML, Lanctot JK, Putta S, Stanton R, Grootenhuis PD. Performance of 3D-database molecular docking studies into homology models. J Med Chem 2004;47(3):764-767.

79. Murray PS, Li Z, Wang J, Tang CL, Honig B, Murray D. Retroviral matrix domains share electrostatic homology: models for membrane binding function throughout the viral life cycle. Structure 2005;13(10):1521-1531.

80. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. Nat Biotechnol 2007;25(10):1171-1176.

81. Junne T, Schwede T, Goder V, Spiess M. The plug domain of yeast Sec61p is important for efficient protein translocation, but is not essential for cell viability. Mol Biol Cell 2006;17(9):4063-4068.

82. Peitsch MC. About the use of protein models. Bioinformatics 2002;18(7):934-938.

83. Tramontano A. The Biological Applications of Protein Models. In: Schwede T, Peitsch MC, editors. Computational Structural Biology: World Scientific Publishing; 2008.

84. Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol 2007;25(9):1051-1056.

85. Kopp J, Schwede T. The SWISS-MODEL Repository: new features and functionalities. Nucleic Acids Res 2006;34(Database issue):D315-318.

86. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P, Kahari A, Kulesha E, Macias JR, Reeves GA, Prlic A. Integrating biological data - the Distributed Annotation System. BMC Bioinformatics 2008;9 Suppl 8:S3.

87. Berman H, al. e. PSI Structural Genomics Knowledge Base. NAR 2008;This issue.

88. Berman HM. Harnessing knowledge from structural genomics. Structure 2008;16(1):16-18.

89. Kopp J, Schwede T. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res 2004;32(Database issue):D230-234.

90. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34(Database issue):D291-295.

91. Huang H, Hu ZZ, Arighi CN, Wu CH. Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. Front Biosci 2007;12:5071-5088.

92. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res 2008;36(Database issue):D281-288.

93. Mulder NJ, Apweiler R. The InterPro database and tools for protein domain analysis. Curr Protoc Bioinformatics 2008;Chapter 2:Unit 2 7.

94. Hartshorn MJ. AstexViewer: a visualisation aid for structure-based drug design. J Comput Aided Mol Des 2002;16(12):871-881.

95. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003;31(13):3381-3385.

96. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11(11):2714-2726.

97. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2006;15(4):900-913.

98. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 2000;28(18):3442-3444.

99. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337(3):635-645.

100. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 2002;30(1):47-49.

101. Goodsell DS, Olson AJ. Structural symmetry and protein function. Annual review of biophysics and biomolecular structure 2000;29:105-153.

102. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 3D complex: a structural classification of protein complexes. PLoS computational biology 2006;2(11):e155-e155.

