PROTEINS: Structure, Function, and Genetics Suppl 3:73–80 (1999)

Addressing the Issue of Sequence-to-Structure Alignments in Comparative Modeling of CASP3 Target Proteins

Česlovas Venclovas,* Krzysztof Ginalski, and Krzysztof Fidelis
Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California

ABSTRACT **During a blind protein structure prediction experiment (the third round of the Critical Assessment of Techniques for Protein Structure Prediction; URL http://PredictionCenter.llnl.gov/casp3/), four target proteins, T0047, T0048, T0055, and T0070, were modeled by comparison. These proteins display 62%, 29%, 24%, and 19% sequence identity, respectively, to the structurally homologous proteins most similar in sequence. The issue of sequence-to-structure alignment in cases of low sequence homology was the main emphasis. Selection of alignments was made by constructing and evaluating three-dimensional models based on series of samples produced mainly by automatic multiple sequence alignments. Sequence-to-structure alignments were correct in all but two regions, in which significant changes in target structures compared with related proteins were the source of errors. Template choice is an important determinant of model quality, and a correct selection was made of a lower homology template for modeling of T0070; however, in the case of T0055, a template with 8% greater sequence homology proved deceptive. Loops and some ungapped template regions were assigned conformations taken from other proteins. Using fragments from homologous structures led to improvement over template backbone more often than cases in which nonhomologous structures were the source. The results also indicate that side-chain prediction accuracy depends not only on sequence similarity but also on accuracy of the backbone. Proteins Suppl 1999;3:73–80.** **Published 1999 Wiley-Liss, Inc.†**

Key words: protein structure prediction; three-dimensional model; model evaluation; low sequence homology; alignment errors

INTRODUCTION

In general, any comparative modeling approach has to face a number of issues, including selection of parent structure (template), identification of structurally equivalent pairs of residues between the target sequence and template structure, modeling of regions not present or significantly different from those in template, and positioning of side-chains. Any of these steps are likely to introduce some errors that, as a rule, will not affect different parts of the structure equally. To be useful for most practical applications, a protein model should carry an indication of which regions are "trustworthy" and which are likely to contain more pronounced errors.

For the third round of Critical Assessment of Techniques for Protein Structure Prediction (CASP3), we have submitted models of four target proteins: 1) major urinary protein α-2u-globulin from rat (target T0047; 62% sequence identity with template closest by sequence), 2) pterin-4-α-carbinolamine dehydratase from *P. aeruginosa* (target T0048; 29% sequence identity), 3) calcium-dependent lectin from tunicate *P. misakiensis* (target T0055; 24% sequence identity), and 4) outer membrane porin Omp32 from *C. acidovorans* (target T0070; 19% sequence identity). Wide sequence identity range (62–19%) makes these proteins quite representative of all CASP3 targets amenable to classical comparative modeling. Because we concentrated on testing our sequence-to-structure alignment procedure, the emphasis was made purposefully on modeling low sequence homology targets, for which sequence alignments with respective parent structures are known to be the major source of errors.

Due to our (Č.V. and K.F.) affiliation with the Prediction Center, where all CASP models were deposited, we took some self-restricting measures to attest to the originality of our predictions. We would submit our model only if it was the first three-dimensional (3D) model deposited for a given target. In addition, at the time of submission, a copy of the model along with the date stamp was sent to the CASP3 independent assessor (Alwyn Jones).

MATERIALS AND METHODS

Selection of Parent (Structural Template)

Related proteins with known structures were identified by searching target sequences against the Protein Data Bank (PDB)[1] using the Smith-Waterman algorithm[2] implemented in the SSEARCH program.[3] The protein that produced the greatest sequence similarity score was selected as the structural template in all but one case. Determining factors for T0070 parent selection were the quality of multiple sequence alignment and the evolutionary distance between species (see Results and Discussion,

Č. Venclovas' permanent address is Institute of Biotechnology, Graičiūno 8, 2028 Vilnius, Lithuania.
K. Ginalski's permanent address is Department of Biophysics, Institute of Experimental Physics, University of Warsaw, Żwirki i Wigury 93, 02-089 Warsaw, Poland.
*Correspondence to: Česlovas Venclovas, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551. E-mail: venclovas@llnl.gov
Received 12 March 1999; Accepted 26 April 1999

below). When multiple PDB entries for the same parent were available, the one with the greatest resolution and the most complete set of atoms was chosen.

