PROTEINS: Structure, Function, and Bioinformatics 63:457–465 (2006)

# Benchmarking of Dimeric Threading and Structure Refinement

**Vera Grimm, Yang Zhang, and Jeffrey Skolnick***
*Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York*

***ABSTRACT*** **The understanding of protein–protein interactions is a major goal in the post-genomic era. The prediction of interaction from sequence and the subsequent generation of full-length dimeric models is therefore of great interest especially because the number of structurally characterized protein–protein complexes is sparse. A quality assessment of a benchmark comprised of 170 weakly homologous dimeric target–template pairs is presented. They are predicted in a two-step method, similar to the previously described MULTIPROSPECTOR algorithm: each target sequence is assigned to a monomeric template structure by threading; then, those templates that belong to the same physically interacting dimer template are selected. Additionally we use structural alignments as the "gold standard" to assess the percentage of correctly assigned monomer and dimer templates and to evaluate the threading results with a focus on the quality of the alignments in the interfacial region. This work aims to give a quantitative picture of the quality of dimeric threading. Except for one, all monomer templates are identified correctly, but approximately 40% of the dimer templates are still problematic or incorrect. Preliminary results for three full-length dimeric models generated with the TASSER method show on average a significant improvement of the final model over the initial template. Proteins 2006;63:457–465.**


**Key words: protein–protein interactions; multimeric threading; template quality; dimer model; TASSER**

## INTRODUCTION

Protein–protein interactions play an essential role in biological processes. In yeast, a lower bound of 30,000 binary interactions (nine partners per protein) has been estimated;[1–3] although, other studies suggest that only approximately 10,000 different types of protein interactions exist in nature, from which only 2000 are known, with a growth rate of 200–300 per year.[4] This shows that our knowledge of protein interactions is sparse, especially when compared to the number of fully sequenced genomes or the known three-dimensional structures of small proteins.[5,6]

The prediction of protein–protein interactions is therefore of enormous interest. Structure-based approaches like docking[7–9] as well as sequence-based methods[10–15] aim to predict protein–protein interactions. With the growing amount of available sequences, the interest in those techniques is increasing. Most sequence-based techniques typically assign a known template structure to a target of unknown structure with sufficient homology, that is, the interaction of a known structure is used to predict a new interaction. A crucial point is whether the homologous proteins will interact in a similar way. Recent studies suggest that above a sequence identity of 25%, pairs of proteins tend to interact in a similar manner, although exceptions are possible.[13] The recently introduced prediction method MULTIPROSPECTOR[1,16] extends this idea to infer interactions between only weakly homologous, but structurally similar, complexes. The approach is based on threading, the attempt to align a sequence to a library of known folds and find the best match. The advantage over comparative modeling is that analogous as well as homologous structures can be recognized (proteins that share a structural but not necessarily an evolutionary relationship).

The logical continuation of the prediction of protein–protein interactions is the generation of full-length models with a low root-mean-square deviation (RMSD) from their native structure. High-quality models of protein–protein complexes could provide essential insights into their structure and function when no experimental information is available, and could be useful for a better understanding of the association process or the development of specific inhibitors targeting protein–protein interfaces. Commonly, the pair of interacting proteins are modeled separately using homology modeling techniques,[17] and empirical potentials are then used to assess how well a pair of protein models fit into the selected homologous complex.[11,12] Here, we intend to construct full-length models directly from a dimeric template structure using the TASSER structure prediction and refinement procedure[2] and focus on the analysis of weakly homologous dimeric target–template pairs, predicted by inferring the interaction from experimentally known dimer templates to pairs

Grant sponsor: the Division of General Medical Sciences of the NIH; Grant numbers: GM-48835, GM-37408.

*Correspondence to: Jeffrey Skolnick, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street NW, Atlanta, GA 30318. E-mail: skolnick@gatech.edu

Received 30 July 2005; Revised 30 September 2005; Accepted 31 October 2005

Published online 3 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20878

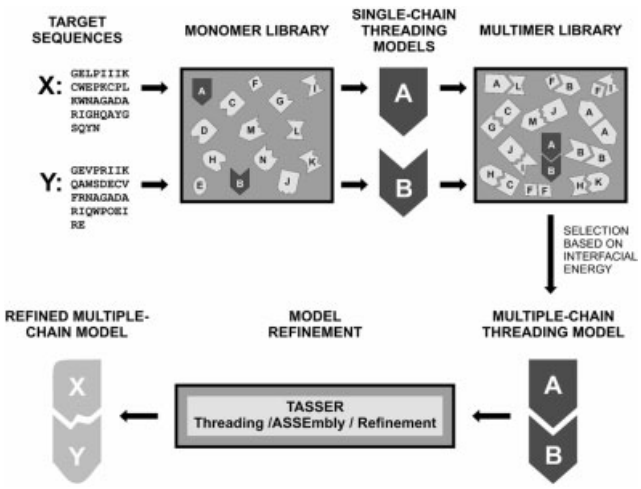


Fig. 1. Overview of the prediction method described in the text.

of target sequences. The prediction methodology consists of two logical steps: first, threading[18] assigns single-chain templates to each target sequence, then pairs of targets where the single-chain templates belong to the same physically interacting dimer are selected. The dimer template of the highest score is considered as the initial model. The subsequent step, model refinement, is strongly dependent upon the quality of the selected templates. We therefore concentrated on four questions: how similar are the native target structures compared to the templates; that is, what percentages of monomeric and dimeric templates are correctly identified? How well can the threading method reproduce the structural alignment (the "gold standard"), with special focus on the interface region and what can one expect from dimeric model refinement using TASSER?

