“Sapienza” university of Rome

Pasteurian Sciences PhD School – Cycle XXV

“New integrative tools for interactive protein structure modeling and function prediction”

CANDIDATE
Alessandro Barbato

SUPERVISOR
Prof. Anna Tramontano

TUTOR
Dr. Jan Kosinski

A.A. 2012-2013

1 - INTRODUCTION

1.1 - The importance of protein structure

The knowledge of the three-dimensional (3D) structure of proteins is fundamental to understand the molecular basis of their function, since it provides a wealth of information that cannot be deduced from the primary sequence alone [1]. Currently, the experimental determination of protein structure is achieved by X-Ray crystallography, nuclear magnetic resonance (NMR) and high-resolution molecular microscopy (EM). However, such experimental techniques are very expensive, time-consuming, and require highly specialized equipment. Even more importantly, they are not always applicable. For instance, the size of protein sequences that can be solved with NMR is limited to approximately 200 amino acids and X-ray crystallography requires high-level of protein expression and efficient purification of the biological sample. These difficulties, together with the increasing number of large-scale genome sequencing and meta-genomics projects, significantly contribute to the discrepancy between the relatively low number of solved structures and the huge number of available protein sequences. In fact, as of July 2012, PDB [2] contains approximately 83,000 structures while Uniref100 [3], the main amino acid sequence repository,

stores ~18,000,000 entries. Due to the intrinsic complexity of protein structure determination, several computational approaches for the prediction of the three-dimensional (3D) structure of macromolecules have arisen in the last decades. In fact, many proteins can be modeled with computational techniques and the 3D models can be useful, depending on their accuracy [4, 5], for several practical applications [6], for example: protein function prediction [7], target inhibitor design [8, 9], rational design of mutagenesis experiments [10, 11], interpretation of disease-related mutations [12], structure-based virtual screening studies [13], and drug discovery [14, 15].

1.2 - Protein structure prediction

The 3D structure of a protein is mainly dictated by its primary sequence [16]. Thus, knowledge of the amino acid sequence and of the environmental conditions of a protein should be sufficient to infer its native structure. However, in spite of the increasing efforts and improvements, protein structure prediction remains an extremely difficult task and still represents one of the greatest challenges in the current biological research.

The computational methods for protein structure prediction fall in two categories: template-based and free modeling (also called *de novo* and *ab*

initio modeling) [17]. The former is based on the idea that the structure of a target protein (i.e. the one we want to model) could be similar to the structure of a template, i.e. a protein the structure of which is known and that can be used for modeling the target. The term template-based modeling refers to both homology modeling and fold recognition: homology modeling is applied when the sequence similarity between target and template is high enough to hypothesize the presence of a homology relationship, which implies structural similarity. Fold recognition includes methods that do not rely on sequence similarity to infer target-template structural similarity. However, the boundary between homology modeling and fold recognition is rather blurred [18] because of the availability of both, more sophisticated methods for sequence comparison [19-21] and more sequence and structure data [3, 22]. In fact, many cases previously considered suitable for fold recognition techniques, are now classified as “hard” homology modeling cases.

Methods that do not make any use of known structures are classified as free modeling. Free modeling approaches, which will not be described here, perform protein structure prediction only relying on the target primary sequence, to which they apply the physical principles of protein folding often in combination with efficient fragment searching techniques [23]. The most

successful free modeling methods are QUARK [24] (best scoring in CASP9 [25]), ROBETTA [26], Rosetta@home [27], Bhageerath [28] and FoldIt [29].

The interest of the scientific community in protein structure prediction is reflected by the growing number of research groups participating to the international CASP [30] (Critical Assessment of techniques for protein Structure Prediction) experiment (Figure 1.1), which registered the participation of 248 groups worldwide in its last round (CASP9, 2010) [31]. The CASP experiment is organized every two years since 1994 and represents an independent mechanism for the assessment of methods of protein structure prediction. The experiment consists of three main stages: first, structures about to be solved by crystallography or NMR are identified, and their sequences are made available to predictors. Second, as the experimental coordinates become available, the models submitted by the participant groups are processed and evaluated by independent assessors. CASP highlighted that homology modeling still remains the most accurate technique for building structural model. Interestingly, besides highlighting the improvements and limitations of structure prediction methods, CASP has fostered the introduction of similar blind tests in other areas of research [32-34].

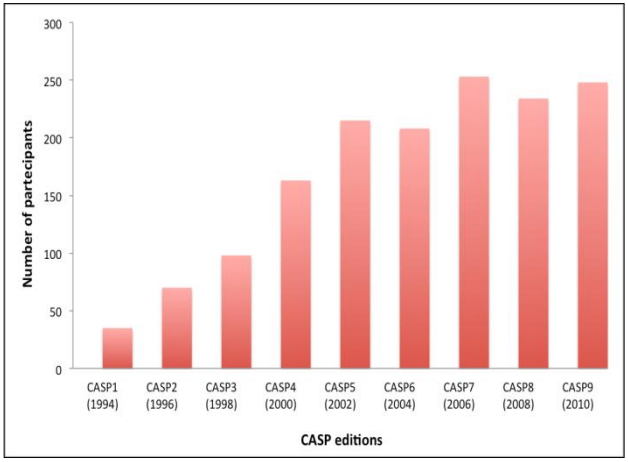


Figure 1.1: Number of participant groups to CASP experiment since its first round (CASP1, 1994).

1.2.1 - Homology modeling: rationale and advantages

As mentioned in section 1.2, homology modeling is based on the principle that evolutionary related proteins (homologs) have similar structures and that, therefore, the atomic coordinates of a suitable template can be used to model the structure of a given target. Such modeling procedure can be summarized in four major steps [35]:

- Template identification/selection;
- Target-template alignment;
- Model building/refinement;
- Model quality assessment.

A detailed description of each step is provided in sections 1.2.2-1.2.5. A reliable indicator of the homology relationship between two proteins is the percentage of their sequence identity (Figure 1.2).

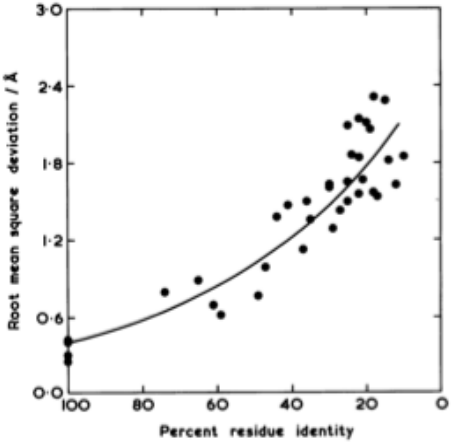


Figure 1.2: Relation between sequence identity and structural diversity (in terms of Root Mean Square Deviation –RMSD) in different pairs of homologous proteins. Extracted from [35].

In fact, as demonstrated by Chotia and Lesk in 1986 [36], the higher the sequence identity, the higher the structural similarity between the conserved regions (core) of target and template (Figure 1.3).

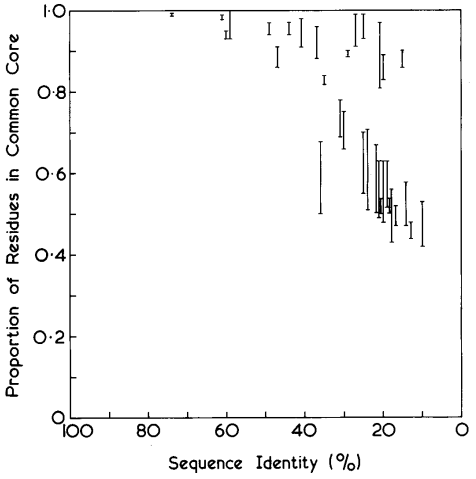


Figure 1.3: Size of common cores as a function of protein homology. If two proteins of length n1 and n2 have c residues in the common core, the fractions of each sequence in the common core are c/n1 and c/n2. Authors plot these values, connected by a bar, against the residue identity of the core. Extracted from [35].

Thus, being homology modeling based on sequence identity between target and template, it follows that: 1) it is the most reliable technique to predict the structure of the conserved region(s) of a protein (such as functional sites), 2) the accuracy of a model can be estimated *a priori* by knowing the degree of

target-template sequence identity. For these reasons, homology modeling is currently the most widely used method for protein structure prediction. In particular, the possibility to know in advance the quality of a 3D model enables researchers to evaluate its usefulness for the biological problem at hand [37]. As rule of thumb, if the target-template sequence identity is low (around, for example, 25%), homology modeling will likely produce “low quality” results. In fact, a low sequence identity reflects a remote homology relationship between target and template, which implies that the two proteins are likely to exhibit different local structures. Although models built on the basis of a higher sequence identity can display very good quality, the manual intervention of the modeler is usually needed to reach the best result [38]. This is particularly true for target-template alignments within the ideal range of 25%-50% sequence identity. In these cases, typical mistakes are inaccuracies in the alignment and inappropriate modeling of loops and side-chains [39-41]. Finally, models derived from templates with very high sequence identity to their targets (approximately > 50%) [42] are usually considered “high-accuracy” predictions, which can be successfully obtained using automatic modeling software.

Large-scale protein structure prediction is increasingly applied and examples of iterative genome-wide homology modeling strategies are available in the

literature [43-45]. Moreover, genomics projects make large use of structural models since, as explained in section 1.1, they cannot keep up with the newly sequenced genes only relying on experimental techniques for protein structure determination [46]. However, the use of homology modeling in genome wide experiments is still limited. In fact, the use of fully automatic modeling methods is necessary because of the large number of proteins to be modeled. However, the automatic models provide reliable models only when a template with high sequence identity is available.

1.2.2 - Template identification and selection

The identification and selection of the optimal template is the first key step in homology modeling. In fact, the detection and correction of errors in 3D models based on wrong templates is a very difficult, if not impossible, task [47]. BLAST [48], PSI-BLAST [20] and HHSearch [49] are the most commonly used algorithms for the identification of templates. All of them rely on target-template sequence similarity. In order to increase the sensitivity of the search, BLAST uses pairwise sequence alignment of each target homolog, PSI-BLAST uses profiles and HHSearch the Hidden Markov Models (HMMs). The results of CASP experiments showed that all these

methods perform well in identifying templates sharing a sequence identity higher than 50% with the target. However, below this threshold, HHSearch outperforms both BLAST and PSI-BLAST, as it is more effective in detecting remote homology relationships [21, 50].

The templates identified by PSIBLAST, BLAST or HHSearch are ranked according to an E-value that reflects the probability of a given template to be homologous of the target. However, the ranking alone does not always ensure that a template is the best one; thus, automatic selection of the best ranking sequence(s) can lead to modeling mistakes. In fact, template selection must be performed also considering structural and functional features, i.e. using a “knowledge-based” approach. Ideally, the structural conformation of a template should be the same as the one of the target. For instance, to model a target protein in a ligand-bound conformation, a template structure solved in complex with a ligand should be preferred to a ligand-free template. Or, if the biologically active form of a target is known to be an oligomer, the best choice would be an oligomeric template with the same number of subunits and symmetry of the target. In this case, better ranking templates exhibiting different quaternary structures might provide less optimal models. Finally, templates solved by X-ray crystallography should be preferred to NMR structures, high-resolution templates to low

resolution ones and structures with no missing coordinates to structures where intrinsic disorder made impossible the determination of some coordinates. If it is not possible to unequivocally select a single best template from a set of alternatives, a model can be also built from multiple templates. This is accomplished either by averaging the coordinates of superposed templates or by modeling different regions of the target based on different templates. The use of more than one template proved to be very effective and provided the basis for several successful protein structure prediction methodologies [51].

1.2.3 - Target-template alignment

Once the optimal template has been selected, a target-template sequence alignment must be performed. This makes it possible to highlight which residues of the target should be modeled using which residues of the template(s). Methods listed in 1.2.2 provide target-template alignments based on the maximization of sequence similarity. This might reflect the evolutionary history of the proteins but not necessarily the best alignment for

modeling. For example, Figure 1.4 shows how the best sequence alignment could not correspond to the structural superposition.

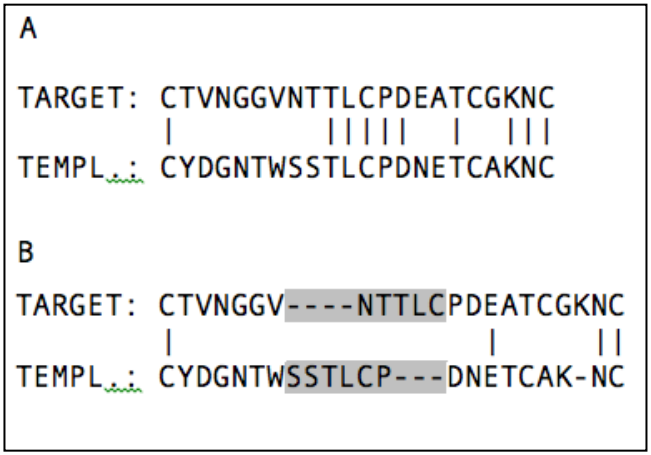


Figure 1.4: Sequence similarity may not reflect structural similarity. The example is extracted from [52]. The target protein is Endoglucanase I, which has a sequence identity of 47% with its best template (PDB 1CELA). (A) The sequence alignment obtained for the region 49-70 of the target protein. (B) The sequence alignment corresponding to the best structural superposition between target and template. Although the sequence similarity results maximized using BLAST to generate the alignment A), the protein shaded regions are not correctly structurally aligned (i.e. their backbones differ of more than 4A).

The discrepancy between sequence and structure-based alignments may be due to the presence of insertions and deletions (indels) in the alignment [53]. Alignment algorithms maximize a score which takes into account both the frequency of substitution of two given amino acids during the evolution and

the penalties for the insertion of gaps. While the score component accounting for the substitution of amino acids is usually highly reliable, gap penalties are only heuristically optimized and thus the indel position may turn out to be inaccurate. For example, it may happen that a sequence-based alignment, places a gap in the model within a secondary structure element rather than in a loop, which would be more appropriate from a structural point of view. The difficulty of automatically aligning two sequences grows with the number of indels, which strongly depends on the sequence identity [37]. In principle, it is not advised to fully rely on models built automatically on templates sharing less than 50% identity with the target. In such cases, the automatic alignment should be used as a starting point for a manual refinement process aimed at localizing and correcting all the misaligned regions. Typically, expert modelers apply the following rules to manually refine a target-template alignment:

Regions aligned with high confidence are firstly identified. As a rule of thumb, regions sharing high sequence similarity and presenting common motifs can be considered as correctly aligned because they tend to be evolutionary conserved [54].

The localization of the hydrophobic residues in the protein core is preserved whereas charged residues are kept exposed to the solvent.

Indels are placed in template solvent-exposed regions, preferably not in secondary structure elements, or in loops that are not conserved in a MSA of target homologs. Moreover, indels introduction is avoided in both the protein core and in the predicted functional sites of the target.

The application of these rules relies on the assumption that, during evolution, the protein core, secondary structure elements and functional sites tend to accumulate fewer mutations. The effect of the manual improvement of template selection and alignment can be observed in Figure 1.5 where, human aided models from CASP7 outperform the automatically built ones in terms of quality.

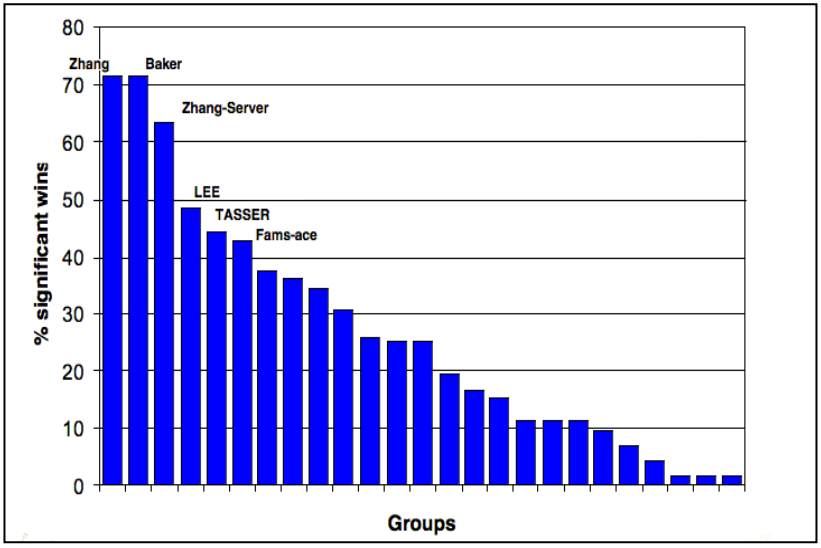


Figure 1.5: Most successful methods in CASP7: In figure it is shown the direct head-to-head comparison of the top 25 participant groups to the TBM category. The bin height corresponds to the counting of the number of statistically significant wins against all other groups on common target. Zhang and Baker ("human aided" predictions) ranked better than the Zhang-Server predictions.

1.2.4 - Model building and refinement

In the model building step of homology modeling, the 3D atomic model is built based on the given target-template alignment and template structures. Nowadays, model building is a fully automated procedure. There are several methods dealing with this task. They can be grouped into two classes: 1)

rigid-fragment assembly and 2) modeling by satisfaction of spatial restraints [55, 56].

The rigid-fragment assembly methods build a model of the target from the structurally conserved regions of the template(s). In addition, such modeling procedure uses the structural fragments obtained either from a structural database, for target regions that are not aligned with any homolog, or from the structure of aligned templates. Eventually, the fully assembled model is optimized to refine the connections between the fragments by reducing their potential steric conflicts. Software tools implementing the rigid-fragment approach are: SWISS-MODEL [57], NEST [58], 3D-JIGSAW [38] and BUILDER [59, 60].

Modeling by satisfaction of spatial restraints relies on a single optimization strategy aimed at building a structural model that optimally satisfies the restraints derived from the target-template alignments, known protein structures, and molecular mechanics force-fields. Such restraints include van der Waals contact distances, chiralities, bond length and angles, side chain rings etc. This modeling approach may not require a further refinement step. The most widely used software implementing the modeling by satisfaction of spatial restraints is MODELLER [61], which allows users to add their own

restraints such as the presence of a ligand or distances between given residues.

Benchmarks of model building software are available in the literature [55, 62] and they highlight that MODELLER performs better on the average. Moreover, the possibility to include both ligand and user-defined restraints made MODELLER the most widely used package.

Subsequently to the model building procedure, it is common to use specialized protocols to enhance the accuracy of non-conserved elements of the models (i.e. loops) and side chains packing. Such operation is important to obtain reliable models useful for an accurate function prediction. In fact, the evolutionary variable elements of a protein often encompass indels or unstructured regions that can play a pivotal role for its biological function as well as the specific orientation of the side chains in the catalytic site of an enzyme.

1.2.5 - Model quality assessment

Model quality assessment constitutes an inseparable part of the homology modeling procedure. In fact, in addition to determine the suitability of a model for biological applications, it indicates whether a model can be built

with higher quality, and then if there is the need to correct the inaccuracies deriving from the previously described steps [63]. The model quality assessment programs (MQAPs) aim at estimating either the overall accuracy of the final model (global quality) or the local accuracy (local quality) of its individual regions. Although a variety of MQAPs, implementing a wide spectrum of methodologies, from physical potentials to knowledge-based scoring functions, are available [45, 64-72], for a matter of consistency with the study presented in this manuscript, here will be only described the QMEAN software [73]. The choice of QMEAN was made by taking into account the results of CASP8 (Figure 1.6) experiment [74], the need of an MQAP able to assess both global and local quality of a model and the ease of integration within the computational framework described in the results.