103. Jones S, Thornton JM. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America 1996;93(1):13-20.

104. Wolynes PG. Symmetry and the energy landscapes of biomolecules. Proc Natl Acad Sci U S A 1996;93(25):14249-14255.

105. Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. Trends Biochem Sci 2004;29(11):618-625.
106. Crick FH, Watson JD. Structure of small viruses. Nature 1956;177(4506):473-475.
107. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. Journal of molecular biology 2007;372(3):774-797.
108. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256(5520):705-708.
109. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng 1997;10(9):999-1012.
110. Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Quarterly reviews of biophysics 2008;41(2):133-180.
111. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 1995;63(1):31-65.
112. Ofran Y, Rost B. Analysing Six Types of Protein–Protein Interfaces. Journal of Molecular Biology 2003;325(2):377-387.
113. Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003;53(3):708-719.
114. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. Journal of molecular biology 2004;336(4):943-955.
115. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332(5):989-998.
116. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. Nature 2008;453(7199):1262-1265.
117. Levy ED. PiQSi: protein quaternary structure investigation. Structure (London, England : 1993) 2007;15(11):1364-1367.
118. Dayhoff JE, Shoemaker BA, Bryant SH, Panchenko AR. Evolution of protein binding modes in homooligomers. Journal of molecular biology 2010;395(4):860-870.
119. Monod J, Wyman J, Changeux JP. On the Nature of Allosteric Transitions: A Plausible Model. J Mol Biol 1965;12:88-118.
120. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI. Structural similarity enhances interaction propensity of proteins. J Mol Biol 2007;365(5):1596-1606.
121. Villar G, Wilber AW, Williamson AJ, Thiara P, Doye JP, Louis AA, Jochum MN, Lewis AC, Levy ED. Self-assembly and evolution of homomeric protein complexes. Phys Rev Lett 2009;102(11):118106.
122. Nishi H, Ota M. Amino acid substitutions at protein-protein interfaces that modulate the oligomeric state. Proteins 2010;78(6):1563-1574.
123. Grueninger D, Treiber N, Ziegler MO, Koetter JW, Schulze MS, Schulz GE. Designed protein-protein association. Science 2008;319(5860):206-209.
124. Malay AD, Allen KN, Tolan DR. Structure of the thermolabile mutant aldolase B, A149P: molecular basis of hereditary fructose intolerance. J Mol Biol 2005;347(1):135-144.
125. Hashimoto K, Panchenko AR. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. Proc Natl Acad Sci U S A 2010;107(47):20352-20357.
126. Elcock AH, McCammon JA. Identification of protein oligomerization states by analysis of interface conservation. Proceedings of the National Academy of Sciences of the United States of America 2001;98(6):2990-2994.
127. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proceedings of the National Academy of Sciences of the United States of America 2005;102(43):15447-15452.
128. Guharoy M, Chakrabarti P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics 2010;11(1):286-286.

129. Hart PE, Nilsson NJ, Raphael B. A formal basis for the heuristic determination of minimum cost paths. Systems Science and Cybernetics, IEEE Transactions on 1968;4(2):100-107.

130. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 2003;52(1):51-67.

131. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 2003;52(1):2-9.

132. Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL. Statistical analysis of interface similarity in crystals of homologous proteins. Journal of molecular biology 2008;381(2):487-507.

133. Chen H, Skolnick J. M-TASSER: an algorithm for protein quaternary structure prediction. Biophysical journal 2008;94(3):918-928.

134. Kittichotirat W, Guerquin M, Bumgarner RE, Samudrala R. Protinfo PPC: a web server for atomic level prediction of protein complexes. Nucleic acids research 2009;37(Web Server issue):W519-525.

135. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8):2444-2448.

136. Juettemann T, Gerloff DL. BISC: binary subcomplexes in proteins database. Nucleic Acids Res 2011;39(Database issue):D705-711.

137. Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics 2007;23(2):e13-16.

138. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125(7):1731-1737.

139. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331(1):281-299.

140. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61 Suppl 7:225-236.

141. Henrick K. PQS: a protein quaternary structure file server. Trends in Biochemical Sciences 1998;23(9):358-361.

142. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. Journal of molecular biology 2001;313(2):399-416.

143. Bordner AJ, Gorin AA. Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces. BMC bioinformatics 2008;9:234-234.

144. Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 2006;7:27.

145. Mitra P, Pal D. Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. Structure 2011;19(3):304-312.

146. Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI. The role of hydrophobic interactions in positioning of peripheral proteins in membranes. BMC Struct Biol 2007;7:44.

147. Xu Q, Dunbrack RL, Jr. The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res 2011;39(Database issue):D761-770.

148. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372(3):774-797.

149. Hubbard SJT, J.M. NACCESS. Computer Program, Department of Biochemistry and Molecular Biology, University College London 1993.

