Constructing templates for protein structure prediction by simulation of protein folding pathways

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · NOVEMBER 2008

Impact Factor: 2.63 · DOI: 10.1002/prot.22073 · Source: PubMed

3 AUTHORS:

Ilona Kifer
Agilent Technologies
18 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Ruth Nussinov
Tel Aviv University
624 PUBLICATIONS **27,983** CITATIONS
SEE PROFILE

Haim J Wolfson
Tel Aviv University
209 PUBLICATIONS **13,801** CITATIONS
SEE PROFILE

Published in final edited form as:
Proteins. 2008 November 1; 73(2): 380–394. doi:10.1002/prot.22073.

Constructing Templates for Protein Structure Prediction by Simulation of Protein Folding Pathways

Ilona Kifer[1], **Ruth Nussinov**[2,3,*], and **Haim J. Wolfson**[1,†]

[1]School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

[2]Basic Research Program, SAIC-Frederick, Inc., Center for Cancer Research, Nanobiology Program, NCI, Frederick, MD 21702, USA

[3]Department of Human Genetics and Molecular Medicine Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Abstract

How a one-dimensional protein sequence folds into a specific 3D structure remains a difficult challenge in structural biology. Many computational methods have been developed in an attempt to predict the tertiary structure of the protein; most of these employ approaches that are based on the accumulated knowledge of solved protein structures.

Here we introduce a novel and fully automated approach for predicting the 3-dimensional structure of a protein that is based on the well accepted notion that protein folding is a hierarchical process. Our algorithm follows the hierarchical model by employing two stages: the first aims to find a match between the sequences of short independently-folding structural entities and parts of the target sequence and assigns the respective structures. The second assembles these local structural parts into a complete 3D structure, allowing for long-range interactions between them.

We present the results of applying our method to a subset of the targets from CASP6 and CASP7. Our results indicate that for targets with a significant sequence similarity to known structures we are often able to provide predictions that are better than those achieved by two leading servers, and that the most significant improvements in comparison with these methods occur in regions of a gapped structural alignment between the native structure and the closest available structural template. We conclude that in addition to performing well for targets with known homologous structures, our method shows great promise for addressing the more general category of comparative modeling targets, which is our next goal.

1 Introduction

The question of how a one-dimensional protein chain folds into a three-dimensional structure is still an unsolved problem. This is despite the tremendous progress in the development of methodologies for protein structure prediction during the last decade.

Targets for protein structure prediction roughly divide into 2 categories. The first category of targets includes the comparative modeling targets, for which at least one structurally related protein with a solved structure is known to exist. Comparative modeling methods usually

†Address correspondence to E-mail: ilonak@post.tau.ac.il or E-mail: wolfson@post.tau.ac.il.

attempt to identify such structural neighbors and model the target protein according to these. In the easier cases, termed Homology Modeling targets, there is often a significant sequence similarity between the target and the template. In such instances modeling the target protein according to its closest neighbors given by the sequence similarity can produce quite accurate results. In the more difficult cases, known as Fold Recognition targets, sequence similarity is not significant and sequence-structure threading techniques are required for identifying templates with similar structures. Here the difficulty lies not only in identifying the correct template, but also in the reduced structural similarity between the target and template proteins. Hence even if a good template is found, it is expected that the 3D model will not resemble the native structure as closely as in homology modeling examples. The second category of targets are the proteins with no known structural neighbor. In such cases ab-initio structure prediction methods are required. These methods often sample the 3D conformational space in an attempt to identify the conformation with the minimal energy, assuming it is the closest to the native structure. Lately however, as an increasing number of protein structures are being solved, the space of native protein folds is becoming more extensively covered. Zhang et al [1] suggested that the space of single-domain proteins is already fully represented by the current version of the PDB. These advancements stress the necessity of developing accurate and efficient methods for comparative modeling. The role of ab-initio methods is becoming less distinct, and they are sometimes used in combination with comparative modeling techniques when constructing templates for difficult cases of comparative modeling (see for example [2]).

In the past few years an increasing number of groups have been integrating the use of structural fragments of various sizes and origins into their methods. Kolodny et al. [3] construct protein decoys by using a dataset of twenty fragments of length five. Their algorithm gradually extends the protein chain by superposing a fragment's first 3 residues on the last 3 residues of the previous fragment in the chain. Despite the small size of their fragment library and not relying on the target sequence for decoy generation, they are able to construct qualitative protein decoys. ROSETTA [4] which is one of the leading methods today, uses 9 residue-length structural fragments to model ab-initio targets using a Monte Carlo (MC) based sampling approach. A more recent method is TASSER [2], which achieved impressive results in rounds 6 and 7 of the CASP experiment. It is based on identifying a template using a threading protocol and then dividing the target sequence into two types of regions: continuous regions that align to parts of the template and gapped ab-initio regions. MC sampling [5] is applied for on-lattice modeling of the gapped regions and off-lattice perturbations of the aligned rigid fragments.

Here we present a novel method for protein structure prediction that uses structural building blocks to construct a 3D model for a given target protein. Within this framework, our current work focuses mostly on constructing good templates for the modeling of a target sequence, which is a main difficulty in structure prediction.

The origin and motivation in the choice of our structural fragments are quite different from existing methods. In general, methods such as those described above (and others), are guided by a "practical" scheme in their choice of fragments: for instance, the Rosetta group [4] has examined many fixed-size fragments and has reached the conclusion that 9-residue pieces together with 3-residue pieces work best, while the Tasser group [2] identifies fragments as the regions with a good alignment of the full template structure to the target sequence.

Our work is based on a premise first suggested by Lesk and Rose, that protein folding is a hierarchical process [6], [7]. According to the hierarchical model, short range interactions are the first to occur, forming substructures of local fragments. Later, the relatively stable local structural units with high population time hierarchically join into larger structural units via local and non-local interactions, referred to as hydrophobic folding units (HFUs) [8], forming the basic unit from which a fold is constructed. The HFUs then associate with each other to

form domains. Although several other theories have been proposed over the years that attempt to explain the protein folding pathway, the pioneering work of Lesk and Rose has led to the publication of numerous works that substantiate the hierarchical folding concept [9], [10], [11], [12]. We note, that the hierarchical model bypasses the known Levinthal paradox by implicitly choosing the preferred folding pathway of a protein.

Based on the hierarchical model, Tsai et al. have devised a procedure for hierarchically dissecting a protein structure into *building blocks* [9], [13]. We define a building block as a highly populated, contiguous fragment in a native protein structure. According to the building block model, if one cuts out a building block fragment from a protein chain and places it in solution, the most highly populated conformation of the fragment is very likely to be similar to the building block when embedded in the native protein.

The building block cutting algorithm developed by Tsai et al. iteratively dissects a protein structure, creating an “anatomy tree” in which a node is a one-segment building block. The building block database consists of structural fragments not shorter than 15 amino acids and usually not longer than 60 (although some can be much longer). The average building block length depends on the levels of the hierarchy included in our analysis. For all 9 levels, this average is around 40 residues. In this work we use only the most populated levels, levels 2–6, in which the average building block length is around 35 residues.

The use of the building block model in our work highlights the motivation behind our choice of structural fragments: our purpose is to create a method that predicts a protein model by mimicking its natural process of folding. In Fragfold [14], David Jones and his colleagues present an ab-initio structure prediction method that employs supersecondary structure motifs. These structural fragments are composed of two or three sequential secondary structures, consistent with experimental evidence that suggest that a supersecondary structural fragment can often behave as an independent folding unit. However the definition of supersecondary structures is not based on their ability to constitute independent folding entities during the process of protein folding. In this work we do not limit ourselves to a specific type of structural fragments, but rather rank all possible fragments according to their ability to be independently stable in solution. For this we use the scoring function developed by Tsai et al. [9]. The conceptual difference between the two algorithms is evident from the space of conformations that they search: Fragfold allows the insertion of fragments at randomly chosen positions of a forming conformation. In a hierarchical process, although a conformational change can take place, it is nevertheless likely that independently folding entities stay relatively unaltered after they fold. Hence our algorithm only searches for the orientations of the matched structural pieces with respect to each other while currently keeping them rigid.

The notion of hierarchical folding has led to the development of two methods in our group. Haspel et al. [15] introduced a new method that attempts to assign a collection of structural fragments to a target sequence according to sequence resemblance. Inbar et al. [16] have devised a novel method for the combinatorial assembly of multiple structural units into a 3D structure, based on pairwise docking considerations. However, no tool has been developed that integrates the two stages. Here we combine the assignment and assembly stages into a single scheme with the goal of creating an automated tool for the prediction of a protein structure from sequence. To achieve this goal, we significantly improve the methodology of the first assignment stage. In the second assembly stage we employ a different technique that uses multiple structure alignment rather than docking considerations (see methods). In later stages of development we intend to explore the combinatorial assembly approach as well. To the best of our knowledge this is the first study that uses independently folding structural entities to mimic the natural folding pathways of proteins, based on the hierarchical folding premise.

The CASP experiment (Critical Assessment of techniques for protein Structure Prediction) has been conducted last year for the seventh time, with the intention of evaluating progress in computational methods for protein structure prediction in all its categories. While developing this method and at the time of writing this paper, the assessments for this last round have not yet been published. Hence we started by performing an extensive analysis on the targets of CASP6. However, target and prediction structures are available from the CASP7 site; thus we also test our conclusions from the analysis of CASP6 on targets from the CASP7 experiment.

2 Methods

Our algorithm consists of three parts - a preprocessing stage and two online prediction stages. Figure 1 presents the basic structure of our algorithm. We first briefly describe our method, and then elaborate on each stage.

The structural units in the building block database are clustered into groups of structurally similar fragments. For each cluster a sequence-based profile is created. Given a target sequence, it is aligned to the building block profiles to select the ones that match well to parts of the target sequence. A graph-theoretic algorithm is then employed to select paths of profiles that span the target sequence. In the second stage of our algorithm we attempt to compute the correct orientation of the path building blocks relative to each other. This results in a set of 3D templates. The modeling stage of our algorithm is optional as we allow the use of any alignment method for finding the best correspondence between the target sequence and the template. We also present a simple alignment approach that we have developed. The above process results in a set of models which serve as our predictions for the tertiary structure of the target protein.