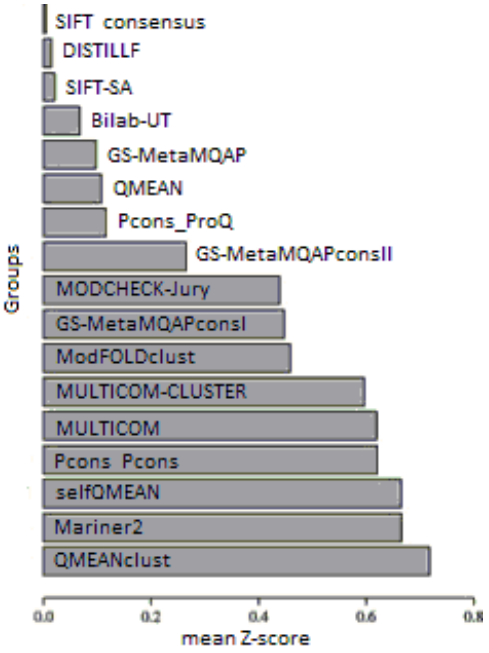


Figure 1.6: Seventeen groups submitted confidence estimates at the residue level, and we evaluated the correlation between such values and the distances in Å̊s between the predicted and observed positions of each Cα after optimal sequence-based superposition of targets and models. It is evident that QMEAN group outperformed all other predictors in this task.

1.2.5.1 - QMEAN

Many MQAPs are designed to give a quality estimation of a structural model by comparing a set of alternative predictions for the same target sequence. Therefore, such relative ranking is not sufficient to determine the suitability of a model for a given biological application. QMEAN is able to provide an

estimation of the absolute global and local quality on the basis of one single model. It is a scoring function composed by the linear combination of six terms called structural descriptors [73]. Long-range interactions are assessed by two different distance-dependent interaction potentials of mean force: one based on Cβ-atoms and another based on all the atoms composing the models. Then, the local backbone geometry is evaluated by a torsion angle potential over three consecutive amino acids whereas a solvation potential is used to analyze the solvent accessibility of the residues. Finally, the agreement between the secondary structure and the solvent accessibility, predicted and evaluated respectively for target and template(s), is also taken into account for the final quality prediction.

1.3 - Aim and contributions of the study

The aim of the study described here is to contribute to the field of protein structure prediction by developing, testing and applying MODORAMA: a web-based platform that facilitates the expert approach to homology modeling of the proteins that cannot be built with automatic procedures. The outcomes from the various CASP meetings have pointed out that human expertise is usually a key component of the success of a homology modeling

experiment (Figure 1.6). For cases, when a template sharing more than 50% of sequence identity is not available, wrong template selection and inaccuracies in the target-template alignment are the major sources of error in the resulting models. Thus, to correctly address the model building procedure, expert modelers tend to consider information that currently cannot be exploited in an automatic fashion. MODORAMA allows manual template selection and target-template alignment refinement and enables the knowledge-based approach typically used by the expert modelers discussed in 1.2.2 and 1.2.3. Although MODORAMA has been designed as a homology modeling platform, the biological annotation displayed by its interface makes the tool also useful for exploring the sequence, structural and functional diversity in a protein family, as it is later shown in the example 2.5. In addition, I performed a study aimed at the assessment of the possibility to implement an automatic target-template alignment refinement method based on the usage of the QMEAN local quality estimation (see 2.4). This part of the work was carried out at the Swiss Institute of Bioinformatics at the Structural Bioinformatics Unit at the University of Basel in collaboration with Dr. Pascal Benkert and Prof. Torsten Schwede, who are the developers of QMEAN. Such study can be considered a step toward the understanding of how to improve the quality of the model built with an automatic procedure.

2 - RESULTS

2.1 - MODORAMA platform overview: MODexplorer and MODalign

MODORAMA is a web-based integrated platform that enables both homology modeling and investigation of the sequence, structural and functional diversity of protein families. The platform is available at the url http://modorama.biocomputing.it (Figure 2.1) and consists of two resources: MODexplorer [75] and MODalign [76] that are described in detail in Appendix A and B, respectively.

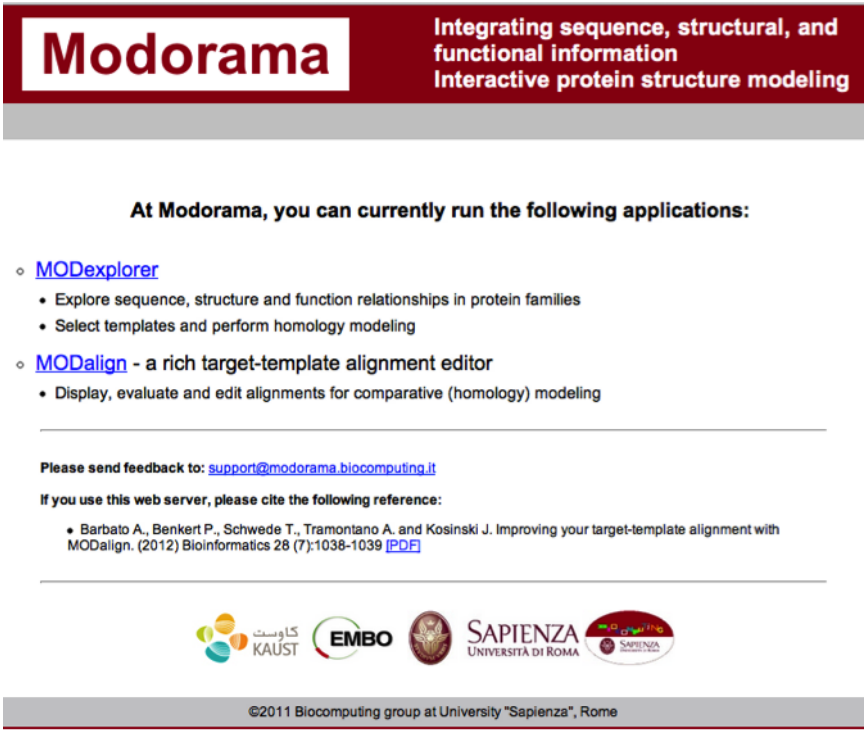


Figure 2.1: Modorama platform home page

MODexplorer takes as input the amino acid sequence or the structure of a protein of interest (target) and builds the multiple sequence alignment (MSA) of the input protein family. Then, with this MSA as input, it uses HHSearch to

retrieve the target homologs with known structure (templates) and to generate the corresponding target-template alignments. Since the database of known protein structures used by HHSearch contains only representative proteins (i.e. for a set of proteins similar in sequence only a representative is included in the database), MODexplorer retrieves and aligns the excluded related structures of the identified representative homologs. Including such related structures is very important for practical applications as they might contain different ligands, represent alternative conformational states or exhibit higher structural quality (see Example 2.3-2.5). Finally, after the computation described above, MODexplorer outputs an interactive interface (Figure 2.2) that displays the MSAs of the families of both target and homologs, along with a variety of structural and functional annotations. The MSAs are displayed as both BLAST-like bar diagrams (Figure 2.2 – red arrow) and as MSAs (Figure 2.2 – blue arrow). Several annotations are graphically depicted on the MSAs, including: ligand and DNA/RNA binding sites, secondary structures, similarity scores, disorder (predicted for the sequence query and estimated by B-factor/missing residue annotations for the structure query and known hit structures), and QMEAN local and global scores of models built based on the alignments. Thus, if the structure of the target protein is known, MODexplorer can serve as an easy-to-use resource

to explore the sequence and structural conservation among similar proteins, to find structures of homologs solved in different conformational states or with different ligands and to transfer functional annotation from the other structures.

Figure 2.2: Snapshot of the MODexplorer interface in “Ligands” display mode, where ligand binding sites are marked in color on the alignments. The interface is composed of three panels. The filtering panel allows filtering the hits by functional and structural annotations. The overview panel displays the hits as a BLAST-like diagram. The detail view panel enables displaying alignment of the query to currently selected hit along with the MSAs of their families. In this example (query: C-terminal domain of PMS2 protein), users can easily find that one of the structures (3KDK) related to one of the two top scoring templates (3KDG) contains metal ions associated with conserved motifs (see detail view panel).

However, if the aim is to predict the structure of the input protein, MODexplorer can be used as a homology modeling tool. In fact, MODexplorer also allows modeling the target protein based on any selected target-template alignment. In particular, the ‘knowledge-based’ approach for template(s) selection (see 1.2.2) is greatly facilitated by the interactive interface that allows the simultaneous browsing of structural and functional annotations of both target and identified templates. For a detailed description of all the features of MODexplorer refer to Appendix A and Methods.

MODalign works as a dashboard for target-template(s) alignment inspection and manual modification. It is designed to enable the application of the rules for target-template alignment refinement described in 1.2.3 and enables building of the three-dimensional models from the alignment that is being analyzed. Thus, MODalign accepts as input a pre-computed target-template(s) alignment that is initially ‘enriched’ with the MSA of representative sequences of both target and template families, and with the sequences derived from the SEQRES and ATOM fields of the PDB entry(ies) of the template (see Methods). Such alignment can be finally displayed and refined using a graphical interface that allows for editing operations such as residue shifts and insertions in the target or the template. In addition, the interface: a) depicts the sequence conservation for each

column of the alignment, b) displays secondary structure and solvent accessibility values for target, templates and respective homologs, and c) upon request computes and display the QMEAN global and local scores of the model implied by the current alignment (Figure 2.3).

Moreover, MODalign, accordingly to what discussed in 1.2.3, highlights potential errors in the alignment, such as: insertions or deletions within secondary structure elements, cases where a hydrophobic or charged residue in the target is aligned to an exposed or buried residue in the template. Finally, it is important to notice that the editing interface automatically displays the changes in all members of the protein families and re-computes all the data described above. For a full description of MODalign refer to Appendix B and Methods.

MODexplorer and MODalign are tightly connected: a target-template alignment selected within the MODexplorer interface can be forwarded to MODalign in order to be refined or more thoroughly investigated. From the point of view of a modeler, such connection makes MODORAMA a platform that can drive the user through all the steps of the homology modeling procedure, from template selection and alignment refinement to model quality estimation.

In the following I will describe some examples of the usefulness and effectiveness of the tools.

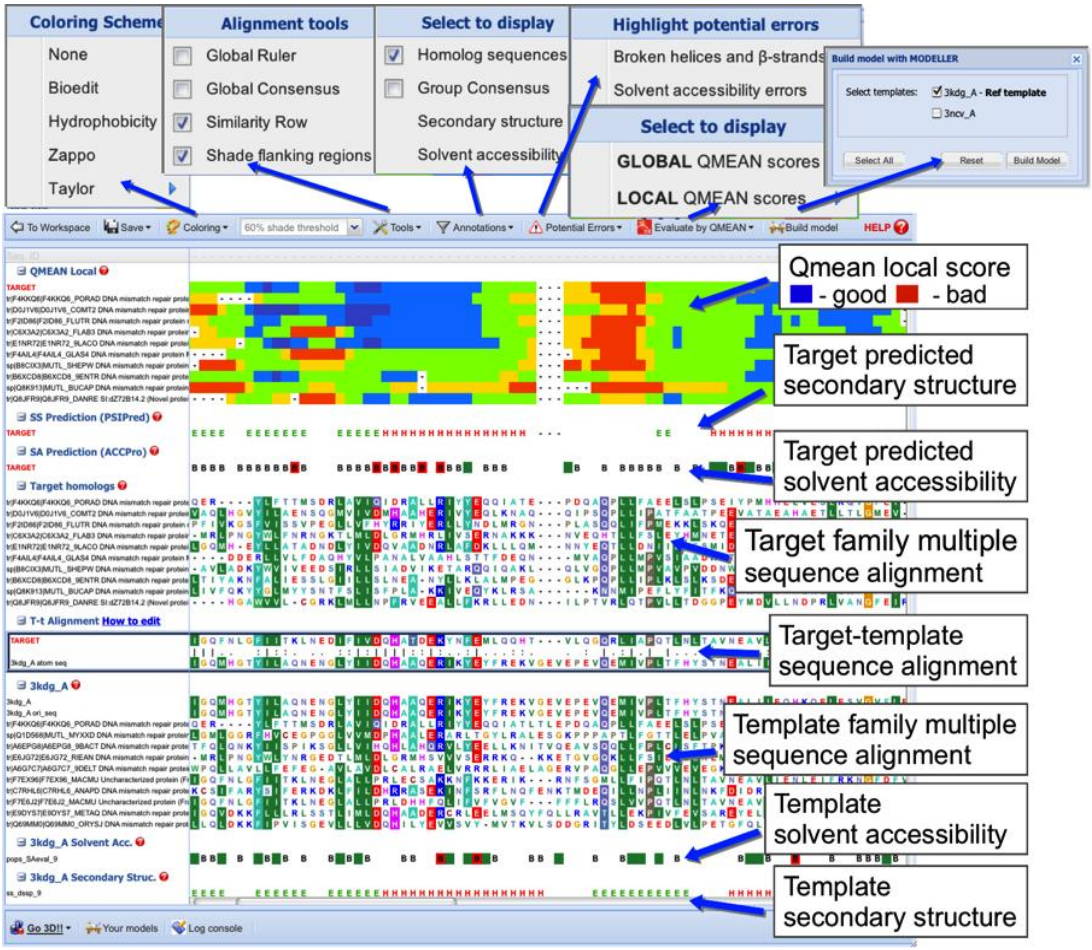


Figure 2.3: The MODalign alignment editing interface. The alignment can be edited in the target-template alignment section (see Appendix B for more details). When the alignment is modified the output, such as coloring by sequence conservation or highlighting of potential errors, is modified in real time, while QMEAN can be re-calculated on request.

2.2 - Modeling of BcnI

First, we tested MODORAMA by modeling the BcnI restriction nuclease, an enzyme that recognizes and cleaves the duplex DNA containing the sequences CC/SGG (where S stands for C or G). The structure of BcnI is known (PDB codes: 2odi, 2odh, 2q10, 3imb) but for the purpose of this study, we assumed that the structure is not available and must be modeled. Finally, we compared the model with the known structure. Such retrospective analysis aimed at a preliminary determination of the platform accuracy and usefulness. Moreover, this example is a good test case for MODORAMA as the most suitable modeling template resulted to share a sequence identity of about 20% with BcnI. In fact, as largely described in the introduction, MODORAMA is intended to facilitate the modeling of those targets that have low sequence identity (lower than 50%) with the detected templates, and for which manual intervention in the modeling procedure can make the difference in term of the prediction quality.

2.2.1 - Template detection

In order to identify the most suitable template for modeling BcnI, we investigated the structural and functional diversity of the families of the

identified homologs with MODexplorer. Being this a retrospective analysis, the best scoring template is the crystallographic structure of BcnI itself (2odi_A). However, if 2odi_A is not considered, the most suitable template, according to the HHSearch probability (see Methods), is 2oa9_A that corresponds to the X-ray structure of MvaI in the absence of DNA. As discussed in 1.2.2, it is a good idea to investigate whether there is a template solved in a desired conformation. For modeling BcnI, ideally we would like to model it in the conformation bound to DNA, which corresponds to the functional conformation of this protein. Thus, using the filtering feature of MODexplorer, we filtered the alternative templates in order to display only those templates solved in the presence of DNA. In fact, amongst the structures related to 2oa9_A, we found 2oaa_A, which is the structure of the restriction endonuclease MvaI-cognate DNA substrate complex. At first glance, the target-template alignment, as provided by HHSearch, required a refinement aimed at proper placing a number of indels (see 1.2.3).

2.2.2 - Manual refinement of the target-template alignment

Subsequently to the identification of 2oaa_A as the most suitable template for modeling the BcnI structure, we forwarded the corresponding target-template alignment to MODalign. In parallel, we performed a structural superposition of the 2oaa_A template with the known structure of BcnI (pdb code: 2odi_A) and derived a BcnI-2oaa_A alignment based on this superposition. Since the structural superposition reflects the best sequence alignment possible, here we will refer to it as the 'correct alignment' and compare the alignment refined with MODalign with this correct alignment.
The refinement process has been performed as follow:

- β-strand (columns 24-27 of the alignment)

In this region there is an insertion placed within a potential secondary structure element (β-strand) of the target. Such situation is detected as a potential error by MODalign and it is depicted at both sequence and structural level (Figure 2.4). To correct it, we placed the insertion in a non-conserved and solvent exposed region (loop). The correctness of the

alignment editing is confirmed by the comparison to the correct alignment (Figure 2.4 - C). Similarly, an insertion within the α-helix at columns 250-256 was placed in the loop nearby.

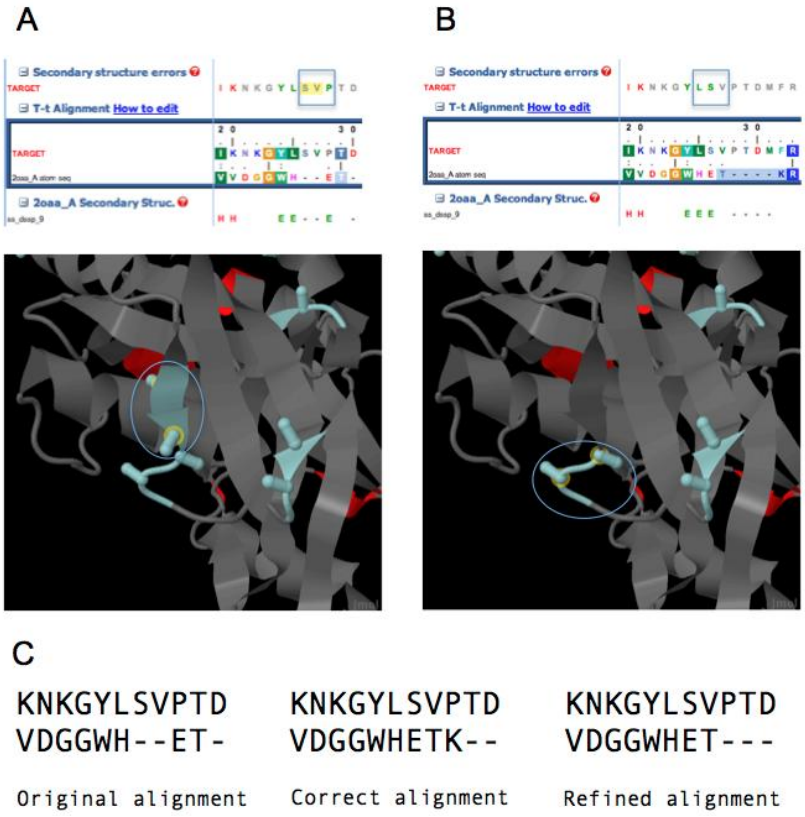


Figure 2.4: (A) MODalign detects a potential error in correspondence of the region within the cyan rectangle. In fact, the HHsearch alignment inserts two gaps in a β-strand of the template, as it is highlighted by the cyan circle. Such an alignment would result in a distortion within a β-strand of the target. The refinement proposed in (B) consisted in moving the insertion in the loop nearby the β-strand. (C) The refined alignment is more similar to the correct alignment than the original one.

- β -strand(columns 131-138 of the alignment):

A truncated β-strand would be built from the original target-template-alignment (Figure 2.5). Thus, the editing was aimed at placing the deletion within the loop nearby the β-strand in order to keep the β-strand intact. Moreover, as it is highlighted by the similarity row, the editing produced a higher sequence similarity between the two shifted amino acids. The resulting alignment turned out to be perfectly correct.

