Protein Structure Prediction of CASP5 Comparative Modeling and Fold Recognition Targets Using Consensus Alignment Approach and 3D Assessment

Krzysztof Ginalski1* and Leszek Rychlewski2
1Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Warsaw, Poland
2BioInfoBank Institute, Poznan, Poland

ABSTRACT For the fifth round of Critical Assessment of Techniques for Protein Structure Prediction (CASP5) all comparative modeling (CM) and fold recognition (FR) target proteins were modeled using a combination of consensus alignment strategy and 3D assessment. A large number and broad variety of prediction targets, with sequence identity between each modeled domain and the related known structure, ranging from 6 to 49%, represented all difficulty levels in comparative modeling and fold recognition. The critical steps in modeling, selection of template(s) and generation of sequence-to-structure alignment, were based on the results of secondary structure prediction and tertiary fold recognition carried out using the Meta Server coupled with the 3D-Jury system. The main idea behind the modeling procedure was to select the most common alignment variants provided by individual servers, as well as to generate several alternatives for questionable regions and to evaluate them in 3D by building corresponding molecular models. Analysis of fold-specific features and sequence conservation patterns for the target family was also widely used at this stage. For both CM and FR targets remote homologs of known structure were clearly recognized by the 3D-Jury system. In the analogous fold recognition subcategory, the correct fold was identified for five out of eight domains. The average alignment accuracy for FR models (48%) was far less than for CM predictions (80%). These finding, coupled with the observation that in the majority of cases the submitted models were not closer to the experimental structure than their best templates, indicate that, especially for difficult targets, there is still ample room for improvement. Proteins 2003;53:410–417. © 2003 Wiley-Liss, Inc.

Key words: protein structure prediction; comparative modeling; fold recognition; Meta Server; 3D-Jury; consensus alignment approach; 3D evaluation

INTRODUCTION

An enormous number of known protein sequences and, in particular, recent advances in genome sequencing, have created a necessity for developing theoretical methods for protein structure prediction, which have now become a mature scientific field with clear applications in molecular biology.1 With the experimental determination of many new protein structures in recent years, and the development of more sensitive remote homolog detection methods, it has become increasingly likely that a protein of biological interest, but unknown three-dimensional structure, will have a homolog of known structure. Comparative modeling and fold recognition are the major protein structure prediction approaches that help to bridge the gap between primary and tertiary structure by allowing the construction of models, which may be used to identify critical residues involved in catalysis, binding, or structural stability; to examine protein-protein or protein-ligand interactions; and to correlate genotypic and phenotypic mutation data. A major new challenge for these methods is their integration with the torrents of data from genome sequencing projects, as well as from functional and structural genomics.

Increased interest in the development of new comparative modeling and fold recognition algorithms has led to a variety of prediction services available on the internet. Since the last CASP42 and CAFASP-23 experiments the number of automated servers has increased two-fold. The latest progress in the automated protein structure prediction is mainly attributed to the development of meta servers, which extract common structural motifs (consensus) from the set of 3D models generated by various independent prediction providers. The resulting final models have a higher chance to be correct than the models produced by any single method.

However, despite continuous progress in the protein structure prediction area, currently available algorithms still introduce significant errors that affect the modeled structure, especially for difficult comparative modeling (CM) and fold recognition (FR) targets representing less than the 20% level of sequence identity with the closest template. Aside from the errors made in different steps of model building (e.g. modeling of variable regions or side chain packing), the major ones are introduced at the earliest stages of template selection and generation of the
alignment of the target sequence with the related known structures. Detection of correct/optimal templates, and alignment accuracy, are still the major limitations for modeling of difficult targets.

For the fifth round of Critical Assessment of Techniques for Protein Structure Prediction (CASP5) experiment, all comparative modeling and fold recognition targets were modeled, based on the results of the 3D-Jury meta prediction system, using a combination of consensus alignment approach, 3D assessment procedure as well as fold characteristics and target family sequence analysis. Importantly, only one prediction for each target was deposited, although submission of up to five models was allowed. Identification of related structures for difficult CM and FR targets, as well as the issue of sequence-to-structure alignment of modeled sequences with their respective templates, was the main emphasis.