150. Breiman L. Random forests. Machine learning 2001;45(1):5-32.

151. Liaw A, Wiener M. Classification and Regression by randomForest. R news 2002;2(3):18-22.
152. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21(20):3940-3941.
153. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589-1591.
154. Krivov GG, Shapovalov MV, Dunbrack RL, Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77(4):778-795.
155. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658-1659.
156. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23(10):1282-1288.
157. Shannon CE. A mathematical theory of communication. SIGMOBILE Mob Comput Commun Rev 2001;5(1):3-55.
158. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57(4):702-710.
159. Poupon A, Janin J. Analysis and prediction of protein quaternary structure. Methods Mol Biol 2010;609:349-364.
160. Bordner AJ, Gorin AA. Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces. BMC Bioinformatics 2008;9:234.
161. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41(1):47-57.
162. Shoemaker BA, Panchenko AR, Bryant SH. Finding biologically relevant protein domain interactions: conserved binding mode analysis. Protein science : a publication of the Protein Society 2006;15(2):352-361.
163. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39(4):331-342.
164. Guharoy M, Pal A, Dasgupta M, Chakrabarti P. PRICE (PRotein Interface Conservation and Energetics): a server for the analysis of protein-protein interfaces. J Struct Funct Genomics 2011;12(1):33-41.
165. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280(1):1-9.
166. Fridovich I. Superoxide radical and superoxide dismutases. Annu Rev Biochem 1995;64:97-112.
167. Wintjens R, Noel C, May AC, Gerbod D, Dufernez F, Capron M, Viscogliosi E, Rooman M. Specificity and phenetic relationships of iron- and manganese-containing superoxide dismutases on the basis of structure and sequence comparisons. J Biol Chem 2004;279(10):9248-9254.
168. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 2011;27(3):343-350.
169. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. Stat Med 1989;8(10):1277-1290.
170. Daily MD, Masica D, Sivasubramanian A, Somarouthu S, Gray JJ. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. Proteins 2005;60(2):181-186.
171. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. Protein J 2008;27(1):59-70.
172. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285(5):2177-2198.
173. Burger L, van Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 2008;4:165.

174. Petsko GA. Dog eat dogma. Genome Biol 2000;1(2):comment1002 1001-1002 1002.

175. Uversky VN, Dunker AK. Understanding protein non-folding. Biochim Biophys Acta 2010;1804(6):1231-1264.

176. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 2008;18(6):756-764.

177. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J 2007;92(5):1439-1456.

178. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 2000;11:161-171.

179. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005;44(6):1989-2000.

180. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323(3):573-584.

181. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19(1):26-59.

182. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins 2003;52(4):573-584.

183. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6(3):197-208.

184. Fink AL. Natively unfolded proteins. Curr Opin Struct Biol 2005;15(1):35-41.

185. Dyson HJ, Wright PE. Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. Methods Enzymol 2001;339:258-270.

186. Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S. Assessing protein disorder and induced folding. Proteins 2006;62(1):24-45.

187. Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins 2006;65(1):1-14.

188. Namba K. Roles of partly unfolded conformations in macromolecular self-assembly. Genes Cells 2001;6(1):1-12.

189. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. Faseb J 2004;18(11):1169-1175.

190. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure 2003;11(11):1453-1459.

191. Oldfield CJ, Ulrich EL, Cheng Y, Dunker AK, Markley JL. Addressing the intrinsic disorder bottleneck in structural proteomics. Proteins 2005;59(3):444-453.

192. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, AK D. Identifying disordered regions in proteins from amino acid sequences. IEEE Internat Conf Neural Networks 1997;1:90-95.

193. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003;53 Suppl 6:561-565.

194. Jin Y, Dunbrack RL, Jr. Assessment of disorder predictions in CASP6. Proteins 2005;61 Suppl 7:167-175.

195. Tress M, Clarke ND, Ezkurdia I, Kopp J, Read RJ, Schwede T. Domain Definition and Target Classification for CASP7. Proteins 2007(This issue.).

196. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36.

197. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.

198. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837-845.

199. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839-843.

200. Shimizu K, Hirose S, Inoue N, Kanai S, T N. POODLE: predicting protein disorder using machine-learning approaches. Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction 2006;7:22-23.