## MATERIALS AND METHODS

Figure 1 presents an overview of the methodology consisting of the assignment of single chain template structures to targets and then the selection of those templates, which belong to the same dimer and finally model refinement.

### Single Chain Assignment

A set of target sequences (described below) is threaded against a nonhomologous template library of the Protein Data Bank (PDB)[19] (on a sequence identity level of 35%) using the previously described threading method PROSPECTOR_3.[18] The template library consists of single chains from monomeric and multimeric proteins. For each target, the 10 best templates with the highest $Z$-scores are selected. The $Z$-score is defined as:

$$Z_k = \frac{E_k - \langle E \rangle}{D}$$

where $E_k$ is the energy of the target sequence in the $k$th template, $\langle E \rangle$ is the average energy, and $D$ is the standard deviation of the energies. The $Z$-score gives the average number of standard deviations between the $k$th template and the average random energy.[20] The target sequence is then threaded against all homologous structures of the template, because the representative template might not be part of a multimeric complex but one of the homologous structures could be.

### Dimer Assignment

To differentiate those threading templates that are part of a complex from those that are not, a dimer template library is generated (described below). Dimer templates sharing more than 35% sequence identity with the target are removed. All pairwise target-template alignment combinations with both single chain templates originating from the same dimer are identified and the interfacial energy is calculated by applying a residue-based statistical potential.[1] If the resulting interfacial energy of a pair of targets belonging to the same physically interacting dimer templates has an interfacial energy below $-10$, then the two targets are assumed to interact. The best dimer template with the highest $Z$-score of the single-chain templates and the lowest interfacial energy is selected and chosen as an initial dimeric model. Only templates with a $Z$-score above 7 are considered because they have a good chance of being correct.[18]

### Dimer Model Refinement

The recently described TASSER methodology[2,21,22] is used for model refinement. TASSER constructs full-length models by reassembling continuous fragments taken from the threading templates whose refinement is driven by an optimized, $C_\alpha$ and side-chain center of a mass-based force field. The unaligned regions are generated from scratch by TASSER and serve to connect the aligned fragments. Currently, TASSER can only handle single-chain proteins; therefore, a flexible linker consisting of 30 Glycine residues between the two chains of the dimeric proteins is introduced. The amino acid Glycine is chosen because of its great flexibility and low tendency to form helices, which could reduce the flexibility of the linker. Thirty residues were found to be long enough to span the spatial distance between terminals of most dimeric chains and also not to increase too much the computational cost of TASSER modeling. Ultimately, we plan to generalize TASSER and remove the need for a linker; here, we are engaged in a preliminary investigation to examine whether such an extension of TASSER would be worthwhile.

### The Dimer Template Library and the Test Set

A library of physically interacting dimers is derived based on the structural data provided by the Protein Quaternary Structure Server (PQS).[23] Because the details of the library construction will be described in detail elsewhere (Grimm and Skolnick, in preparation), here, we only briefly summarize the procedure. All multichain entries from PQS are downloaded and processed according to the following rules: protein chains with less than 30 amino acids, protein–protein complexes with less than 30 interface contacts, DNA/RNA constructs and lyzosymes,

which are most often monomeric, are discarded. Interfacial residues are defined as those that are part of different protein chains and that have at least one heavy atom in contact with one or more heavy atoms of any other residue on the other chain (pairwise distance less than 4.5 Å). At present, we restrict ourselves to dimeric interactions between proteins. We find 4313 homodimers and 647 heterodimers; they consist of 1029 nonredundant homodimers and 352 nonredundant heterodimers (at a sequence identity cut off level of 35%). Four hundred forty-two nonredundant dimers (406 homodimers and 36 heterodimers) are randomly from chosen this dimer library as targets with known native structure to test the methodology described above.

### Structural Alignments Form the "Gold Standard"

Structural alignments are viewed as the "gold standard" in sequence alignment and threading. They define which proteins can be aligned at all and also allow the assessment of the quality of a threading alignment.

First, we perform structural alignments of the single chains of the native target structure and the template to estimate the percentage of correctly identified folds. We are then interested in the percentage of target pairs whose native structures correspond to a structurally related but only weakly homologous dimer template in the dimer library. This defines the maximum number of predictable dimers from this benchmark set by any threading algorithm. Therefore, structural alignments between every native target structure and each template with respect to the protein–protein complex as an entity are employed in the structural alignment procedure. The two chains of each dimer are concatenated, because the used alignment method can only deal with single chains. The percentage of identifiable templates and the fraction of correctly identified templates are determined. We are also interested in the extent of interface similarity between the target and template. Thus, the extracted interface residues from the native target and the template are structurally aligned as well.

All the aforementioned alignments originate from structural superposition and are thought to be optimal. To assess the performance of the threading algorithm, these structural alignments are compared to the threading alignments. Special attention is paid to the interface region. The fraction of correctly aligned interface residues in the threading alignment with respect to this "perfect" structural alignment based interfacial alignment is then calculated.

