Homology-based modeling of 3D structures of protein–protein complexes using alignments of modified sequence profiles

Petras J. Kundrotas\textsuperscript{a}, Marc F. Lensink\textsuperscript{b}, Emil Alexov\textsuperscript{a,∗}

\textsuperscript{a} Computational Biophysics and Bioinformatics, Department of Physics, Clemson University, Clemson, SC 29634, United States
\textsuperscript{b} Université Libre de Bruxelles, Center for Structural Biology and Bioinformatics, Boulevard du Triomphant, CP.263, B-1050 Brussels, Belgium

\begin{abstract}
Customary practice in predicting 3D structures of protein–protein complexes is employment of various docking methods when the structures of separate monomers are known \textit{a priori}. The alternative approach, i.e. the template-based prediction with pure sequence information as a starting point, is still considered as being inferior mostly due to presumption that the pool of available structures of protein–protein complexes, which can serve as putative templates, is not sufficiently large. Recently, however, several labs have developed databases containing thousands of 3D structures of protein–protein complexes, which enable statistically reliable testing of homology-based algorithms. In this paper we report the results on homology-based modeling of 3D structures of protein complexes using alignments of modified sequence profiles. The method, called HOMology-BAsed COmplex Prediction (HOMBACOP), has two distinctive features: (I) extra weight on aligning interfacial residues in the dynamical programming algorithm, and (II) increased gap penalties for the interfacial segments. The method was tested against our recently developed ProtCom database and against the Boston University protein–protein BENCHMARK. In both cases, models generated were compared to the models built on basis of customarily protein structure initiative (PSI)-BLAST sequence alignments. It was found that existence of homologous (by the means of PSI-BLAST) templates (44% of cases) enables both methods to produce models of good quality, with the profiles method outperforming the PSI-BLAST models (with respect to the percentage of correctly predicted residues on the complex interface and fraction of native interfacial contacts). The models were evaluated according to the CAPRI assessment criteria and about two thirds of the models were found to fall into acceptable and medium-quality categories. The same comparison of a larger set of 463 protein complexes showed again that profiles generate better models. We further demonstrate, using our ProtCom database, the suitability of the profile alignment algorithm in detecting remote homologues between query and template sequences, where the PSI-BLAST method fails.
\end{abstract}

© 2008 Elsevier B.V. All rights reserved.
recently several labs have developed large databases of 3D structural templates for modeling protein–protein complexes [4–7]. Some of these databases have more than 10,000 entries, a number that seems to be large enough to provide a good pool of structural templates for a significant number of cases. These latest developments became possible because of several reasons. First, the expansion of the structural information deposited in the Protein Data Bank (PDB) [8] already provides a significant number of structural templates. Second, it has become evident that protein–protein complexes can be modeled on structural templates that are not necessarily complexes of two polypeptide chains. Thus, structural templates can be delivered both from mono- and multimeric proteins, which significantly increase the number and diversity of the available templates and provide a good starting point for large-scale template-based structure prediction of protein–protein complexes.

Several groups pioneered the template-based modeling approach. Using a scoring function based on pairing propensities that were determined from the interfaces of existing 3D complexes, Sali et al. [9] evaluated putative complexes comprised of domains found to be homologous to the query sequences. The predicted domain interfaces were integrated in the PIBASE database [7]. Aloy and Russell [10] used 3D structures of protein complexes to deliver empirical potentials and then used these potentials to score putative protein pairs within two protein families. The InterPreTS resource [11] uses a BLAST2 search protocol to reveal pairs of homologous sequences and then employs statistical potentials to validate the putative domain pairs in terms of Pfam [12,13]. This method was recently extended to combine experimental and numerical techniques to predict the domain–domain complexes in yeast [14]. Skolnick and coworkers use threading techniques to predict the structures of full-length protein complexes [15]. MULTIPROSPECTOR [16] threads the two target sequences onto each structure in a representative, non-redundant library of folds. If the query sequences score high with respect to structures forming a dimer, then in all likelihood the target sequences would also form a dimer. The putative model is further validated by a second threading, within the structure of the complex. Thus, in most of the above mentioned methods further verification of the models is obtained either through screening of the models with statistical pairing potentials [11] or by additional threading [16]. Further development resulted to the M-TASSER [17] method that recently has been tested on a benchmark set comprising 241 dimers having templates with weak sequence similarity and 246 without multimeric templates in the dimer library. Of the total of 207 targets predicted to interact as dimers, 165 (80%) were correctly assigned as interacting with a true positive rate of 68% and a false positive rate of 17% [17]. Such procedures are well established in benchmarking docking algorithms by ranking decoys, and on a small scale one could use also full energy functions instead of statistical potentials (see Ref. [18] and the papers cited therein).

Most of the energy functions used to rank decoys and native structures perform well on bound structures, but have much less success in ranking unbound monomers of the complexes where the decoys have quite often an energy that is lower (i.e. more favorable scoring) than that of the native complex. The principle advantage of using homology-based approaches is that a complex is modeled by using the template of another complex, thus taking implicitly into account possible conformational changes [19] occurring during the protein binding. It was also pointed out in review [20] that the homology-based modeling approach is useful in structure determination of macromolecular assemblies on the cell level. The template-based methodology was also used to predict protein interaction networks [21].