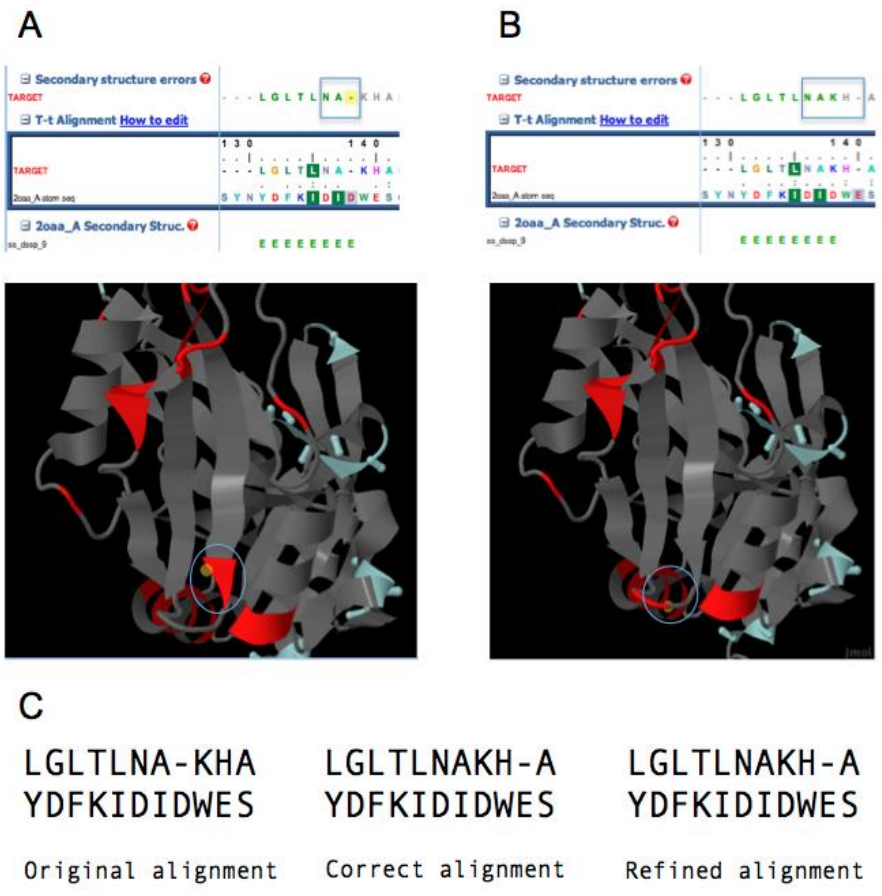


Figure 2.5: (A) The presence of a deletion, corresponding to the last residue of a β-strand (cyan circle), is the cause of the potential error detected by MODalign. Such an alignment would result in a target model containing a disrupted β-strand. The refinement proposed in (B) consisted in placing the deletion within the closest loop (cyan circle). (C) The identity between the refined and the correct alignments witnesses the correctness of the editing.

- β -strand-loop β -strand (columns 143-158):

Two potentially distorted β-strands would be built in the segment of the model corresponding to this region of the target-template alignment (Figure 2.6). To avoid this situation and to facilitate the subsequent connection of these elements by the modeling software, we edited the alignment such that both the two β-strands have shorter distortions and the deletion entirely falls within the connecting loop. Moreover, the distance between the two residues flanking the deletion only in the refined version is close to the expected Cα-Cα distance (Figure 2.6 - B).

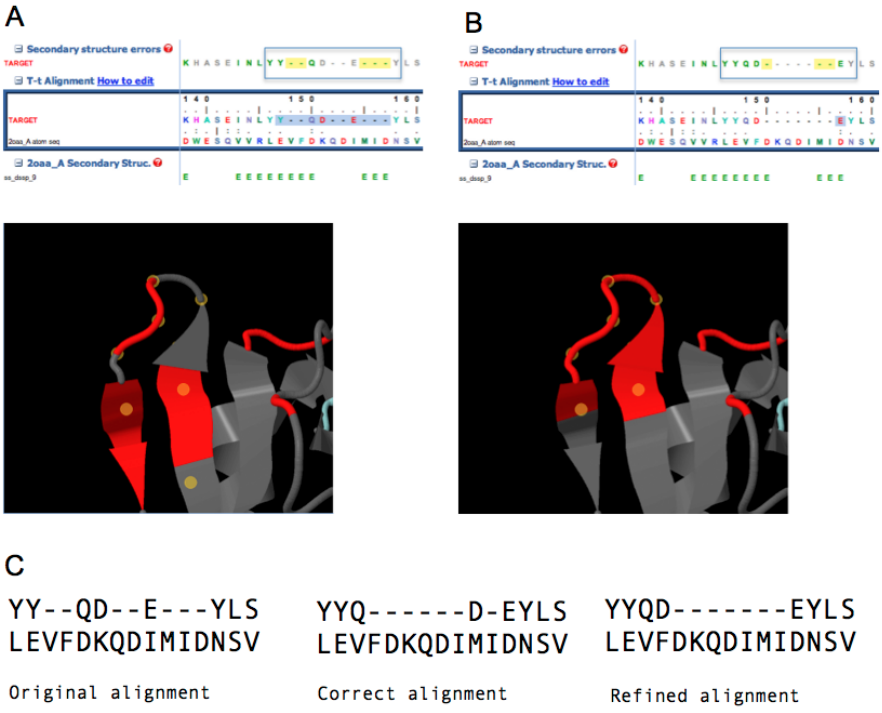


Figure 2.6: (A) The error detected by MODalign corresponds to the region in the cyan square. A deletion affects an entire β-strand composed of three residues, two amino acids in a loop and two amino acids in another β-strand. (B) The refinement consisted in placing the greatest part of the deletion within the connection loop. Moreover the distance between the residues flanking the loop has kept short in order to facilitate the connection of the backbone by the modeling software. (C) The similarity between the refined and the correct alignments, is an evidence that the refinement was correctly done.

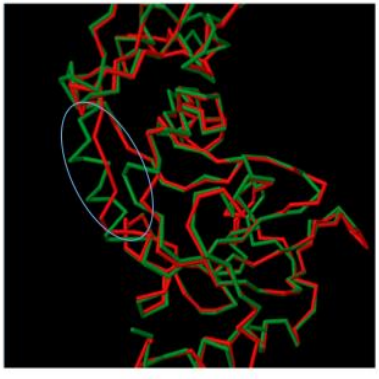
- Helix (columns 164-174):

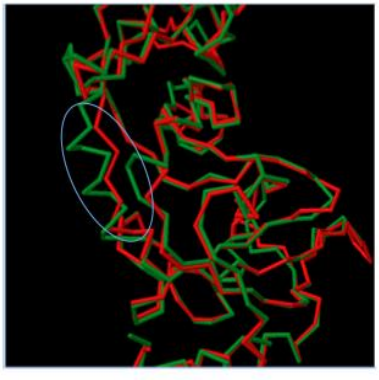
Here we observed a deletion of an entire helix. Usually deletions are placed outside the conserved regions, but in this case, relying on the annotation displayed in the MODalign interface, we could not move the deletion since the alignment of the strands on both sides of the helix seemed to be correct (there was high sequence identity in the strands). Moreover, the secondary structure prediction for BcnI did not include any helix in this region highlighting that it is more likely that the secondary structure element is really absent in the target protein. However, it is not possible to leave the alignment as it is, since later the modeling software can experience problems in connecting the backbone since the deletion affects an extended region (Figure 2.7). Moreover, MODalign indicated that this alignment would result in modeling a hydrophobic residue (a leucine) as exposed to the solvent. Thus, accordingly to what stated above and also considering the placement of the leucine at position 172, within the hydrophobic core of the protein, we edited the alignment as shown in Figure 2.7-B. Several alternative alignment shifts were possible but only a single solution led to the alignment in which all hydrophobic residues are buried in the protein core and distances between residues flanking the deletions are close to expected

Cα-Cα distance. The alignment editing also led to much better local QMEAN scores in the edited part and in the flanking regions. The final alignment very well agreed with the alignment from structural superposition and the protein backbone of the resulting model was almost perfectly overlapping with the native BcnI structure.

A

B





C

```
TWDLN------LSK
YWSFDSLQNQLDKK
```

Original alignment

```
TW---DL-NLS-KI
YWSFDSLQNQLDKK
```

Correct alignment

```
TWD---L-NL-SKI
YWSFDSLQNQLDKK
```

Refined alignment

Figure 2.7: (A) The potential error detected by MODalign consists in both a six-residues long deletion falling in a α-helix and in the alignment of a leucine of the target with a solvent-exposed aspartate of the template. The potential misalignment is also reflected by the QMEAN local score profile (yellow residues have lower quality than the green ones). Moreover, from the superposition of the target model (red), obtained from the original alignment, with the template (green) it is evident that the region of the target within the cyan circle contains a 'not natural' distortion of the backbone. (B) The refinement consisted, firstly in shifting the leucine 172 of the target in order to align it with another leucine of the template that is not exposed to the solvent. Then, as for the step 3 of the refinement, to make the backbone connection easier for MODELLER, we shifted residues inside the deletion, by keeping them 'spaced' in such a way that the distances between residues flanking all deletions are close to the expected Cα- Cα distance. The resulting QMEAN local profile does not present any 'yellow' residue, reflecting a higher quality of the refined alignment with respect to the original one. In the superposition of the template with the target model deriving from the refined alignment, the target region in the blue circle nearly 'follows' the coordinates of the template, without presenting any distortion. (C) Finally, also in this situation, the correct alignment and the refined alignment almost correspond.

2.2.3 - Assessing the quality of the results

The higher similarity of the refined alignment with respect to the one deriving from the structural superposition of BcnI native structure and MvaI, indicates that MODalign is very useful in refining the alignment. To further assess the importance of the alignment refinement we performed an additional analysis by building the model from the refined alignment and comparing its quality with of a model built using a non-refined alignment as follows.

The root-mean-square deviation (RMSD) is a measure of the average distance between the atoms of superimposed structures and it is widely used

to assess a structural similarity between proteins. In particular, we evaluated the effective improvements in the model built after the manual refinement of the target-template alignment by comparing its RMSD to the one of the native BcnI structure. Moreover, we calculated the RMSD of the same model but built using SWISS-MODEL, which is a widely used automatic modeling method.

The structural superposition between the native structure of BcnI (2odi_A) and the model for which the alignment has been refined as described above, showed a global RMSD of 4.68 Å whereas the structural superposition between 2odi_A and the model built from the not refined alignment showed a global RMSD of 4.82 Å. The difference is very small and this is because the core region of BcnI is big and identically modeled in both models. In fact, accordingly with TM-score (the measure provided by the server used to superpose the structures), 199 out of 238 residues of the target protein are considered part of the core.

If we extend the RMSD analysis focusing on those regions affected by the editing, we can better appreciate the usefulness of the refinement process. In fact, the RMSD improvement in the local regions (Table 2.1) around the refined ones is ~1 Å (3.68 Å for the superposition of 2odi_A with the non-

refined model and 2.78 Å for the superposition of 2odi_A with the refined one).

Region	**Original alignment RMSD (Å)**	**Refined alignment RMSD (Å)**	**RMSD improvement (Å)**
β-strand (region: 23-34)	4.77	3.43	1.30
β-strand (region: 131-144)	3.68	3.38	0.30
β-strand-loop-β-strand (region: 145-159)	2.82	2.59	0.23
Helix (region: 161-177)	2.45	1.31	1.14

Table 2.1: table showing the RMSD difference between every single region refined with MODalign as described in 2.2.2

2.3 - Inferring the potential MSH6 ATP-bound conformation

2.3.1 - Human MutSα DNA lesion recognition

The MutSα complex from human is a heterodimer composed by two subunits, MSH2 and MSH6, and it is a crucial component of the DNA mismatch repair (MMR) pathway [77]. The complex recognizes and binds those DNA strands containing errors such as mispairs or insertion/deletion loops. In addition to its DNA recognition activity, MutSα acts as an ATPase thanks to the two ATP binding sites placed on the C-termini of MSH2 and MSH6 [78]. There are also evidences that the DNA recognition modulates the ATP hydrolysis and vice-versa. Moreover, defects in MMR pathway are correlated with hereditary non-polyposis colorectal cancer (HNPCC), and a number of HNPCC causative mutations have been identified on the two genes encoding for MSH2 and MSH6 subunits [79, 80]. Thus, given the high scientific interest for such a complex, in collaboration with the group of Prof. Jiricny, we decided to assess the feasibility for selective inhibition of the human MutSα complex via chemical genetic approach. A prerequisite for

designing this approach the structure of MutSα in complex with ATP is necessary but such structure has not been yet solved experimentally.

2.3.2 - Chemical-genetic and classical genetic approaches

Classical and chemical genetic approaches are considered effective tools for protein function investigation thanks to their ability to affect the activity of a single protein in a whole pathway [81]. In particular experimental systems that incorporate the advantages of both of these approaches are powerful tools to identify which proteins regulate different biological processes or to understand the molecular details of how a given protein performs its biological functions. With the classical genetic approach a phenotype of interest (e.g. as a cell presenting a given pathway in which a particular gene is silenced) is created by mutating a specific gene in a cell or animal (e.g. a knock-out mouse). However, the usage of the classical genetic approach to alter a protein function might be a not-ideal process, since protein levels respond very slowly to changes at the gene level and the effect of the mutation is not reversible. On the other hand, the chemical-genetic approach overcomes such limitations since to alter the protein function it affects the

protein and not the gene. A self-explicative example of what effectively is the chemical genetic approach can be derived from the pioneering studies of Shokat et al. in which the authors engineered specific kinases to interact with synthetic inhibitors and substrates which were not recognized by the wild-type kinases [82, 83]. In particular, in their study, Shokat and coworkers first identified an highly conserved bulky-residue (named gatekeeper), then discovered that the mutation of the gatekeeper residue in a smaller amino acid, such as glycine or alanine, creates a ‘hole’ in the binding site that makes the mutant kinase able to accept inhibitors and substrates with steric hindrance substituents that cannot enter the wild-type kinase ATP-binding pocket. In this manner, in absence of the inhibitor, the mutated kinase can work as the wild-type one, conversely in presence of the inhibitor, the activity of the mutated kinase is inhibited since the ATP binding pocket is occupied and the ATP cannot enter the catalytic site (Fig 2.8). Moreover, since the inhibitor can be designed to specifically bind the artificial binding site with the "hole", it will not bind to other ATPases in the cell. This in turn will make it more likely that any phenotypic effects are due to inhibiting the ATPase under study. Finally, it is evident how the inhibition accomplished via chemical genetic approach is fully reversible: it is solely dependent by the concentration of the inhibitor itself.

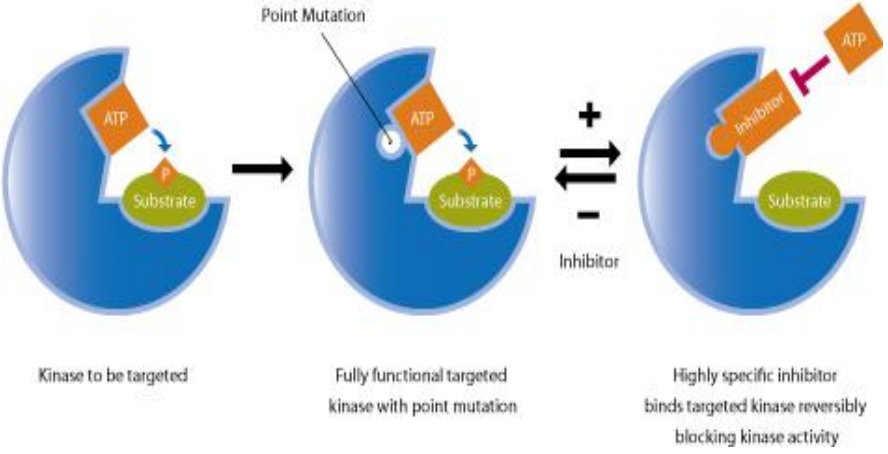


Figure 2.8: Schematic description of chemical genetic approach proposed by Shokat

In our case, a reversible inhibitor allowing to ‘switch on/off’ the targeted complex, would make easier the understanding of the tumorigenic mechanisms of the human MutSα complex in the context of MMR pathway. At the time when this thesis was written, the study was still ongoing. However, preliminary results of the research suggested that the drug bicyclomycin can bind the complex and then it can be potentially used for chemical genetic approach purposes. Below I describe how MODexplorer helped in addressing this specific problem.

2.3.3 - Bicyclomycin: a known inhibitor that can bind MutSα

Experimental results show that the ATP binding site of the MSH6 subunit has a more important role in modulating the DNA recognition activity of the MutSα complex than the site in MSH2. In order to search for available compounds proven to be effective inhibitors of other ATPases, we first extracted the structure of MSH6 ATP binding domain (according with the PFAM annotation [84]) and then we used this structure to retrieve all the PDB entries with a structurally similar ATP domain using DALI server [85]. Subsequently, we filtered out all those structures that have not been solved

with an inhibitor. Since we are looking for a compound that should be both selective and susceptible to the chemical genetic approach, we focused our attention on one particular molecule: the antibiotic bicyclomicin (BCM) [86]. Such a drug is an inhibitor of the prokaryotic Rho protein, and it replaces the water molecule necessary to carry out the ATP hydrolysis within the ATP binding site [87].

2.3.4 - MutSα ATP-bound conformation

Since there were not available structures of the MutSα complex (PDB id 2o8b/f) solved in presence of ATP we used MODexplorer to predict whether the MSH6 ATP binding site maintains the same geometrical shape if bound to ATP or ADP.

1. We submitted the MSH6 structure to MODexplorer;
2. We listed, within the MODexplorer interface, all the retrieved homologs of MSH6 that have been solved in presence of ATP, using the available filtering options;
3. The structure with both highest HHSearch probability and maximum coverage resulted to be the chain A of the E-Coli MutS complex bound to ATP (PDB code 1w7a_A). In order to find this protein, the

related structures to 1wb9_A had to be inspected since 1w7a_A does not appear in the list of the main homologs found;

4. We selected 1w7a_A. Subsequently, we inspected the superposition of MSH6 and 1w7a_A using the JMOL viewer embedded in MODexplorer;
5. The superposition of the two protein chains did not suggest any substantial shape differences between the MSH6 protein, that is bound to the ADP, and its prokaryotic homolog bound to the ATP;

The MODexplorer interface greatly facilitated this analysis: without MODexplorer one would not see in the HHPred server output a 1w7a_A structure since it is not in the non-redundant dataset used by the tool. Only by manually browsing the PDB entry of 1wb9_A one would have been aware of the availability of an ATP bound form of the MutS prokaryotic complex.

2.3.5 - Assessment of bicyclomycin binding capabilities

We superposed the structure of MSH6, in its hypothetical ATP-bound form, with the structure of the BCM as it is in the chain C of the Rho protein

(selected according with the DALI ranking). Then we observed that the compound is potentially able to enter the ATP binding cleft of MSH6 and create a binding pattern similar to the one of Rho. However no clashes with residues that can be mutated are highlighted. Thus, this analysis suggests that bicyclomycin is a potential good compound for the chemical genetic approach. The Prof. Jiricny's group is currently performing experiments to verify bicyclomycin binding to MSH6.

2.4 - Modeling of PMS2 C-terminal domain

PMS2 belongs to the MutL protein family. The MutL synergistically works with MutS in the DNA mismatch repair (MMR) pathway. It is recruited after the recognition of the DNA mismatches by MutS protein and then, together with MutS, it initiates DNA repair. MutL is a dimeric protein composed by two conserved domains and it exists as a homodimer in prokaryotes and as a heterodimer in eukaryotes. Moreover, the dimerization is primarily mediated by the C-terminal domains (CTD) that is present in the monomers. In human there are at least three different MutL complexes: MutLα, MutLβ and MutLγ. Among those human complexes, MutLα is the one involved in the MMR pathway: it is composed by the proteins hMLH1 and hPMS2. The structure of the N-terminal domain (NTD) of the complex has been solved for the *E. coli* homodimer and for the human PMS2 monomer. In the case of dimeric CTDs, only few prokaryotic MutL homologs have an experimentally determined structure. For the human PMS2-CTD a homology model has been built [88] using as template a structure of MutL from *E. coli*, which was the only MutL-CTD structure available at the time, and a structure of a distance homolog with another function, which allowed to model a zinc ion-binding site. Since then, new prokaryotic structures of the MutL CTD have

been solved. Thus, we decided to see whether it was possible to improve the quality of the available model of PMS2-CTD by using MODORAMA and the new structures.