Preprocessing Stage - Creation of building Block Profiles

The purpose of the preprocessing stage is to remove dataset redundancy and to increase the efficiency and sensitivity of the first stage of our algorithm. The preprocessing stage was conducted as follows: The building block cutting procedure as described in [9] was applied to the non-redundant PDB database (up to 95% sequence identity). We used two versions of the PDB that correspond to the dates of the CASP6 and CASP7 experiments in order not to use structures that have not been published at the time the experiments took place. The building blocks in each SCOP [17] family were clustered into groups according to a pairwise distance matrix constructed using the SSGS structural alignment program [18]. Our definition of similarity for the clustering is that at least 80% of the residues of the longer of the two building blocks are at a distance of at most 1Å from the corresponding aligned residue in the second fragment. The distance is measured between the C_α atoms of the corresponding residues. This strict criterion is necessary due to two reasons. First, small structural differences may be significant for a short fragment. Second, our clustering procedure is based on pairwise structural comparisons versus a pivot structure. This implies that any two structures in that cluster may be twice as dissimilar to each other than to the pivot structure. So to ensure an adequate similarity between the cluster structures, the similarity threshold has to be strict. As the pivot is guaranteed to be most similar to all other structures in the cluster, it was chosen as the cluster representative. The secondary structure of the representative was determined using the DSSP program [19].

Next a FASTA search [20] was performed against a filtered Swissprot database for every building block inside each cluster. Hits were considered sequence segments which are at least 90% the length of the building block, and have at least 80% sequence similarity but are not identical to the building block sequence. All hits from building blocks within a cluster were combined and redundancy was removed to a level of 95% sequence identity. Each cluster now consisted of structural building blocks as well as sequence segments without structural information. Using STACCATO [21], a structure-based multiple sequence alignment was

constructed for each cluster. Based on this alignment, an HMM model was built and calibrated using the HMMER package [22]. The HMM was configured to the *hmms* mode that searches for the single best local alignment with respect to the sequence, and global with respect to the model.

The result of this preliminary processing is a set of building block profiles that provide the setting for the first stage of our algorithm, in which we assign structural building blocks to segments of the target sequence.

Stage One - Assignment of Profiles to the Target Sequence

Given an input target sequence we apply an assignment algorithm to match building blocks to parts of the sequence. An outline of the assignment stage is presented in figure 2. The following filtering procedure is carried out on all building block profiles in our dataset: Secondary structure of the target sequence is predicted using PSIPRED [23]. The sequence is then aligned against every profile HMM, producing an optimal alignment of the HMM to the target sequence and an e-value that measures the significance of the match. Models with insignificant e-values are discarded, as well as models for which there is a large difference between the secondary structure of the model representative and the predicted secondary structure in the appropriate region of the target sequence.

The clusters that pass the filtering are scored according to two criteria. One is the e-value of the alignment and the other is its length. Shorter building blocks tend to reach higher (less significant) e-values since shorter random matches are more probable than longer ones. Hence we compensate for differences in fragment length in our scoring formula:

$$score = log(evalue * length(alignment)) \quad (1)$$

The result of the above procedure is a set of scored profiles that map to different segments of the target protein. Naturally, there may be several candidates for overlapping regions, and also regions to which no HMM aligned well.

We now move to represent each profile HMM by the representative building block of the corresponding cluster. To choose an assignment of building blocks to the target sequence we employ a graph-theoretic algorithm. We construct the following graph: every profile that passed the filtering is considered as a node. The weight of each node is as given by equation 1. Two virtual nodes are added, a start node and an end node. The edges in the graph are directed and forward only. We add an edge from the start node to every node that represents an assignment of a building block that starts at the first third of the target sequence, and likewise - there is an edge from every node whose assignment ends in the last third of the sequence, to the end node. This graph structure was used to ensure that building block paths span the length of the target sequence, yet to allow the edges of the sequence not to be fully covered by the path as they tend to be less conserved. We do not add edges between two completely overlapping nodes as it is pointless for them to appear in the same building block path. In our current implementation we have also limited the number of edges such that they exist only between members of the same SCOP superfamily. Naturally, this limits our algorithm to proteins with known structural neighbors, such as is the case for homology modeling targets.

An edge between two nodes is scored by the following scheme: if the regions to which the two building blocks are assigned do not overlap and are at least three residues apart, the weight on the edge is the number of residues separating them. If they assign to regions that are between (−2) and 2 residues apart [1], the edge is assigned a fixed weight of 2. If there is an overlap of at least 3 residues between the assigned regions, the edge weight is given by:

$$weight = RMSD(overlap) - 0.01 * length(overlap) \quad (2)$$

Equation 2 reflects that a high RMSD (Root Mean Square Deviation) value between the overlapping parts of two building blocks is an indication that they should not be assigned to the same model structure. The equation also shows a slight preference towards longer fragment overlaps since a given RMSD value grows more significant with the length of the match.

We note that the scheme by which we have chosen to assign edges in our graph is more appropriate for the specific purpose of homology modeling addressed in this paper. It is very likely that when we apply this method to targets where the building block coverage of the target sequence is expected to be smaller, a more relaxed definition of edge existence and weighting will be required.

To calculate paths of building blocks we employ an algorithm for finding the K-shortest-paths in an acyclic graph [24]. This algorithm is implemented using Dynamic Programming and runs very efficiently. We currently use a K value of 10,000. This results in a large number of paths of building blocks, each representing a structural assignment to the target sequence. The paths are clustered according to structural similarity to avoid redundancy. The top scoring paths, each consisting of a set of individual structural fragments, are passed on to the next stage of the algorithm to be assembled into a full structural template.

Stage Two - Assembly of Fragments into a 3D Model

An outline of the assembly stage is presented in figure 3. To calculate the orientation between building blocks in a chosen path we considered several methods. We speculated that for homology modeling targets, a combinatorial docking algorithm such as mentioned in the introduction [16] may not be the best choice. This is because in such targets there is a tendency for the building blocks in a path to originate from structures that are closely related to the target protein and to each other. Hence the relative building block orientation can be more accurately deduced from a set of rigid transformations given by the structural alignment of those proteins.

To obtain these rigid transformations we align the structures from which the building blocks originated by applying MultiProt [25], a multiple structural alignment algorithm. Each structure in the multiple alignment is associated with a transformation relative to a pivot structure. This transformation is applied to the corresponding building block(s), thus providing its orientation with respect to the rest of the fragments.

Applying this transformation procedure to a path of building blocks produces a 3D template. However this template is often incomplete: First, building blocks in the path are frequently assigned to overlapping regions, causing multiple residues in the model to correspond to the same residue in the target sequence. Second, loops are often missing between the building block fragments. Third, a building block is occasionally assigned to a location in the target sequence which does not correspond to its true location in the structure. This is a result of using small structural fragments that may align well to several segments on the target protein, especially when sequence identity to the correct segment is not very significant.

To deal with the structure incompleteness we use the following scheme: For two consecutive overlapping fragments, the common region is removed from the less significantly-matching fragment. In the case of missing loops between two building blocks the better matching fragment is extended by the number of residues missing in the alignment to the target sequence.

[1] a negative distance means an overlap

The third case where a building block is transformed to a wrong location may lead to two scenarios: In the first the edges of the fragment are far from the edges of the previous and following building blocks and prolonging the fragment does not bring it within a distance of 1Å of those edges. In the second the fragment completely overlaps with other building blocks which are assigned to their correct positions. A fragment which satisfies one of those two conditions is removed from the model. Loops caused by the removal of such fragments, and also a lack of an assignment at the beginning or at the end of the target sequence are compensated by extending the relevant fragments. The above procedure results in a 3D template that can be used for modeling the target sequence.

The final stage required to produce a 3D model of the target protein is constructing an alignment between the target sequence and the generated template, and then using a modeling software to induce a 3D model from the target sequence, template and alignment. In this work we have not focused on an alignment approach for calculating the optimal target-template correspondence. Hence we consider the modeling part of our tool to be optional and allow the use of other alignment and modeling tools.

Our alignment approach follows the following scheme: we use STACCATO [21] to obtain a structure-based multiple sequence alignment of the structures from which the building blocks originated, together with that of the template. STACCATO can also align structure-less sequences to the structure consensus, an option which we use to add the target sequence to the alignment. Other SCOP family members are added to the alignment if they cause more residues to be aligned between target and template.

The resulting alignment of the target sequence to the template structure is given to MODELLER [26] for construction of a complete 3D model of the target protein.

3 Results

As we have stressed throughout the paper the novelty of our method is mainly in the algorithm for constructing templates for comparative modeling. Hence we believe that it is more appropriate to assess the quality of our templates, than to assess the final 3D models which also incorporate the quality of the alignment between the target sequence and the template. Several available methods exist for aligning a sequence to a structure and they can be used in combination with our template construction procedure. In this section we show that the templates created by our method are at least as close to the native structure as the best single template defined by CASP assessors [27] [28], and closer to it than the final models of two leading servers for comparative modeling. Hence, assuming a good alignment and refinement procedure, our final models are expected to be closer to the native structure as well.

Throughout most of this section we focus on the easier targets of comparative modeling of the CASP6 and CASP7 experiments. These are targets for which a structural homolog can be identified using standard sequence alignment tools such as BLAST or PSI-BLAST. We have also performed some analysis on more difficult targets and have had some success as well. We discuss these results at the end of this section.

The section is organized as follows: we first present a comparison of our CASP6 templates to the single best templates as published by CASP assessors [27]. At the time of writing this paper, CASP6 assessments and analysis have already been published in a special issue of PROTEINS, but the special issue of CASP7 was not available yet. Hence it was more convenient for us to first analyze our results on CASP6 data. However, predictions are not available anymore from the CASP6 site, and hence for CASP6 we do not present a comparison with other groups' achievements. We then discuss our observations from analyzing CASP6 results and test our conclusions on the High Accuracy (HA) targets of CASP7. This category consists of a subset

of the TBM (Template-Based Modeling) targets which are characterized by having at least one known structural neighbor with GDT ≥ 80. In all but two cases these neighbors can be identified using PSI-BLAST. We presently exclude the two targets from our analysis since our method for identifying structural matches is currently sequence-based. CASP7 targets are compared both to the single best template as defined by the assessors [28] and to two leading structure prediction servers - nFold [29] and ROBETTA [30].