The most prominent and widely used method in detecting homology between query and template sequences is the iterative PSI-BLAST alignment algorithm [22]. In our earlier work [23] we demonstrated the capability of that alignment method to predict 3D structures of the protein–protein complexes with good accuracy. However, during benchmarking of the algorithm on our ProtCom [4] database (for details see [23]) we noticed that in many cases alignments produced by PSI-BLAST covered only those parts of the protein sequences that are far away from the complex interface. Moreover, PSI-BLAST fails in many cases to find homologous sequences for both components of the query complex, whereas there are good alignments for one of the components. In this paper we address the above issues by utilizing a profile-to-profile [24] dynamic programming technique to align query and template sequences. The choice of profile-to-profile alignment for the present study was inspired by earlier studies [24–28] on modeling the 3D structures of single proteins which have shown superiority of profile-to-profile alignment methods compared to customarily PSI-BLAST sequence searches, since the profiles may contain additional information that assists in detecting weak relationships between query and template sequences. The additional information used in this study was sequence positions (experimentally determined or predicted by other methods [29]) of residues on the complex interface and we modified the standard dynamic programming algorithm by emphasizing on alignment of the interfacial residues between the query and template sequences. We tested our approach against two different datasets, our own ProtCom40 database and the Boston University benchmarking set [30], and demonstrated that inclusion of the additional information into profiles improves quality of the models in terms of predicting both location of binding pocket and mutual orientation of the complex components. We also show that in absence of close homology relations between query and template sequences, i.e. homology not detectable through PSI-BLAST, the models built still are of good enough quality to determine the location of the binding sites of the complex in question.

2. Methods

2.1. Sets of protein used in calculations

Protein–protein complexes studied in this paper were extracted from the previously developed ProtCom [4] database (release of November 2006) (http://www.ces.clemson.edu/compbio/protcom), which contains more than 3000 entries. To avoid a bias towards overrepresented structures, the entries were purged with the CD-hit program [31] at 40% sequence identity level (note that this requirement automatically removes all homo-complexes). This resulted in 463 protein–protein complexes out of which 123 structures are original PDB files and the rest of two-chain structures were taken from PDB files of multi-chain complexes. This subset of the ProtCom database (hereafter referred to as the ProtCom40 dataset) was used to test the profile-to-profile algorithm described above. Note that in this study artificial “complexes” in the ProtCom database, created from domain–domain structures, were excluded from the dataset for complex benchmarking although those structures were considered as possible templates.

The second set of proteins used in the present paper was extracted from the Boston University benchmark set (protein–protein docking benchmark set), Version 2.0 (http://zlab.bu.edu/zdock/benchmark.shtml). Note that this benchmarking set was developed for testing docking algorithms and it therefore does not necessarily contain proteins for which homologues can be easily found. For the sake of computational efficiency, we excluded from the full dataset all complexes with more than two polypeptide chains, which resulted in 43 cases.
The models of complexes in the Boston University benchmark set were built using the ProtCom database purged at 95% sequence identity (ProtCom95). It was done to increase the pool of available templates. During the modeling the sequences of the complexes in the Boston University benchmark set were excluded from the list of templates.

2.2. Computational details

A flowchart of the HOMology-BAsed COmplex Prediction (HOM-BACOP) [23] method is shown in Fig. 1. A pair of query sequences was separately subjected to the PSI-BLAST search algorithm [22] by means of the program BLASTPGP. The search was carried out against all sequences in the non-redundant database of sequences (http://www.ncbi.nlm.nih.gov/) with the substitution matrix BLOSUM62 [32] and five iterations. The outcome of this step, the position-specific substitution matrices (PSSMs), is enriched with an extra element indicating if the position corresponds to the interfacial or not. The modified profiles were then used in a dynamic programming algorithm (see below). The resulting statistically significant alignments were searched for common hits, i.e. such alignment pairs where template sequences belong to the same complex in our dataset. After creating a list of possible templates, the structures of the components were modeled separately by the program NEST [33] and resulting PDB files were merged into a single file to produce the anticipated three-dimensional arrangement of the modeled complex, which was then further evaluated (see below).

We also attempted to build the models using alignments obtained from the customary PSI-BLAST sequence homology search algorithm, in order to compare the quality of the models (see below) for the same query sequence obtained by means of two alignment protocols. In this case, unmodified PSSM’s were used in sequential BLAST runs against all non-redundant sequences in the PDB data bank.

For remote homologue detection we used slightly modified HOMBACOP protocol where a model is built on the basis of alignments irrespectively whether the condition for statistical significance was satisfied or not. This was done to test the possibility of building reasonable models on low-homology alignments. In all cases we have removed the query complexes from the list of the templates.