Sequence-to-Structure Alignment

This procedure had the heaviest weight in our model-building scheme and included both sequence-based and structure-based methods to produce and evaluate alignments.

Defining alignable regions of the template

The goal of this step was to identify conserved structural regions in the selected parent, in which alignment is meaningful. The FSSP database[4] was inspected to determine whether the protein selected as a structural parent had other closely related structures. If other similar structures were found, then alignable regions were derived from 3D superpositions of all related proteins. Regions were considered alignable if they were structurally conserved throughout this set of proteins with pairwise distances between corresponding residues typically not exceeding 2.5 Å. If no related structures for the parent protein were found, then secondary structure elements were used as a first approximation of alignable regions.

Multiple sequence alignments

Multiple sequence alignment was used as the initial step to derive sequence-structure mapping of target and template. Homologous sequences that matched the target with E-values $< 10^{-5}$ were collected from a nonredundant protein sequence database using the gapped BLAST algorithm.[5] To avoid to some extent the domination of a group of very similar sequences in the alignment, the resulting set of proteins was subjected to pairwise sequence similarity checks. Only those sequences that were $< 80\%$ identical to any other sequence were considered further. Alignment of multiple sequences was performed with the PILEUP program (GCG Inc., Madison, WI) using the Blosum50 substitution matrix.[6] In each case, 14 alignment variants were produced by gradually lowering the gap opening penalty from 12 (default) to 6 and using extension penalties of both 2 and 1. All of the obtained alignments were inspected for both variability and violation of structural integrity of the alignable regions in the template, such as insertions or deletions within secondary structure elements. If the resulting alignments for such regions were greatly dependent on gap penalties, or if they were inconsistent structurally, then a number of manual alignments were derived by using the PHD[7] secondary structure prediction as a guide.

Selecting final sequence-to-structure alignment

All plausible alternative sequence-to-structure alignments obtained as described above were evaluated by building 3D models and testing fitness of the resulting residue mapping with the structural scaffold of the template. Models were built using the Homology module of InsightII (MSI Inc., San Diego, CA). For regions that were defined as alignable, backbone conformation was taken from the template structure, and only side-chains were substituted. As a rule, modeling of most loops was skipped for the 3D structures that were built to test the fitness of a particular alignment variant. After the coordinates were assigned to the target sequence, side-chains were rebuilt by using a backbone-dependent rotamer library.[8] Along with visual inspection, structural consistency of the models was evaluated primarily by using the ProsaII[9] energy profiles under the assumption that, if correct alignment was among those tested, then it should have produced the lowest energy values. For detailed checks of model quality, a structure verification module of the WHATIF program [10] was used. It should be noted that structural evaluation was a decisive factor in finalizing the sequence-to-structure alignments.

Decorating 3D Model

After selection of final alignment, backbone conformation of the model was revisited and substituted with chain fragments from other related structures if these fragments represented local sequence better and/or improved residue-residue interactions. All variable regions (mainly loops) were assigned conformations derived from the fragments of known protein structures. The search for a suitable fragment to represent the conformation of a particular loop was performed first among homologous structures and, subsequently, among nonhomologous structures in the PDB. The goodness of fit for any candidate loop was estimated mostly from the deviation of the flanking regions, but interactions with the structural environment also were taken into account.

To either remove remaining side-chain clashes or preserve apparently conserved contacts, some side-chain rotamers were set manually. Final models were subjected to 100 or fewer steps of steepest descent minimization by using the Discover module of InsightII to improve stereochemistry without significantly changing position of atoms.

Error Estimates

Estimation of expected deviations between model and experimental target structure was derived from structural superposition of template with the homologous structure, which best represented the level of target-template sequence similarity. Deviations up to a 3Å cut-off between corresponding Cα atoms were tabulated and assigned without changes to the backbone and Cβ atoms of corresponding residues in the model. Model regions (mostly loops) for which automatic assignment was not made were assigned error estimates manually. For side-chains, error estimates increased with the number of bonds separating a particular atom from Cα and with the side-chain solvent accessibility as calculated in the model structure.