The comparison was conducted by aligning our template to the native structure using the LGA structure comparison tool [31]. We used the structure analysis mode with a distance restraint of 3.8Å. The same procedure was performed for the best template and for the top ranking models submitted for these targets by the nFold and ROBETTA servers.

CASP6 analysis

CASP6 homology modeling targets were divided into two categories - *easy* targets for which a template can be found using sequence alignment tools, and *hard* targets that require more sophisticated methods for successful detection of a good template. There were 24 target domains in the easy category of which we predicted the structures of 18. The six other targets were excluded from our analysis because the domain consisted of two or more sequentially non-consecutive segments currently not handled by our method. Table 1 presents the performance of our method on the 18 targets[2], in terms of the number of C_α atoms of the template that align to corresponding C_αs in the native structure within the accepted distance of 3.8Å The table also presents this statistic for the corresponding best templates. On all presented targets we use the top ranking building block path as our prediction.

On eight of the targets our algorithm chose a path of building blocks all originating in the best template. This in itself is quite significant, since identifying the best template by which to model a target sequence is a very difficult task. For three other targets the quality of our template is similar to that of the best template. On five of the seven remaining targets our method produced a template which is more accurate than the best template. The two other, less accurate templates constructed by our method were the outcome of using a building block with a missing piece[3]. This artifact can be avoided by excluding such building blocks, and it is also possible that a good loop modeling program will be able to compensate for the missing part. The results suggest that for easy homology modeling targets our templates are at least as good as the best structural templates available. Such an outcome is important since best templates are defined a-posteriori according to structural similarity and are thus hard to identify given only the target sequence.

Conclusions from CASP6 analysis

In order to better understand our results and to identify the cases for which it would be beneficial to use our method, we examined the cases on which the quality of our template differs from that of the best template. As we pointed out previously, the two cases in which the quality of our template was lower are the ones where we used a building block that was incomplete. For the targets which we were able to improve, two main reasons for our advantage stood out: one is a reduced accuracy in the orientations of secondary structure elements in several best templates, and the other is our better modeling of gap regions.

For example, in target T0246 (figure 4a) there are two regions where it is evident why our template is more accurate. Region **A** shows a beta sheet that is longer than needed in the best

[2]target T0229 consists of two domains and is counted as two targets. However, the domains were modeled as one because one of them was too short for us to model by itself.

[3]By missing piece we refer to a segment of one or more amino acids that is missing from the PDB structure.

template and although our template is not accurate for this sheet, it is still closer to the native. Region **B** shows a helix-loop segment that is shifted in the best template, and reduces the accuracy of this template in that region. Target T0269_1 (figure 4b) shows a slight improvement achieved by our template over an insertion generated by the best template. For target T0276 (figure 4c) our success is evident in two regions: one is the C-terminus, where the best template is a little shorter than the target protein while our method prolongs the template to cover the entire sequence. This is a less interesting outcome. The second region is in residues 84–92 of the native structure, where there is an anti-parallel beta sheet that is completely missing from the best template yet appears clearly in ours. This beta sheet is missing several residues in the native structure, yet we were able to successfully reconstruct it in our template using a building block from a homologous structure. It is hard to measure how exact its reconstruction is due to the missing native structure residues. For target T0282 (figure 4d) our template was constructed from four structural fragments, none of which originated in the best template. Three regions diverge here between the native structure and at least one of the two templates. There are two fairly large best-template insertions that are non-existent in our template structure (see figure 4d – A and B). We achieved this by using an alternative template 1wohA in these regions instead of the best template 2cevB. The third deletion which is also evident in figure 4d (C), is common to both templates. We were not able to improve this deletion due to the lack of a more suitable homolog in this region.

Our examination of the results on the CASP6 dataset raises several issues. First, our method generally shows promise for comparative modeling targets as already at this stage it is able to construct templates that are often at a higher level than the best single structural templates available for easy homology modeling targets. Locating the best structural template is still a major problem for current homology modeling methods, and hence it is an important achievement to be able to provide a template that is at least as good as the best available template. Second, the examples presented above suggest two interesting directions in which we can contribute to the improvement of structural templates for comparative modeling. First, it seems that our use of local similarity measures allows us to improve certain shifts in secondary structure orientations in the modeled protein (figure 4a and b). Second, often if there is an insertion or deletion in the best template when aligned with the native structure, and an alternative structural homolog exists that is closer in the specific region to the native structure, our method manages to improve upon the best template in this region.

The benefits of these improvements are clear: our method shows potential in improving the field of comparative modeling in providing better templates, and in that - may be able to assist in a better understanding of the connection between protein sequence and function. However, before we can conclude on the importance of this outcome, we must first verify that our results can be generalized to other than CASP6 targets. In addition, we must also make sure that our improvements upon the best template are not trivial: if applying a loop modeling software upon the best template can bring it as close to the native structure as our templates, then our achievements are not as significant. We explore these points by conducting an analysis of CASP7 targets, and comparing our hybrid templates not only to the best templates but also to the final models of two leading servers.

CASP7 analysis

We chose our subset of TBM targets according to the following criteria: our current focus is the High Accuracy (HA) targets. We exclude target T0324_1 because it consists of two sequentially non-consecutive segments. Targets T0324_2 and T0367 are more difficult since their best templates cannot be identified using PSI-BLAST. Hence our method is currently unable to score their building blocks above other, irrelevant structural fragments. Target T0326

has only one homologous structure and in that does not provide an interesting test case; it is apparent that all predictors will use this one parent template for modeling.

Table 2 lists our results on CASP7 targets, and compares the template quality of our method to the best template and to the final models of two leading methods - nFold [29] and ROBETTA [30]. For each predicted structure, the table presents the number of residues of the native structure that align to corresponding residues in this prediction. We note that this number does not entirely measure the extent to which a native structure is modeled correctly. Template insertions do not align to residues of the native structure and hence do not affect the counts, but still cause the template to differ from the native structure. A possible correction for this partial measure could be to list the lengths of the predictions to which we are comparing the native structure. If their residue number is larger then it can point to a template insertion. However, it is not very indicative to list template lengths as they sometimes reflect unimportant differences at the edges of the homologous protein. It is also not indicative to list model lengths as their sequence is similar to that of the target, implying equal lengths. Given the above limitations, we present our results as the number of residues predicted correctly, and maintain that it gives a sufficient measure to the success of the predictors.

Our aim, in addition to examining the general performance of our method, was to check our hypothesis from the CASP6 results. Hence we were especially interested in our performance on targets for which the best template (provided on the CASP7 website [28]) contains an insertion or deletion when structurally aligned to the native structure. We found 11 such targets, two of which are targets T0324_2 and T0367 and were excluded from our analysis as stated previously. The remaining nine targets are presented in table 2 separated from the other targets, to highlight their importance to our analysis. The average percentage of correctly modeled residues is also presented for each group separately. From here on, we refer to the group of targets with the gapped best template as the *gapped* group and to the rest of the targets as the *regular* group.

Several things stand out from table 2. First, that the general prediction accuracy on the regular group of targets is higher than that achieved on the gapped group. This indicates that the latter group of targets is more difficult to predict accurately. Second, on the regular group of targets our performance is comparable to that achieved by the best templates and by ROBETTA. However on the gapped group, our method outperforms both ROBETTA and nFold by more than two percent and is similar to the accuracy achieved by the best template. This outcome suggests that identifying a good template to model a target is more difficult when structural gaps between homologs exist.

Table 2 appears to point to another observation: the quality of our templates is on average similar to that of the best structural templates, but is not better. We stress that this is a worthy achievement in itself, since the best templates are chosen a-posteriori according to structural knowledge of the target sequence, and also in light of the two leading methods not reaching this accuracy even after a refinement procedure. Additional notes on this resemblance in template quality are presented in the discussion. We must also bear in mind that this apparent similarity of accuracy between us and the best template can be misleading. This is because in cases of a best template insertion with respect to the native structure our template's improvements are not evident. Such is the case for target T0332 (figure 5). The left figure shows that the best template exhibits an insertion with regard to the native structure which our template is able to fix. Figure 6 shows target T0313, where circles A and C focus on two template insertions that are common to both the best template and the top ranking model of nFold, and that do not appear in our template.

We wish to highlight another important issue that should be taken under consideration. It is possible that unessential gaps between targets and template structures can be eliminated by using modeling and refinement procedures. In such a case our hybrid template approach may seem redundant. However, as we show below this is not necessarily the case. Target T0332 (figure 5, left) exhibits an insertion common to both the best template and nFold's parent template (1j85A), and the server is not able to eliminate this insertion during the modeling stage. The figure on the right side presents a reduced accuracy in the orientation of two helices in the nFold model, which is also the result of using a less accurate template. Figure 6 presents target T0313, for which table 2 suggests that our template is less favorable than the best template and nFold's top ranking model. The reason for this is evident from the left part of the figure, circle B, where a bad join of two building blocks has caused a large insertion to the template. However, circles A and C clearly show two template insertions, common to both the best template and the top ranking model of nFold, that our template manages to avoid. These gaps also appear in the parent template of nFold, indicating that nFold's modeling stage could not eliminate it. On target T0339_2 (figure 7) our template and nFold achieve similar results, while ROBETTA produces a seemingly better model. This target was modeled by ROBETTA using the best template, yet circle A shows that ROBETTA improves the modeling of a loop that is missing in the best template and in our hybrid template. However, this modeling procedure causes the insertion shown in circle B where our template and the best template are much closer to the native structure. In target T0303_1 (figure 8), both the best template and the template used by nFold vary from the native structure in a large loop that contains a small alpha helix. nFold is unable to recreate this loop in the modeling stage, while in our template this helix is clearly visible although not accurate.