**MATERIALS AND METHODS**

The critical steps in modeling: selection of template(s), and generation of sequence-to-structure alignment, were based mainly on the results of secondary structure prediction and tertiary fold recognition carried out using the Meta Server (http://BioInfo.PL/Meta/). The enormous number of models and corresponding alignments originating from the numerous primary servers (including those participating only in the CAFASP3 experiment) was narrowed down by the 3D-Jury system. The 3D-Jury meta predictor that is able to indicate the most abundant models was the central part of the modeling procedure. It was used both for template/fold identification, as well as for selection of the most confident alignment variants.

In most cases the 3D-Jury system was run with default setting that included eight primary servers used for consensus building: ORFeus, SamT02, FFAS03, mGenTHREADER, INBGU, RAPTOR, FUGUE-2, and 3D-PSSM. Detailed evaluation of this meta-predictor in the latest LiveBench-6 program (presented in this journal issue) has shown that the highest specificity is obtained when 3D-Jury is operating in this default mode. However, for some hard FR targets, additional servers were also included in consensus building.

The detailed description of the modeling procedure used for modeling of CASP5 target proteins is presented below.

**Selection of Template(s)**

For trivial and easy comparative modeling targets, related proteins with known structures were identified with PSI-BLAST searches performed against the non-redundant protein database starting with the stringent expectation (e) value cutoff ($10^{-30} - 10^{-20}$) and in subsequent iterations gradually increasing to 0.001.

For difficult comparative modeling, as well as fold recognition targets, template(s)/fold identification was based on the 3D-Jury results for the target and its homologs submitted (also by other groups) to the Meta Server. Where appropriate, both full protein sequences, as well as separate domains alone, were used as queries. Multiple predictions, indicating the same fold with a relatively high 3D-Jury score, were considered reliable at this stage; while compatibility of target family specific features (including predicted secondary structure) with the characteristic features of the template/fold was a final indicator of correct fold identification. For difficult FR targets, where the highest 3D-Jury score was significantly low, several distant homologs were also submitted to the Meta Server. In such cases this analysis was also coupled with the iterative sequence searches performed against different databases, including unfinished genomes.

Finally, FSSP and SCOP databases were inspected to determine whether selected templates had other closely related structures. All identified structures selected as plausible templates for building a 3D model of the target protein were then gathered and subjected to further analysis. When multiple PDB entries for the same template protein were available, the one with the highest resolution, most complete set of atoms, and similar ligand (where appropriate) was chosen. Final selection of template(s) used for model building was done after generating final sequence-to-structure alignment (see below).

**Analysis of Template(s) and Target Proteins**

To produce sequence-to-structure alignment consistent with the general architecture of the identified fold, structural determinants of the fold were analyzed. Previously identified structures representing a given fold, and corresponding structural alignment extracted from the FSSP database, were inspected for both conservation and variability of the structural elements. If compared structures were highly diverged or displayed significant differences due to local conformational changes (shift and/or rotation of part of the structure), the structure-based alignment was adjusted manually. Conservation of specific residues and contacts responsible for maintaining tertiary structure, and critical for substrate binding and/or catalysis, were also established.

Additionally, homologous sequences that matched the target were collected with PSI-BLAST searches performed against the non-redundant protein sequence database until profile convergence. When only a few hits were found, further searches were carried out against unfinished genomes. Sequences more than 90% identical to any other sequence were then filtered out, and remaining entries shortened to consider only domain(s) in the target protein. Finally, the CLUSTAL W program was used to generate a multiple sequence alignment to identify conserved residues within the target family. Some manual adjustments were introduced in the resulting alignment, based on careful analysis of sequence conservation and results of secondary structure prediction, as well as tertiary fold recognition carried out for the target protein. Occasionally, the literature was searched for any available biochemical information (function, mutations, catalytic residues, etc.) for target protein and/or its close homologs.