2.3. Alignment of modified sequence profiles

We utilize a standard profile-to-profile method for global alignment of sequence profiles using a dynamic programming technique. Initial sequence profiles for both query and template sequences are $N \times 20$ matrices ($N$ being the number of amino acids in a sequence) consisting of the position-specific scoring matrices (PSSM) [22] obtained in a regular PSI-BLAST search. In order to make the method more suitable for the modeling of 3D structures of protein–protein complexes (where one of the major goals is to identify interfaces between two protein molecules) we added a 21st row in the profile matrices (denoted hereafter as $P_x$) containing the following flags:

$$P_x(i, 21) = \begin{cases} 1/\sqrt{|i - i_{\text{int}}| + 1}, & \text{if } |i - i_{\text{int}}| < 3, \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $i_{\text{int}}$ is the position of an interfacial residue (we define a residue as interfacial if any of the heavy atoms of the residue is less than 6Å away from a heavy atom of another complex component) nearest to the residue $i$ and $x$ equals Q or T for query and template sequence, respectively. In this paper we aimed at testing the method on a set of known structures and therefore we adopt here a pragmatic approach and flag interfacial residues of a query complex using the corresponding X-ray structure. However, a natural question arises at this point, as it is straightforward to identify interfacial residues for the template complex, but how to do this for the query complex, the structure one wants to model? There are many methods for predicting interfacial residues [29,34–38], among them our recent study [39] where we present a method for inferring putative interfacial residues by scoring pairs of continuous residue segments using empirical statistical potentials and/or evolutionary-derived substitution matrices (CIIS method). This method yields rather high accuracy (up to 60%) and testing of the method for modeling protein complexes presented in this paper with interfacial residues on the query side predicted by the
CIIS, rather than taken from experimental structures, is the study in progress.

A dynamic programming matrix was constructed using the Needleman–Wunsch algorithm [40] with the affine gap penalty [41] using the following matching matrix:

\[
A(i, j) = \exp[P_Q(i, 21) \times P_T(j, 21)] \sum_{k=1}^{20} P_Q(i, k) \times P_T(j, k),
\]

where indices \( i \) and \( j \) number residues in the query and template sequences, respectively. Note that the actual matching matrix used in the calculations is Z-normalized, i.e.,

\[
A'(i, j) = \frac{A(i, j) - \langle A \rangle }{\sigma},
\]

where \( \langle A \rangle \) and \( \sigma \) are, respectively, the average value and standard deviation of the elements in the matrix \( A \) (2). Such normalization allows using unified parameters for the dynamic programming algorithm for all dataset sequences. The affine gap penalty for deletion (on query side) or insertion (on template side) of \( m \) residues was in our case also defined as a position-specific function

\[
\begin{align*}
\delta_{\text{open}}(i) &= \delta_{\text{open}} + (m - 1) \delta_{\text{ext}} \times \exp(P_Q(i, 21)), \\
\delta_{\text{ext}}(j) &= \delta_{\text{open}} + (m - 1) \delta_{\text{ext}} \times \exp(P_T(j, 21)),
\end{align*}
\]

where \( \delta_{\text{open}} \) and \( \delta_{\text{ext}} \) are gap opening and gap extension penalties, respectively (the choice of values for these parameters is described below in Section 3).

2.4. Statistical significance of sequence alignment

Evaluating statistical significance of the sequence alignment is not a trivial problem [42]. Usually this is done through parameters of the extreme value distribution which are obtained by fitting this distribution to a set of raw scores of all possible alignments in a dataset [43]. This approach performs well for the local-to-local sequence alignments where no negative raw scores are allowed in the algorithm and the resulting raw-score distribution is thus highly asymmetrical. However, in the case of global-to-global alignments this distribution has a large tail towards negative raw scores and therefore a fit to the extreme value distribution (which by definition is performed only for the right tail of the distributions, towards the positive raw scores) artificially enhances the statistical significance of the alignments with small (or even large negative) absolute raw scores.

Therefore, we adopted here a simple pragmatic approach. For each alignment we calculated normalized raw score,

\[
r = \frac{\sum_{k=1}^{L_{ij}} s(Q_k, T_k)}{L_{ij}},
\]

where \( L_{ij} \) is the length of the alignment, \( k \) runs over alignment positions and

\[
s(Q_k, T_k) = \begin{cases} 
A'(i_k, j_k), & \text{if } Q_k \neq \text{‘-’} \text{ and } T_k \neq \text{‘-’} \\
\delta_{\text{open}}, & \text{if } (Q_k = \text{‘-’} \text{ and } Q_{k-1} \neq \text{‘-’}) \text{ or } (T_k = \text{‘-’} \text{ and } T_{k-1} \neq \text{‘-’}) \\
\delta_{\text{ext}}, & \text{if } (Q_k = \text{‘-’} \text{ and } Q_{k-1} = \text{‘-’}) \text{ or } (T_k = \text{‘-’} \text{ or } T_{k-1} = \text{‘-’})
\end{cases}
\]