Target T0308 is a special case, yet it also demonstrates the importance of choosing a good template. Here, our hybrid template is further from the native structure than both the best template and the nFold top ranking model, as we were unable to identify the best structural template by sequence similarity. This target protein contains a loop in residues 24–38 that is modeled in our template using a fragment from structure 1r8sA. This structure has a 99% sequence identity to the best template 1o3yB (they differ by a single amino acid in a different part of the sequence). However, 1r8sA is a bound conformer while 1o3yB was not solved as part of a complex. The interaction which involves residues 24–38 causes a change in this loop in 1r8sA, which is what causes our template to be less accurate. We note, that ROBETTA chose a template in which this loop is even further from the native structure than our template, while nFold chose the best template as its parent. Their choices are evident from the figures in table 2.

The above examples clearly emphasize the limitations of current refinement and loop modeling tools, and stress the necessity of beginning the homology modeling process with templates that are as close to the native structure as possible.

To summarize this section, we wish to that while ROBETTA shows a superior performance over nFold on the regular group of targets, it is less accurate on the gapped target group. Our examination shows that the reason for this is that on regular targets, ROBETTA could choose the best template as its parent template in more cases than nFold. nFold, however, seems to provide better accuracy templates on the harder targets at the expense of the accuracy of the regular ones. In comparison, our results are at least as good as the two servers and the best templates on both target groups, suggesting that our method is able to improve upon problematic cases of High Accuracy modeling without compromising on the accuracy of the easier targets.

Quality analysis

An important question to ask is how significant are the improvements that we were able to achieve with respect to other templates. For example, improved residues at the ends of the target protein are less meaningful in terms of predicting its fold. Since our prediction performance is fairly similar to that of the best template, it is possible that our improvements in predictions concentrates at the edges of the target protein while on other, more significant residues our predictions are not as accurate.

To analyze this issue, we looked at the target residues that were modeled well by our hybrid template but not by the best template. For each residue, we noted its location on the target sequence. Figure 9 presents the distribution of the locations of these residues on the target sequence. Each target sequence was divided into 20 bins of equal length, each representing a twentieth part of the target sequence. The bin to which each improved residue belongs was noted.

According to figure 9, roughly 23% of the residues we were able to improve belong to sequence edges. Although this is a significant portion of the residues, it is not discouraging: the total number of improved residues at the edge amounts to only 17 residues, while the opposite analysis showed that the best template improved 22 residues at the edges of the target protein over our templates. Taking our bins to be one-tenth of the sequence length gives similar results. More importantly, over three quarters of the residues we managed to improve are not at the edge. Hence our method shows its advantage in improving the prediction of residues at the heart of the target protein, which can be significant for structure prediction purposes.

summary

To conclude this section, we wish to highlight several observations that stand out from our results. First, we point out that in general our hybrid templates are at least as close to the native structure as both the best single template and the top ranking models of two leading servers, and often much better. Some work still needs to be done in further improving our choice of building blocks and in joining them together. By developing an alignment procedure that is optimized for our templates and employing a modeling and refinement procedure, it is expected that our final models could outperform other models. Second, it stands out that choosing a good template is often more important for creating a good model than the modeling stage; The above examples show that ROBETTA managed to perform well on several targets where they chose a better parent template than nFold. In other cases, when there was a structural gap between the native structure and the template chosen by nFold or ROBETTA, the modeling procedure these methods employed could seldom help in reconstructing this gap correctly. We deduce that creating templates which are as accurate as possible, even in loops, is important to the progress of methods for comparative modeling. Third, our analysis shows that our observation regarding CASP6 targets indeed holds - for the set of gapped targets, we manage to outperform both the nFold server and ROBETTA. This indicates to us that we have captured the essence of devising complete templates from structural fragments. It is expected that in cases of lower sequence similarity where structural resemblance is lower as well, our approach will prove itself even more.

CASP6 Hard Target Analysis

There were a total of 17 hard homology modeling domain targets in the CASP6 homology modeling category. For these targets, our 3D models were able to predict at least 70% of the C_α atoms within a distance of 5Å of their corresponding residue for all but three predictions. This outcome suggests that we were generally able to predict the conformations of most of the target structures. The three structures for which we were not able to provide a good prediction

were the ones where the best template structure had a very low sequence identity to the target sequence - an identity of ~ 10% which is much lower than the homology threshold.

When comparing ourselves to leading servers in the CASP6 homology modeling category such as ROBETTA, our results were not as conclusively good as on the easy targets, yet showed some success. Our prediction on these targets was made difficult by limitations of our template selection procedure. Our main obstacle was in predicting the correct template on some targets with very low sequence identity to good structural templates. While other methods used sophisticated fold recognition tools, we presently employ a sequence-based profile-to-sequence alignment tool which is not always sensitive enough in subtle cases. We are currently working to further develop our abilities in detecting matches of structural fragments to segments of a given sequence. Despite this, our fragment-based approach allowed us in some cases to recognize templates by local resemblance, where other methods required profile-to-profile alignment tools or threading tools for this purpose. Examples for this are targets T0199_1 and T0226_1 for which ROBETTA required the 3D-jury fold-recognition tool [32] in order to detect a suitable template while our algorithm hit a good template using the top ranking path in both cases.

4 Discussion and Conclusions

In this work we present a novel, fully automated method for creation of templates for protein structure prediction. The notion behind our research is that the protein folding process is hierarchical in nature. Under this assumption we have attempted in this work to mimic the natural folding process by first matching parts of the target sequence with independently folding structural building blocks, and then assembling them into a full 3 dimensional structure.

In spirit, our method can be considered as a fragment-based or multi-template method. Such methods attempt to construct a 3D model using multiple templates where each template matches a certain region of the target sequence. Multi-template methods allow these regions to overlap, and use several templates as input to a modeling tool. To the best of our knowledge almost none of the above methods employ the use of fragments with biological significance. An exception is David Jones' FRAGFOLD which uses supersecondary structures. As we pointed out in the introduction, it cannot be said that the group of independently folding units is covered well by supersecondary structural elements. Hence our method is unique in its predetermined choice of fragments according to biological motivation. In this sense our work is closer than many others in the field to mimicking, and perhaps eventually better understanding the natural process of protein folding.

In this work we do not focus primarily on creating 3D models of the target sequence, but rather on constructing potential templates for modeling the native structure. Although we do provide a simple method for finding a correspondence between the residues of our template and those of the target sequence, we do not claim to its superiority over other methods and our attention is focused mainly on providing improved templates as a basis for modeling the target sequence. As stated in [33], the question of generating a good sequence-template alignment has been studied extensively. Yet the question of selecting or constructing a good template for the purpose of modeling has received less attention. Hence a method that can produce templates which are of similar if not higher quality than that of the best template available can prove very beneficial. Although we believe that it is advantageous for us to develop an alignment method which is optimized for the way we construct our templates, we have shown in this paper that the choice of a good template is in itself an important task and in which there is considerable room for improvements. In many cases, loop modeling and refinement techniques employed by leading servers were not able to correct a mis-structured piece in their chosen parent template. We note that it is quite natural to extend our own template-construction method to

provide a target-template sequence correspondence, as the alignment between building blocks and parts of the target sequence is already calculated as part of the method. It is still not straightforward though, since the alignment needs to be extended to parts where building blocks are not assigned. We plan to implement this shortly. For now, any alignment tool can be used to accompany our template construction method.

To test the performance of our method we have compared it to the quality of the corresponding single best templates provided by the assessors of CASP6 and CASP7 [27], [28], and to two leading structure prediction servers - nFold and ROBETTA. In this paper we mainly address the targets with a more significant sequence similarity to known protein structures. On these targets we are able to show a comparable prediction performance to the best single templates and to templates devised by two leading servers. Our results suggest that among HA targets, cases of significant template insertions and deletions are the harder targets and the choice of a good template is more difficult there. Nevertheless we were able to outperform leading servers on these targets. It is also important to point out why our templates do not seem to improve on average over the best single templates. There are two reasons for this. First, as shown in the CASP7 analysis section, the measure we present is not indicative of better modeling of best template insertions, since it accounts only for residues that are correctly modeled in the native structure. Second, although our improvements are evident in many regions of gapped alignments between the best template and the native structure, other more subtle structural differences between the native and best template were difficult for our method to detect. Since the best template is chosen a-posteriori with knowledge of the native structure, it naturally minimizes such structural differences. However, as these are small divergences from the native structure they have a lesser chance of having a major effect on the functionality of the protein than do large insertions or deletions. Hence, we believe that our method's improvements upon gapped regions can provide a contribution to better understanding the connection between a protein sequence and its function, and consequently improve drug design abilities - one of the top priority goals in structural biology today.

Our aim in this research is mainly to produce improved templates for high accuracy targets. Hence it could be interesting to speculate on the impact of using fragment-based techniques for high-resolution modeling of target proteins. It is apparent that such methods have a great advantage in identifying local matches over methods that search for a single best template. This is evident in our work from the many insertions and deletions we were able to model better than the structurally closest single template. However, several difficulties may arise when combining the fragments that originate in different structures. First, there could be a shift in the orientation of the fragments to be joined because of differences in the orientations of the fragments in their originating structures. Second, there are sometimes gaps in the target sequence where structural fragments are not assigned. Third, the combination of several templates may cause increased steric clashes when a refinement procedure is applied, which may require extensive rearrangements to the final model. In order to complete template gaps, either the fragments on the sides of the gap could be prolonged, or an attempt can be made to match smaller fragments. The first approach can easily bring to bad modeling of such regions (as shown in one of our examples) while the second technique may allow the matching of unrelated fragments to the target as it is extremely difficult to match very short protein segments. Hence, there is an apparent tradeoff between the accuracy achieved by increasing local specificity, and the accuracy lost when joining unrelated fragments to each other. However, the lack of assignment to certain sequence parts can be used to our advantage: these segments of the target sequence are likely to display a reduced conservation, and hence can easily indicate regions that should be modeled differently, using loop modeling techniques or sampling procedures. Hence, they may in the end contribute to a construction of an improved template, even in non-conserved regions. However, the effects of a high resolution refinement procedure on such hybrid templates still needs to be examined.