**Sequence-to-Structure Alignment**

The aim of this most significant step of the modeling was to obtain reliable sequence-structure mapping using the consensus alignment strategy and 3D assessment procedure.
Initial alignment building with consensus approach

All alignments produced by different servers interacting with the Meta Server were inspected for both variability and violation of structural integrity. Initial (consensus) alignment was obtained by taking the most common alignment for each region (mainly for each secondary structure element) within the context of the structural constraints, taking into account the structure-based alignment of template proteins. These target-template alignments producing the most abundant models, as detected with the 3D-Jury system, were mainly used for consensus alignment building. Consequently, only the regions with single dominant alignment variant were considered as reliably aligned. To confirm this assumption they were assessed in 3D at a later stage. In some cases alignments obtained for close homologues submitted to the Meta Server were also evaluated to improve sensitivity of the consensus building in questionable regions.

Generating plausible alignment variants for questionable regions

All alignment regions that displayed low stability, i.e. were highly dependent on the server and no single dominant alignment variant could be observed, were considered of lower confidence. For these regions all alignment alternatives provided by the 3D-Jury system were taken as equally possible at this stage if consistent with general requirement of the fold and secondary structure predictions for target sequence. In many cases the several plausible alignment variants were also derived manually, guided mainly by secondary structure predictions. Exact positions of gaps in the insertion/deletion regions were adjusted manually to satisfy the structural scaffold of the template(s).

3D evaluation procedure

All plausible alternative sequence-to-structure alignments were tested using the 3D evaluation procedure described previously but with some minor modifications. To increase sensitivity of this procedure, corresponding alignment variants were assessed in 3D by building molecular models not only for the target sequence, but in some cases also for its close homologues. These auxiliary 3D models not only for the target sequence, but in some alignment variants were assessed in 3D by building molecular models. All the resulting models were then subjected automatically with the MODELLER program. In contrast to auxiliary models generated in the 3D evaluation procedure more than one template was used if possible. Specifically, both templates and fragments extracted from homologous structures were selected in this way to minimize the number of insertion and deletion regions in the final sequence-to-structure alignment. In general, for trivial and easy CM targets, only proteins identified with relatively high E-value scores in PSI-BLAST searches were used as parent structures. For some difficult CM and FR targets, a set of more diverse structural templates was chosen to provide considerable variation outside the structurally conserved regions of the fold, as well as variation in relative orientation of core secondary structure elements. In some cases suitable fragments, taken from other distantly homologous structures for loops or variable regions modeling, were added to a selected set of templates. In the absence of such fragments, both loops as well as longer variable regions were generated with the MODELLER program using predicted secondary structure information (manual consensus of the results of several methods available through the Meta Server). Each target domain was modeled separately unless these domains were present simultaneously in one or more template proteins. Finally, all target domains were assembled into a single 3D model avoiding any steric clashes, but with no special effort to reproduce relative orientation of the domains.

Additionally, for some trivial comparative modeling targets, side-chains were rebuilt using the SCWRL program22 with a backbone conformation-dependent rotamer library. In some sporadic cases models were subjected to energy minimization in the AMBER forcefield23 to remove remaining steric clashes and improve stereochemistry. The overall quality of each modeled structure was checked in detail with the WHAT_CHECK program24.

RESULTS AND DISCUSSION

Out of 67 CASP5 target proteins, 50 comparative modeling and 16 fold recognition domain structures, available on time before the Asilomar Meeting, have been used for evaluation and assessment of predictions quality. A considerable variety of prediction targets, with the sequence identity between each modeled domain and the related known structure ranging from 6 to 49%, represented all difficulty levels in these modeling categories. The results, describing main successes and failures of the applied modeling approach, are presented below. The reasons for these successes or failures are illustrated in detail for a few modeling examples.

Comparative Modeling

An overall summary of comparative modeling predictions is presented in Table I. More than half the submitted models have comparable or significantly better quality than the best of any primary server predictions used for consensus building by the 3D-Jury system. Moreover, only for targets T0179_2 and T0192, the number of correctly predicted residues is lower than the corresponding mean

Building Final 3D Model

The final 3D model of the target protein was built automatically with the MODELLER program. In contrast to auxiliary models generated in the 3D evaluation procedure more than one template was used if possible. Specifically, both templates and fragments extracted from homologous structures were selected in this way to minimize the number of insertion and deletion regions in the final sequence-to-structure alignment. In general, for trivial and easy CM targets, only proteins identified with relatively high E-value scores in PSI-BLAST searches were used as parent structures. For some difficult CM and FR targets, a set of more diverse structural templates was chosen to provide considerable variation outside the structurally conserved regions of the fold, as well as variation in relative orientation of core secondary structure elements. In some cases suitable fragments, taken from other distantly homologous structures for loops or variable regions modeling, were added to a selected set of templates. In the absence of such fragments, both loops as well as longer variable regions were generated with the MODELLER program using predicted secondary structure information (manual consensus of the results of several methods available through the Meta Server). Each target domain was modeled separately unless these domains were present simultaneously in one or more template proteins. Finally, all target domains were assembled into a single 3D model avoiding any steric clashes, but with no special effort to reproduce relative orientation of the domains.