### The TM-Score and TM-Align

To assess the quality of two aligned structures with an a priori specified alignment, we use the previously described scoring function, TM-score,[24] which is defined as

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_{\text{Target}}}\sum_{i=1}^{L_{\text{ali}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{Target}})}\right)}\right]^2$$

**TABLE I. Templates with the Highest TM-Score Obtained by a Structural Comparison of Every Native Target Structure with Each Template from the Dimer Library**

| | Templates with highest TM-score | Templates with highest TM-score and nic > 0 |
|---|---|---|
| <seq. id.>[a] | 11% ± 6% | 14% |
| <RMSD>[b] | 5.1 Å | 4.8 Å |
| <coverage> | 65% | 76% |
| <TM-score> | 0.54 | 0.60 |
| $N_{\text{TM-score}>0.5}$[c] | 177(40%) | 159(36%) |
| $N_{\text{TM-score}>0.4}$ | 355(80%) | 239(54%) |
| $N_{\text{TM-score}>0.3}$ | 442(100%) | 279(63%) |
| <nic>[d] | 13% | 22% |

[a]Average sequence identity (seq.id.).
[b]Average root-mean-square deviation (RMSD) to native in the aligned region.
[c]Number of target–template pairs with a TM-score below or above the given threshold.
[d]Average percentage of native interface contacts (nic).

Where $L_{\text{Target}}$ is the length of the native target structure; $L_{\text{ali}}$ is the number of aligned residues; $d_i$ is the distance between the $i$th pair of aligned residues, and $d_0\,(L_{\text{Target}}) = 1.24(L_{\text{Target}} - 15)^{1/3} - 1.8$. The TM-score has values between 0 and 1, with more similar structures or better templates having higher TM-scores and with values below 0.17 corresponding to a random prediction. For all target–template pairs found by PROSPECTOR_3, we calculate the TM-score in the aligned region between the native target structure and the template.

For the structural alignments, we use the TM-align[25] method. TM-align aims to find the best structural alignment between two proteins. It is based on the idea that the optimal superposition of two structures corresponds to its maximal TM-score that is identified by a heuristic implementation of the TM-score rotation matrix. This TM-score is used to assess the quality of the resulting structural alignments.

## RESULTS

### Structure Comparison: Native Target Structures versus Dimer Template Library

We are interested in the percentage of targets for which only weakly homologous dimer templates with structural similarity can be found in the dimer library. Table I and Figure 2 show the results for the dimer templates with the highest TM-score obtained by a structural comparison of every native target structure with each template. The dimers are compared by TM-align as entities with the residue numbers of two chains being reordered sequentially. The average RMSD is 5.1 Å, with an average coverage of 65%. The average TM-score is 0.54, with all templates having a TM-score above 0.17.

Because TM-align does not distinguish between different protein chains, sometimes the structural alignments of the complex sometimes show a tendency where one chain from the native target structure is aligned to parts of both chains belonging to the template. Although both dimers share a certain degree of structural similarity and belong

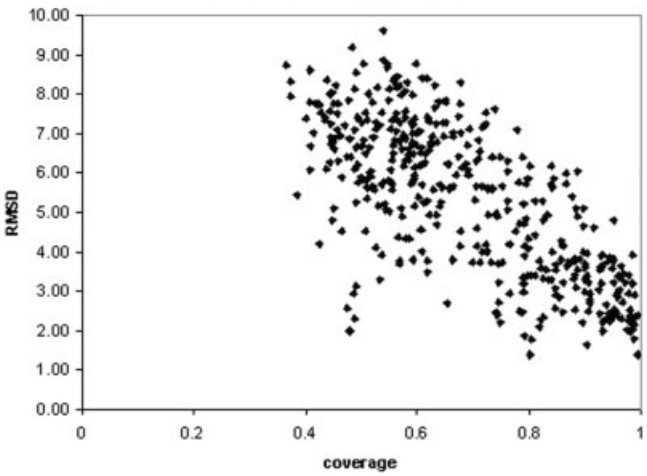


Fig. 2. RMSD to native of the templates identified by the structural alignment of every native target structure with the selected each dimer template versus the alignment coverage.

**TABLE II. Summary of the Native Interface Contact Results for the Templates with the Highest TM-Score Obtained by a Structural Comparison of Every Native Target Structure with Each Template from the Dimer Library**

| | Templates with highest TM-score |
|---|---|
| $N_{nic>0}$‡ | 278(63%) |
| $N_{nic>10\%}$ | 154(35%) |
| $N_{nic>20\%}$ | 112(25%) |
| $N_{nic>0,TM\text{-}score>0.5}$ | 159(36%) |
| $N_{nic>20\%,\ TM\text{-}score>0.5}$ | 110(25%) |

[a]Number of templates with the specified criteria.

to the same family, the chain orientations and the interfaces are different. We therefore also analyze the percentage of native interface contacts, assuming this to be a better measure of the quality of the structural interface. Target–template pairs with no single native interface contact are most likely to have no common interface. Only 278 templates (63%) have more than zero native interface contacts. Excluding templates without native interface contacts lowers the average RMSD to 4.8 Å with an average coverage of 76%. The average TM-score is 0.6. Nevertheless, only 110 templates have a very confident TM-score above 0.5 and more than 20% native interface contacts (compare Table II).

## Overall Prediction Results

One hundred seventy dimers of the benchmark set (39%) are predicted as interacting by threading. But assuming that at most 278 of all benchmark dimers could have been predicted; over 60% of the possible number of interacting partners is identified correctly based on weakly homologous templates. Of the 170 dimers, 160 are homodimers and 10 are heterodimers, in total belonging to 180 different protein chains. The list of proteins used in this study as well as corresponding modeling results can be found at http://www.bioinformatics.buffalo.edu/dimers/.