RESULTS AND DISCUSSION

An overall summary of the models based on a comparison with the currently available experimental structures is given in Table I.

TABLE I. Overall Summary of Models Submitted to CASP3†

	Parent			Model	
Target	PDB code	Cα RMSD	Sequence identity (%)	Cα RMSD	All atom RMSD
T0047 (158)	1MUP (157)	1.2/157	62	1.3/158	3.5/1,287
				1.2/157	3.5/1,279
T0048 (116)	1DCP (99)	1.5/97	29	6.6/116	7.4/923
				2.0/99	3.8/800
T0055 (123)	1ESL (157)	2.2/112	24	3.8/123	4.7/966
				3.8/122	4.7/958
T0070 (332)	2OMF (340)	2.0/247	17	3.2/230	4.6/1,744
				—	—

†Numbers in parentheses next to the target and parent identifiers correspond to the number of residues in the 3D structures. Root mean square deviation (RMSD) values are presented together with the number of atom pairs used to derive a given value. Values for the parents were obtained from Dali superposition of target and parent structures.[11] Percentage of sequence identity both in this table and throughout the text is given for structurally equivalent residues. For the models, the first row corresponds to the whole model structure, the second row corresponds to the model structure after excluding N- and/or C-terminal residues that did not have counterparts in parent structure and had to be modeled de novo. The apparently high RMSD value for the complete T0048 model is due mainly to the contribution of 17 such residues. PDB, Protein Data Bank.

Sequence-to-Structure Alignments

Out of the four targets predicted, only the T0047 (62% sequence identity to the parent 1mup) sequence-to-structure alignment could be obtained trivially without any insertions or deletions. To evaluate the performance of our sequence-to-structure alignment protocol for the remaining three targets, first, we obtained a structure superposition of each target and its respective parent using Dali[11] and selected the regions in which assignment of structurally equivalent pairs of residues was straight forward. Such regions, defined as a "core," consist of at least three contiguous residues for which distances between corresponding Cα atoms of target and parent do not exceed 2.5 Å, and are complementary to the "loop" regions.[12] Next, target and model structures were superimposed in a sequence-independent manner with Dali,[11] and the resulting alignments in "core" regions were inspected for errors. Models of two targets had one region each where errors could be attributed to wrong sequence-to-structure mapping. In the first model (T0055), a sequence fragment corresponding to an α-helix was shifted by three residues (Fig. 1a). In the other model (T0070), a short β-strand was misaligned (Fig. 2a).

Our alignment procedure can be considered as a two-step process: generating a number of candidate alignments and then selecting one that produces the best 3D model according to the evaluation criteria. The question is, which one of the steps failed in these two cases and why?

In the case of T0055, multiple sequence alignment procedure suggested two major alignment variants for the considered α-helix, and both required structural modification in the preceding region that was conserved in all available parent structures (Fig. 1). One of these variants corresponded to the correct mapping (insertion of two residues) and the other corresponded to the wrong mapping (one-residue deletion), as in the submitted model. Surprisingly, evaluation at the 3D level showed that structural changes associated with both variants were unfavorable within the framework of the template. A one-residue deletion resulted in placing the polar serine (S31) side-chain in a hydrophobic environment and, at the same time, exposing most of the bulky hydrophobic methionine (M33) side-chain. On the other hand, a two-residue insertion could not be accommodated without extensive steric clashes. It is noteworthy that the sequence pattern of the adjacent helix seemed to favor the sequence mapping, which turned out to be incorrect (Fig. 1a). Because models based on both alignment variants had significant structural flaws, final selection of the alignment was made arbitrarily. It turned out that, as the experimental structure of T0055 became available, a two-residue insertion was followed by extensive changes both in the immediate vicinity as well as in more distant parts of the protein chain (Fig. 1b). One of these was a significant change in the orientation of the adjacent helix. Two loops that were close in 3D in the parent structure moved farther away, changing dramatically the environment of the newly formed short helix in the target structure.