Despite the apparent problems, our work suggests that there is much to gain from using fragment-based methods for the modeling of high-accuracy targets. However, we intend our method to address the broader range of comparative modeling targets with various difficulties. For this purpose, it is clear that there is still much work to be done. First, we intend to employ the profile-sequence alignment which is part of the output of the first stage of our method in order to build an alignment between the target sequence and template structure. Since the templates we were able to build were often more accurate than the best structural template, it is reasonable to assume that the alignments deduced from our assignment stage will also be quite accurate. This will enable us to construct complete 3D models of the target sequence. Second, it is evident that we need to strengthen our ability to identify structural building blocks when sequence identity between target structure and good templates is low. We intend to devise a more sensitive scheme for identifying such building blocks, where structural considerations will be brought into account.

Considering the results presented in this paper, we believe that our strategy will provide high quality 3D models for the prediction of protein structures from templates.

Acknowledgements

We thank Dr. Chung-Jung Tsai and Yuval Inbar for useful insights during this work. IK is a fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. The research of HJW has been supported in part by the Israel Science Foundation (grant no. 281/05) and by the Hermann Minkowski- Minerva Center for Geometry at TAU. The research of HJW and RN has been supported in part by the NIAID, NIH (grant No. 1UC1AI067231) and by the Binational US-Israel Science Foundation (BSF). This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

1. Zhang Y, Skolnick Y. The protein structure prediction problem could be solved using the current pdb library. Proc. Natl. Acad. Sci. USA 2005;102:1029–1034. [PubMed: 15653774]
2. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA 2004;101:7594–7599. [PubMed: 15126668]
3. Kolodny R, Levitt M. Protein decoy assembly using short fragments under geometric constraints. BioPolymers 2003;68(3):278–285. [PubMed: 12601789]
4. Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol 1997;268:209–225. [PubMed: 9149153]
5. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: Parallel hyperbolic monte carlo sampling of protein folding. Proteins 2002;48:192–201. [PubMed: 12112688]
6. Rose G. Hierarchic organization of domains in globular proteins. J. Mol. Biol 1979;134(3):447–470. [PubMed: 537072]
7. Lesk A, Rose G. Folding unit in globular proteins. Proc. Natl. Acad. Sci. USA 1981;78:4304–4308. [PubMed: 6945585]
8. Tsai C, Nussinov R. Hydrophobic folding units at protein-protein interfaces: Implications to protein folding and to protein-protein association. Protein Sci 1997;6:1426–1437. [PubMed: 9232644]
9. Tsai C, Maizel J, Nussinov R. Anatomy of protein structure: Visualizing how a 1d protein chain folds into a 3d shape. Proc. Natl. Acad. Sci. USA 2000;97:12038–12043. [PubMed: 11050234]
10. Przytycka T, Srinivasan R, R G. Recursive domains in proteins. Protein Sci 2002;11:409–417. [PubMed: 11790851]
11. Chikenji G, Fujitsuka Y, Takada S. Protein folding mechanisms and energy landscape of src sh3 domain studied by a structure prediction toolbox. Chemical Physics 2004;307(2):157–162.

12. Hockenmaier J, Aravind KJ, D K. Routes are trees: The parsing perspective on protein folding. Proteins 2007;66:1–15. [PubMed: 17063473]
13. Tsai C, Ma B, Sham Y, Kumar S, Wolfson H, Nussinov R. A hierarchical, building-block-based computational scheme for protein structure prediction. IBM J. Res & Dev 2001;45(3):513–522.(3)
14. Jones D. Successful ab initio prediction of the tertiary structure of NK-Lysin using multiple sequences and recognized supersecondary structural motifs. PROTEINS 1997;(S1):185–191. [PubMed: 9485510]
15. Haspel N, Tsai C, Wolfson H, Nussinov R. Reducing the computational complexity of protein folding via fragment folding and assembly. Protein Sci 2003;12:1177–1187. [PubMed: 12761388]
16. Inbar Y, Benyamini H, Nussinov R, Wolfson H. Protein structure prediction via combinatorial assembly of sub-structural units. Bioinformatics 2003;19:158–168.
17. Murzin A, Brenner S, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol 1995;247:536–540. [PubMed: 7723011]
18. Wainreb G, Haspel N, Wolfson H, Nussinov R. A permissive secondary structure-guided superposition tool for clustering of protein fragments toward protein structure prediction via fragment assembly. Bioinformatics 2006;22(11):1343–1352. [PubMed: 16543273]
19. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637. [PubMed: 6667333]Online available at www.cmbi.kun.nl/gv/dssp/article.html
20. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc. Natl. Acad. Sci. USA 1988;85:2444–2448. [PubMed: 3162770]
21. Shatsky M, Dror O, Schneidman-Duhovny D, Nussinov R, Wolfson H. BioInfo3D: A suite of tools for structural bioinformatics. Nucleic Acids Res 2004;32:W503–W507. [PubMed: 15215437]
22. Eddy S. Profile hidden markov models. Bioinformatics 1998;14:755–763. [PubMed: 9918945]
23. Jones D. Protein secondary structure prediction based on position specific scoring matrices. J. Mol. Biol 1999;292:195–202. [PubMed: 10493868]
24. Cormen, T.; Leiserson, C.; Rivest, R. Introduction to Algorithms. MIT Press; 1990.
25. Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins 2004;56:143–156. [PubMed: 15162494]
26. Sali A, Blundell T. Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol 1993;234:779–815. [PubMed: 8254673]
27. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the casp6 comparative modeling category. Proteins. 2005;(S7)27-4
28. http://predictioncenter.org/casp7/meeting/talks.html.
29. Bryson K, McGuffin L, Marsden J, Sodhi J, Jones D. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 2005;61(S7):143–151. [PubMed: 16187356]
30. Chivian D, Kim D, Malmstrom L, Schonbrun J, Rohl C, Baker D. Prediction of casp6 structures using automated robetta protocols. Proteins 2005;S7:157–166. [PubMed: 16187358]
31. Zemla A. A method for finding 3-d similarities in protein structures. Nucleic Acids Res 2003;31(13): 3370–3374. [PubMed: 12824330]
32. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3d-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19(8):1015–1018. [PubMed: 12761065]
33. Sadowski M, Jones D. Benchmarking template selection and model quality assessment for high-resolution comparatibe modeling. Nucleic Acids Res. 2007In Print