**TABLE III. Summary of Results from TM-Align (Structural Alignments of the Single Chains) and PROSPECTOR_3 (Threading Alignments)**

| | TM-align | PROSPECTOR_3 |
|---|---|---|
| <seq. id.>[a] | 22% ± 7% | 24% ± 10% |
| <RMSD>[b] | 2.7 Å | 6.2 Å |
| <coverage> | 84% | 84% |
| $N_{RMSD<6.5}$[c] | 180 (100%) | 119 (66%) |
| $N_{RMSD<6.0}$ | 179 (99%) | 112 (62%) |
| $N_{RMSD<5.0}$ | 176 (98%) | 97 (53%) |
| $N_{RMSD<4.0}$ | 167 (93%) | 68 (38%) |
| $N_{RMSD<3.0}$ | 129 (72%) | 38 (21%) |
| $N_{RMSD<2.0}$ | 29 (16%) | 3 (5%) |
| <TM-score> | 0.72 | 0.62 |
| $N_{TM\text{-}score>0.5}$[d] | 170 (94%) | 136 (76%) |
| $N_{TM\text{-}score<0.17}$ | 1 (0.7%) | 8 (4%) |

[a]Average sequence identity (seq.id.).
[b]Average root-mean square deviation (RMSD) to native in the aligned region.
[c]Number of target–template pairs with RMSD below the specified distance (Å).
[d]Number of target–template pairs with a TM-score below or above the given threshold.

## Assignment of Monomers to Single-Chain Templates

### *Structural alignments (single chains)*

We first analyze the percentage of correctly identified folds by structural alignments of the single chains between the native target structure and the threading identified template. Table III shows the results for the 180 monomers where the target dimers have been predicted as interacting by PROSPECTOR_3. All structural alignments have a RMSD below 6.5 Å; the average RMSD is 2.7 Å, and the average coverage (fraction of residues aligned relative to the target protein's length) is 84%. The folds of the target structure and the template are on average very similar, with the sequence identities between the target and template ranging from 10 to 35%. However, the RMSD value can be a misleading measure; it is dependent on the length of the aligned region.[27] The TM-score, which balances coverage and accuracy, is therefore used. 94% of the structural alignments of the single chains have a TM-score above 0.5; the average is 0.72 (see Table II and Fig. 3). Only one template shows a TM-score below 0.17 (random), indicating that for this case the fold was not correctly identified despite the relatively low RMSD of 4.6 Å over 39 aligned residues.

### *Threading alignments*

Table III also shows a summary of the threading results from PROSPECTOR_3 for monomers. 66% of the templates have a RMSD below 6.5 Å to native in the aligned region, a result consistent with the benchmark.[18] The average RMSD is 6.2 Å, with an average coverage of 84%. The sequence identities between target and template are between 12 and 35%. In contrast to the structural alignment, only 76% of the target–template pairs have a TM-score above 0.5 in the regions aligned by PROSPEC-

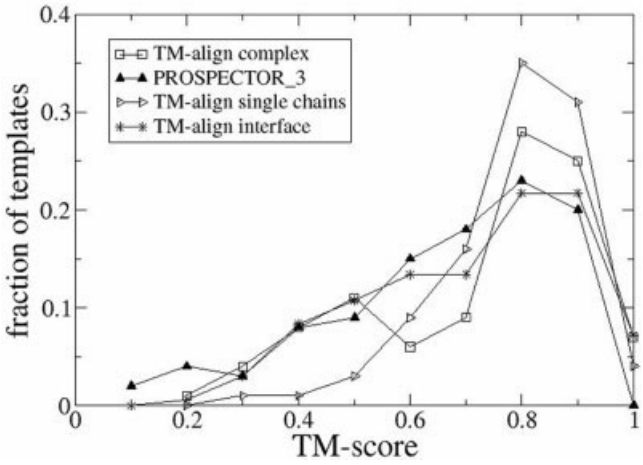


Fig. 3. Distributions of the TM-score for structural alignments of the interface, the complex, the single chains and from the threading alignment provided PROSPECTOR_3 for monomers.

**TABLE IV. Summary of Results from TM-Align (Structural Alignments of the Complex) and PROSPECTOR_3 (Threading Alignments)**

| | TM-align | PROSPECTOR_3 |
|---|---|---|
| <coverage> | 77% | 84% |
| <RMSD>[a] | 3.5 Å | 10.8 Å |
| $N_{RMSD<6.5}$[a] | 167 (98%) | 94 (55%) |
| $N_{RMSD<6.0}$ | 170 (100%) | 88 (52%) |
| $N_{RMSD<5.0}$ | 156 (92%) | 65 (38%) |
| $N_{RMSD<4.0}$ | 128 (75%) | 43 (25%) |
| $N_{RMSD<3.0}$ | 65 (38%) | 29 (17%) |
| $N_{RMSD<2.0}$ | 8 (5%) | 8 (5%) |
| <TM-score> | 0.66 | 0.56 |
| $N_{TM\text{-}score>0.5}$[c] | 126 (74%) | 93 (55%) |
| $N_{TM\text{-}score<0.17}$ | 1 (0.6%) | 12 (7%) |

[a]Average root-mean-square deviation (RMSD) to native in the aligned region.
[b]Number of target–template pairs with RMSD below the specified distance (Å).
[c]Number of target–template pairs with a TM-score below or above the given threshold.

TOR_3; the average TM-score is 0.62 (see Table III and Fig. 3). In summary, PROSPECTOR_3 could identify all but one fold correctly, although the alignment accuracy shows room for improvement in approximately one-third of the cases.