Additionally, for some trivial comparative modeling targets, side-chains were rebuilt using the SCWRL program22 with a backbone conformation-dependent rotamer library. In some sporadic cases models were subjected to energy minimization in the AMBER forcefield23 to remove remaining steric clashes and improve stereochemistry. The overall quality of each modeled structure was checked in detail with the WHAT_CHECK program24.
value derived from the first models of these servers. These findings suggest that the consensus approach, based on selection of the most common alignment variants, combined with the 3D assessment procedure and information about sequence variability in the target family, as well as fold characteristic features, is able to improve on average primary server predictions. As the modeling strategy enables to accurately combine local regions correctly predicted only by some of the primary servers, resulted model has a chance to be significantly better than any of the models originated from these servers.

Target T0192 is an excellent example showing that lack of one of the modeling components can significantly deteriorate the quality of the prediction. For this target two secondary structure elements were misaligned: one β-strand and one α-helix. These were the only regions where servers provided several alignment variants, but none of them was assessed in 3D to test the fitness of the resulting residue mapping with the structural scaffold of the template.

The applied modeling scheme resulted in relatively high alignment accuracy (fraction of correctly aligned residues within structurally equivalent residues between model and target) with an average of 92, 73, and 57% for trivial, easy and difficult comparative modeling targets, respectively. The majority of alignment errors are located in questionable regions which were assigned lower confidence in the final sequence-structure mapping. The 3D assessment procedure failed in these cases, especially for difficult CM targets, mainly due to anticipated changes in secondary structure by minimizing violations of the spatial restraints derived from the reference proteins, which are weighted according to the degree of local sequence similarity between the templates and the modeled sequence, this was the main reason for using this program. Additionally, it