Unlike T0055, in the case of target T0070, multiple sequence alignment studies did not provide distinct alternatives in regions that turned out to be misaligned (Fig. 2a) due to an unexpected structural change. Subsequent model-building studies also did not indicate any structure-related violations in the misaligned region. It is interesting to note that analysis of the experimental T0070 structure and its comparison with other related porin structures did not provide an immediate answer why this conserved strand was interrupted by insertion of a four-residue loop (Fig. 2b). One possible explanation is that two hydrophobic residues (Tyr 27 and Leu 29) that were predicted to be part of a long loop (and, thus, were not modeled), in fact, are determining structural factors for this chain fragment. The side-chain of Leu 29 contributes to the nonpolar core within the trimeric structure, whereas

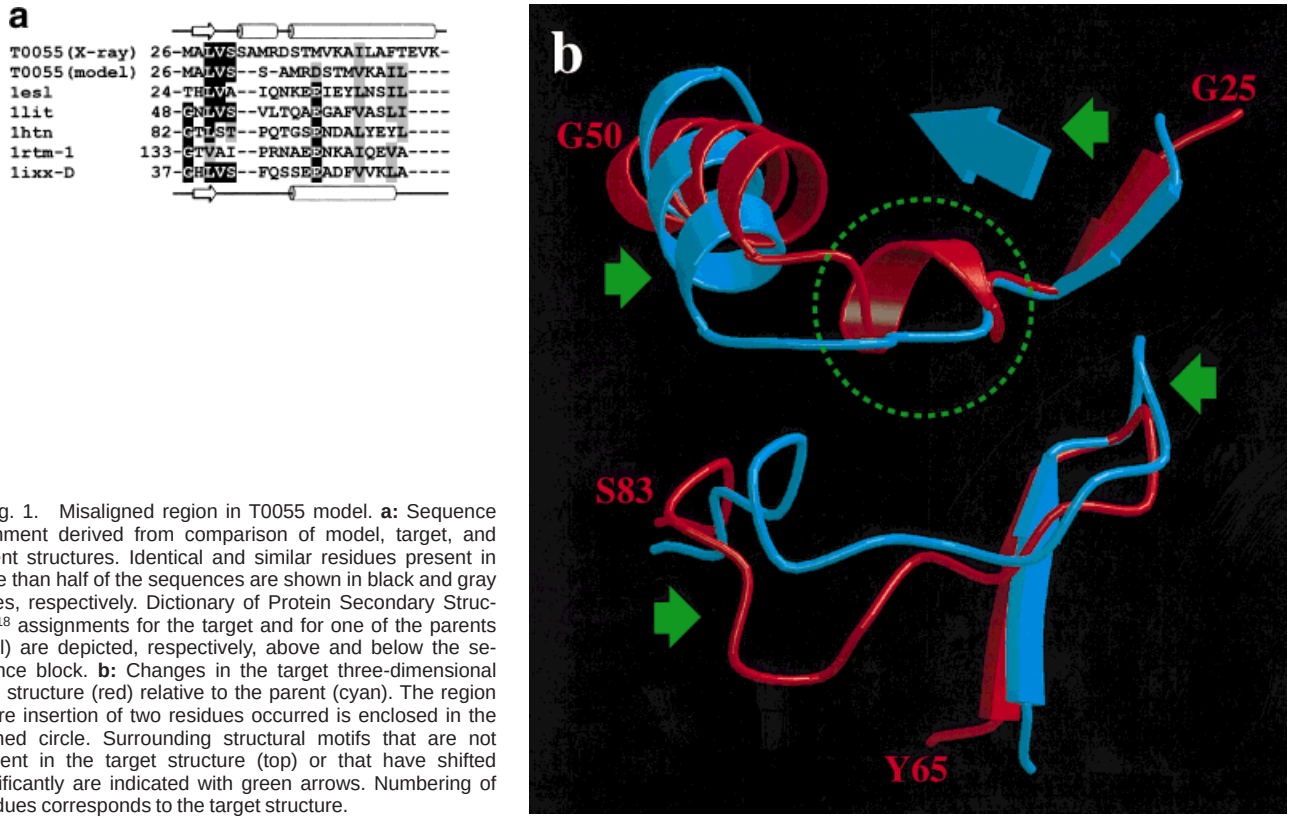

Fig. 1. Misaligned region in T0055 model. **a:** Sequence alignment derived from comparison of model, target, and parent structures. Identical and similar residues present in more than half of the sequences are shown in black and gray boxes, respectively. Dictionary of Protein Secondary Structure[18] assignments for the target and for one of the parents (1esl) are depicted, respectively, above and below the sequence block. **b:** Changes in the target three-dimensional (3D) structure (red) relative to the parent (cyan). The region where insertion of two residues occurred is enclosed in the dashed circle. Surrounding structural motifs that are not present in the target structure (top) or that have shifted significantly are indicated with green arrows. Numbering of residues corresponds to the target structure.