Here \( Q_k \) and \( T_k \) stand for the content of the \( k \)-th position in the alignment (gap or not-gap) on the query and template side, respectively, \( i_k \) and \( j_k \) denote the positions in the original (non-gapped) query and template sequences, corresponding to the \( k \)-th alignment position, and ‘-’ stands for the gap in the \( k \)-th position of the alignment. Thus, the raw score of the alignment was normalized with respect to the alignment length, allowing us to compare results for alignments of different lengths. We aligned each of the 850 \times 2 = 1700 sequences in the ProtCom40 [4] dataset against each sequence in the dataset, apart from itself. The analysis of the resulting distribution of normalized raw scores suggested that from a pragmatic point of view all the alignments having \( r > 1 \) can be considered as statistically significant, since it would correspond to the case of Z-score, \( Z > 3 \) for the symmetrical Gaussian-like distribution since only \( \sim 0.5\% \) of cases fall into the region \( r > 1 \). Note, that due to the complicated shape of the raw-score distribution we were not able to make any definite correspondence between the normalized raw score and expectation values used in PSI-BLAST alignments. However, comparison of the alignments produced by both PSI-BLAST and profiles algorithm gives a rough estimate that \( r \sim 1 \) corresponds to the PSI-BLAST expectation value \( 10^{-3} \) to \( 10^{-7} \).

2.5. Assessing quality of the models

The predicted models were compared against the 3D structures of protein–protein complexes either in our ProtCom database [4] or in the Boston University benchmark set [30]. Quality of the models was evaluated by calculating the percentage of correctly predicted interfacial residues (IRs)

\[
f_{\text{IR}}(\%) = \frac{\text{number of IR in both model and PDB}}{\text{number of IR in PDB}} \times 100,
\]

and the percentage of correctly predicted pairs of residues interacting across the complex interface (IP)

\[
f_{\text{IP}}(\%) = \frac{\text{number of pairs in both model and PDB}}{\text{number of pairs in PDB}} \times 100
\]

The latter quantity corresponds to the fraction of native contacts used in the evaluation of models in the CAPRI experiment. We also calculated and analyzed root-mean-square deviations (RMSDs) between various sets of atoms: (i) global RMSD over all heavy atoms of the model and X-ray structure, after superimposition of the structures using the SKA program (http://trantor.bioc.columbia.edu/); (ii) interface RMSD (IRMSD) calculated over the heavy atoms of the interface residues of the complex, and (iii) ligand RMSD (LRMSD) calculated over the heavy atoms of the ligand, after superimposing the receptor of the model and X-ray structure with minimal RMSD. The last two criteria conform the CAPRI assessment criteria [44]. All RMSD values for a given protein complex have been calculated over the same set of residues to allow a fair comparison between the PSI-BLAST and profile-to-profile approach.

All these quantities are important characteristics of the quality of the models, but they have different emphases. A correct prediction of the interfacial pairs indicates that the model can be used to study the interactions across the interface and to some extent to estimate the role of the individual residues in the binding. However, it does not necessarily indicate a correct prediction of the global 3D structure of the protein–protein complex. The overall RMSD provides this measure and a low overall RMSD assures that the model resembles the overall structure of the experimentally determined complex.
monomer B. In this context, the RMSD’s as defined in CAPRI, I-RMSD and L-RMSD, reflect better the quality of the model in terms of predicting the interface of the complex and the mutual orientation of the complex components. Thus, in assessing the quality of the models one should take into account all characteristics mentioned above. It should be noted that \( f_{IP} \) and \( f_{IR} \) do not reflect amount of false positive predictions made by a model. For this purpose one should use total number of interacting residues/pairs in a model, rather than in PDB in (Eqs. (7) and (8)). We calculate the latter quantities for the models discussed in the paper and found they were close to corresponding \( f_{IP} \) and \( f_{IR} \) values and therefore we do not present them separately.

### 3. Results and discussions

In general, there are two approaches of modeling 3D structure of protein complexes. If one’s goal is to create a high-quality model of the interface for a particular complex and structures of unbound monomers are known, then the superior choice is one of the numerous \textit{ab initio} docking algorithms. However, all docking algorithms are computationally expensive and therefore if one aims at performing computational studies of protein binding on a large scale, say on the level of predicting all interactions in a particular metabolic pathway, then docking algorithms would be inapplicable. In such case, probably the best choice is the homology-based approach. This methodology is capable of producing models in a computationally efficient manner and of good enough quality to grasp the essentials of the protein binding mode. These two approaches complement each other well in balancing both high quality of the docking models and low computational cost of the homology models. Below we will concentrate on the homology-based approach and will show its suitability for large-scale predictions of 3D structures of protein–protein complexes.

In this section, we first test the sensitivity of the profile-to-profile alignment to the gap opening/insertion penalties and find the appropriate value for these parameters. Then, we report the results of benchmarking profile-to-profile alignment (hereafter referred to as PROFILES models) and comparison with PSI-BLAST results on the commonly used Boston benchmark set \[30\] of protein–protein complexes. Next, we investigate the quality of models built with the profile-to-profile method on a larger set of 463 protein–protein complexes. Further, we compare the quality of the common models built with profile-to-profile method and PSI-BLAST, taking into consideration all models built with high confidence by both algorithms. At the end, we investigate the quality of models built for each of the complexes (463 cases or 100% recovery rate) irrespective of the level of statistical confidence.