### Assignment of Pairs of Chains to Dimer Templates

#### *Structural alignments (complexes)*

Next, we are interested in the percentage of correctly *assigned* dimer templates. Structural alignments of the native target structure and the selected dimeric template were performed with the results summarized in Table IV. An average coverage of 77% and an average RMSD of 3.5 Å are found. The number of templates with a RMSD below 6.5 Å to native is 167 (98%). We again examine the TM-score to better estimate the number of incorrect templates. The structural alignments of the complexes result in an average TM-score of 0.66 with 126 (74%) templates having a TM-score above 0.5. Only one dimer template has a TM-score below random (0.17); this corresponds to the same template for which the monomer template is incorrectly identified. Ninteen percent of all dimer templates show no single native interface contact; these templates are most likely incorrect. An additional 15% of the structural complex alignments have less than 20% native interface contacts, which indicate small common interfaces; theses cases could be incorrect (see Table V). The three main reasons for native interface contacts between 0 and 20% are different interface topologies, different chain orientations [see Fig. 4(a)–(b)] and multidomain dimers that can fall in both categories. For the latter, frequently, only one domain is aligned, and no or only small parts of a common interface are observed. One quite unusual multidomain dimer example can be seen in Figure 4(c)–(d). The target consists of two chains, each with two domains. The template has also two chains, but only one contains two domains. Parts of two separate domains from the native target structure seem to have fused to result in only one domain in the template. Only 5% native interface contacts are observed, although both structures share structural similarity.

**TABLE V. Native Interface Contacts (nic) for the Structural Alignments of the Complex, the Interface, and Comparison to the PROSPECTOR_3 Alignment**

| | Complex | Interface | PROSPECTOR_3 |
|---|---|---|---|
| <nic>[a] | 32% | 33% | 28% |
| $<nic>_{nic>0}$[b] | 40% | 40% | 35% |
| $<nic>_{nic>20\%}$ | 47% | 45% | 45% |
| $N_{nic=0}$[c] | 32 (19%) | 2 (1%) | 52 (30%) |
| $N_{nic>20\%}$ | 110 (65%) | 110 (65%) | 84 (50%) |

[a]Average percentage of native interface contacts (nic).
[b]Average percentage of native interface contacts (nic) calculated only for those templates within the specified threshold.
[c]Number of target–template pairs with the specified criteria.

#### *Threading alignments*

Table IV also shows the PROSPECTOR_3 alignment results for the 170 dimers that are predicted to have interactions. The average coverage is 84%, with an average RMSD of 10.05 Å. Ninety-four templates (55%) have a RMSD below 6.5 Å to native. In comparison to the structural alignments, only 55% of the PROSPECTOR_3 results have a TM-score above 0.5, with an average TM-score of 0.56.

### Native Interface Contacts

#### *Structural alignments (complexes)*

The percentage of native interface contacts reflects the alignment quality in the interface region. Table V summarizes the results for all alignments. The average percentage of correct native interface contacts is 32% when the dimer complex is treated as a single entity in the structural alignment. Excluding all target–template pairs with zero native interface contacts raises the average percentage of correctly predicted native interface contacts to 40%.

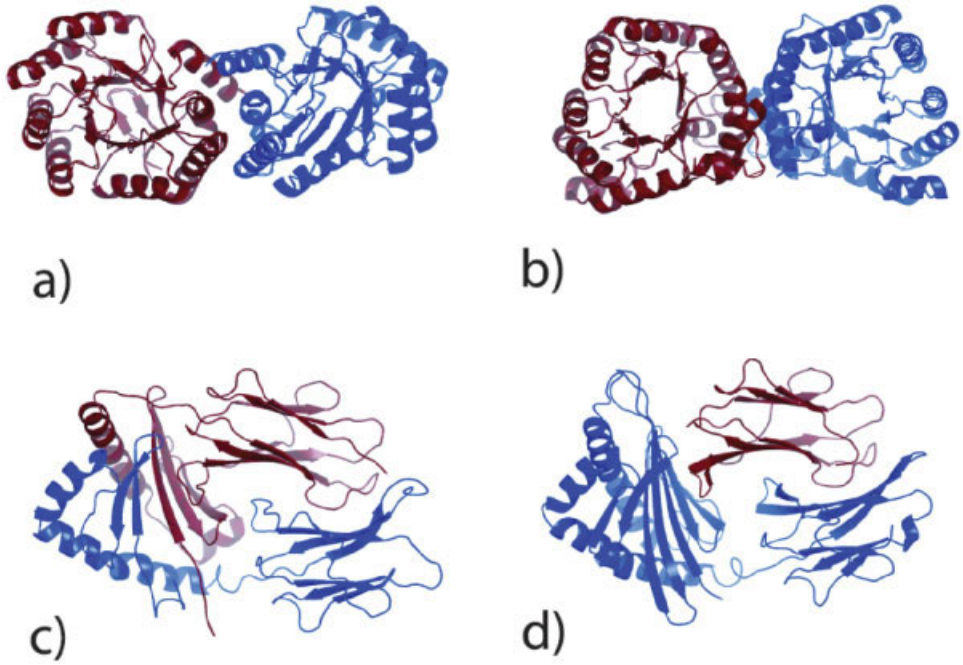


Fig. 4. Problematic cases: native target structure (**a**), and template (**b**), do not share the same interface; one is twisted with respect to the other. Multidomain dimers: native target structure (**c**), and template (**d**), have evolved differently. Parts of two domains from different chains seem to have fused in the template structure. Only a few common interface residues are found. Chain A is colored in red and chain B in blue, respectively.