### Table I. Summary of Comparative Modeling Predictions

<table>
<thead>
<tr>
<th>Target</th>
<th>Category</th>
<th>Length</th>
<th>Equivalent</th>
<th>Correct</th>
<th>Target</th>
<th>Category</th>
<th>Length</th>
<th>Equivalent</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0137</td>
<td>CM(B)</td>
<td>133</td>
<td>133/132</td>
<td>133 (100%)</td>
<td>T0133</td>
<td>CM(P)</td>
<td>293</td>
<td>235/219</td>
<td>179 (76%)</td>
</tr>
<tr>
<td>T0140</td>
<td>CM(B)</td>
<td>57</td>
<td>56/65</td>
<td>51 (91%)</td>
<td>T0141</td>
<td>CM(P)</td>
<td>187</td>
<td>115/113</td>
<td>52 (45%)</td>
</tr>
<tr>
<td>T0142</td>
<td>CM(B)</td>
<td>280</td>
<td>254/248</td>
<td>228 (90%)</td>
<td>T0149</td>
<td>CM(P)</td>
<td>201</td>
<td>131/151</td>
<td>72 (55%)</td>
</tr>
<tr>
<td>T0143</td>
<td>CM(B)</td>
<td>216</td>
<td>204/202</td>
<td>198 (97%)</td>
<td>T0152</td>
<td>CM(P)</td>
<td>198</td>
<td>145/134</td>
<td>102 (70%)</td>
</tr>
<tr>
<td>T0150</td>
<td>CM(B)</td>
<td>96</td>
<td>95/96</td>
<td>94 (99%)</td>
<td>T0165</td>
<td>CM(P)</td>
<td>318</td>
<td>210/224</td>
<td>127 (60%)</td>
</tr>
<tr>
<td>T0151</td>
<td>CM(B)</td>
<td>106</td>
<td>96/100</td>
<td>79 (82%)</td>
<td>T0169</td>
<td>CM(P)</td>
<td>156</td>
<td>138/136</td>
<td>116 (84%)</td>
</tr>
<tr>
<td>T0153</td>
<td>CM(B)</td>
<td>134</td>
<td>126/128</td>
<td>113 (90%)</td>
<td>T0172</td>
<td>CM(P)</td>
<td>192</td>
<td>153/156</td>
<td>126 (82%)</td>
</tr>
<tr>
<td>T0154</td>
<td>CM(P)</td>
<td>185</td>
<td>179/176</td>
<td>176 (98%)</td>
<td>T0176</td>
<td>CM(P)</td>
<td>100</td>
<td>73/75</td>
<td>55 (75%)</td>
</tr>
<tr>
<td>T0154_2</td>
<td>CM(B)</td>
<td>103</td>
<td>93/96</td>
<td>75 (81%)</td>
<td>T0184</td>
<td>CM(P)</td>
<td>72</td>
<td>65/68</td>
<td>94 (98%)</td>
</tr>
<tr>
<td>T0155</td>
<td>CM(B)</td>
<td>117</td>
<td>117/117</td>
<td>117 (100%)</td>
<td>T0185</td>
<td>CM(P)</td>
<td>101</td>
<td>93/96</td>
<td>66 (71%)</td>
</tr>
<tr>
<td>T0160</td>
<td>CM(B)</td>
<td>125</td>
<td>117/119</td>
<td>100 (85%)</td>
<td>T0185_3</td>
<td>CM(P)</td>
<td>130</td>
<td>109/114</td>
<td>71 (65%)</td>
</tr>
<tr>
<td>T0167</td>
<td>CM(B)</td>
<td>180</td>
<td>160/161</td>
<td>146 (91%)</td>
<td>T0186_1</td>
<td>CM(P)</td>
<td>77</td>
<td>69/66</td>
<td>89 (99%)</td>
</tr>
<tr>
<td>T0177_1</td>
<td>CM(B)</td>
<td>57</td>
<td>56/56</td>
<td>53 (95%)</td>
<td>T0186</td>
<td>CM(P)</td>
<td>250</td>
<td>169/189</td>
<td>82 (49%)</td>
</tr>
<tr>
<td>T0177_2</td>
<td>CM(B)</td>
<td>88</td>
<td>88/87</td>
<td>88 (100%)</td>
<td>T0189</td>
<td>CM(P)</td>
<td>319</td>
<td>258/274</td>
<td>180 (69%)</td>
</tr>
<tr>
<td>T0177_3</td>
<td>CM(B)</td>
<td>75</td>
<td>70/70</td>
<td>60 (88%)</td>
<td>T0192</td>
<td>CM(P)</td>
<td>170</td>
<td>135/144</td>
<td>106 (79%)</td>
</tr>
<tr>
<td>T0178</td>
<td>CM(B)</td>
<td>219</td>
<td>214/213</td>
<td>214 (100%)</td>
<td>T0195</td>
<td>CM(P)</td>
<td>290</td>
<td>234/238</td>
<td>204 (87%)</td>
</tr>
<tr>
<td>T0179</td>
<td>CM(B)</td>
<td>56</td>
<td>56/56</td>
<td>56 (100%)</td>
<td>T0130</td>
<td>CM/FR</td>
<td>100</td>
<td>72/85</td>
<td>36 (50%)</td>
</tr>
<tr>
<td>T0179_2</td>
<td>CM(B)</td>
<td>218</td>
<td>206/209</td>
<td>197 (96%)</td>
<td>T0132</td>
<td>CM/FR</td>
<td>147</td>
<td>119/119</td>
<td>97 (82%)</td>
</tr>
<tr>
<td>T0182</td>
<td>CM(B)</td>
<td>249</td>
<td>248/248</td>
<td>248 (100%)</td>
<td>T0136</td>
<td>CM/FR</td>
<td>256</td>
<td>147/157</td>
<td>67 (46%)</td>
</tr>
<tr>
<td>T0183</td>
<td>CM(B)</td>
<td>247</td>
<td>217/217</td>
<td>217 (100%)</td>
<td>T0136_2</td>
<td>CM/FR</td>
<td>264</td>
<td>151/161</td>
<td>80 (53%)</td>
</tr>
<tr>
<td>T0184_1</td>
<td>CM(B)</td>
<td>165</td>
<td>137/142</td>
<td>133 (82%)</td>
<td>T0159</td>
<td>CM/FR</td>
<td>167</td>
<td>97/106</td>
<td>76 (78%)</td>
</tr>
<tr>
<td>T0185_2</td>
<td>CM(B)</td>
<td>197</td>
<td>174/175</td>
<td>151 (87%)</td>
<td>T0159_2</td>
<td>CM/FR</td>
<td>142</td>
<td>81/87</td>
<td>28 (35%)</td>
</tr>
<tr>
<td>T0188</td>
<td>CM(B)</td>
<td>107</td>
<td>105/102</td>
<td>96 (91%)</td>
<td>T0168</td>
<td>CM/FR</td>
<td>170</td>
<td>95/132</td>
<td>54 (57%)</td>
</tr>
<tr>
<td>T0190</td>
<td>CM(B)</td>
<td>111</td>
<td>110/110</td>
<td>107 (97%)</td>
<td>T0168_2</td>
<td>CM/FR</td>
<td>141</td>
<td>72/86</td>
<td>24 (33%)</td>
</tr>
<tr>
<td>T0191_2</td>
<td>CM(B)</td>
<td>143</td>
<td>113/121</td>
<td>74 (65%)</td>
<td>T0193</td>
<td>CM/FR</td>
<td>130</td>
<td>101/125</td>
<td>79 (78%)</td>
</tr>
</tbody>
</table>