#### 3.1. Gap opening and gap extension penalties

The results of aligning two sequences using dynamical programming algorithm is in general sensitive to the values of the gap opening and gap extension penalties and the rule of thumb here is that the gap opening penalty should be considerably larger than elements of the matching matrix and that the gap extension penalty should be much smaller compared to the gap opening penalty (usually \( g_{\text{open}} \approx 3–4 \) and \( g_{\text{ext}} \approx 0.1–0.4 \) \[45\]). In order to find the optimum value of \( g_{\text{open}} \) for our dataset we performed alignments with \( g_{\text{open}} \) ranging from \( 1 \) to \( 10 \) and found that the alignments do not depend much on \( g_{\text{open}} \) when \( g_{\text{open}} > 2 \). Therefore, the rest of the results presented in this paper have been obtained on the basis of alignments with \( g_{\text{open}} = 3 \). This is related to additional weights (Eq. (2)) and enhanced penalties (Eq. (4)) which forcefully align interfacial residues in a gapless manner irrespectively of \( g_{\text{open}} \) values and then there is no much freedom left in aligning the rest of the sequences. Note that in this study we simplify the task by fixing \( g_{\text{ext}} \) value with respect to the \( g_{\text{open}} \) \((g_{\text{ext}} = 0.08 \times g_{\text{open}}) \) as suggested in Ref. [26].

#### 3.2. Modeling the Boston University benchmark set

The 43 pairs of queries selected from the Boston University benchmark set, Version 2.0 (http://zlab.bu.edu/zdock/benchmark.shtml) were subjected to both model building protocols, the quality of the resulting models were evaluated in terms of the quantities described above and the results are presented in Table 1. ProtCom95 was used as template pool and self-hits, if any, were excluded. As can be seen, the PSI-BLAST protocol has built only six models (14% recovery rate) even when the expectation value \( (E\text{-value}) \) was set to 10. By contrast, our profile-to-profile HOMBACOP protocol, at a cut-off normalized raw score of 1, generated 19 models recovering 19 complexes in the Boston University benchmark set (44% recovery rate). All query sequences recovered by the PSI-BLAST protocol were also recovered by the profile-to-profile HOMBACOP protocol and the quality of all six common models is good with \( f_{IR} \) ranging from 75 to 96% and \( f_{IP} \) from 53 to 96% and with

### Table 1

<table>
<thead>
<tr>
<th>Query</th>
<th>Template</th>
<th>BLAST</th>
<th>PROFILES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1acb</td>
<td>1azz</td>
<td>Residues (%)</td>
<td>Native contacts (%)</td>
</tr>
<tr>
<td>1axx</td>
<td>1shy</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1ay7</td>
<td>1b27</td>
<td>92.1</td>
<td>89.19</td>
</tr>
<tr>
<td>1bdc</td>
<td>2e9f</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1cji</td>
<td>1y0c</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1dfr</td>
<td>1azz</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1dfj</td>
<td>1a4y</td>
<td>84.1</td>
<td>53.42</td>
</tr>
<tr>
<td>1eow</td>
<td>1y0c</td>
<td>74.6</td>
<td>68.49</td>
</tr>
<tr>
<td>1e2g</td>
<td>1w1w</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1f2g</td>
<td>1agt</td>
<td>89.1</td>
<td>88.33</td>
</tr>
<tr>
<td>1gcq</td>
<td>10th</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1ghq</td>
<td>1w1i</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1he8</td>
<td>1njr</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1kaa</td>
<td>2j12</td>
<td>91.3</td>
<td>80.00</td>
</tr>
<tr>
<td>1k9</td>
<td>2bo</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1u0i</td>
<td>1i3q</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2s1c</td>
<td>1w6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2smi</td>
<td>1w6</td>
<td>90.2</td>
<td>95.77</td>
</tr>
<tr>
<td>7cei</td>
<td>1xx2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
I-RMSD 1–3 Å. All of these models are classified as medium-quality models in terms of the CAPRI experiment. Remarkably, the HOMBA-COP protocol produced 13 additional models, all of them being of reasonable quality with $f_{IR}$ ≥ 50% (one exception having $f_{IR}$ = 43%) and with two models with good predictions of mutual monomer orientation ($f_{IP}$ > 70%). Thus, results of our benchmarking on this dataset substantiate the potential of the sequence-homology-based approach using alignments of modified sequence profiles for modeling 3D structures of protein–protein complexes.

3.3. Benchmarking of PROTCOM40 dataset with profile-to-profile HOMBACOP

Out of 463 pairs of sequences in the dataset, only 74 produced 113 statistically significant pairs of global alignments (both alignments of query sequences to sequences of a template complex have raw scores (Eq. (5)) larger than 1). This rather low, ∼16%, recovery rate (we define recovery rate here as ratio of query sequences for which statistically significant alignments were found to the total number of query sequences) should not be considered as discouraging in practical usage of homology modeling for predicting 3D structures of protein complexes since it just reflects the existence of homologous protein complexes in the ProtCom. In this context it is interesting to note that in our previous study[23] on a much smaller and restricted set of protein complexes (92 structures) the recovery rate was only about 9%. Since then our database of protein complexes underwent two major updates and now consists of a much more diverse pool of possible templates and the recovery rate is now almost twice larger. This suggests that further elaboration of the database of protein complexes is of great importance for a successful practical application of the sequence-homology-based concept to predicting structures of protein complexes on a large scale.