2.4.1 - Template identification

The presence of metal ion binding sites on the PMS2-CTD structure was already known when the model was built for the first time, but at the time there were not MutL templates solved with the ions. Here, by using MODexplorer we have been able to find templates enabling better modeling of the target structure using the following procedure:

1. The sequence of PMS2-CTD was submitted to MODexplorer;
2. By investigating the results, I noticed that there were PMS2-CTD homologs more similar to PMS2 than MutL from *E. coli* (*1x9z)*;
3. In order to select only those templates solved in presence of metal ions, I filtered the retrieved hits by the following ligands: Zn (Zinc), Co (Cobalt), Mn (Manganese) and Ni (Nickel);
4. The best scoring template that was not filtered out corresponded to a structure related to 3kdg: 3kdk_A, which is the structure of the MutL-CTD from *B. Subtilis* bound to Zn^{2+} ion;

2.4.2 - Alignment editing and model building

After selection of the most suitable template for modeling, I decided to inspect the target-template alignment between PMS2-CTD and 3kdk_A with MODalign. From the MODexplorer interface, I forwarded the alignment to MODalign in order to see if the original target-template alignment could be improved:

1. In the MODalign interface, I highlighted the potential errors relative to α-helices and β-strands. A broken helix was present at the C-terminal of the model and then this potential error was corrected by shifting the corresponding residues of the template;
2. A deletion falling in an α-helix of the target was highlighted as a potential error and, in order to avoid building of a model with a truncated helix, I shifted this deletion to the closest loop;
3. The insertions seemed to be correctly placed in the original target-template alignment, thus I did not edit them in the alignment;

2.5 - Insights about detecting alignment errors with QMEAN local scores

As described in Appendix A and B, local QMEAN scores are widely used in MODORAMA to assess global and local quality of the alignments and structural models. In the MODalign alignment editing interface, the local QMEAN scores seem to help detecting those regions within the current target-template alignment that are likely to cause error in the target model. However, the local QMEAN scores are calculated based on the full atom models so that such predictions do not directly reflect the alignment quality. Moreover, building a model is computationally costly as after every step of alignment refinement, the users must wait from two to five seconds to get the updated QMEAN scores making it inconvenient to analyze many alternative alignments. In fact, as shown in Appendix B, the current implementation of MODalign displays and computes the QMEAN scores only upon request. To overcome this limitation, it is useful to develop and use a fast version of QMEAN, which can predict the local quality from a target-template alignment: this would allow real-time updating of the local QMEAN and enhancement of the error detection capability of the MODalign interface. Moreover, there are no available MQAPs for detecting local errors given a

target-template alignment and therefore a fast alignment-based QMEAN would be very useful also in other applications, such as the evaluation of the alignment accuracy in automatic methods aimed at alignment refinement.
For these reasons, I performed a thorough analysis to evaluate the possibility of using QMEAN local scores for detecting local errors at the alignment level, without building a model.

2.5.1 - Selection of the QMEAN version to use

QMEAN is a linear combination of eight scoring functions (referred to as terms). It predicts both global (i.e. for the entire structure) and local (i.e. per residue) error estimates based on one single model (see Introduction). Moreover, there are three versions of QMEAN that may work relying on a target-template sequence alignment. These versions of QMEAN are named according to the number of the embedded terms: QMEAN3, QMEAN4 and QMEAN5 (Table 2.2).

QMEAN3	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, salvation Cβ
QMEAN4	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED
QMEAN5	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED, ACCpro

Table 2.2: Combination of the terms used in the QMEAN3, QMEAN4, QMEAN5.

The predictions of QMEAN5 resulted to be the most accurate if compared with QMEAN3 and QMEAN4 [73].

Since QMEAN5 lacks the 'all_atom' term, it does not need the complete target structure to perform the prediction. In fact, it could output the quality estimates based on pre-calculated tables of values extracted from the target sequence and the template structure(s). However, since the implementation of an alignment-based version of QMEAN requires a large rewriting of the

QMEAN software and several optimizations, we used a re-parametrized QMEAN5 that can work with models composed only by backbone and Cβ atoms without modeled insertions (later referred as raw models). Being composed of the same terms, the accuracy of the re-parametrized QMEAN5 based on Cβ models directly reflects the accuracy of the hypothetical alignment-based version of QMEAN5, since the two versions only differ for details relative to the computational implementations. Thus, we investigated the feasibility of the re-parametrized QMEAN5 local scores (Q5LS) in detecting the local errors on raw models. I perfomed the experiment in collaboration with Dr. Pascal Benkert and Prof. Torsten Schwede, who are the developers of the QMEAN software.

2.5.2 - Dataset building: the ‘optimal’ alignments

First, we created a non-redundant list of target structures, which constitutes a dataset for re-parameterizing and testing QMEAN. The list consists of 1657 monomeric structures solved by X-ray crystallography at a resolution lower than 2 Å, and having amino acid sequences longer than 150 residues with sequence identity lower than 25% between each pair.

Next, we defined the ‘optimal’ target-template alignments to be used as reference to define the local errors in the low quality (‘suboptimal’) alignments. Initially, we retrieved via the HHSearch software the homologs with known structure (templates) for all the targets and obtained a set of 940,579 templates. To avoid redundancy we applied the following two filters: we removed templates that, with respect to the corresponding targets, showed a coverage lower than 90% and sequence identity higher than 60%. Thus, the list of templates was reduced to 75,397 elements. However, a number of low quality template structures were still present. Then, we discarded the templates not solved by X-Ray crystallography or with a resolution lower than 3Å, obtaining a list of 40,603 protein structures.

For ease of data manipulation and since the number of templates per target was different, we firstly subdivided the templates based on their targets. Next, we ranked the template structures according to their X-Ray resolution

and grouped them in six bins depending up on the identity of each template sequence with the corresponding target. Thus, we considered the five best ranking structures per bin allowing a maximum of 30 templates per target. The final number of templates was 4,086.

Since the most accurate sequence alignment possible is reflected by the structural superposition, we superposed the structures of each target with that of each template. Finally, from the structural superpositions, we extracted the corresponding 4,086 target-template sequence alignments forming the part of the dataset containing the ‘optimal’ alignments (table 2.2, row 1).

2.5.3 - Dataset building: the ‘suboptimal’ alignments

The ‘suboptimal’ alignments could have been built following different strategies. One could use a computational procedure to randomly perform modifications aimed at affecting the quality of a correct target-template alignment, like in Tosatto et al. [89]. Such approach, however, produces ‘unrealistic’ cases that would have not been even generated with the less accurate sequence alignment software.

To avoid this situation we used a different approach. First, we repeated the search of the target homologs by using four methods with decreasing

accuracy: HHSearch, PSIBLAST, BLAST and CLUSTALW. Thus, for each target/template pair from ‘optimal’ target-template alignment set we obtained up to four alternative target-template alignments. Notably, given the lower accuracy of PSIBLAST and BLAST with respect to HHSearch, for some target/template pairs PSIBLAST and BLAST did not detect the template and as a result the number of alignments from these methods is lower than the number of optimal alignments (Table 2.2 column 3). Moreover, given that CLUSTALW cannot be directly used to query a database, we generated CLUSTALW alignments for each target/template pair from the optimal alignment set. . The final dataset, also containing the ‘suboptimal’ target template alignments, was composed as shown in table 2.2.

Alignment 'quality'	Alignment Method	Total #aln	Training set #aln	Testing set #aln
Optimal alignments	TMAlign	4086	3010	1076
Suboptimal alignments	HHSearch	4086	3010	1076
	PSIBLAST	3652	2691	961
	BLAST	2975	2192	783
	CLUSTALW	4086	3010	1076
		18885	13913	4972

Table 2.3: Composition of the dataset. The dataset used for this study was composed by 'optimal' and 'suboptimal' alignments. It was further split in training and testing set for QMEAN5 re-parameterization. The training set consisted in the raw models built for 3/4 (13,913) of the elements of the dataset. The remaining 1/4 (4,972) were used as test set to assess the quality of the prediction.

2.5.4 - Local QMEAN5 scores correlation with Cα deviations

Next, the set composed of 3/4 of the full alignment dataset (13,913 alignments) was used to re-parameterize QMEAN5 (re-parametrization has been performed by Dr. Pascal Benkert). Next, the remaining 1/4 (4,972 alignments) were used as test set to assess the quality of the QMEAN5 predictions.

First, we superposed the native target structures with the corresponding raw models, built from the 'optimal' alignments, and evaluated the Cα deviations. Next, being both values expressed as RMSD, we verified the correlation

between Cα deviations and the Q5LS (QMEAN5 local scores) of the raw models built from the ‘suboptimal’ alignments. A positive correlation would imply that QMEAN5 correctly predicts the local quality of a raw model and that it can distinguish the local errors from the structural deviations between target and template.

We used the 13,913 raw models of the training set to perform the correlation analysis and we obtained the results shown in figure 2.1.

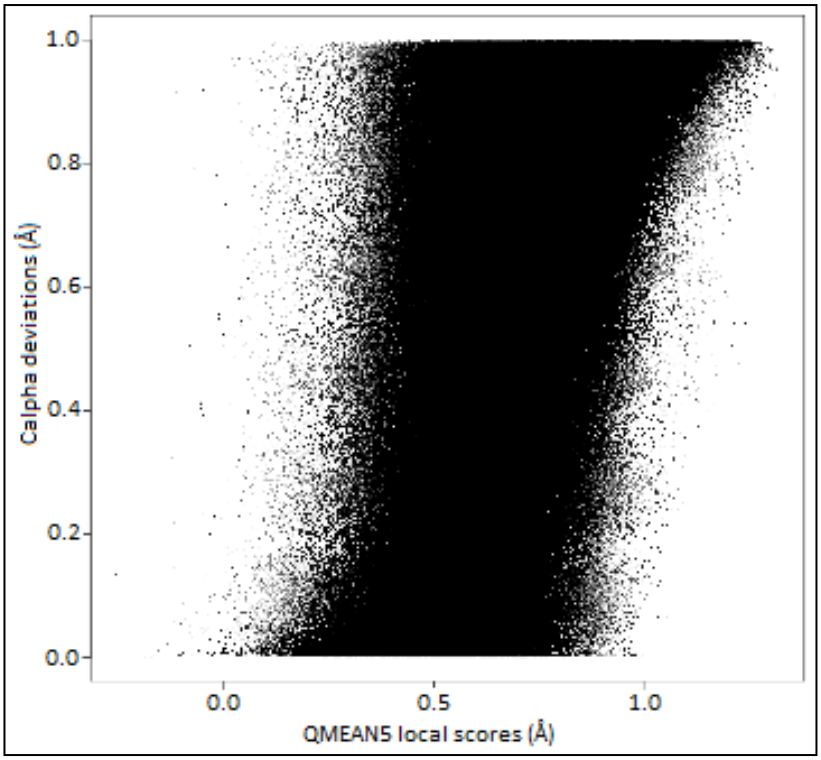


Figure 2.9: Correlation between Q5LS of the models built from the 'suboptimal' alignments and the Cα deviations between the native structures and the corresponding raw models built from the 'suboptimal' alignments.

The resulting correlation coefficient was low (0.2) and the signal was very noisy. Thus, to work with 'cleaner' data, we considered a small subset of the 'suboptimal' target-template alignments obtained as following. We isolated

67 alignments presenting only one shift, of a minimal length of 3 residues with respect to the corresponding ‘optimal’ alignments. Then, we compared the Q5LS of the 67 raw models with the Cα deviations of the respective models built from the ‘optimal’ alignments. A signal of correlation was detected since for the ~30% of the cases we observed higher values of Q5LS for the residues corresponding to the shifted segments than residues in the correct regions. As shown in fig. 2.10 the Q5LS difference resulted clearly shifted toward scores higher than zero. However, in order to work with a subset composed by both a larger number of elements and less noisy data, we used another strategy to continue the experiment.

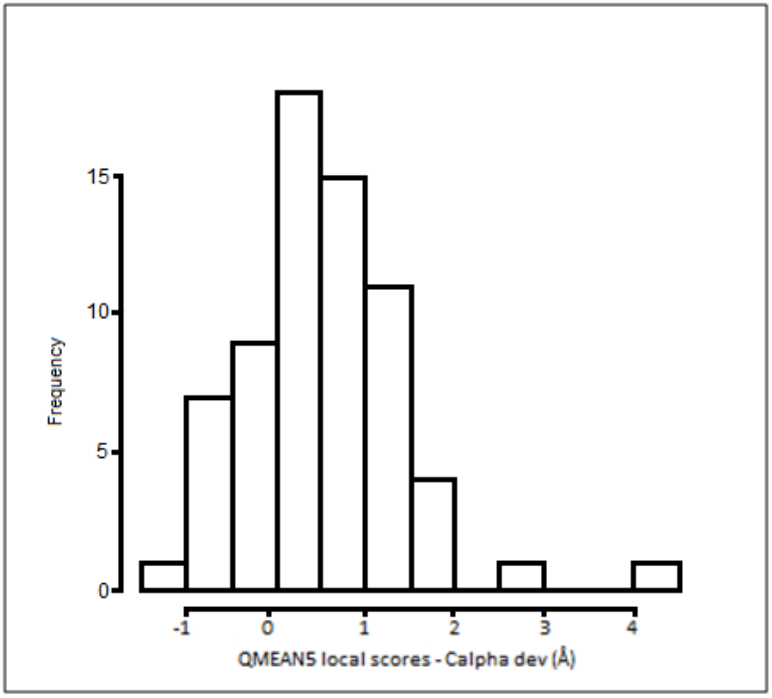


Figure 2.10: Distribution of the differences between Q5LS and Cα deviations for the subset of 67 target template alignments.

2.5.5 - A subset of ‘suboptimal’ raw models of high quality

We created a sub-dataset containing raw models in which we wanted the errors deriving from the structural deviations between target and template structures to be minimized. It contains 115 raw models built from HHSearch alignments in which the sequence identity between target and template was higher than 50% as discussed in Rost, 1999 (Table 2.3 last line). Subsequently we repeated the correlation analysis using the new dataset (Fig 2.10).

GROUP	SEQ ID %	#ALNs
1	< =18	991
2	19 – 30	1229
3	31 – 49	675
4	**>= 50**	**115**

Table 2.3: ‘Suboptimal’ alignments obtained via HHSEARCH. The number of alignments used to create the sub dataset of high-quality raw models is highlighted in red.

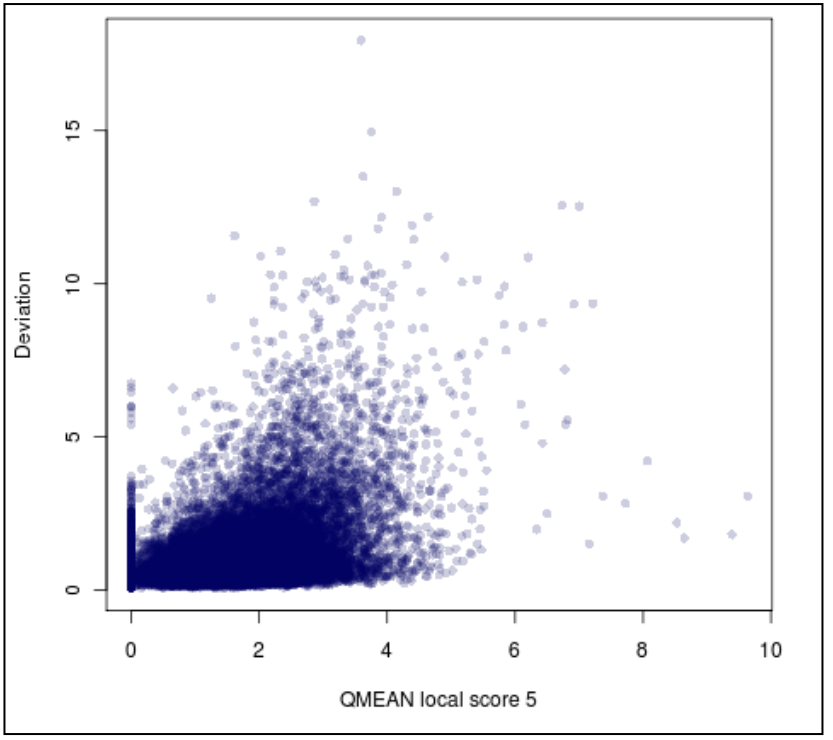


Figure 2.11: Correlation between Q5LS and Cα deviations of raw models built from the high quality model dataset.

A correlation, higher than the one achieved in 2.5 (0.43 vs 0.20), is obtained between the Cα deviations and Q5LS. However, such value was still too low to demonstrate that Q5LS can detect errors in the alignments.

Notwithstanding the high sequence identity between target and template, there are regions of the raw models wrongly modeled by our automatic procedure. In fact, the corresponding Cα deviations calculated with respect to the native structure resulted higher than expected (5Å). Thus, to reduce the noise caused by such modeling errors, we performed the correlation analysis twice: first by cutting of all those residues showing a Cα deviation strictly higher than 5 Å and then lowering such threshold to 2.5 Å (fig 2.3, 2.4).

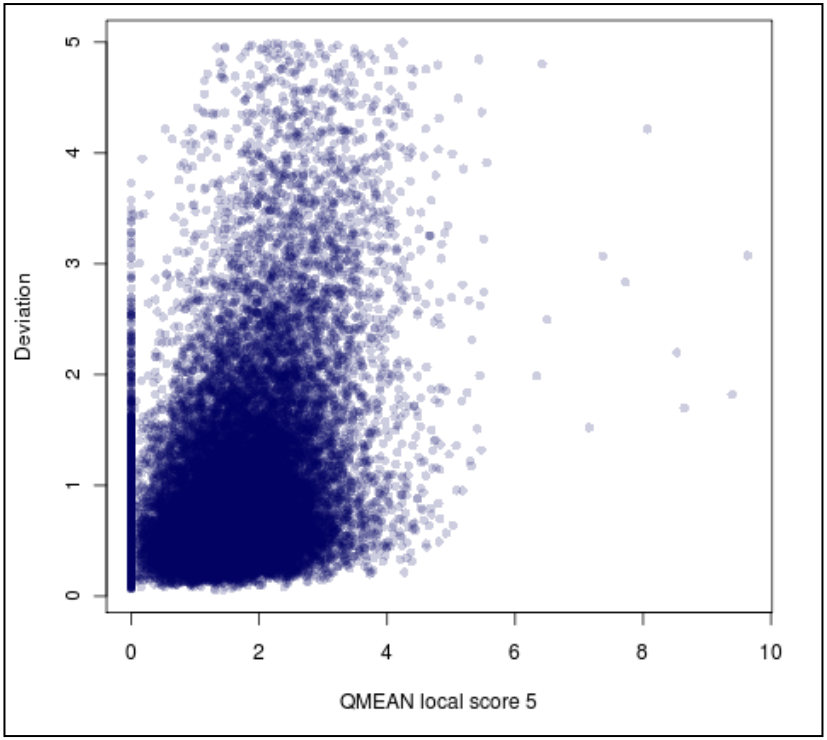


Figure 2.12: Correlation between Q5LS and Cα deviations (threshold 5 Å) of raw models built from the high quality model dataset.