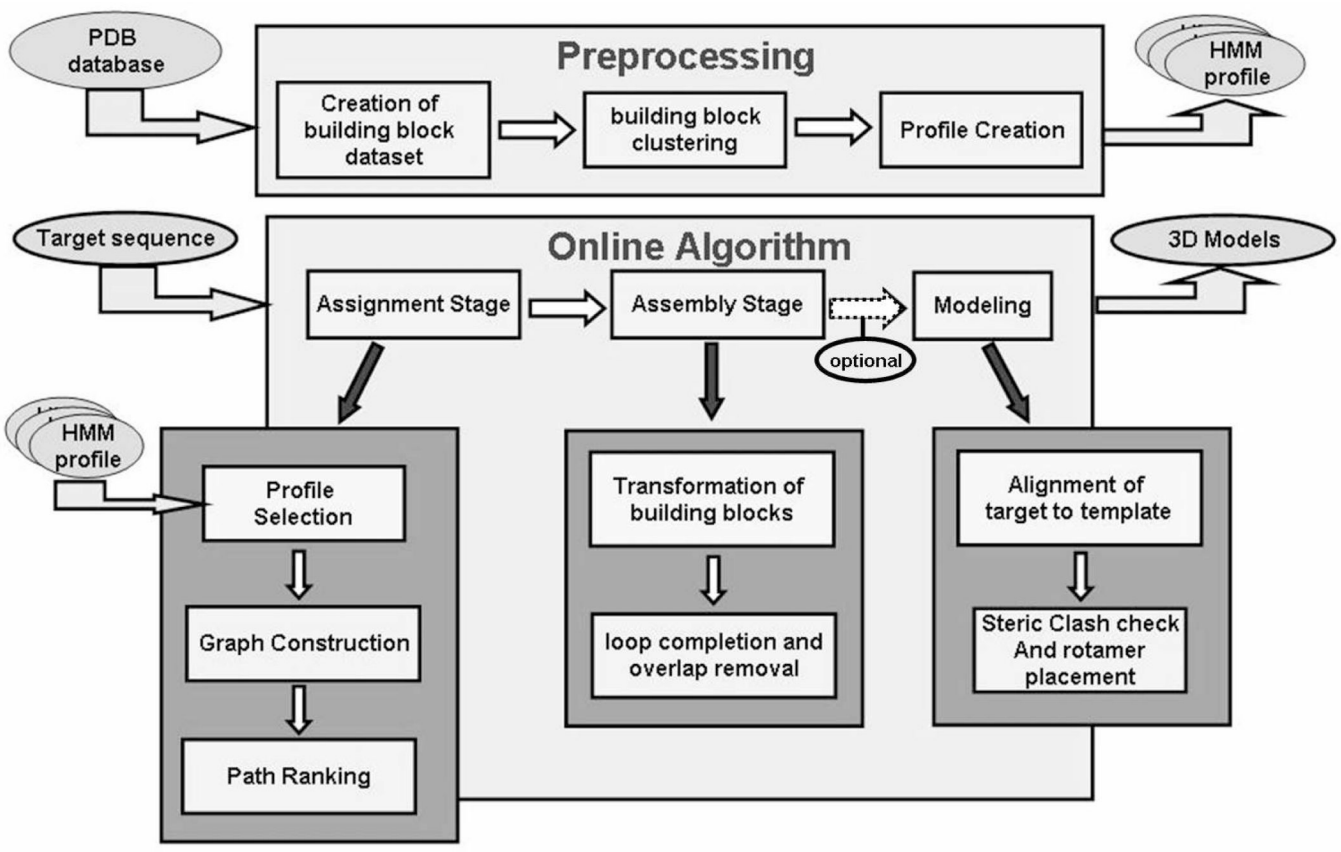

Figure 1.
An outline of our algorithm

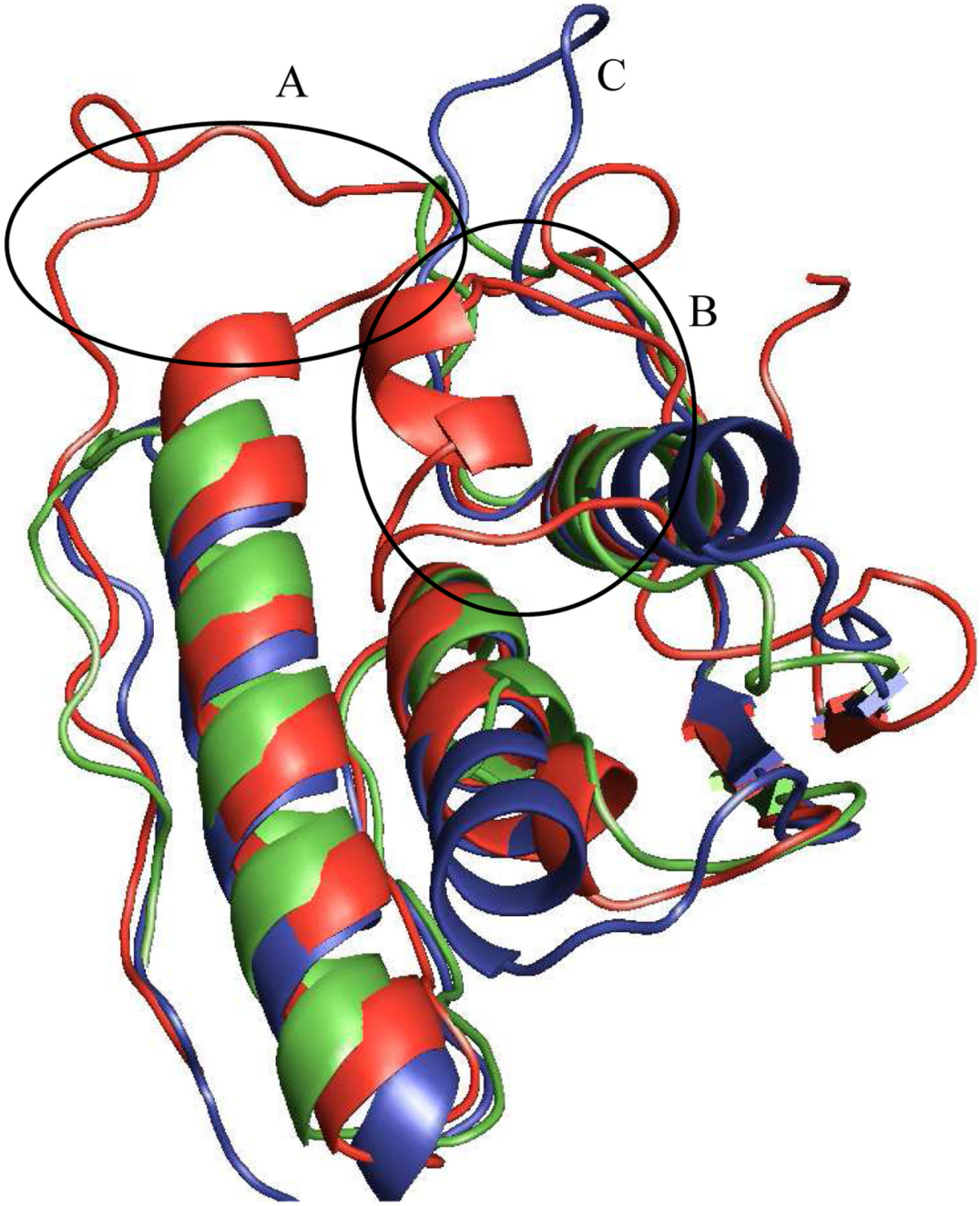

Figure 2.
An outline of the assignment stage. (A) After a filtering procedure on building block profiles, the remaining building blocks are aligned against the target sequence. (B) Each aligned HMM is represented as a node in a graph and positioned according to the region to which it was aligned in the target sequence. Weighted edges are added between nodes as explained in the text. (C) A K-shortest-paths algorithm is run to retrieve the most likely paths of building blocks for constructing the 3D template.

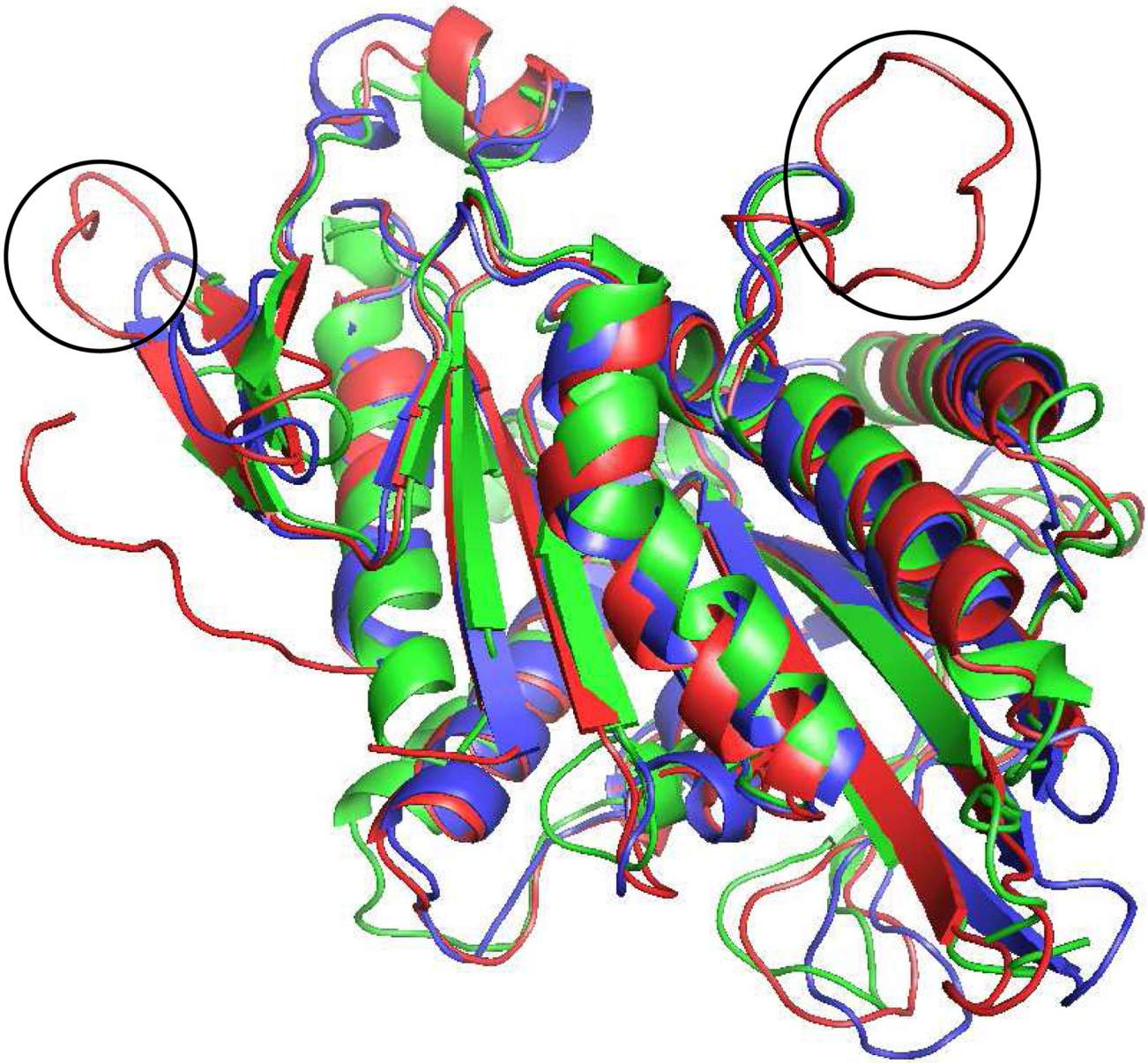

Figure 3.
An illustration of the assembly stage on target T0233_1 - 1vquA_1. The input from the assignment stage is a path of building blocks that are matched to consecutive parts of the target sequence. The PDB structures from which the building blocks originated are retrieved and their multiple structure alignment is computed. The resulting 3D transformations are then applied to the assigned building blocks to form a final 3D template.