The 113 models (hereafter referred to as the set of PROFILES models) were built and evaluated by means of the procedure described above and the results are presented in Fig. 2 and in the rightmost column of Table 2. For the vast majority of the models (87 models) the percentage correctly predicted sites is larger than 50% (Fig. 2A) and for 60 models $f_{IR}$ was even larger than 80%. This is comparable to the best results achieved by means of machine learning algorithms (∼50% at 50% coverage) reported by Zhou et al. [35,36,46,47], but we would like to stress that the starting point of our protocol is a pair of sequences without any a priori use of three-dimensional structure information. Panel B in Fig. 2 demonstrates that the mutual orientation of complex components is satisfactorily well predicted as well. For 56 models or almost 50% of cases, $f_{IP}$ is larger than 50%, which is one of the conditions for the high ranking of a model in the CAPRI experiment.

It can be noted that there is quite a large fraction of the models (29 structures, or 26%) with almost no interacting pairs predicted while the proportion of the models with bad prediction of interfacial residues is much smaller (only eight models or 7%) and thus, it would be interesting to correlate $f_{IP}$ and $f_{IR}$. The correlation is presented in Fig. 3 and it is clearly seen from the figure that our protocol starts correctly predicting mutual orientation of the complex components.

### Table 2

Percentage of models for the PROTCOM40 dataset in different CAPRI quality classes obtained by means of BLAST and PROFILES alignment algorithms

<table>
<thead>
<tr>
<th>CAPRI classification</th>
<th>% of total number of models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLAST</td>
</tr>
<tr>
<td>Medium</td>
<td>20.6</td>
</tr>
<tr>
<td>Acceptable</td>
<td>45.8</td>
</tr>
<tr>
<td>Incorrect</td>
<td>33.6</td>
</tr>
</tbody>
</table>

Fig. 2. Percentage of correctly predicted interfacial residues (panel A) and percentage of correctly predicted interacting pairs (panel B) for 116 models obtained by PROFILES algorithm on the PROTCOM40 dataset.

Fig. 3. Correlation between percentages of correctly predicted interacting pairs and correctly predicted interfacial residues for 116 models obtained by PROFILES algorithm on the PROTCOM40 dataset.
ponents only if more than half of binding site is correctly predicted in the model. Also, \( f_{IR} \) is always larger than \( f_{IP} \) and this perhaps is due to the fact that at that stage of our studies we did not make any attempts to perform the computationally costly refinement of the obtained models since one of our primary goals is to create a computationally efficient algorithm for large-scale modeling. Thus, very often models contain small imperfections like perturbed orientation of secondary structure elements. This is illustrated in Fig. 4 where a model of the human aspartylglucosamidase complex is shown superimposed on its original PDB structure. This model has 91.5% correctly predicted interfacial residues, but only 58.2% accurately predicted interacting pairs. This is due to a slight displacement of two helices in the larger component (red and green ribbons in Fig. 4), which still leaves them on the complex interface, but results in them making “wrong” contacts. Note that despite such imperfections our models possess almost no steric clashes, which would lead to a considerable distortion of the secondary structure elements during the refinement procedure. We expect that a small rotation of the predicted monomer, keeping its 3D structure rigid, would result in much better prediction of the interfacial contacts.

The ability of the used protocol in predicting interfaces of protein–protein complexes is further illustrated in Fig. 5 where distributions of RMSD’s between interfacial atoms are plotted for all 113 models separately for the backbone and side-chain atoms. In the majority of cases both backbone and side-chain orientations are predicted with reasonable accuracy: 89 models (~78%) have I-RMSD smaller than 10 Å for the backbone atoms and 84 models (~74%) have I-RMSD smaller than 10 Å for the side-chain atoms. Note, that an I-RMSD inferior to 10 Å is one of the conditions in CAPRI to consider a model of acceptable quality.

The distribution of L-RMSD’s (Fig. 6), however, does not show the same good trends as those observed for I-RMSD. Although there is a large fraction of the models (67 or 59%) with L-RMSD less than 10 Å, a significant number of models (31 or ~27%) have very large RMSD, larger then 30 Å (see insert in Fig. 6). It indicates that in these models the position of the ligand with respect to the receptor is wrongly predicted, although the overall fold and structure of individual components may be predicted reasonably well. This is illustrated in Fig. 7 where the model of peptide:N-Glycanase-Rad 23 complex is shown superimposed on its original structure. This model has only 2% of correctly predicted interacting pairs and a
relatively large L-RMSD of 16 Å. As can clearly be seen from the figure, the model of the receptor almost perfectly superimposes on the experimental structure (red and green ribbons in Fig. 7). Overall tertiary and secondary structures of the ligand (blue and cyan ribbons in Fig. 7) are also predicted correctly, however, ligand orientation with respect to the receptor is completely wrong. Surprisingly enough this model has 46% of correctly predicted interfacial residues and a relatively low backbone atoms I-RMSD of 8.1 Å. This again demonstrates the potential of homology-based approaches in modeling complex interfaces at least as a starting point for more elaborate (but time-consuming as well) computational techniques. For instance, we argue that the quality of the model shown in Fig. 7 may be improved significantly by keeping the overall structures of the monomers as predicted, but sliding their interfaces by an offset of 10–20% from the original positions. Note, however, that refinement of the obtained models is a separate large field of study and out of the scope of the present investigation.