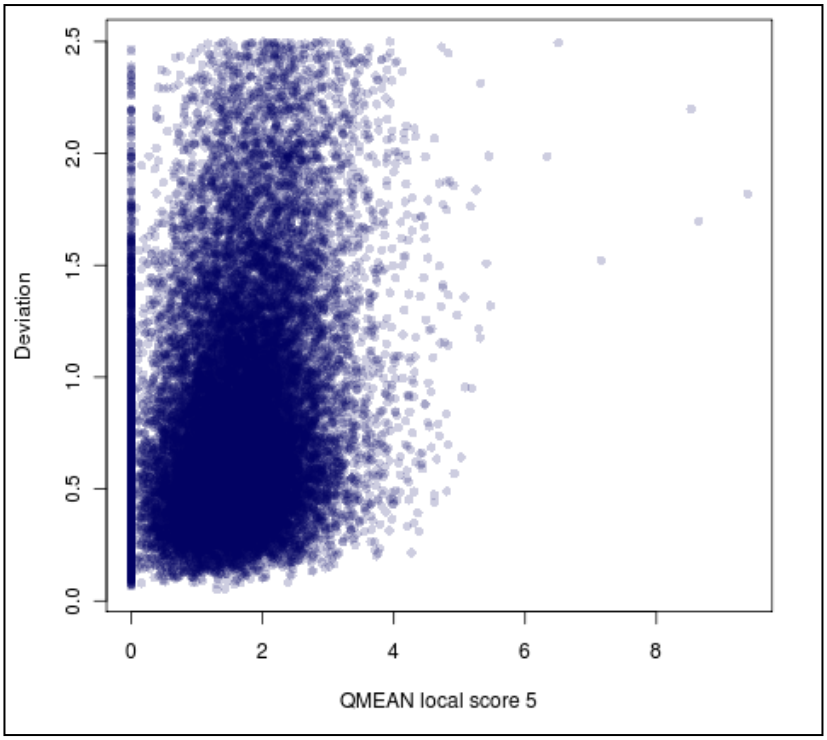


Figure 2.13: Correlation between Q5LS and Cα deviations (threshold 2.5 Å)of raw models built from the high quality model dataset.

We obtained the following correlation values: 0.41 for the 5Å and 0.35 for the 2.5Å thresholds. They highlight a poor local correlation between Cα deviations and Q5LS. In fact, the smaller the Cα deviations, the lower their

correlation with Q5LS. These results suggest a limited ability of Q5LS to reflect fine errors in the raw models.

2.5.6 - Grouping Cα deviations

The above analysis indicated that Q5LS is not able to consistently predict the Cα deviations of the models. However, for alignment refinement it would be sufficient to predict if the residue is aligned correctly (Cα deviation below 1-2 Å) or incorrectly (Cα deviation above 1-2 Å). Thus, we verified the correlation between Q5LS and Cα deviations grouped by into three bins (Table 2.4). First, the raw models, built from the 'optimal' alignment subdataset described in 2.5.2, were used to calculate their Cα deviations from the corresponding native structures.

GROUP NAME	Cα DEVIATION BIN (Å)	# OF RESIDUES
VERY GOOD	0 – 1	5281
GOOD	1 – 2	1023
BAD	2 – 4	320

Table 2.4: The result of the Cα deviation values grouping. The number of alignment residues in the dataset for each bin is shown in the table.

Then, we made a boxplot analysis to verify the correlation of the Q5LS predicted for the raw models built from the 'suboptimal' alignments and the binned Cα deviation. Results are shown in fig 2.14.

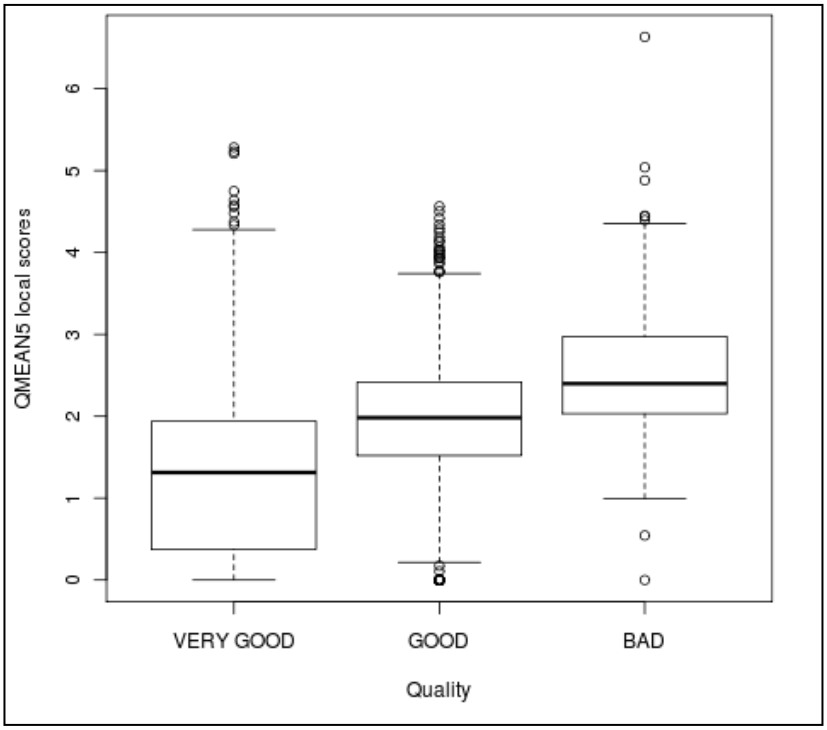


Figure 2.14: Box plot analysis to evaluate correlation of Q5LS with Cα bin obtained.

Clearly, the QL5S scores correlate with the deviation bins, and this we implemented a method to predict the binds based on local raw models quality and then we assessed its performance.

2.5.7 - A predictor based on boxplot analysis

The predictor was based on the distribution of the relative frequencies of the Cα deviations contained into the three bins (fig. 2.15-2.17).

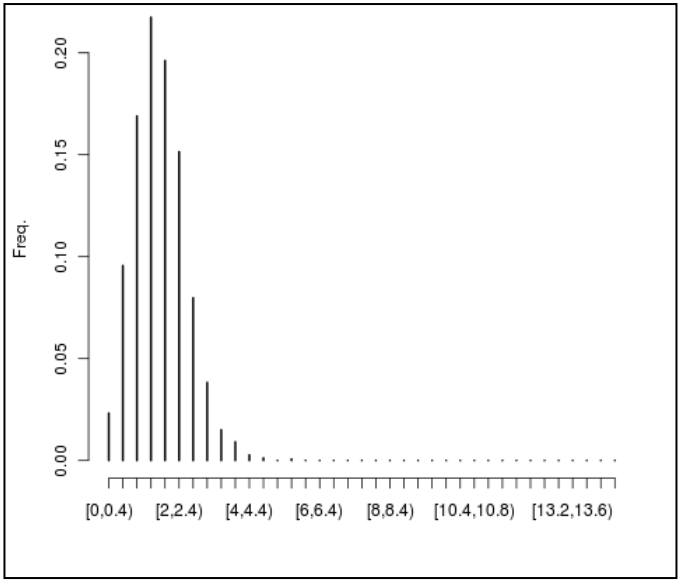


Figure 2.15: Distribution of the relative frequencies of the Cα deviations in the VERY GOOD bin. Notice that in this bin the values equal to zero are omitted in order to remove a bias in the distribution.

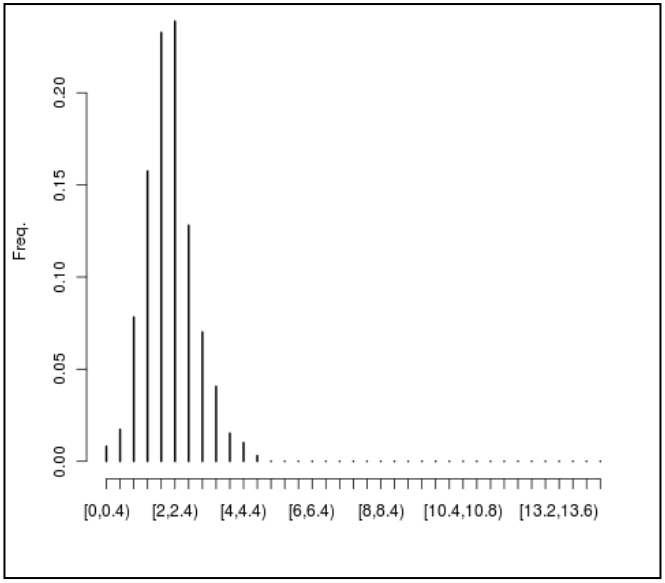


Figure 2.16: Distribution of the relative frequencies of the Cα deviations in the 'GOOD bin.'

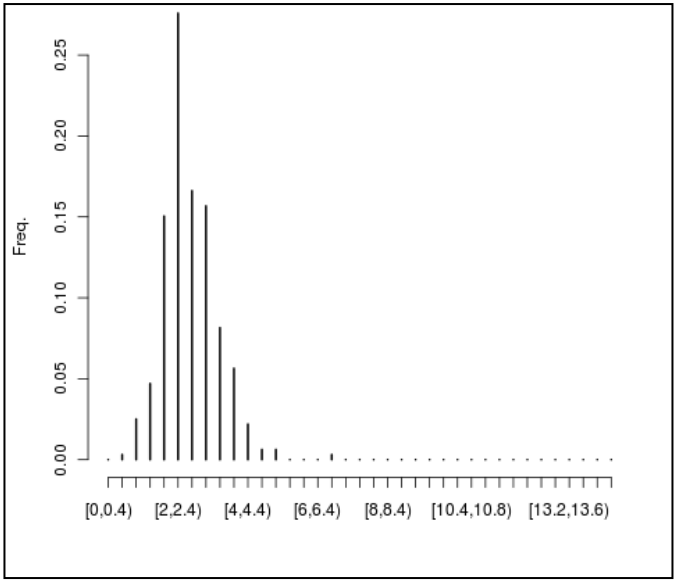


Figure 2.17: Distribution of the relative frequencies of the Cα deviations in the 'BAD' bin.

From the distribution shown in figures 2.15, 2.16 and 2.17, the predictor calculates the probability *P* of a residue with Q5LS equal to *x* to belong to the group G_n according to the formula 2.1:

$$P(x \in G_n) = \frac{Rel.Freq\ x\ in\ G_n}{\sum_{i=1}^{3} Rel.Freq\ x\ in\ G_i} * 100$$

Where:

$$G_1 = 'VERY\ GOOD' bin$$

$$G_2 = 'GOOD' bin$$

$$G_3 = 'BAD' bin$$

Formula 2.1

Moreover, relying on the observation of the distributions on figures 2.6-2.8, Q5LS equal to zero were assigned to have a probability to belong to the VERY GOOD bin (pVGOOD) equal to 100%. Moreover, the Q5LS higher than 5 were considered to have a probability to belong to BAD bin (pBAD) equal to 100%.

The predictor evaluates if a residue is an error relying on the ratio between its pBAD and its pVGOOD. The final prediction does not consider the probability of a given Q5LS to belong to the GOOD group (pGOOD). In fact,

as described for the example 2.5.9, pGOOD did not seem to enhance the quality of the final prediction.

2.5.8 - Testing the predictor

Finally, the predictor was tested on the sub-dataset composed by the 115 raw models built from the 'suboptimal' alignments described in 2.5.3. In order to find the threshold that minimizes the number of false positives, we tried several pBAD/pVGOOD thresholds. To evaluate the quality of the predictions we created the confusion matrices assuming the following definitions:

- TP (True Positive): the alignment is incorrect and the Q5LS is equal or higher than the pBAD/pVGOOD threshold (i.e. Q5LS the predicts alignments also as incorrect);
- FP (False Positive): the alignment is incorrect and the Q5LS is lower than the threshold;
- TN (True Negative): the alignment is correct and the Q5LS is lower than the threshold;
- FN (False Negative): the alignment is correct and the Q5LS is equal or higher than the threshold.

From the confusion matrices we calculated (tables 2.5) accuracy, specificity, sensitivity and Matthews correlation (refer to 3.3.1). The evaluation of the predictor performances shows that the Q5LS cannot efficiently detect local errors of the raw models.

Thresholds	Matthews corr.	Accuracy	Sensitivity	Specificity
0.2	0.00	0.96	0.01	0.99
0.4	-0.01	0.94	0.01	0.94
0.6	0.11	0.35	0.33	0.94
0.8	0.10	0.55	0.54	0.70
1.0	0.10	0.76	0.46	0.77
1.2	0.10	0.88	0.27	0.90
1.4	0.05	0.92	0.11	0.95
Always FALSE	0.00	0.96	0.00	1.00

Table 2.5: Confusion tables thresholds from 0.2 to 1.4. In addition also the ‘always false’ predictor was tested. In the right part: accuracy, sensitivity, specificity and Matthews correlation values are shown.

2.5.9 - EXAMPLE – Local error detection with QMEAN5

An example of successful prediction is demonstrated on the example of the alignment between 1amf_A and 3gzg_B. The ‘suboptimal’ alignment obtained via HHSearch presents only one shift with respect to the ‘optimal’ one derived from the structural superposition (figure 2.18).

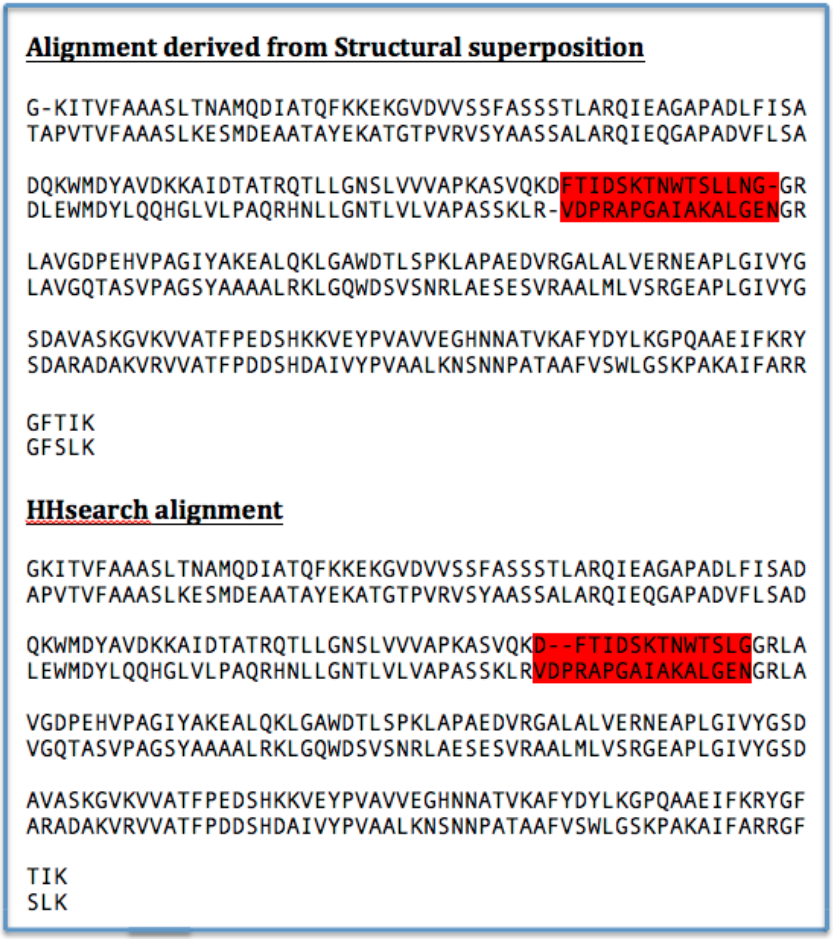
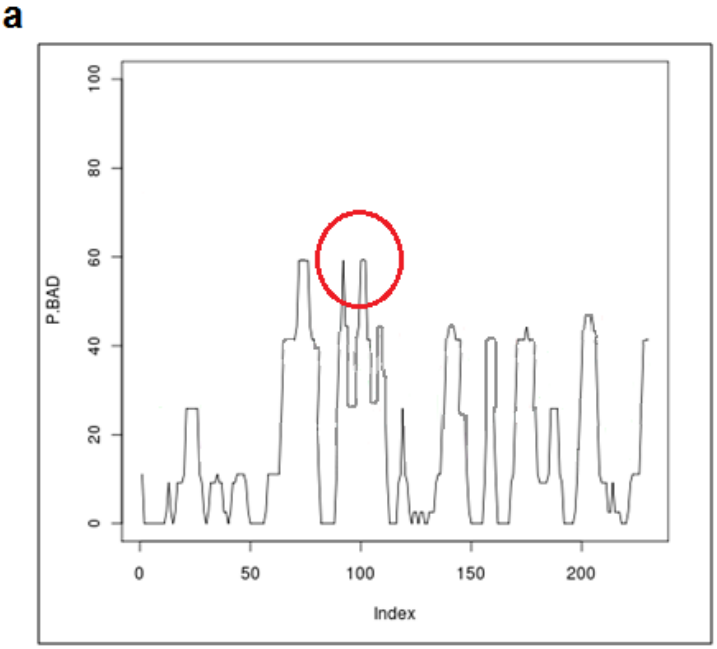


Figure 2.18: Sequence alignments for 1amf_A (target) and 3gzg_B (template). The region that differs from the optimal alignment is highlighted in red.

The predictor, using the re-parametrized QMEAN5, clearly detects the region containing the shift; as evident from figure 2.19 and 2.20.



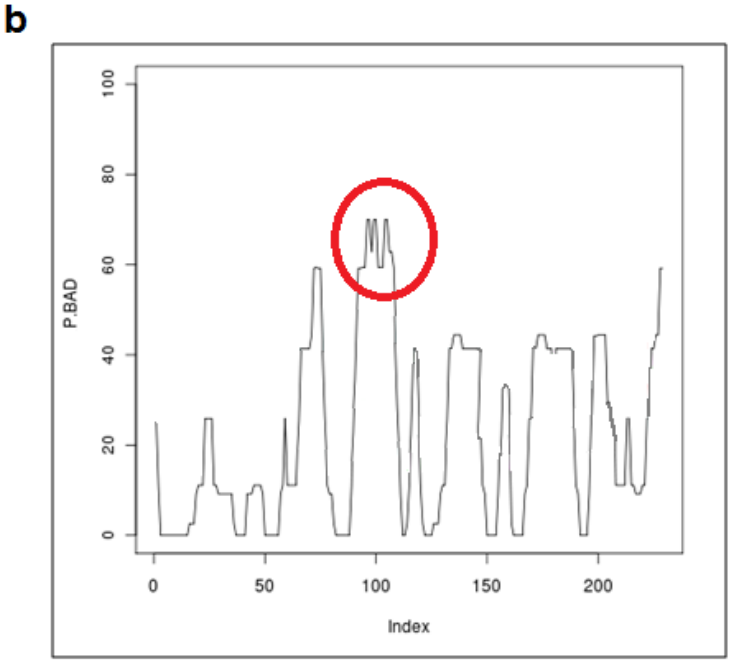


Figure 2.19: Plot of the pBAD for the 1amf_A-3gzg_B alignments (a. 'optimal', b. 'suboptimal'). The red circles highlight the region with the highest pBAD in both cases.

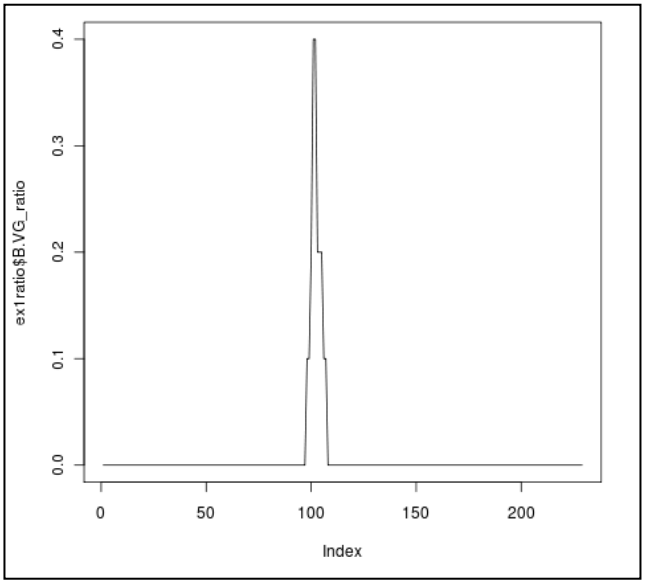


Figure 2.20: Plot of the ratio pBAD/pVGOOD for the 1amf_A-3gzg_B suboptimal alignment. The error in the region nearby the amino acid with index 100 is clearly highlighted by the peak. In this case a threshold of the ratio equal to 0.4 is optimal to detect the local error.

We analyzed the two alignments (fig 2.21) with MODalign and the QMEAN local scores accurately pointed out the errors in the ‘suboptimal’ alignment.