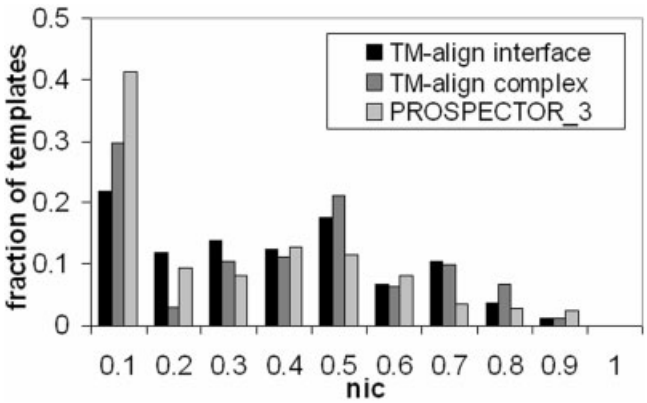


Fig. 5. Histogram of the fraction of native interface contacts (nic) identified from structural alignments of the interface and the complex in comparison with the threading alignment.

### *Threading alignments*

On average, 28% of the native interface contacts are found for the PROSPECTOR_3 alignments; excluding those target–template pairs with zero native interface contacts raises this to 35%, on average. Figure 5 shows a histogram of native interface contacts for the threading and the structural alignments of the interface and the complex. The distributions are quite similar, but the threading method tends to generate approximately 15% more alignments with less than 20% native interface contacts. The percentage of native interface contacts is also strongly dependent on the sequence identity cutoff employed during threading. Including homologous templates increases the average percentage of native interface contacts to 65% (data not shown).

## Accuracy of Interface Alignment

### *Structural alignment (interfaces)*

Next, we restrict ourselves to structural alignments only involving interfacial residues, and the resulting structural interface alignment is assumed to be "perfect." The fraction of correctly aligned interface residues with respect to this "perfect alignment" is determined. The average percentage of correctly aligned interface residues is 45% for the structural superposition of the single chains. Discarding templates with no native interface contacts raises the average to 62%.

### *Threading alignment*

The average percentage of correctly aligned interface residues for the threading results is 34% and rises to 54% if only cases with correct templates are considered. Figure 6 shows a histogram of the correctly aligned interface residues in the threading alignment in comparison with the structural alignment of the single chains. The threading method results in ~15% more alignments with less than 20% correctly aligned interface residues compared to the structural alignment. We are not only interested in whether the interface residues are aligned exactly to the correct position but also if the interface alignment is approximately correct. This information could be enough for subsequent refinement using TASSER. We therefore shift the alignment by one residue in each direction and recalculate the percent of correctly identified native inter-

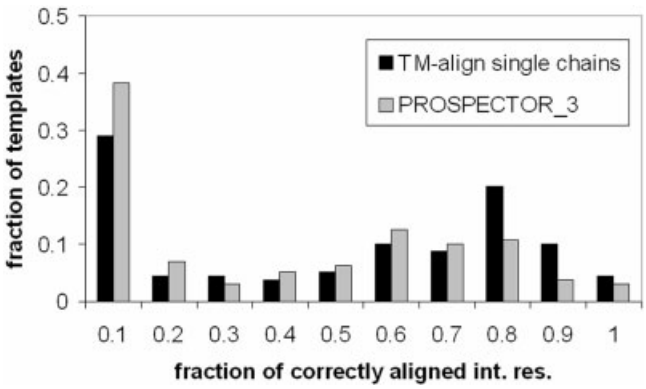


Fig. 6. Histogram of the fraction of correctly aligned interface residues (int.res.) with respect to the structural interface alignment, which is assumed to be "perfect." Shown are the structural alignments for the single chains and the threading alignments.

face contacts. The average percentages of native interface contacts for a simple sequence shift within ±1 residues are 46%. This rises to 65% if only cases with correct templates are considered.

### Preliminary Results: TASSER Refinement of Dimeric Models

For a first, very preliminary test of the refinement of dimeric models, three examples with optimal conditions are chosen. The selected target-template pairs have overall good threading alignments, sequence identities between 34 and 23%, similar chain orientations, and similar interface topologies in the native target structure and the template. Table VI summarizes the results for the dimer model refinement of these three examples. Each of the refined three models shows a lower RMSD to native in the same aligned region than the best template. The RMSDs to native calculated over the entire complex are 2.4, 3.3, and 4.8 Å, respectively.

The first target originates from the homodimeric human Grancalcin protein (1alw). The template is the dimeric Calpain from *Sus scrofa* (1f40). Both belong to the penta-EF-hand protein family and have a sequence identity of 34% and a high percentage of native interface contacts (75%). The RMSD of the best TASSER model is very good, with an interface RMSD of 1.49 Å and a TM-score of the interface of 0.63. The global RMSD of the complex, evaluated over the threading aligned residues, drops from 7.9 to 2.4 Å. Figure 7(a) shows a comparison of the best TASSER model to the native structure.

The next case is the homodimeric malate dehydrogenase from *Aquaspirillum acticum* (1b8v) as the target and a malate dehydrogenase from the same family from *Sus scrofa* (1mld) as the template. The sequence identity is 27%, and 36% of the native interface contacts are found. In this case, the RMSD of the model is 4.93 Å compared to the best template (5.01 Å in the same aligned regions). The introduced Glycine linker [not shown in Fig. 7(b)] distorts the interface. Figure 7(b) shows the best TASSER model and the native structure.

The third target is the homodimeric cytochrome *c′* from the denitrifying bacterium *Alcaligenes xylosoxidans* (1e84), with the template being the homodimeric cytochrome *c′* (1bbh) from the purple phototrophic bacterium *Cromatium vinosum*. They share a sequence identity of 25%, 65% of native interface contacts and belong to the same family. The final model has an RMSD of 3.26 Å compared to native. In this case, TASSER improves the RMSD of the template aligned residues from 8.61 to 3.26 Å. Figure 7(c) shows superposition of the best TASSER model onto the native structure.