3.4. Comparison of the common models produced by PROFILES and BLAST alignment protocols on ProtCom40 dataset

Since PSI-BLAST alignment is one of the most widely tools used by the scientific community to produce sequence alignments, it would be interesting to compare the effectiveness of the alignment protocol used in the present study with the performance of the PSI-BLAST protocol. For this purpose we built the models for the complexes in the ProtCom40 dataset using the same algorithm as described in Section 2, but based on alignments produced by runs of the BLASTPGP [22] program with its default parameters, using the ProtCom40 database as a template pool (with a sequence alignment on itself excluded). Due to the differences in the alignment technique, the PSI-BLAST protocol recovers a different set of query sequences (90 pairs) and produces a different set of the models (155) compared to the set of PROFILES models. Fortunately, a statistically significant set of 99 common models was present in both model sets (by a common model we mean a model of the same pair of query sequences built by means of both protocols on the same pair of template sequences). The quality of those PSI-BLAST models was evaluated and compared to the quality of corresponding models in the PROFILES set and the results are shown in Figs. 8 and 9 and Table 2.

Table 2 displays a comparison of the model quality obtained by the two alignment protocols in terms of the CAPRI classification. It can be seen that both protocols perform practically equally well with about two thirds of the model being in the acceptable and medium-quality categories. However, Fig. 8 shows that the PROFILES protocol produces models with a higher percentage of both correctly predicted interfacial residues, and interacting pairs (majority of the points are in the upper triangle in both panels of the graph). The difference in quality is sometimes strikingly large. For instance, the PROFILES model of Sumo E1 activating enzyme (PDB ID 1y8q), built on the basis of the structure for the NEDD8 activating enzyme (PDB ID 1yov), has $f_{IR} = 86.9\%$, while the corresponding number for the PSI-BLAST model is only 24.6\% (a point marked by an arrow in Fig. 8A). This observation can be explained by analyzing the query-template alignments produced by both alignment protocols. While the PROFILES procedure aligns whole sequences, PSI-BLAST aligns sequences only partially. In this case a significant portion of the interface in the query complex is formed by those parts of the sequences, which are missing in the BLAST alignments. The better-quality tendency for the PROFILES models is not so clearly pronounced when considering percentage of correctly predicted
interacting pairs, but the majority of the PROFILES models nevertheless have $f_{IP}$ larger than that for the corresponding BLAST models (Fig. 8B).

Fig. 9 shows comparison of the quality of the models with respect to I-RMSD for backbone atoms (panel A in Fig. 9) and L-RMSD (panel B). Plot for I-RMSD and L-RMSD of the side-chain atoms is similar and hence it is not presented and discussed separately. As it is seen there is no clear tendency which models are of better quality with respect to RMSD’s. Almost equal amounts of points are in lower (PROFILES models are better) and in upper triangle (PSI-BLAST models are better) triangle in both panels of the figure. Remarkably, this tendency holds both at small and large L-RMSD values (see insert in Fig. 9B), i.e. both protocols fail to predict correct orientation of complex components to the same extent. There are several exceptions, most remarkable being model “reversed” to that discussed in the previous paragraph, namely, model of the NEDD8 activating enzyme (PDB ID 1yov) modeled on the structure of Sumo E1 activating enzyme (PDB ID 1y8q). PROFILE model for this query has L-RMSD 10.6 Å while BLAST model 56.6 Å (point marked by the arrow in Fig. 9B). We again argue that this due to partial alignments of the sequences in the BLAST protocol and in this case absence of the significant part of the interface in the alignment leads to completely wrong orientation of monomers. Interestingly the PROFILES model for this query has $f_{IR} = 81.4\%$ and $f_{IP} = 45.5\%$ compared to $f_{IR} = 27.3\%$ and $f_{IP} = 16.4\%$ for the corresponding PSI-BLAST model.

There are two main factors influencing the difference in outcome between the BLAST and modified PROFILES alignment protocols. First, we use the global alignment algorithm while BLAST alignment aligns sequences only partially and therefore PROFILE models always contain the binding site, while BLAST model may not contain the binding site or contain it only partially. This is especially true if the binding site is located on terminal parts of polypeptide chain(s) since local sequence alignment very often does not comprise terminal residues. Secondly, extra weight on interfacial residues in our algorithm ensures absence of gaps in the alignment for interfacial region, i.e. no segments have to be built from scratch on or close to complex interface. This also improves the quality of resulting models, since flexible loops build from scratch in the interracial region have high probability to penetrate another complex com-

![Fig. 9](image.png)

**Fig. 9.** Comparison of I-RMSD calculated for backbone atoms (panel A) and of L-RMSD (panel B) for 99 models produced by both PROFILES and BLAST alignment protocols. Insert to the panel B shows zoom of the main graph at small L-RMSD values. The solid lines in both panels represent the case of equal performance for both algorithms. The arrow in panel B marks the data point for the model of NEDD8 activating enzyme (PDB ID 1yov) built on Sumo E1 activating enzyme (PDB ID 1y8q).