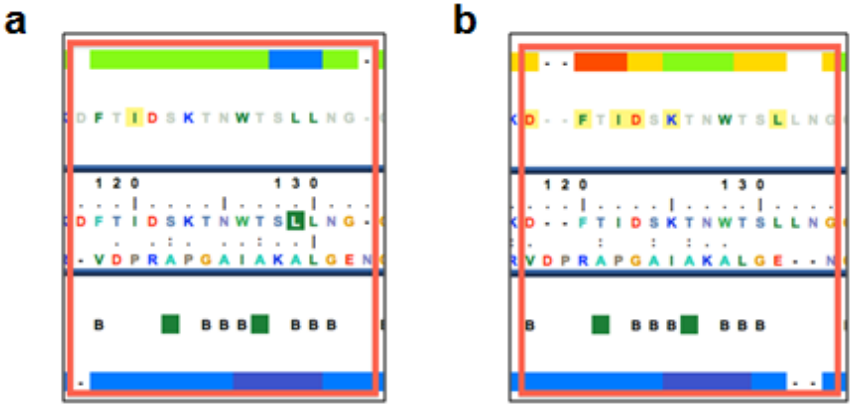


Figure 2.21: a. and b.: The ‘optimal’ and ‘suboptimal’ alignments displayed in MODalign. By comparing the regions zoomed in the red frame it is evident that with QMEAN local scores are able to point out misaligned regions.

The low score of the region ranging from residues 117 to 134 is due to the misalignment of five residues, which is indeed inconsistent with their solvent exposure (Table 2.6). This situation is easily detectable using the MODalign interface.

Suboptimal Alignment			Optimal Alignment		
Target res	**Solvent accessibility**	**Templ res**	**Target res**	**Solvent accessibility**	**Templ res**
D	Buried	V	D	-	-
F	Exposed	R	F	Buried	V
I	Exposed	P	I	Exposed	P
D	Buried	G	D	Buried	R
K	Buried	I	K	Buried	P
L	Exposed	E	L	Exposed	L

Table 2.6: The potential errors, according to MODalign in both 'optimal' and 'suboptimal' alignments, are highlighted in red. The low QMEAN local scores are due to bad placement of charged aminoacids in buried regions of the protein or polar amino acid placed in the exposed.

3 - METHODS

3.1 - MODORAMA: embedded software and databases

The MSA of the target protein family is constructed using two iterations of HHblits [19] with default parameters (which can be modified using the input page), and using the specially formatted non-redundant UniProt database [3] available from ftp://toolkit.lmb.uni-muenchen.de/HH-suite/databases/hhsuite_dbs/.

MODexplorer searches the potential homologs with known structure in the PDB (PDB filtered at 70% sequence identity downloaded from ftp://toolkit.lmb.uni-muenchen.de/HH-suite/databases/hhsearch_dbs/) using as query the MSA constructed as described above, through HHSearch version 2.015 [49] with default parameters (which include -realign, -mact 0.3, -sc 5). Subsequently MODexplorer, for every HHSearch hit, retrieves the related PDB chains from MMBD non-redundant PDB chain set (ftp://ftp.ncbi.nih.gov/mmdb/nrtable/) using 10e-80 redundancy level.

In both MODexplorer and MODalign, PDB annotations such as full SEQRES sequences, experimental method or crystallographic resolution are retrieved

from a local PDB database implemented in MySQL using a custom Python library, LocalPDB (http://biocomputing.it/LocalPDB), and utilizing DB-LOADER (http://sw-tools.pdb.org/apps/DB-LOADER/source/source.html) for creating the database. Then, for every HHSearch hit or related PDB chain, the corresponding MSAs are obtained from the HHSearch alignment database weekly downloaded from ftp://toolkit.lmb.uni-muenchen.de/HH-suite/databases/hhsearch_dbs/. If for a given PDB chain the MSA is not present in this database (it contains MSAs only for representative PDB chains), the PDB chain sequence is aligned to the MSA of the related representative chain using HHAlign from the HHSuite.

Secondary structure and solvent accessibility for the target are predicted using PSI-PRED [90] and ACCpro [91], respectively. The values for the templates are calculated using DSSP [92] and POPS version 1.5.3 [93].

In MODexplorer, the ligand binding sites are retrieved from a local database of ligand binding sites extracted every week from the PDBeMotif database [94]. DNA and RNA binding sites are retrieved from a local copy of NPIDB [95]. We have selected these databases, among the other available, since they are: 1) regularly updated, 2) comprehensive (e.g. including all chains from PDB and not only the representative ones, 3) downloadable as a local copy.

Then, disordered regions for the query are predicted using DISOPRED [96] using the HHblits MSA as input. The B-factor values and information about missing residues for the hits and are derived from the original PDB records. In the structure input mode, if the PDB code is provided, B-factor values and information about missing residues are retrieved from the PDB record, while if PDB file is provided, only B-factors are retrieved.

The sequence similarity scores displayed in MODexplorer are extracted from the HHSearch output. These values are displayed as a color gradient mapped on the hit panel. Both in MODexplorer and MODalign, the MSAs are colored with a background shading that reflects the conservation of a given amino acid in either the whole MSA or the single family. In the similarity row of MODalign, which shows similarity of the target to template, residues are indicated as identical ('|'), highly similar (':'), similar ('.') or not similar (' ') and is derived based on Blosum62 matrix, while in MODexplorer the similarity row is a copy of a similarity row from HHSearch output.

In MODExplorer, the displayed global and local QMEAN [63] scores are calculated based on models constructed from HHSearch alignments relying on the QMEAN implemented in the OpenStructure library [97]. For performance reasons, in MODexplorer only the scores for up to 30 HHSearch hits are automatically calculated. However, users can calculate

QMEAN scores for the remaining hits using the graphical interface. In MODalign, the global and local QMEAN scores are computed and then displayed for the model, template and representative homologs implied by the current alignment (without modeling the insertions). In both MODexplorer and MODalign, models are built using Modeller [61] in 'very_fast' modeling mode with default options. However, within the interface of both MODexplorer and MODalign, users can additionally choose to avoid or not the additional model optimization, the modeling of insertions and tails, and the modeling of the regions aligned with residues missing in templates as chain breaks rather than as insertions.

The structure superposition of models with their templates is performed using Theseus based on the corresponding target-template alignment. Finally, in MODexplorer the structure superposition of two PDB chains selected within the interface are generated using SALIGN [98] from the Modeller Python library with default options and all feature weights set to 1.

3.2 - MODORAMA: implementation details

The server is built in Python (using the Django web framework), HTML and JavaScript (using ExtJS library). Apache Solr and Django Haystack are used for filtering the PDB chains and HHSearch hits by annotations. Django Celery, Celery and RabbitMQ are used for managing job tasks. APE Server is used for Comet client-server communication. For manipulation of biological data, the server utilizes PyCogent [99] and to lesser extent BioPython [100] and Modeller Python library.

3.3 - QMEAN local scores for alignment errors detection

The initial non-redundant list of 1657 monomeric target structures has been created by using PISA [101] that is an interactive tool allowing a) the exploration of macromolecular interfaces, b) prediction of probable quaternary structures and database searches of structurally similar interfaces and assemblies, and c) searches based on various assembly and PDB entry parameters. The monomeric PDB structures derived from PISA have been later filtered using the PISCES server [102]. PISCES enabled the

removing those proteins which had: a) X-ray resolution higher than 2A, b) more than 25% of sequence identity with another sequence in the dataset, c) an amino acidic sequence shorter than 150 residues.

The homologs with known structure (templates) used for building of the optimal alignment dataset were retrieved using HHSearch version 2.015 with a local version of the PDB database. The HHSearch output filtering, described in the Results and aimed at reducing the redundancy, was performed using the Openstructure framework. For each target, the homologs which passed the filter were grouped in six bins based on their sequence identity: the bins were <10%, 10%-20%, 20%-30%, 30%-40%, 40%-50% and 50%-60%. Then, for ease of data manipulation we considered, for each bin, only the five best ranking structures (when available) accordingly to their X-ray resolutions and we obtained 4,086 target-template pairs. The TMalign software [102], embedded in Openstructure, was used to superpose all identified target-template pairs. The sequence alignments corresponding to the structural superpositions were extracted in order to create the optimal alignment dataset. To generate the suboptimal sequence alignments for the 4,086 target-template pairs, we used HHSearch version 2.015, CLUSTALW version 2.1, and PSI-BLAST

and BLAST released in the version 2.2.25 of the blast+ executables (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.25/).
Finally, all the statistical analyses, such as boxplot and correlation analyses were performed using the R software (http://www.r-project.org/).

3.3.1 - Predictor performance measures

While describing the measures used to evaluate the performance of the aforementioned predictor, the following annotation will be used: TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

3.3.1.1 - Sensitivity

Sensitivity describes the ability of the predictor to identify true positives.

$$Sensitivity = \frac{TP}{TP + FN}$$

3.3.1.2 - Specificity

Specificity describes the ability of the predictor to identify true negatives.

$$Specificity = \frac{TN}{TN + FP}$$

3.3.1.3 - Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) in the sample.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3.3.1.4 - Matthews correlation

The Matthews correlation coefficient (MCC) is used as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure, which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the MCC is generally regarded as being one of the best such measures.

$$\text{MCC} = \frac{(TP x TN) - (FP x FN)}{\sqrt{(TP+FP) x (TP+FN) x (TN+FP) x (TN+FN)}}$$

4 - CONCLUSIONS AND OUTLOOK

Homology modeling has demonstrated to be useful in several applications because of its reliability. In addition, the number of targets suitable for this modeling technique is larger than it was in the past thanks to the development of increasingly more accurate methods for remote homology detection and thanks to the increase of the number of solved structures that could be used as templates. An indication of the difficulties that might be encountered during a homology modeling procedure can be obtained by sequence identity between the target and template, since it reflects their evolutionary distance. In fact, the automatic homology modeling methodologies generally provide accurate predictions when they can identify templates sharing high sequence similarity with the target (more than 50%). Otherwise, if the aim is to model a target based on templates having lower sequence similarity, the manual intervention of the modeler is mandatory to achieve reliable structural models. In fact, both the inaccuracies in target-template alignment and the selection of the most suitable templates, cannot be usually handled by the automatic procedures.

The examples in the “Results” chapter demonstrate that MODORAMA represents a homology modeling platform of great help to the users who

need to manually customize the modeling procedure according to their biological expertise. In particular, MODexplorer allows the “knowledge based” approach to template selection (previously described in 1.2.2), and MODalign enables the inspection and refinement of the target-template alignment, facilitating the manual procedure used by the expert modelers (described in 1.2.3). The example 2.6 is a remarkable demonstration of how an integrated usage of MODexplorer and MODalign can lead to building a homology model that is more likely to be similar to the hypothetical crystallographic structure. Specifically, although a structural model of the PMS2-CTD protein was already available, we first, succeeded in identifying a “better” template structure with MODexplorer, and later with MODalign, we placed the indels within those regions, of the target protein, which are more indicated to “accommodate” them.

Moreover, MODexplorer is a useful resource also when the target structure is known. In fact, MODexplorer is extremely useful for exploring the sequence, structural and functional diversity between, and within, the families of the retrieved homologous proteins. In particular, the example 2.5 shows how, by investigating the protein family of the MutS prokaryotic complex, we inferred the hypothetical structural identity between the ADP-

and ATP-bound conformations of the MSH6 monomer from MutSα complex, a key component of the human DNA mismatch repair pathway.

As largely discussed in the introduction, homology modeling is widely used within structural genomics projects. However, because of the huge amount of data to manage and in order to be sure to have high quality predictions, only those proteins, that can be automatically modeled, are considered for such large-scale modeling studies. However, it is known that, assuming a correct template selection, the inaccuracies in target-template alignment are the main source of errors when modeling a target on a template with low sequence similarity. Thus, an automatic procedure that emulates the modeler approach for target-template alignment refinement (see 1.2.3) would greatly help the overcoming of the afore-mentioned scope limitation of automatic homology modeling methodologies. Thus, the study described in 2.3 aimed at understanding whether the intuitive approach of the human modeler to the target-template alignment refinement could be automatized. In collaboration with Prof. Torsten Schwede and Dr. Pascal Benkert we developed and tested a re-parameterized version of their QMEAN and later we assessed its feasibility in discriminating the erroneous regions of a target-template alignment. We observed cases in which QMEAN local scores lead to the detection of misaligned regions (see example 2.4.9). However, the

average accuracy of QMEAN in accomplishing such task appeared to be poor since the current implementation of the QMEAN scoring function is not accurate enough to make the local scores unambiguously descriptive of the misaligned regions. Additionally we observed that the analysis of the QMEAN scores of the target, combined with other information such as target-template sequence similarity as well as match of secondary structure, residue accessibility and QMEAN scores of the template families, often enables to foresee whether the alignment is effectively correct (see example 2.5.9). In fact, such successful “integrated approach” is the basis of the MODalign editor interface.

In conclusion, the study described in this manuscript highlights that MODORAMA is a platform that undoubtedly facilitates the manual intervention of the modeler in those steps that are known source of errors in the models built on low sequence similarity templates.

In the future we aim at rendering such process completely automatic by reducing the manual intervention at minimum. In particular, the automation of the manual target-template alignment refinement, as it is performed by the homology modeling experts, is far to be reached and it represents a challenge of great importance for the structural biology.

5 - ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Prof. Anna Tramontano. I could not have asked for better role models, each inspirational, supportive, and patient. I could not be prouder of my academic roots and hope that I can in turn pass on the research values and the dreams that she has given to me.

I would also like to thank Jan Kosinski, who supported me during the whole PhD and taught me about homology modeling, DJANGO, ExtJS and computational programming.

My PhD has been founded by a fellowship paid on the finds of a KAUST awards to Anna Tramontano so I would like to thank the KAUST for its generous support. As a member of Biocomputing group I have been surrounded by wonderful colleagues; such community has provided a rich and fertile environment to study and explore new ideas. At Biocomputing, in particular I would like to thank, Dr. Paolo Marcatili, Dr. Allegra Via and Dr. Domenico Raimondo, who have both been extremely supportive in allowing me to participate in lab activities whilst pursuing my PhD studies.

To Prof. Torsten Schwede I am grateful for the chance to visit and be a part of his lab at the Swiss Institute of Bioinformatics at University of Basel

(Switzerland). Thank you for welcoming me as a friend and helping to develop the ideas in this thesis.

I would not have contemplated this road if not for my parents, Pino and Fausta, who instilled within me a love of creative pursuits, science and language, all of which finds a place in this thesis. To my parents, thank you.

This thesis would also not be possible without the love and support of Valentina. I love you.

6 - APPENDIX A - MODexplorer description

6.1 - Input computation

After the input submission, MODexplorer computes the output through the following steps (detailed description of software, databases and parameters is provided in the Methods):

1. If the user does not provide its custom multiple sequence alignment (MSA), a MSA of the input protein (query) family is created with HHblits. This step can be optionally skipped. However, it is known that comparing the MSAs of the target and template families results in an alignment more accurate than the one obtained using just target and template sequences. In fact, with MSAs it is easier to spot conserved patterns in the families, which in turn, help to find the best alignment.
2. The homologs with known structure (templates) of the input protein are retrieved by starting from the MSA of the query family (or its single sequence) using HHSearch. The template search is made against a PDB database filtered at 70% of sequence identity.

3. For each template, the related PDB chains are retrieved based on MMDB non-redundant chain set and aligned to the corresponding template. This enables accessing structures which are filtered out by HHSearch but that can be informative. For example, they might contain different ligands, represent alternative conformational states or exhibit high structural quality.

If we do not consider the time eventually spent in queue, this pipeline usually takes from 5 to 25 minutes, depending on the length of the query protein and the number of homologs.

The output from the computation above can be opened in a graphical interface that:

- Displays the sequence alignments to each template and their related PDB chains. They are shown both as a BLAST-like bar diagram and as detailed MSAs.
- Graphically highlights annotations on the alignments. The annotations include ligand and DNA/RNA binding sites, secondary structure (predicted for the query, calculated for the templates), disorder information, HHSearch similarity scores and QMEAN scores of models built from the alignments.

- Allows for filtering of the templates based on the presence in their structures of nucleic acids or other ligands, the HHSearch similarity scores, the experimental techniques used to solve the structures and the crystallographic resolution.
- Enables modeling the structure of the query protein via Modeller, based on any selected alignment. The quality of the models is then evaluated via QMEAN.
- Provides visualization of structural superposition of PDB chains based on pairwise alignment inferred from the alignment to the query protein. This allows a user to verify if a low scoring template in the aligned region has a structure similar to the one of an higher scoring template.
- Allows the alignment editing and inspection via MODalign.

6.1.1 - Description of the web interface

6.1.1.1 - Input form

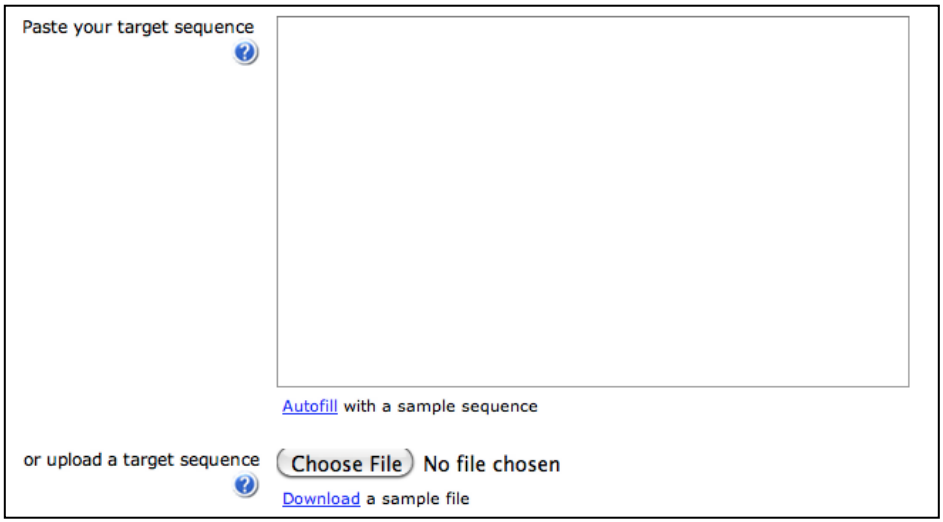
The protein input form is composed by the following fields:

- ‘Choose input type’: It allows selection between sequence and structure input mode

If the ‘Sequence’ input mode is selected the following fields will be displayed:

- ‘Paste your target sequence’: It is a textbox where the input sequence can be pasted. Both FASTA formatted sequences and strings without header are accepted. The ‘Autofill with a sample sequence’ link below the textbox, allows the filling of the field with a correctly formatted input sequence.
- ‘or upload a target sequence’: It allows the uploading of a FASTA formatted file containing only one sequence. The ‘Download a

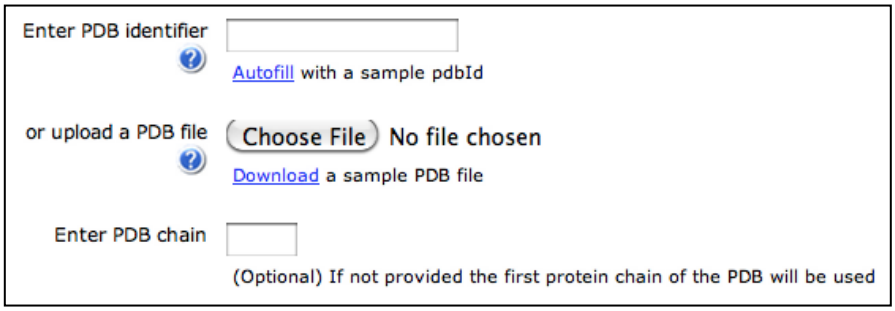
sample file' link below the file-uploading field, allows the download of a FASTA file containing the PMS2-CTD sequence.