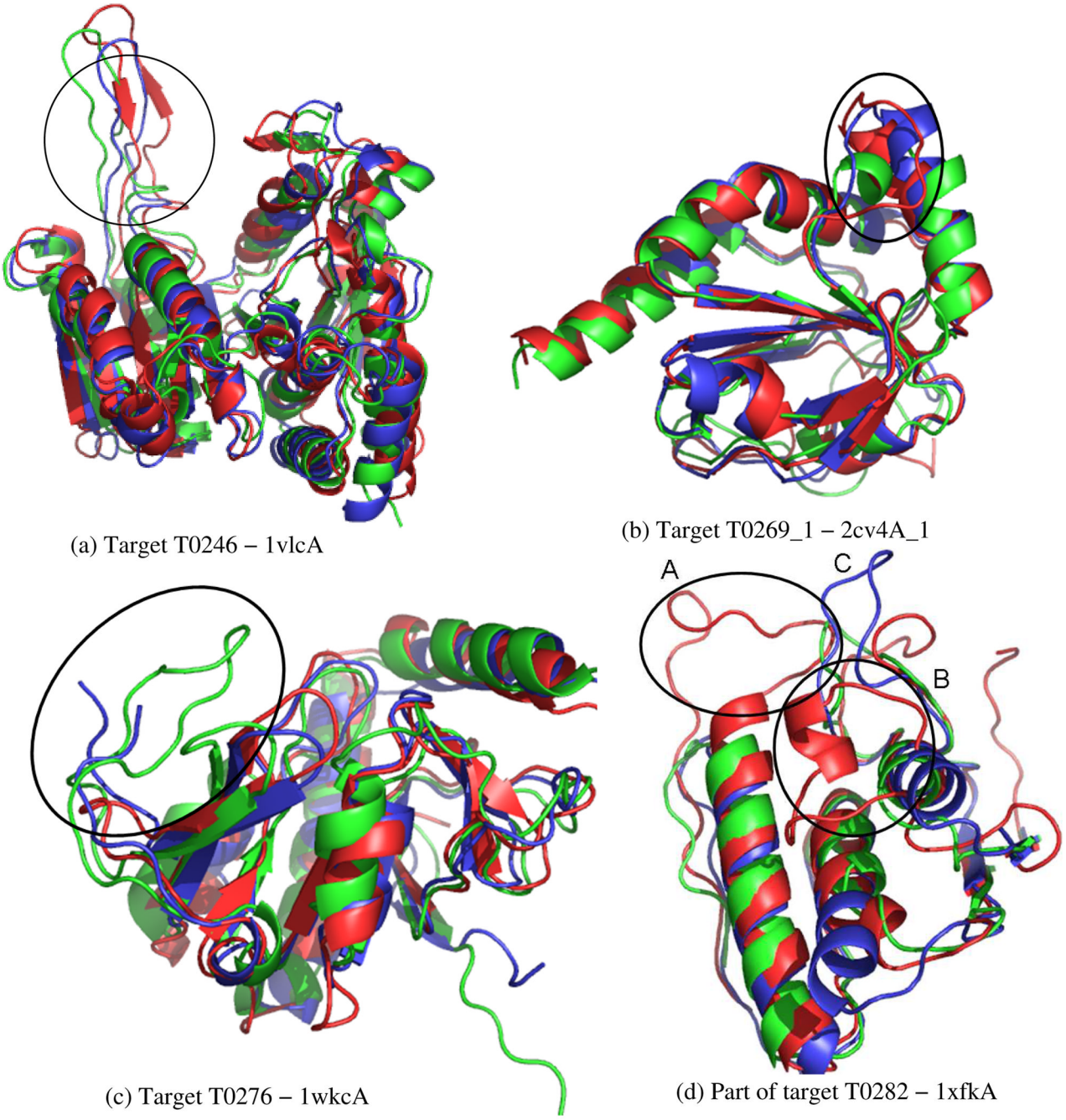

Figure 4.

Several successful predictions on targets from CASP6. Native structure is shown in blue, our template is shown in green and best template is shown in red. (a) Target T0246. Region **A** shows a beta hairpin that was extended by the best template. Region **B** shows a helix-loop segment with a shifted orientation towards the rest of the structure in the best template. (b) Target T0269_1. A loop inserted by the best template is circled. (c) Target T0276. A beta hairpin missed by the best template which clearly appears in our template. Since several residues are missing from the native structure, it is hard to know how well we modeled this beta sheet. (d) Part of Target T0282 – 1xfkA. Example of two best-template insertions that our

method eliminates. The insertion areas are labeled as **A** and **B**. A third deletion which we are not able to correct appears as **C**

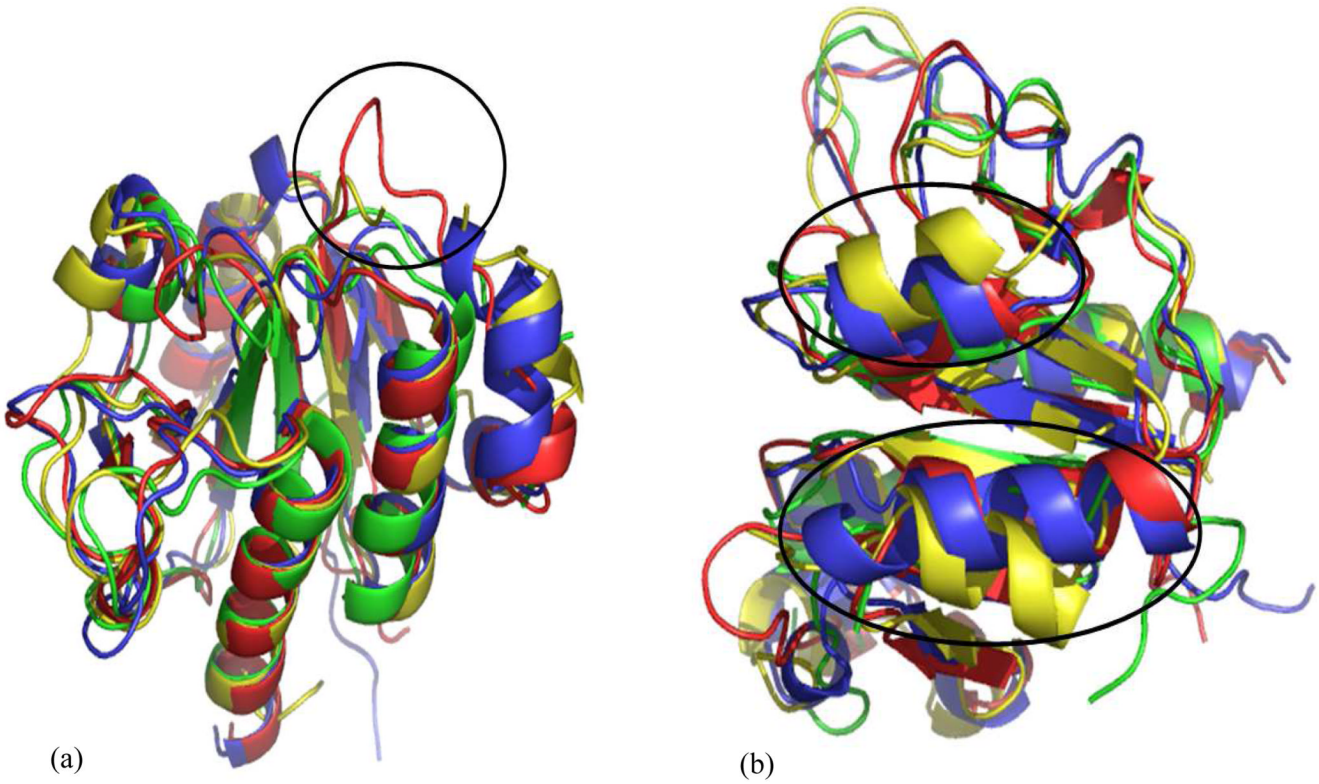

Figure 5.
Target T0332 - 2ha8A. Native structure is shown in blue, our template is shown in green, best template is shown in red and nFOLD model is shown in yellow. (a) Example of another best-template insertion eliminated by our method. Our template models the insertion much better than both best template and nFold model. Insertion area is circled. (b) The orientation of some of the helices has been compromised in the nFOLD model during its modeling process.

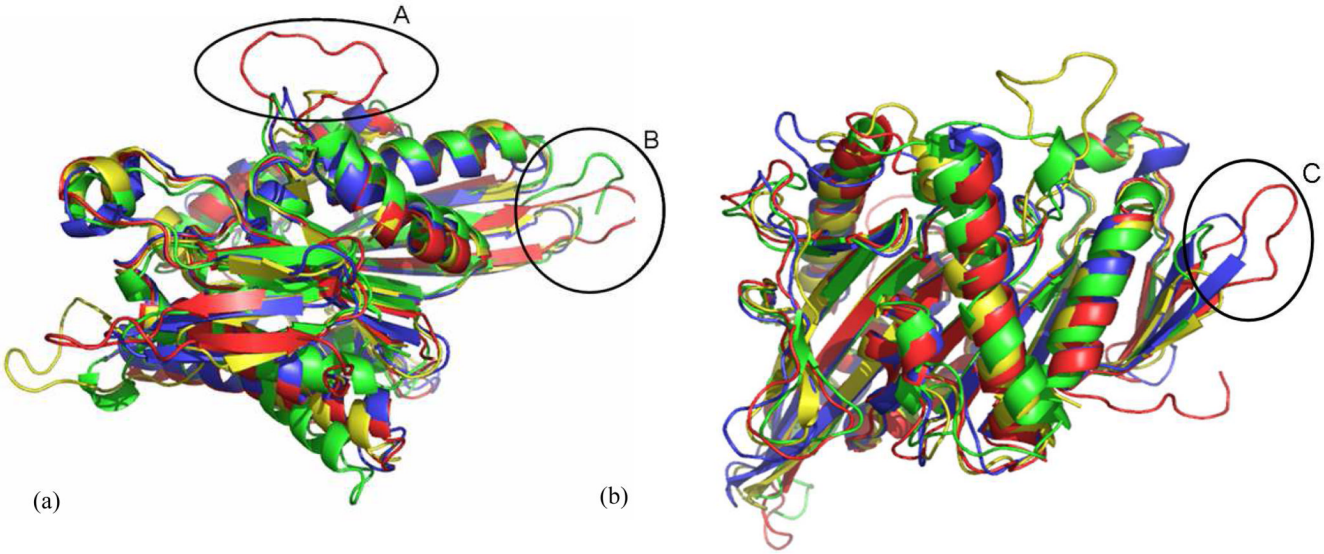

Figure 6.

Two views for target T0313 - 2h58A. Left picture shows (A) an insertion eliminated by our method and (B) a mismodeled part of our template due to an inaccurate join of two building blocks. Right figure shows another insertion eliminated by our method. It is evident in the two insertion regions how much closer our template is to the native structure than both the best template and the model submitted by nFOLD. Native structure is shown in blue, our template is shown in green, best template is shown in red and nFold model is shown in yellow. Areas of interest are circled.

Figure 7.
Target T0339_2 - 2hdyA_2. (A) a deletion of the best template which was modeled well only by ROBETTA (B) an insertion created by ROBETTA which is non existent in our and best templates. Native structure is shown in blue, our template is shown in green, best template is shown in red and ROBETTA model is shown in yellow.