![Fig. 10](image.png)

**Fig. 10.** Percentages of correctly predicted interfacial residues (panel A) and correctly predicted interacting pairs (panel B) as function of the average raw score of sequence alignment on basis of which the model was built. The data points shown in the figure were obtained for 427 models obtained by means of top-raw-score-alignment protocol.
ponent thereby reducing dramatically quality of the model. Which factor is the most essential for improving quality of a particular model depends on specifics of each case and for the sake of brevity we omit here detailed analysis of this phenomenon and present rather a statistical picture of difference between the two methods.

3.5. Models produced on alignments between low-homology sequences

All the results described and discussed above were obtained by means of the PSI-BLAST and the PROFILES protocols, both of which were obtained at stern limitation that only statistically significant alignments are considered for the further model building. This implies existence of highly homologous sequence in the template database and this restricts severely the number of the models we were able to build for our ProtCom40 pool of query sequences. While the PSI-BLAST search often cannot align a pair of query sequences to a pair of sequences belonging to the same template structure (even at $E$-value cut-off equal to 10), the profile technique always generates an alignment. Therefore, we tested a possibility that even pair of alignments, one or both of which do not exhibit “good” raw score, can be basis of a reasonable model. For this purpose, we built models for all 463 complex structures in the ProtCom40 dataset using for each query that pair of alignments which has the highest raw score among all alignments for that particular query irrespectively of whether condition for statistical significance is satisfied or not. Note that in this protocol each query produces strictly one model (the top model) while the PSI-BLAST and PROFILES protocols may produce no model or generate several models for a particular query.

In 36 cases the NEST program was not able to build a model due to presence of long gap stretches in the alignment. Models constructed on remaining 427 alignments were evaluated in terms of percentages of correctly predicted interfacial residues and correctly predicted interacting pair and results are shown in Fig. 10A. A tendency is observed that with increasing statistical significance of an alignment (i.e. increasing average raw score) the quality of the model also increases, but this observation is not very interesting since all models in the right part of the graph were obtained by means of the “ordinary” PROFILES protocol described above. However, it is clearly seen from Fig. 10A that even models built on the basis of low-homology alignments have in some cases very high portion of the binding site predicted correctly, although orientation of the complex components usually is wrong (low $f_{IP}$, see Fig. 10B). Thus, we extracted from the common pool of models only the models constructed from low-homology alignments (222 cases) and presented $f_{IP}$ distribution for them separately in Fig. 11. As it is seen in ~40% of cases binding pocket was predicted with more than 50% confidence level. This suggests an interesting and promising suggestion that for purposes of computationally fast search for binding pocket the low-homology models may be used as a starting point for further docking or other elaborated calculations.

4. Conclusions

Homology-based approach of predicting 3D structures of protein–protein complexes complements $ab$ $initio$ docking methods, since it can be used to predict high-quality 3D structures of complexes within families for which representative structure exits. In addition, obligatory complexes and many other complexes are formed between monomers that are unstructured in isolation. Such complexes cannot be modeled $ab$ $initio$. Therefore, testing and further development of homology-based methods for prediction of 3D structure of protein–protein complexes, seems to us, is an important task and in this study we addressed some aspects of this issue.

We performed comprehensive studies of homology-based approach to modeling of protein–protein complexes using conventional PSI-BLAST alignment protocol and alignments of modified sequence profiles. The studies were carried out on two different large and diverse datasets which ensures statistical reliability of the reported results. We clearly demonstrated that given existence of homologous template, the methodology used in the present study is capable of producing models satisfying CAPRI definitions for acceptable models. We stress that existence of large and diverse database of putative templates is crucial for successful homology modeling and work now is progress to further enrich and expand our ProtCom database.

We have also shown that usage of sequence profiles with additional information about interfacial residues included increases recovery rate of the computational protocol and improves quality of the models. Earlier studies [24–28] on modeling the 3D structures of single proteins have also shown that profile-to-profile methods of alignment are superior to the pure PSI-BLAST sequence searches. The profiles contain additional information that assists in detecting weak relationships between query and template. Our implementation of the profile-to-profile alignment technique includes structural information on the interfacial residues of the templates and predicted interfacial residues of the queries and we argue that including further structural information into the sequence profiles will further enhance the computational protocol. Note, that our model building protocol aligns interfacial residues not between complex components, but between one of the components of the query complex and one of the components of the template complex. We assume that such alignments made for both components of the query complex will lead to building of model interface similar to that of the homologous template complex. Perhaps inclusion of correlations between complex monomers will further improve the algorithm.

We also demonstrated that using profile-to-profile alignment technique reasonable models can be built even for a query sequences with no highly homologous templates found in the database. Those models can be used as an initial guess about binding site location and thus they are good starting point for further more elaborated, but computationally costly calculations. Our computational protocol is fast and computationally efficient, which enables its use for large-scale computational predictions.

![Fig. 11](image-url) Percentage of correctly predicted interfacial residues for 222 models constructed from low-homology alignments.
Acknowledgements

We are grateful to Dr. Brian Dominy and Dr. Matthew Saltzmann for the computer assistance.

References