a In bold are indicated targets for which the number of correctly predicted residues (identified from sequence-dependent superposition) was equal or greater than for any model from primary servers selected for consensus building by the 3D-Jury system. Numbers in parentheses correspond to the number of the first models of these servers that have more correctly predicted residues. Targets in italics denote those two cases for which the number of correctly predicted residues was lower than the corresponding mean value derived from the first models of these servers.

b Number of residues in the experimental structure.

CM(B), trivial comparative modeling (template structure detectable by BLAST); CM(P), easy comparative modeling (template structure detectable by PSI-BLAST); CM/FR difficult comparative modeling (template structure detectable by transitive PSI-BLAST).

Number of structurally equivalent residues between the model and the target, and between the structurally closest template and the target, derived from sequence-independent superposition generated with LGA program26 with 5 Å equivalence cutoff.

*Number of residues in the model that were correctly aligned as derived from sequence-independent superposition. Numbers in parentheses (alignment accuracy) correspond to fraction of correctly aligned residues within structurally equivalent residues.
**Fig. 1.** Sequence-to-structure alignment for target T0143. 

**a:** Experimental structure of the target colored by the efficacy of the consensus alignment (green) and 3D assessment (violet and orange) approaches in obtaining correct residue mapping by taking as a consensus of server results the most common alignment variant in each region and by evaluating corresponding 3D models for several alignment alternatives, respectively. The remaining parts of the target for which the majority of models generated by primary servers had correct alignment with the respective parent structures are shown in white.

**b:** Sequence alignment for the target family including several close homologs and templates. Identical and similar residues present in more than 50% of all sequences are highlighted in red and blue, respectively. Positions of conserved hydrophobic residues, shown in (a) as sticks, critical for local hydrophobic packing of the protein, are denoted by asterisks. Secondary structure assignment at the top corresponds to the target structure.

**Fig. 2.** Sequence-to-structure alignment for target T0138. 

**a:** Experimental structure of the target. Coloring scheme is the same as in Figure 1.

**b:** Sequence alignment for the target family and several related structures possessing flavodoxin-like fold. Positions of the conserved hydrophobic residues, shown in (a) as sticks, critical for packing of the protein core, are marked with asterisks. Secondary structure assignment at the top corresponds to the target structure.
allows for building 3D protein models without the time-
consuming separate stages of core region identification,
and loop region building or searching, that are inherent
in manual homology modeling schemes. In most case the
main differences between models and targets were located
outside the protein core, where parental structures differ
significantly, indicating that modeling of variable regions
should be improved.