## DISCUSSION AND CONCLUSIONS

One hundred seventy predicted dimeric target–template pairs with less than 35% sequence identity have been analyzed as a prequel to a full-length model refinement with TASSER. Preliminary results for three dimeric models from TASSER are shown as well. All but one of the monomeric templates are correctly identified by the threading method and the quality of the alignments is comparable to the structural alignments of the single chains for two-thirds of the cases. The one incorrectly assigned fold is comprised of a mainly beta, extracellular domain of a tumor necrosis factor from *Homo sapiens* (1ext, 158 residues) as the target and an antifreeze, all-beta protein (1ezg, 82 residues) forming right-handed beta helices as the template. Motif searches in PROSITE[26] for the target as well as for the template results in the same Cysteine rich profile with a high score. This similar sequence profile could be the reason why PROSPECTOR_3's scoring function incorrectly assigns this particular fold to the target, because, among other terms, it is based on sequence profiles.

Nineteen percent of the dimeric templates provided by threading are most likely incorrect, and an additional 15% of the templates have less than 20% of the native interface contacts correctly identified. Nevertheless, the preliminary results from the TASSER model refinements are encouraging. For the three selected cases with optimal conditions, the final models show clear improvement over the best templates. In one case, the introduced Glycine linker between the chains did influence the interface topology. A binding loop is pulled away from the interface to make room for the linker. This indicates that a more sophisticated method has to be developed to explicitly introduce multichain proteins to the TASSER methodology; this is currently underway. This target–template pair has only 36% of the native interface contacts recovered by the threading alignment; nevertheless, a medium resolution (4.8 Å to native) is recovered in spite of the distorting Glycine linker. This is encouraging as well, indicating that around 35% correctly identified native interface contacts can be sufficient for TASSER refinement. Because these three cases share the same chain orientation in the native target structure and the template, the capability of the TASSER method to correct chain orientations has to be analyzed in subsequent work.

It is to be expected that at least for half of the target–template pair's full-length dimeric models with low or medium RMSD to native will be predicted. Although no detailed study has been done for homologous pairs, the

**TABLE VI. Analysis of the Best TASSER Models for Three Target–Template Pairs**

| | Overall structure | | | | Interface | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Best template | | Best TASSER model | | Best template | | Best TASSER model | | | |
| Target–template pair | TM-score | RMSD (Å)[a] | TM-score | RMSD (Å) | TM-score | RMSD (Å) | nic (%)[b] | TM-score | RMSD (Å) | nic (%)[b] |
| 1alw-1f4o | 0.51 | 7.9 | 0.9 | 2.41 | 0.49 | 3.0 | 66 | 0.63 | 1.49 | 70 |
| 1b8v-1mld | 0.65 | 5.01 | 0.81 | 4.93 | 0.55 | 4.3 | 31 | 0.57 | 3.48 | 37 |
| 1e84-1bbh | 0.67 | 8.61 | 0.79 | 3.26 | 0.53 | 3.7 | 62 | 0.53 | 1.01 | 68 |

[a]Average root-mean-square deviation (RMSD) to native in the same aligned region.
[b]Native interface contacts (nic) for the interface of the best TASSER model.