Otherwise, if the 'Structure' input mode is selected the following fields appear:

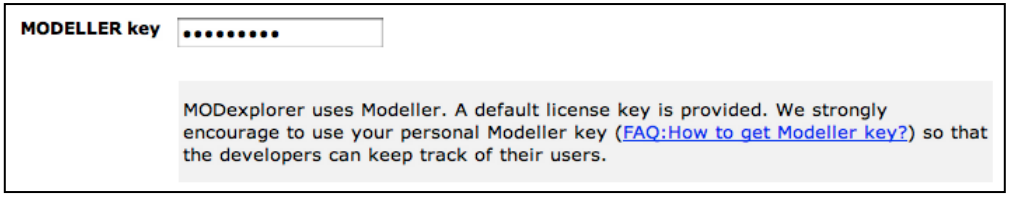
- 'Enter PDB identifier': it allows the specification of the PDB code of the input protein (ex. 1amf). 'Autofill with a sample id' link fills the field with a sample PDB code.

- ‘or upload a PDB file’: it allows to upload a PDB file which contains the atomic coordinates of the input protein structures. ‘Download a sample PDB file’ link allows downloading a sample PDB file.
- ‘Enter PDB chain’: The chain of interest of the input protein may be specified here. The filling of the field is optional, if the chain is not provided, the first protein chain present in the structure will be used.



Both sequence and structure input modes have in common all the following fields:

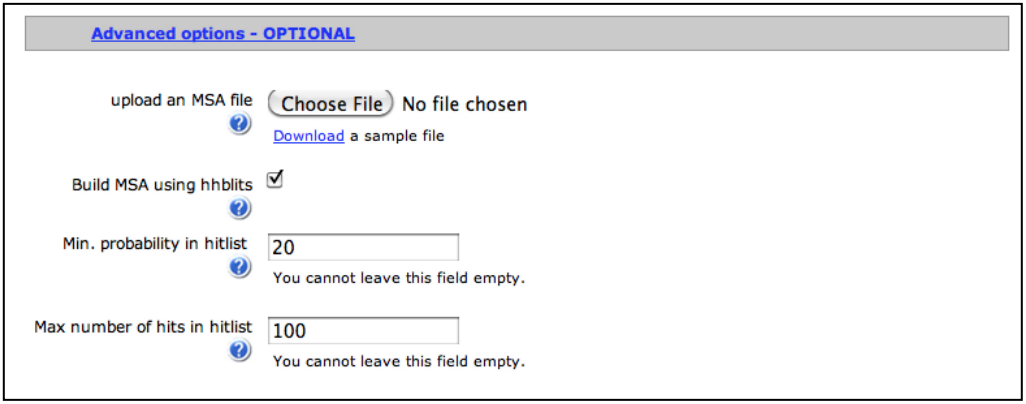
- ‘MODELLER key’: MODexplorer uses MODELLER software for several tasks. Thus, it is not possible to submit a job without input a valid MODELLER license key. Although it is provided by default, we strongly encourage using a personal MODELLER license key, so that the MODELLER developers can keep track of their users.



The advanced options section of the form is collapsed by default. When expanded, it displays the following fields:

- 'Upload an MSA file': It allows the uploading of a file containing a custom MSA of the target protein. *FASTA, CLUSTAL, PHYLIP and MSF formats are accepted.* An example alignment of the PMS2-CTD family is downloadable by clicking the 'Download a sample file' link.
- 'Build MSA using hhblits': If unchecked, MODexplorer will not build the MSA of the query protein family and the template search will be made with the query sequence alone.
- 'Min. probability in hitlist': It allows defining a threshold value below which the homologs retrieved by HHSearch will not be considered. It must be a number between 0 and 100.
- 'Max number of hits in hitlist': It allows setting the maximum number of the templates in the HHSearch output interface. It must be a number ranging from 1 and 500.

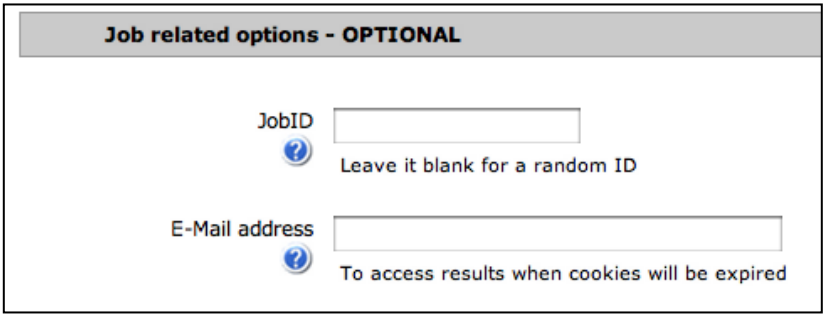
The last two options are useful to limit the size of the output especially when the query is a protein that have many homologs because of the presence of highly conserved domains.



Finally, in the form there are the ‘Job related options’:

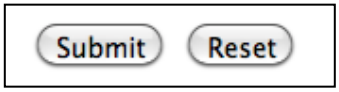
- ‘JobID’: by filling this field a user can specify a job name, every character but the ‘/’ (front-slash) is allowed. The job name specification is optional = a random set of characters will be used if the field is left empty.
- ‘E-Mail address’: It allows for the specification of an e-mail address to which a message will be sent when the computation is complete. The

e-mail will also contain a link from which it will be possible to access the workspace.



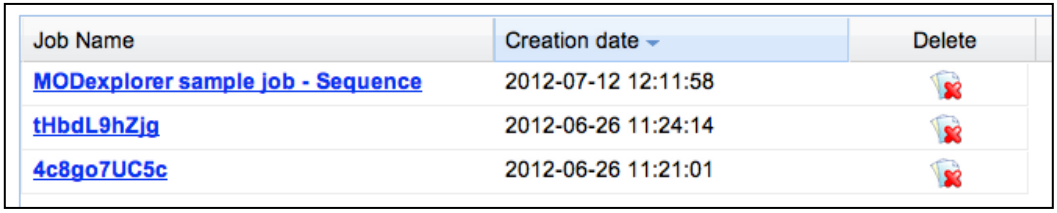
Finally, in order to facilitate the usability of the resource, the possibility to automatically fill the form with default values is offered as well.

Once the protein input form is filled, it can be either submitted for the computation or cleared for filling it again by clicking one of the two buttons in the figure.



6.1.1.2 - Recent job section

Each submitted job will be displayed for 30 days after the last access to its workspace in a table in MODexplorer home page. From the table a user can retrieve the name of the job and the submission date. Moreover, the job can be deleted by clicking the icon in the column ‘Delete’.



6.1.1.3 - The workspace

If the window of the browser is kept opened, just after the submission, the user is automatically redirected to his workspace. It can be accessed in several ways:

- By using the links in the e-mail (only if provided) sent to the address specified during the submission.
- By cutting and pasting the job id (retrievable from the computation log) within the 'Retrieve an old job by ID' field in the MODexplorer home page.

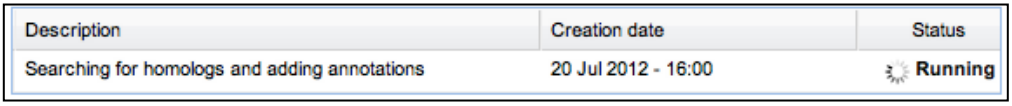
- Finally by entering the url in a web browser: http://modorama.biocomputing.it/modexplorer/workspace/?folder=<job_id>

Just after the submission and while the computation is ongoing, in the workspace a user can:

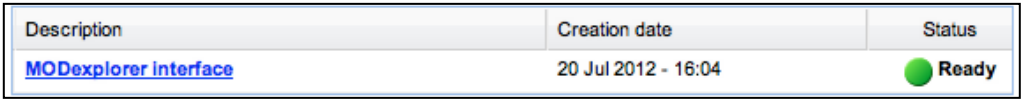
- Monitor the status of the submitted job:

    a. ‘Queued’ - if there are several other jobs already running on the server and this job is waiting in the queue;
    b. ‘Running’ - if the computation is ongoing;
    c. ‘Error’ - if during the computation an error arises and makes the whole process crash;

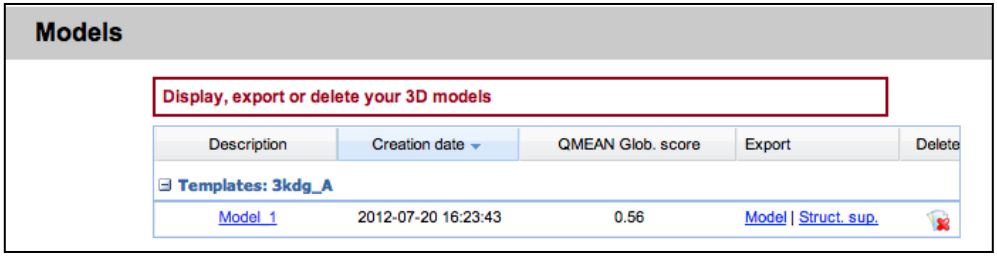
- Check the computation log. This information is automatically displayed if the job status is running. However, in the other cases the user has to click on ‘Display log’ button;



Then, when the computation is complete, and when errors did not occurred, from the workspace it is possible to enter the MODexplorer results page by clicking on the ‘MODexplorer interface’ link.



Then, if a user builds models in the MODexplorer interface they will be displayed also in the workspace.



Additionally, since MODexplorer is integrated with MODalign, if there are edited target-template(s) alignments referring to the same MODexplorer job, these are displayed in the same workspace and links to the relative MODalign sessions are listed as well.

6.1.2 - Results display

MODexplorer interface mainly consists of three panels: i) the top one containing the filtering options, ii) the central one that displays the annotation and the structures retrieved, iii) the bottom one displaying sequence alignments.

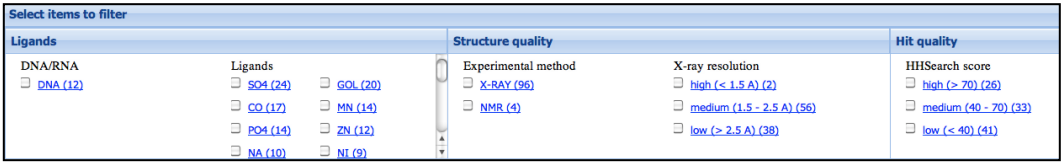
6.1.2.1 - Top panel: filtering options

The filtering of the homologous structures by different criteria:

- 'Ligands': DNA/RNA and 'other' ligands such as ions, inorganic and organic compounds;
- 'Structure quality': experimental method and X-ray resolution;

- ‘Hit quality’: it allows filtering by HHSearch score, a value reflecting the probability of a given hit to be homologous to the query protein.

The top panel can be collapsed and expanded to maximize the ease of usability of the interface.



6.1.2.2 - Central panel: annotation

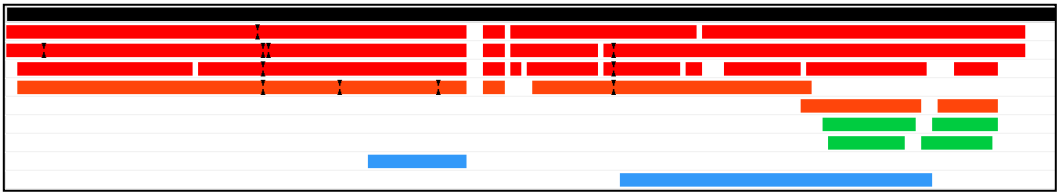
This panel is the core of the MODexplorer interface and contains the list and all the annotation of the homologs (or hits) of the query protein.

In the upper part of this panel, the user can decide the annotation that will be displayed on the hit bars by clicking on specific buttons.

The hit bars will assume a coloring scheme relative to the user choice:

- ‘Score’: HHScore ranges from 100 (good) to 0 (bad) and it is represented on the hit bars as a color gradient from red to blue;
- ‘Similarity’: Shows per-residue HHSearch scores between each hit and the query. The scores are provided by HHSearch (HHSearch column score) and it ranges from greater than 1.5 to lower than -0.5. The color gradient is from deep blue for highly similar residues to light cyan for dissimilar ones;
- ‘Secondary structure’: This annotation appears as colors on the gray hit bars. α-helices are depicted as red squares whereas β-strands as green ones;
- ‘Ligands’: When ‘ligands’ annotation is selected, a new frame becomes available: it contains all the compounds (with exception of nucleic acids) bound to each retrieved hit that is present in the interface. In this way one can arbitrary restrict to a particular molecule (or set of molecules) the ligands that are displayed on the hit bars. The residues interacting with the ligand are highlighted in different colors corresponding to the bound ligand.
- ‘DNA/RNA’: Nucleic acid binding sites are highlighted in red;

- ‘QMEAN Global’: It is a value estimating the global quality of a model of a query built based on the corresponding alignment. The gradient goes from deep blue for good models to red for the poor models;
- ‘QMEAN Local’: This is a per-residue score reflecting how a given amino acid is properly ‘placed’ in the model. The color gradient is the same used for QMEAN Global: from deep blue (very good scores) to red (very bad scores);



6.1.2.3 - Central panel: hit list

The names of the retrieved homologs are listed in the left part of the central panel. They are in the same row of the corresponding hit bar and their names are in the format <pdb code>_<chain> (ex. 1nq9_A). By clicking on the hit name its relative PDB entry is opened in a new tab of the browser.

A star picture is always present just after the hit names: by clicking on it one can make that hit as favorite. This functionality is useful when a user wants to have his list of hits.

Sometimes, close to the star picture, the '+' symbol is available: this means that MODexplorer was able to find related structures and they can be displayed by clicking the '+' button.

In the same column of all the elements above, for each hit, there is a square that enables its selection: this is necessary when one wants to use the selected structures to build the model of the query protein or to see their structural superposition.

Finally, in the right part of each hit bar there is a clickable icon of a trash-bin that enables the temporary removing of the correspondent hit from the output.

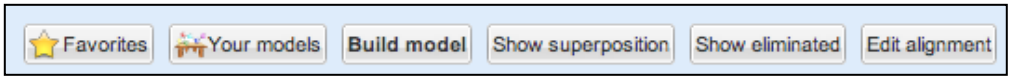
Additionally, only for the related sequences there is the possibility to make them always displayed by clicking on the green icon on the right side of the trash-bin icon. Noticeable is that both removing and making always displayed actions are completely reversible.

6.1.2.4 - Other elements of the central panel

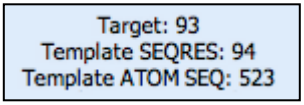
Within the central panel, over the hit list, there are two elements: a group of buttons allowing a number of operations with the structures in the interface and a frame containing the so-called ‘ruler’.
The buttons are (from left to right):

- ‘Favorites’: It allows displaying only the hits marked as ‘favorites’;
- ‘Your models’: It opens a new window, containing the list of all the models built in the current workspace;
- ‘Build model’: When one or more hits are selected, they can be used as modeling templates for building the structures of the input protein. With this button a basic interface for the software Modeller is displayed and the possibility to build the model is provided;
- ‘Show superposition’: If two hits are selected, the superposition between them can be analyzed by clicking this button;
- ‘Show/Hide Eliminated’: If there are eliminated hits in the interface, this button makes them visible so that if wanted, they can be put back in the interface;

- ‘Edit alignment’: This button gives to the user the opportunity to edit the alignment between the query protein and a number of selected ones (maximum 5) with MODalign editor;



The ruler consists of a central frame where it is displayed and automatically updated the position of the amino acid that the user is currently pointing with the mouse arrow. It provides the residue numbers of the target, template SEQRES and template atom sequences.

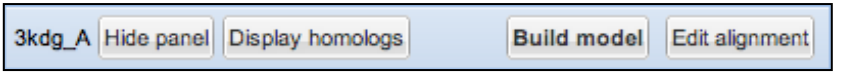


6.1.2.5 - Bottom panel

The last panel contains all the information corresponding to the sequence alignment between the query protein and a hit that the user clicked on.
By default only the sequences of the query, aligned to a hit of interest is displayed with the sequence conservation through the MSA.

In the upper part of the bottom panel, from left to right:

- The name of the selected template;
- ‘Hide panel button’: this element enables the hiding of the whole panel;
- ‘Display/Hide homologs button’: by clicking this element one can decide to display or hide the sequences of the protein families of both target and template;
- ‘Build model’: the same functionality of the button in the central panel;
- ‘Edit alignment’: as above;


3kdg_A
Hide panel
Display homologs
Build model
Edit alignment

7 - APPENDIX B - MODalign description

7.1 - Input computation

MODalign performs the following operations with the submitted target-template(s) alignment (detailed description of software, databases and parameters is provided in the Methods – 3.1):

4. Creates a MSA for both target and template(s) families. The MSA of the target family is built by using HHblits software, whereas the template one(s) is (are) retrieved from the HHSearch alignment database. Like in MODexplorer, the possibilities to skip the creation of the target family MSA and to input a custom one are provided.
5. Detects the representative sequences for the MSAs above using the hhfilter software from the HHSearch package.
6. Aligns the template sequence(s) provided in the input alignment, with the SEQRES and ATOM sequences extracted from the corresponding PDB file(s).
7. Calculates the solvent accessibility and the secondary structure values for target and template families. For the target, the predictions

are made by ACCpro and PSI-PRED, respectively. For the template the solvent accessibility and the secondary structure are derived from corresponding PDB structures with POPS and, DSSP respectively.

8. Merges all the sequences into one single MSA that after the computation contains:
    a. Original target and template sequences;
    b. ATOM an SEQRES sequences template(s);
    c. Homologs of target and template(s);

The MSA can be displayed and edited in an interactive interface allowing:

- Highlighting the residues with no coordinates in the template structure (missing residues). They are displayed as lowercase characters.
- Analyzing of the sequence conservation for each column of the alignment. This parameter is reflected by the background shading of the residues.
- Displaying the solvent accessibility and secondary structure values.
- Highlighting potential errors in the alignment such as:
    - Insertions or deletions within secondary structure elements;

    - Hydrophobic or charged residues of the target aligned with exposed or buried residues of the template(s), respectively;

- Computation of the QMEAN global and local scores for all the models (built without modeling insertions) and the structures of the alignment.
- Evaluating the ‘fitting’ of a template and its family to the target. In fact, all the computation above is also performed on the representative homologs of both target and template(s).
- Editing the alignment by shifting the residues and insertions in templates or target. The system performs the changes in all the proteins of the family of the edited sequence and automatically computes again all the data described above.
- Building the model of the target protein via Modeller;
- Exporting the alignment in PIR or FASTA format;
- Analyzing the template and model structures in Jmol. Additionally, the mapping of insertions and deletions present in the target-template alignment are highlighted on the template structures.

7.1.1 - Description of the web interface

7.1.1.1 - Input form

A number of fields of the input form play the same role in MODalign and MODexplorer. To avoid a redundant description in this thesis, for such fields a reference to the corresponding MODexplorer element is provided here. However, the input form of MODalign consists of:

- ’Paste your target-template alignment’: the field is a text-box allowing the manual specification of the input. The ‘Autofill with a sample TT alignment’ link provides an example of input.
- ‘or upload a Target-template alignment’: the field allows the uploading of a file containing an alignment. The ‘Download a sample TT alignment’ link enables the downloading of a file containing a correctly formatted alignment.
- ‘Alignment format’: a drop-down menu in which the user must specify the format of the input target-template alignment.

The alignment can contain up to 4 templates and FASTA, PIR, CLUSTAL, PHYLIP, or MSF format are accepted. The target name specification format

is arbitrary while the template names should be provided in the format: *pdbcode_chain* (e.g. *1xxx_A*).

- ‘Modeller key’: the same as described in 6.1.1.1.

The advanced section contains:

- ‘Upload an MSA file’: the same as described in 6.1.1.1.
- ‘Build MSA using hhblits’: the same as described in 6.1.1.1.
- ‘Include template MSAs’: If this field is not selected, the MSAs of the template families will not be considered in the computation.

The ‘job related option’ and ‘Your recent job’ sections can be referred to 6.1.1.2.
The form can be filled with a working example by using the ‘Autofill it with an example’ link.