T0143 (V8 protease from S. aureus)

Staphylococcus aureus V8 serine proteinase belongs to the
peptidase family SB2 that includes also the structurally
characterized exfoliative toxins A and B from the same
species. This protein is composed of two structurally
similar domains, each with a central six-stranded antipar-
allel β-sheet folded into a β-barrel. Although this target
displays a relatively high sequence identity to the closest
templates, i.e. exfoliative toxins A (28%) and B (31%),
that was used as alignment anchor. Since two aspartic acid
residues (Asp 79 and Asp 82) were found in the correspond-
ing region of the target and, unfortunately, only a few
homologues sequences were available in the non-redun-
dant protein database, a sequence search for other ho-

mologs through unfinished genomes, followed by multiple
sequence alignment revealed which aspartic acid is con-
served within the target family.

Importantly, the total number of correctly predicted
residues for this target was significantly higher than for
any model submitted by servers. Nevertheless the result-
ing alignment quality was quite low with two α-helices
misaligned by several residues. What was the reason for
this? Firstly, correct alignment for one of these helices was
not among the variants produced by servers. Secondly, the
relative orientation of these α-helices was changed signifi-
cantly in the target structure compared to principal parent
(Poly(A) polymerase). Due to variations in the local struc-
tural environment, evaluation of 3D models based on the
principal template structure could not discriminate correct
and erroneous alignments. Probably using a different
template for this region with similar orientation of corre-
spending α-helices would help to improve the final results.

Fold Recognition

Analysis of fold recognition models was carried out in
the same way as for comparative modeling predictions. An
overall summary of FR predictions submitted to CASP5 is
presented in Table II.

Importantly, all targets that were assigned to a specific
fold with relatively high confidence by the 3D-Jury system
(T0134, T0138, T0147, T0157, T0174 and T0193_1) turned
out to be correctly predicted. Additionally, guided by
3D-Jury results for distant homologs submitted to the Meta Server, fold identification for T0135, T0148, T0156 and T0191_1 was also successful. As a result for all FR(H) targets remote homologs of known structure were recognized. However, it should be noted that targets in this fold recognition subcategory are easier to predict and align correctly as they share conserved key residues with their relative structures. The average alignment accuracy of 54% obtained for these targets is comparable to that achieved in the hard comparative modeling subcategory with two outstanding predictions for targets T0134_2 and T0138. In the analogous fold recognition (FR(A)) subcategory, embracing those modeled proteins that display strong structural similarity to known folds, but potentially analogous relationship, the correct fold was identified for five out of eight domains. On the other hand these targets were modeled with significantly low average alignment accuracy of 38% with only one exception for T0148 where the quality of submitted prediction achieved a more or less satisfactory level.

The main reason for the low alignment quality was the relatively small number of correct server predictions that could be used for consensus alignment building. This resulted in only a few regions, with single dominant alignment variant, that could be assigned a high confidence level. Importantly, in majority of cases these regions turned out to be predicted correctly. Additionally, for the remaining regions, in many cases the 3D evaluation procedure failed to distinguish between correct and erroneous alignments due to large local structural variations in the target protein and/or nonoptimal template selection. Use of the structurally closest template in the 3D assessment procedure could possibly overcome some of these problems. Moreover, for many of questionable regions, the correct alignment was even not among any of the variants produced by the servers. This suggests that other alternatives, apart from those provided by individual servers alone, have also to be tested.

On the other hand it should be noted that the applied methodology combining consensus alignment approach with 3D assessment resulted in models that had higher overall quality than an average first model from eight selected primary servers used by the 3D-Jury system. In addition, almost half of the submitted predictions were better than any of the models originating from these servers. Importantly, at this extremely low level of sequence identity applying the 3D assessment procedure not only for the target but also for its close homologs seems to be necessary. This may enable detection of significant alignment errors, which manifest themselves in 3D models only for some family members.

Optimal template selection is an arduous problem in modeling of distantly related proteins. Although in general sequence homology correlates with structural similarity for fold recognition targets the structural parent closest by sequence does not always have the most similar structure. Moreover, for targets sharing little sequence similarity with the related known structure, fold recognition methods can sometimes assign the correct fold, but are less successful in identifying the optimal template. As a consequence, only for two FR targets the structurally closest templates were among those used in the modeling.