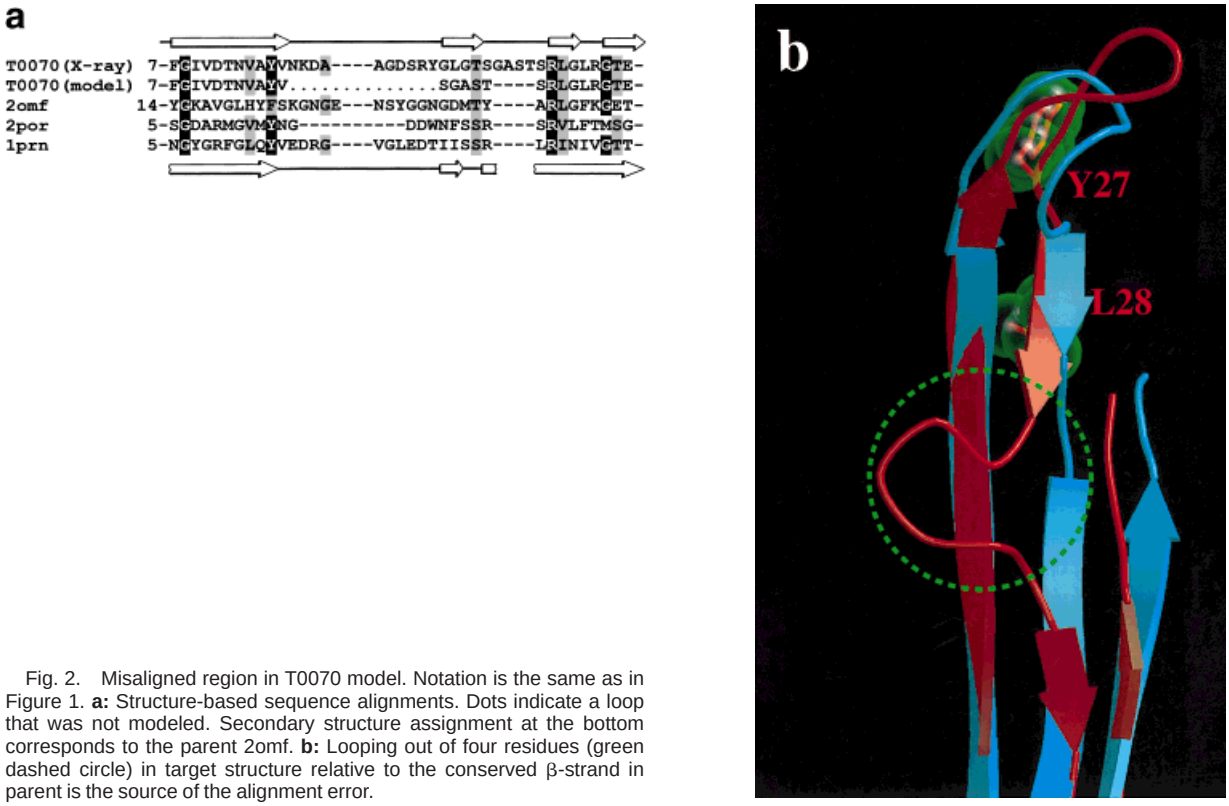

Fig. 2. Misaligned region in T0070 model. Notation is the same as in Figure 1. **a:** Structure-based sequence alignments. Dots indicate a loop that was not modeled. Secondary structure assignment at the bottom corresponds to the parent 2omf. **b:** Looping out of four residues (green dashed circle) in target structure relative to the conserved β-strand in parent is the source of the alignment error.