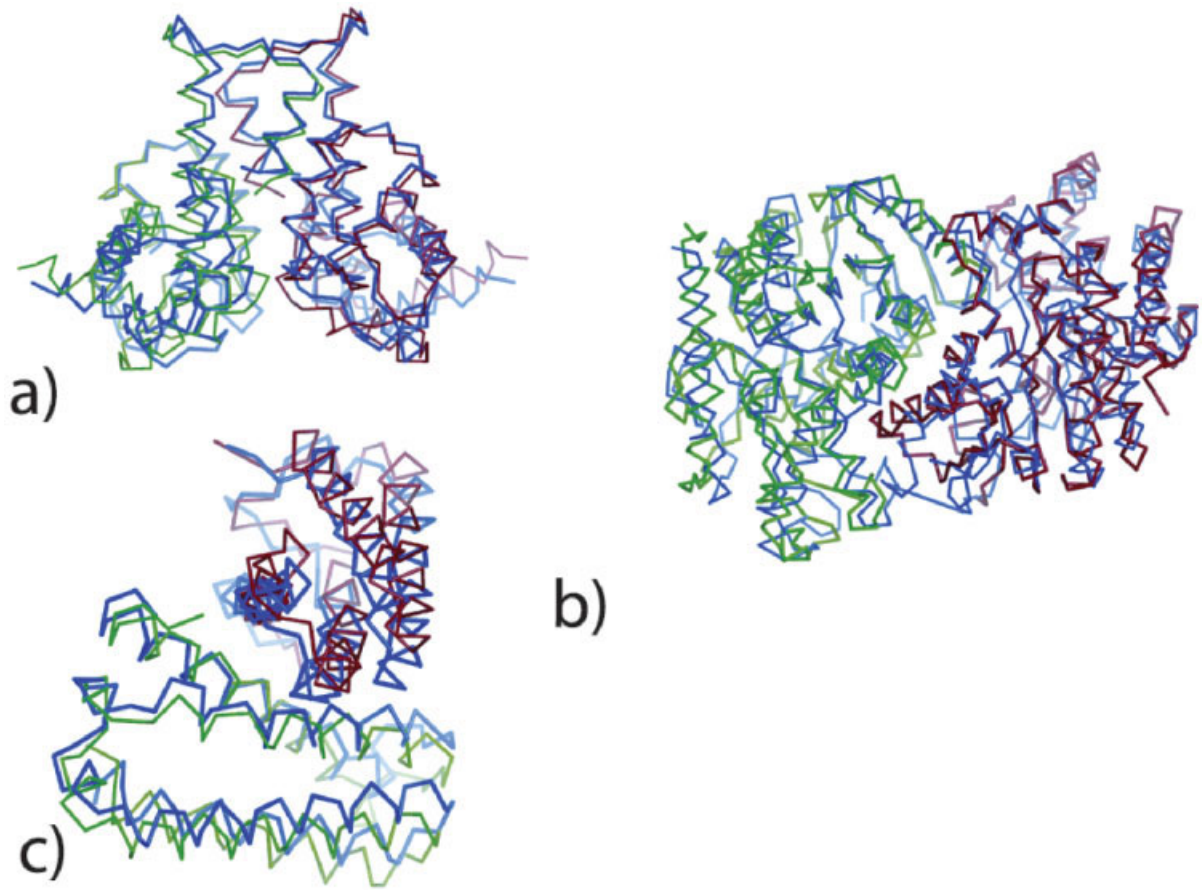


Fig. 7. Best TASSER models (colored in blue) in comparison to the native target structure (chain A in green and chain B in red, respectively). Shown are the backbones only. (**a**) Target: human Gancalcin (1alw), template: Calpain from *Sus scrofa* (1f4o); (**b**) target: malate dehydrogenase from *Aquaspirillum acticum* (1b8v), template: malate dehydrogenase from *Sus scrofa* (1mld); (*c*) target: cytochrome *c′* from the denitrifying bacterium *Alcaligenes xylosoxidans* (1e84), template: cytochrome *c′* (1bbh) from the purple phototrophic bacterium *Cromatium vinosum.*

increase of the average native interface contacts from ~30 to ~65% indicates that even better models when homologous target–template pairs exist in the PDB can be expected.

Some points should be critically addressed. First, the structural alignment method, TM-align may not be entirely suitable for the situation here, because (1) TM-align does not distinguish the chains separately when dimers are aligned; (2) for structural alignments of interface region, the small collections of residues are not always consecutive in sequence, a situation for which TM-align was not designed.

Another important point is the question of why target–templates pairs with less than 10% or even no native interface contact results in an interfacial energy below the interfacial energy threshold of −10. Several things could cause this. The threshold and/or the interfacial potential could be incorrect. Furthermore, it seems possible that the target and template sequence in the interface region are similar, and that the native target structure (or the template) is only twisted. This could lead to good interfacial energies because the calculation is based on the interfacial contacts in the template and the sequence-structure alignment from PROSPECTOR_3. Additionally,

even homologous protein complexes do not necessarily have to share the same mode of interaction.[28] The template (or the target) could also build multiple interfaces at different positions of the surface. The occurrence of a variety of different oligomeric states and interfaces within one protein family is well known,[29] for example, the DJ-1 family.[30] We should also not ignore the possibility that these problematic cases could be caused by incorrectly predicted quaternary structures from PQS. The automatic method to predict the quaternary structure and especially the empirical rules applied to distinguish biological interfaces from crystal artifacts are far from being perfect, with an error rate of approximately 15%.[31]

Future work will include the development of a better refinement procedure for multichain proteins, the extension of the structural alignment method to multichain proteins, and the further development of interfacial potentials.

## ACKNOWLEDGMENTS

We would like to thank Adrian Arakaki for his help with the figures and Olav Zimmermann for stimulating discussions.

## REFERENCES

1. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. Proteins 2002;49(3):350–364.
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.
3. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein–protein interactions. Nature 2002;417:399–403.
4. Aloy P, Russell RB. Ten thousand interactions for the molecular biologist. Nat Biotechnol 2004;22:1317–1321.
5. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;334:793–802.
6. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 2005;102:1029–1034.
7. Janin J. Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci 2005;14:278–283.
8. Szilágyi A, Grimm V, Arakaki A, Skolnick J. Prediction of physical protein–protein interactions. Phys Biol 2005:1–6.
9. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A. A structural perspective on protein–protein interactions. Curr Opin Struct Biol 2004;14:313–324.
10. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999;285:751–753.
11. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99:5896–5901.
12. Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 2003;19:161–162.
13. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–998.
14. Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res 2002;30:255–259.
15. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. Genome Res 2004;14:1107–1118.
16. Lu L, Arakaki AK, Lu H, Skolnick J. Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. Genome Res 2003;13:1146–1154.
17. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
18. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 2004;56:502–518.
19. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 2002;58:899–907.
20. Zhang L, Skolnick J. What should the Z-score of native protein structures be? Protein Sci 1998;7:1201–1207.
21. Zhang Y, Skolnick J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 2004;87:2647–2655.
22. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
23. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
24. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
25. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
26. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 2002;3:265–274.
27. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. Biopolymers 2001;59:305–309.
28. Park SY, Beel BD, Simon MI, Bilwes AM, Crane BR. In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. Proc Natl Acad Sci USA 2004;101:11646–11651.
29. Ponstingl H, Kabir T, Gorse D, Thornton JM. Morphological aspects of oligomeric protein structures. Prog Biophys Mol Biol 2005;89:9–35.
30. Bandyopadhyay S, Cookson MR. Evolutionary and functional relationships within the DJ1 superfamily. BMC Evol Biol 2004;4:6.
31. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.