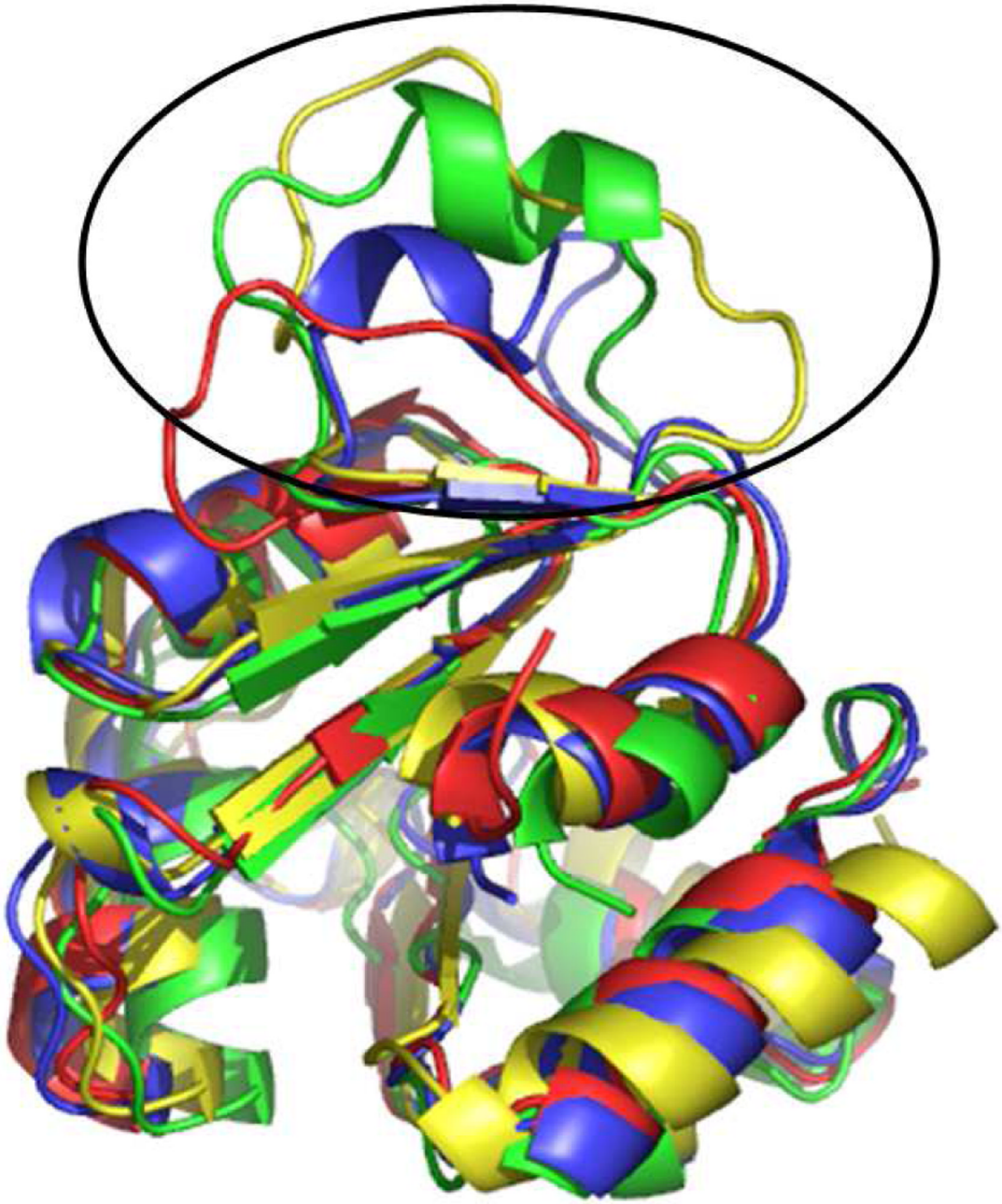

Figure 8.
Target T0303_1 - 2hszA_1. An example of a large best template deletion, partially an α helix, which our method was able to identify and improve with respect to both the best structure and nFOLD model. Native structure is shown in blue, our template is shown in green, best template is shown in red and nFOLD model is shown in yellow. Deletion area is circled.

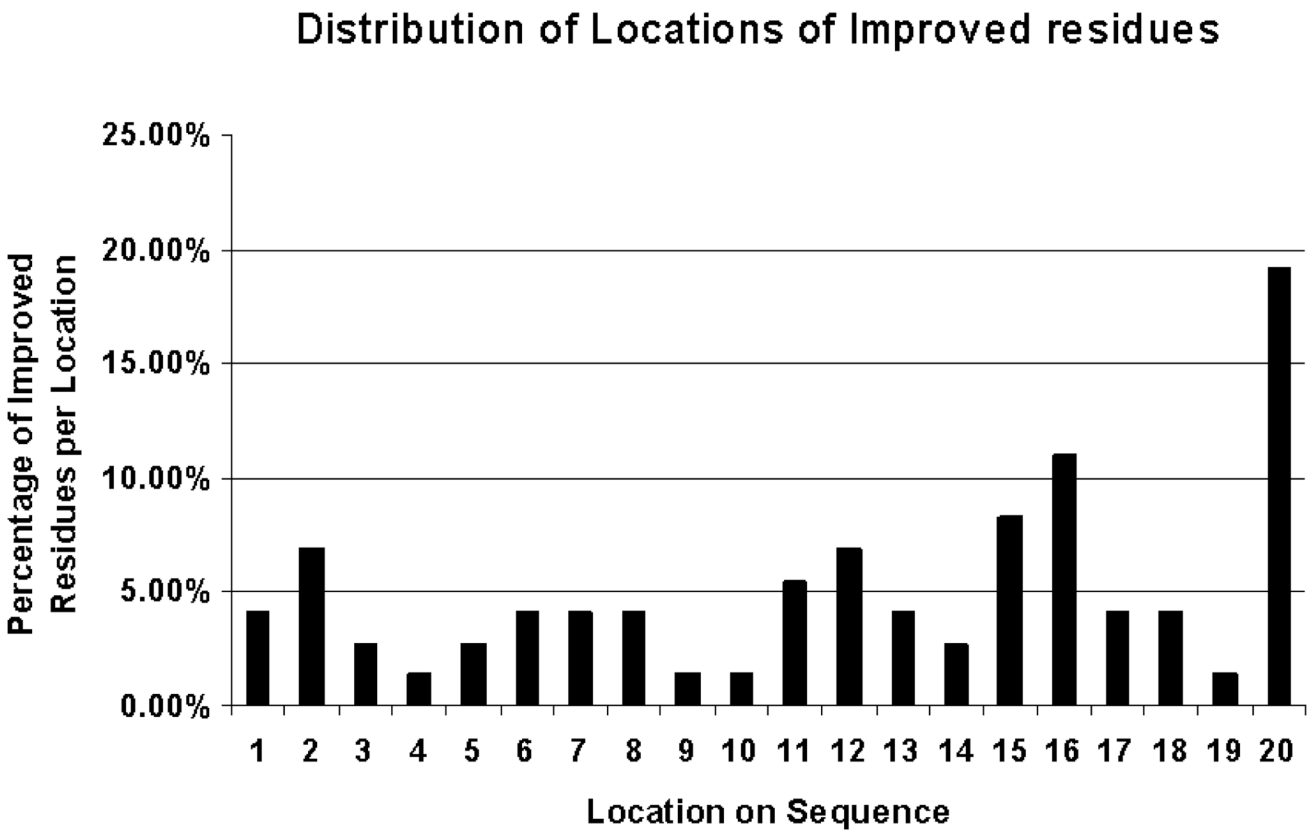

Figure 9.
Distribution of location of improved residues on the target sequence. Target sequences were divided into 20 bins of equal size. Each bin represents a 20-ieth part of a target sequence. The graph presents the percentage of residues improved for this location window out of the total number of improved residues.

Table 1

Results on the easy comparative modeling targets of CASP6

Target	Difficulty Ranking	Length	#Cα under 3.8Å	
			Our Method	Best Template
T0204	43.33	297	best temp. used	267
T0229	12.5	138	best temp. used	121
T0231	4.83	137	best temp. used	130
T0233_1	11.5	66	66	66
T0233_2	15.33	265	241	245
T0246	8.67	354	338	332
T0247_3	40.0	76	68	68
T0264_1	25.67	116	102	101
T0266	31.67	150	best temp. used	144
T0268_2	15.83	109	103	103
T0269_1	24.0	158	145	143
T0271	43.33	161	best temp. used	134
T0274	42.33	159	best temp. used	143
T0275	35.33	135	106	122
T0276	50.00	168	152	149
T0277	19.17	117	best temp. used	112
T0282	67.67	324	263	245

Table 2

Results on CASP7 HA targets. Targets with gaps between native structure and best template appear before the rest of the targets, in order to separate the two groups. For each subgroup, the average percentage of correctly modeled residues is presented. Targets that were modeled using only the best template are indicated by an asterisk

Target	Length	#Cα under 3.8Å			
		our template	best template	nFold top ranking model	ROBETTA top ranking model
T0292_2	171	152	150	144	151
T0295_2*	95	94	94	94	88
T0303_1	129	122	115	117	116
T0308	165	148	160	155	142
T0313	316	289	293	291	283
T0328*	295	281	281	281	275
T0332	159	146	145	131	146
T0334*	527	507	507	507	501
T0339_2	247	224	221	224	228
average performance	**-**	**93.1%**	**93.1%**	**91.6%**	**91.0%**
T0288	91	87	82	84	84
T0290	173	172	172	162	172
T0291	281	266	264	239	267
T0292_1	80	78	78	78	78
T0295_1*	176	175	175	175	175
T0302	134	128	128	128	128
T0305	278	267	270	252	268
T0315	248	246	245	245	246
T0317	149	144	142	140	140
T0340	90	85	85	85	83
T0345	185	182	180	180	183
T0346	172	165	172	162	172
T0359	93	82	87	84	82
T0366	92	84	84	76	87
average performance	**-**	**95.9%**	**96.0%**	**93.3%**	**95.9%**