7.1.1.2 - The workspace

To keep the interface consistency in the whole MODORAMA platform, and to facilitate the users who work with both the resources, I used the same style

for the workspace of MODalign and MODexplorer. For the description of the MODalign workspace refer to 6.1.1.3.
The workspace contains all the saved versions of the original target-template alignment and all the models built from them.

7.1.2 - Alignment editor interface

After the minimization of a window with a log, the interface becomes available. By default, it is composed of three sections:

- ‘Target homologs’: it contains the sequence alignment of up to ten representative target homologs. The insertions in homologs, relatively to the target sequence, are not shown and the section is not editable.
- ‘T-t alignment’: It displays the target and template sequences along with the similarity row, ruler and global consensus. In this section, the user can edit the alignment
- ‘SEQRES, original user sequence, and sequence alignments of template homologs’:
    - ‘SEQRES’: canonical SEQRES derived from the PDB file.

- ‘ori_seq’: original sequence submitted by the user.
- ‘homologs’: representative homologs of the template selected via hhfilter. Like the ‘Target homologs’ section it is not editable, it contains up to ten representative sequences and the insertions in homologs relatively to the template sequence are not shown.

7.1.2.1 - Analyzing sequence conservation

MODalign provides several tools to analyze sequence conservation between target and templates:

- ‘Similarity row’: It depicts pairwise target-template sequence similarity between target and reference template. It is displayed as a row between the two sequences.
- ‘Color shading’: it depicts sequence conservation within or between target and template families. If a given amino acid type is regarded as conserved in a given column, its background is shaded with a color corresponding to its type. An amino acid type is considered as conserved if its amino acid type is more frequent than a given adjustable threshold.

- ‘Shading threshold’: this threshold is used for calculating conservation in color shading. Using high values, strongly conserved residues are highlighted, while using low values results useful to show less apparent conservation.
- ‘Coloring schemes’: the coloring scheme can be adjusted in the menu: they are based on BioEdit and Jalview editors.
- ‘Similarity shading’: The conservation for color shading is calculated based on the full alignments of target and template homologs - not only the representative ones that are displayed. There are two coloring modes reflecting sequence conservation within (Group mode) or between (Global mode) target and template families.

7.1.2.2 - Displaying potential errors in the alignment

A user can display potential errors mapped on the target sequence and, optionally, on the sequence of its representative homologs. By selecting one of the options below, new sections are added in the interface:

- ‘Broken helices and β-strands’: in this section, the target and its homologs sequences are colored according to secondary structure

implied by the current alignment to the reference template (helices, strands). The insertions and deletions within the helices and strands are marked with yellow background as potential errors.

- ‘Solvent accessibility errors’: buried charged residues and exposed hydrophobic residues are highlighted in this section.

7.1.2.3 - Comparing secondary structure and solvent accessibility

With MODalign it is possible to compare secondary structure and solvent accessibility, predicted for target and its homologs, and calculated for the templates.

- ‘Secondary structure’: it adds new sections to the interface for target and template(s). Within these sections, secondary structures are depicted as HHHHH (helices) and EEEEE (β-strands, "E" from "extended conformation").
- ‘Solvent accessibility’: as for above, new sections for target and templates are added in the interface in which their respective solvent accessibility are depicted.

7.1.2.4 - Accessing flanking regions

"Flanking regions" are columns of the alignment before and after the first and the last residue of the target. They contain regions of template sequences (derived from PDB SEQRES records) even if the original target-template alignment did not contain those regions. By default, the flanking regions are shaded, but the shading disappears as soon as a previously shaded column or residue becomes aligned with a target residue. The shading can be toggled on/off using the Tools menu.

7.1.2.5 - Changing the reference template

The reference template is the sequence that in the target-template alignment section is just below the target. It can be changed by moving itself, in the ‘target-template alignment’ section, with the Up and Down keyboard keys. The reference template is used to calculate possible secondary structure and solvent accessibility errors and as a default template for calculating QMEAN scores.

7.1.2.6 - Editing the alignment

The selection of the residues to edit is possible in the target-template section by left-click on them. Selection of multiple residues, also in different sequences, is allowed. Selected residues can be shifted using left and right arrows keyboard keys. The gaps after the selected residue can be added by hitting the space bar. Gaps can be removed by selecting the residue after the gap and pressing backspace key.

7.1.2.7 - Assessing the alignment quality with QMEAN

Both global and local QMEAN scores for target and templates as well as their representative homologs can be displayed. For target and homologs, MODalign will take the current alignment, build quick full atom models in background (without modeling insertions, see details about modeling), and return the results.

- 'Local scores': The local QMEAN scores will be displayed as a separate section above the target and below the templates. The colors correspond to predicted residue error in Å (from blue – small or no error to red – big error).

- ‘Global scores’: The global scores appear next to the labels of the target, template and their homologs.

7.1.2.8 - Analyzing the alignment in 3D

A user can click on the ‘Go 3D!!’ button and select a template, then a Jmol window will open. The ‘Indels highlighting’ button enables analyzing the position of insertions and deletions in 3D. The insertions will be highlighted in cyan, the deletions in red. If the Jmol window is opened and the alignment is being edited, the coloring will instantly change to reflect these changes.

7.1.2.9 - Saving and exporting the alignment

The alignment can be exported from MODalign using the menu. There are two options:

- ‘Export all sequences alignment – FASTA’: the exported alignment will contain all sequences currently present in the alignment, including the original user template sequences, the ATOM and SEQRES sequences and all the homologs and secondary structure

and solvent accessibility predictions and assignments. The output format is FASTA.

- ‘Export your t-t alignment’: the exported alignment will contain only the target and template sequences. The output format may be FASTA or PIR, accordingly with the user choice.

7.1.2.10 - Building the models

The ‘Build model’ button allows building a 3D model from the current alignment: a window will be opened from which a user can select the templates for modeling. There are several options allowing the customization of the modeling procedure:

- ‘Do not model tails’: N-terminal and C-terminal residues of the target that are not aligned with any template will not be modeled.
- ‘Do not model insertions’: insertions will not be modeled.
- ‘Breaks on missing residues’: at the position of missing residues in the template, the backbone will be “kept interrupted" as in the template. This is useful to avoid distorting the model structure due to connection of distant residues flanking the missing residues.

By clicking ‘Your models’ button, the list of models built for the target-template alignment can be accessed.

7.1.2.11 - Other features

- ‘Global ruler’: this buttons will add a ruler displaying the numbering of the residues in the target-template alignment section.
- ‘Global consensus’: it adds a consensus in the target-template alignment section based on all the sequences that are in the MSA (not only the representatives that are shown).
- ‘Display homologs’: if it is unchecked, the homologs are not displayed in the target-template alignment section.
- ‘Group consensus’: it adds, in the target and template homologs sections, a consensus reflecting the sequence conservation of the target and template(s) families.

8 - REFERENCES

1. Hegyi, H. and M. Gerstein, *The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.* J Mol Biol, 1999. **288**(1): p. 147-64.
2. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.
3. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
4. Zhang, Y., *Protein structure prediction: when is it useful?* Curr Opin Struct Biol, 2009. **19**(2): p. 145-55.
5. Carbajo, D. and A. Tramontano, *A resource for benchmarking the usefulness of protein structure models.* BMC Bioinformatics, 2012. **13**(1): p. 188.
6. Schwede, T., et al., *Outcome of a workshop on applications of protein models in biomedical research.* Structure, 2009. **17**(2): p. 151-9.
7. Hermann, J.C., et al., *Structure-based activity prediction for an enzyme of unknown function.* Nature, 2007. **448**(7155): p. 775-9.
8. Sievers, S.A., et al., *Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation.* Nature, 2011. **475**(7354): p. 96-100.
9. Fleishman, S.J., et al., *Computational design of proteins targeting the conserved stem region of influenza hemagglutinin.* Science, 2011. **332**(6031): p. 816-21.
10. Baud, O., et al., *The mouse eugenol odorant receptor: structural and functional plasticity of a broadly tuned odorant binding pocket.* Biochemistry, 2011. **50**(5): p. 843-53.

11. Raffa, G.D., et al., *Verrocchio, a Drosophila OB fold-containing protein, is a component of the terminin telomere-capping complex.* Genes Dev, 2010. **24**(15): p. 1596-601.
12. Kosinski, J., et al., *Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair.* Hum Mutat, 2010. **31**(8): p. 975-82.
13. Tang, H., et al., *Do crystal structures obviate the need for theoretical models of GPCRs for structure-based virtual screening?* Proteins, 2012. **80**(6): p. 1503-21.
14. Hillisch, A., L.F. Pineda, and R. Hilgenfeld, *Utility of homology models in the drug discovery process.* Drug Discov Today, 2004. **9**(15): p. 659-69.
15. Liu, T., G.W. Tang, and E. Capriotti, *Comparative modeling: the state of the art and protein drug target structure prediction.* Comb Chem High Throughput Screen, 2011. **14**(6): p. 532-47.
16. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.* Proc Natl Acad Sci U S A, 1961. **47**: p. 1309-14.
17. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
18. Ginalski, K., *Comparative modeling for protein structure prediction.* Curr Opin Struct Biol, 2006. **16**(2): p. 172-7.
19. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. **9**(2): p. 173-5.
20. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
21. Dunbrack, R.L., Jr., *Sequence comparison and protein structure prediction.* Curr Opin Struct Biol, 2006. **16**(3): p. 374-84.

22. Tramontano, A. and V. Morea, *Exploiting evolutionary relationships for predicting protein structures.* Biotechnol Bioeng, 2003. **84**(7): p. 756-62.
23. Jothi, A., *Principles, Challenges and Advances in ab initio Protein Structure Prediction.* Protein Pept Lett, 2012.
24. Xu, D. and Y. Zhang, *Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.* Proteins, 2012. **80**(7): p. 1715-35.
25. Kinch, L., et al., *CASP9 assessment of free modeling target predictions.* Proteins, 2011. **79 Suppl 10**: p. 59-73.
26. Kim, D.E., D. Chivian, and D. Baker, *Protein structure prediction and analysis using the Robetta server.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W526-31.
27. Das, R., et al., *Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home.* Proteins, 2007. **69 Suppl 8**: p. 118-28.
28. Jayaram, B., et al., *Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins.* Nucleic Acids Res, 2006. **34**(21): p. 6195-204.
29. Cooper, S., et al., *Predicting protein structures with a multiplayer online game.* Nature, 2010. **466**(7307): p. 756-60.
30. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Curr Opin Struct Biol, 2005. **15**(3): p. 285-9.
31. Moult, J., et al., *Critical assessment of methods of protein structure prediction (CASP)--round IX.* Proteins, 2011. **79 Suppl 10**: p. 1-5.
32. Hirschman, L., et al., *Overview of BioCreAtIvE: critical assessment of information extraction for biology.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S1.

33. Janin, J., et al., *CAPRI: a Critical Assessment of PRedicted Interactions.* Proteins, 2003. **52**(1): p. 2-9.
34. Fleishman, S.J., et al., *Community-wide assessment of protein-interface modeling suggests improvements to design methodology.* J Mol Biol, 2011. **414**(2): p. 289-302.
35. Qu, X., et al., *A guide to template based structure prediction.* Curr Protein Pept Sci, 2009. **10**(3): p. 270-85.
36. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-6.
37. Tramontano, A., *Protein structure prediction. Concepts and applications*, ed. Wiley-VCH. 2006, Weinheim: Springer-Verlag. 208.
38. Bates, P.A., et al., *Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM.* Proteins, 2001. **Suppl 5**: p. 39-46.
39. Venclovas, C., *Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance.* Proteins, 2003. **53 Suppl 6**: p. 380-8.
40. Koh, I.Y., et al., *EVA: Evaluation of protein structure prediction servers.* Nucleic Acids Res, 2003. **31**(13): p. 3311-5.
41. Bordoli, L., et al., *Protein structure homology modeling using SWISS-MODEL workspace.* Nat Protoc, 2009. **4**(1): p. 1-13.
42. Rost, B., *Twilight zone of protein sequence alignments.* Protein Eng, 1999. **12**(2): p. 85-94.
43. Tosatto, S.C. and S. Toppo, *Large-scale prediction of protein structure and function from sequence.* Curr Pharm Des, 2006. **12**(17): p. 2067-86.
44. Reese, M.G., et al., *Genome annotation assessment in Drosophila melanogaster.* Genome Res, 2000. **10**(4): p. 483-501.

45. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.
46. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
47. Sadowski, M.I. and D.T. Jones, *Benchmarking template selection and model quality assessment for high-resolution comparative modeling.* Proteins, 2007. **69**(3): p. 476-85.
48. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
49. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
50. Mariani, V., et al., *Assessment of template based protein structure predictions in CASP9.* Proteins, 2011. **79 Suppl 10**: p. 37-58.
51. Kosiński, J., Tkaczuk, K. L., Kasprzak, J. M. and Bujnicki, J. M. , ed. *Template Based Prediction of Three-Dimensional Protein Structures: Fold Recognition and Comparative Modeling, in Prediction of Protein Structures, Functions, and Interactions* ed. J.M. Bujnicki. 2008, John Wiley & Sons, Ltd: Chichester.
52. Martin, A.C., M.W. MacArthur, and J.M. Thornton, *Assessment of comparative modeling in CASP2.* Proteins, 1997. **Suppl 1**: p. 14-28.
53. Thorne, J.L., *Models of protein sequence evolution and their applications.* Curr Opin Genet Dev, 2000. **10**(6): p. 602-5.
54. Tress, M.L., D. Jones, and A. Valencia, *Predicting reliable regions in protein alignments from sequence profiles.* J Mol Biol, 2003. **330**(4): p. 705-18.
55. Wallner, B. and A. Elofsson, *All are not equal: a benchmark of different homology modeling programs.* Protein Sci, 2005. **14**(5): p. 1315-27.

56. Fiser, A., *Protein structure modeling in the proteomics era.* Expert Rev Proteomics, 2004. **1**(1): p. 97-110.
57. Kiefer, F., et al., *The SWISS-MODEL Repository and associated resources.* Nucleic Acids Res, 2009. **37**(Database issue): p. D387-92.
58. Petrey, D., et al., *Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling.* Proteins, 2003. **53 Suppl 6**: p. 430-5.
59. Koehl, P. and M. Delarue, *Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy.* J Mol Biol, 1994. **239**(2): p. 249-75.
60. Koehl, P. and M. Delarue, *A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling.* Nat Struct Biol, 1995. **2**(2): p. 163-70.
61. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.
62. Dalton, J.A. and R.M. Jackson, *An evaluation of automated homology modelling methods at low target template sequence similarity.* Bioinformatics, 2007. **23**(15): p. 1901-8.
63. Benkert, P., M. Biasini, and T. Schwede, *Toward the estimation of the absolute quality of individual protein structure models.* Bioinformatics, 2011. **27**(3): p. 343-50.
64. Eramian, D., et al., *How well can the accuracy of comparative protein structure models be predicted?* Protein Sci, 2008. **17**(11): p. 1881-93.
65. McGuffin, L.J., *The ModFOLD server for the quality assessment of protein structural models.* Bioinformatics, 2008. **24**(4): p. 586-7.
66. Pettitt, C.S., L.J. McGuffin, and D.T. Jones, *Improving sequence-based fold recognition by using 3D model quality assessment.* Bioinformatics, 2005. **21**(17): p. 3509-15.

67. Randall, A. and P. Baldi, *SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs.* BMC Struct Biol, 2008. **8**: p. 52.
68. Melo, F. and E. Feytmans, *Assessing protein structures with a non-local atomic interaction energy.* J Mol Biol, 1998. **277**(5): p. 1141-52.
69. Samudrala, R. and J. Moult, *An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.* J Mol Biol, 1998. **275**(5): p. 895-916.
70. Tosatto, S.C., *The victor/FRST function for model quality estimation.* J Comput Biol, 2005. **12**(10): p. 1316-27.
71. Wallner, B. and A. Elofsson, *Can correct protein models be identified?* Protein Sci, 2003. **12**(5): p. 1073-86.
72. Zhou, H. and Y. Zhou, *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): p. 2714-26.
73. Benkert, P., S.C. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins, 2008. **71**(1): p. 261-77.
74. Cozzetto, D., A. Kryshtafovych, and A. Tramontano, *Evaluation of CASP8 model quality predictions.* Proteins, 2009. **77 Suppl 9**: p. 157-66.
75. Kosinski, J., A. Barbato, and A. Tramontano, *MODexplorer: an integrated tool for exploring protein sequence, structure and function relationships.* Bioinformatics, 2013.
76. Barbato, A., et al., *Improving your target-template alignment with MODalign.* Bioinformatics, 2012. **28**(7): p. 1038-9.
77. Warren, J.J., et al., *Structure of the human MutSalpha DNA lesion recognition complex.* Mol Cell, 2007. **26**(4): p. 579-92.

78. Jiricny, J., *The multifaceted mismatch-repair system.* Nat Rev Mol Cell Biol, 2006. **7**(5): p. 335-46.
79. Kolodner, R.D., *Mismatch repair: mechanisms and relationship to cancer susceptibility.* Trends Biochem Sci, 1995. **20**(10): p. 397-401.
80. Peltomaki, P., *Role of DNA mismatch repair defects in the pathogenesis of human cancer.* J Clin Oncol, 2003. **21**(6): p. 1174-9.
81. Mayer, T.U., *Chemical genetics: tailoring tools for cell biology.* Trends Cell Biol, 2003. **13**(5): p. 270-7.
82. Bishop, A.C. and K.M. Shokat, *Acquisition of inhibitor-sensitive protein kinases through protein design.* Pharmacol Ther, 1999. **82**(2-3): p. 337-46.
83. Bishop, A.C., et al., *A chemical switch for inhibitor-sensitive alleles of any protein kinase.* Nature, 2000. **407**(6802): p. 395-401.
84. Punta, M., et al., *The Pfam protein families database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D290-301.
85. Holm, L. and P. Rosenstrom, *Dali server: conservation mapping in 3D.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W545-9.
86. Miyoshi, T., et al., *Bicyclomycin, a new antibiotic. I. Taxonomy, isolation and characterization.* J Antibiot (Tokyo), 1972. **25**(10): p. 569-75.
87. Skordalakes, E., et al., *Structural mechanism of inhibition of the Rho transcription termination factor by the antibiotic bicyclomycin.* Structure, 2005. **13**(1): p. 99-109.
88. Kosinski, J., et al., *The PMS2 subunit of human MutLalpha contains a metal ion binding domain of the iron-dependent repressor protein family.* J Mol Biol, 2008. **382**(3): p. 610-27.
89. Sommer, I., et al., *Improving the quality of protein structure models by selecting from alignment alternatives.* BMC Bioinformatics, 2006. **7**: p. 364.

90. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* J Mol Biol, 1999. **292**(2): p. 195-202.
91. Cheng, J., et al., *SCRATCH: a protein structure and structural feature prediction server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W72-6.
92. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
93. Cavallo, L., J. Kleinjung, and F. Fraternali, *POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level.* Nucleic Acids Res, 2003. **31**(13): p. 3364-6.
94. Golovin, A. and K. Henrick, *MSDmotif: exploring protein sites and motifs.* BMC Bioinformatics, 2008. **9**: p. 312.
95. Spirin, S., et al., *NPIDB: a database of nucleic acids-protein interactions.* Bioinformatics, 2007. **23**(23): p. 3247-8.
96. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
97. Biasini, M., et al., *OpenStructure: a flexible software framework for computational structural biology.* Bioinformatics, 2010. **26**(20): p. 2626-8.
98. Braberg, H., et al., *SALIGN: a web server for alignment of multiple protein sequences and structures.* Bioinformatics, 2012. **28**(15): p. 2072-3.
99. Knight, R., et al., *PyCogent: a toolkit for making sense from sequence.* Genome Biol, 2007. **8**(8): p. R171.
100. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.

101. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J Mol Biol, 2007. **372**(3): p. 774-97.
102. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.