**T0134_2 (C-terminal subdomain of delta-adaptin appendage domain from H. sapiens)**

Target T0134 shares domain structure with its remote homologs belonging to the clathrin adaptor appendage family: α-adaptin AP2 and 2-adaptin AP2. This relationship was correctly identified with high confidence by the 3D Jury system. C-terminal subdomain (T0134_2) has a three-layered α-β-α fold composed of five-stranded antiparallel β-sheet flanked by three α-helices. The model for this subdomain appeared to be essentially correct and very close to the experimental structure, achieving the quality of a comparative modeling prediction. The predicted alignment was almost perfect with only one out of five β-strands that form a central β-sheet misaligned by one residue, resulting in an RMSD between the model and the experimental structure of only 1.83 Å for all Co atoms. However, it should be emphasized that the correct mapping of residues was not trivial for approximately half the target protein. Three out of five β-strands were correctly aligned using consensus strategy and structure-based evaluation of several alignment variants. On the other hand sequence-structure mapping for the external β-strand was less successful with the sequence of this secondary structure element being shifted by one residue. For this region most of the servers suggested two major alignment variants. Because none of them could be dismissed by the structure-based selection, the dominant one was taken and turned out to be incorrect. Retrospective analysis shows that the target and the template 1qts display greatest local sequence similarity in the second, correct alternative. This finding might probably enable selection of the correct alignment variant.

**T0138 (KaiA N-terminal domain, S. elongates)**

This target is related to several two-component-type receiver domains embracing those from CheY, FixJ, DctD and NarL and AmiR, as clearly indicated by the 3D-Jury system. These proteins belong to CheY-like superfamily with a flavodoxin-like fold composed of an α-β-α sandwich built around a five-stranded parallel β-sheet. The results obtained for this target demonstrate that structure-based assessment of candidate alignments, combined with detailed analysis of fold characteristics and sequence conservation within the target family, proved to be effective in some very difficult cases. Importantly, no alignment errors were introduced in the secondary structure elements, and the only mistake was the prediction of an α-helix for a region that appeared to be a long loop in the experimental structure. For this target the most difficult alignment regions were located in two C-terminal secondary structure elements. Importantly, correct residue mapping for these C-terminal β-strand and following α-helix was not among variants produced by servers. Because no dominant alignment was observed for this region, several alternatives were generated manually, guided by secondary structure assignment derived from NMR chemical shifts available in the BMRB (BioMagResBank, University of
CONSENSUS ALIGNMENT APPROACH IN CASP5

Wisconsin-Madison) database. Moreover, detailed examination of more than 10 template structures possessing this fold, combined with target family sequence analysis, revealed that conservation of hydrophobic residues at specific positions is critical for hydrophobic packing of the protein core (Fig. 2). This finding, followed by building and evaluation of several models for target and close homologues, led to selection of the correct alignment variant.

CONCLUSIONS

The methodology developed and used for modeling CM and FR targets was based on a combination of the consensus alignment and the 3D assessment approaches. The general idea behind the consensus approach is similar to that applied by other consensus algorithms, which extract common structural motifs from the set of 3D models generated by various independent prediction servers. The results show that the applied methodology can improve the quality of models generated by primary servers used for consensus building.

The application of the 3D-Jury method resulted in the identification of the correct fold for almost all difficult targets. However, the alignment accuracy remained below satisfactory level. The main reasons for the alignment errors in the structurally conserved regions included structural variations of the surrounding environment in the target protein, nonoptimal template selection in the 3D assessment procedure or insufficient sampling of possible alignment variants. These findings indicate that for difficult targets there is still a lot of room for further improvement both in optimal template selection, as well as in the quality of the sequence-to-structure alignments that are particularly error-prone in cases of low sequence similarity. Additionally, comparative modeling and fold recognition predictions could also benefit from a modified ab inito protocol for modeling of variable regions.

ACKNOWLEDGMENTS

The authors would like to thank Dr. D. Shugar for critical reading of the manuscript. Furthermore, the authors are grateful to CASP5 organizers and assessors for their hard work as well as for providing extensive evaluation data. The authors also acknowledge experimental groups for providing prediction targets and their structures.

REFERENCES