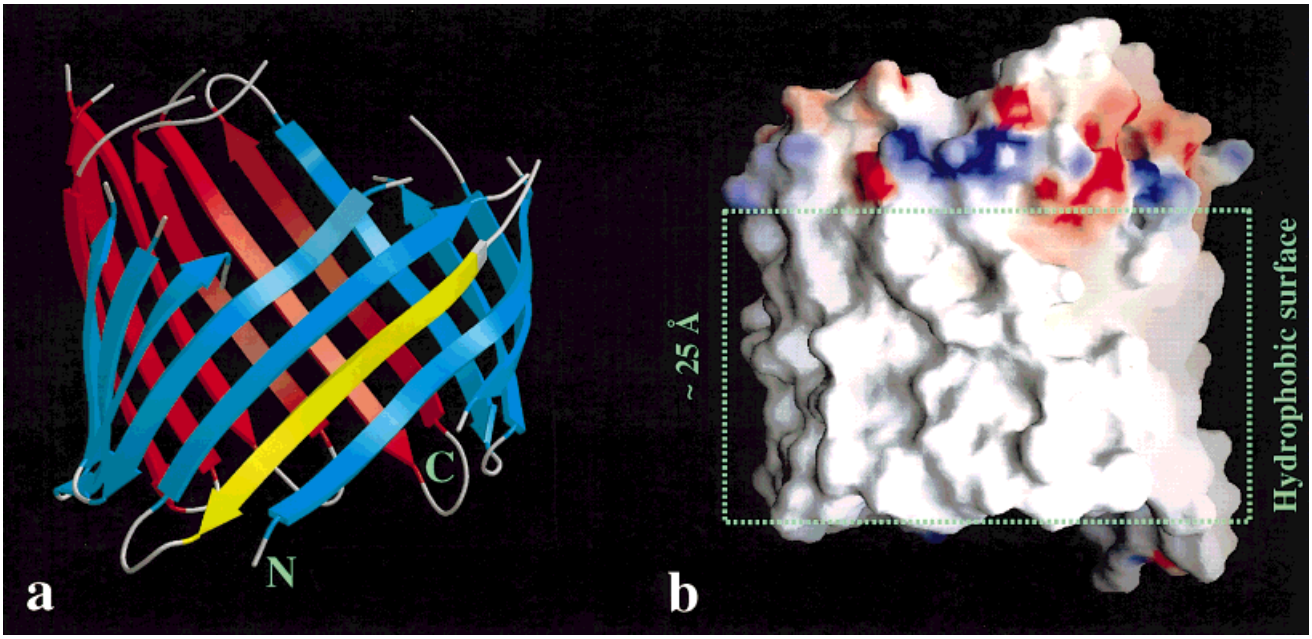

Fig. 3. Sequence-to-structure alignment of structurally conserved β-barrel in T0070. **a:** T0070 model structure color coded by the efficacy of the three approaches to produce correct alignment: from only one strand obtained with a pairwise alignment (yellow), to additional nine strands (cyan) gained with a multiple sequence alignment, and to the complete set, including six remaining strands (red), obtained by evaluation of 3D models. White color at the beginning of the yellow strand indicates the misaligned region. **b:** GRASP[19] representation of the molecular surface electrostatic potential for T0070 experimental structure. The polar surface is colored blue (positively charged) and red (negatively charged), and the apolar surface is white. The rectangle with its height corresponding roughly to the thickness of a membrane, denotes surface region facing the membrane lipid environment.

Tyr 27 is likely to be anchored at the membrane interface in the native environment. This may be sufficient to force the observed shift in the strand register together with subsequent looping-out of four hydrophylic residues into the solvent-exposed barrel interior.

Do these errors diminish the importance of the 3D evaluation of alignments? It appears that, in cases like those described above, some alteration of the template structure and/or incorporation of variable regions into the 3D model is necessary. However, we believe that even when the template is kept rigid and loops are omitted, model construction followed by evaluation is the most effective way to avoid alignment errors. One of the advantages of such an approach over sequence methods is that, at the 3D level, a larger number of different constraints can be applied, including both general properties of protein structure as well as structural features that are specific to a given protein family. Retrospective analysis of T0070 modeling (Fig. 3) provides an excellent example of the efficacy of structure-based methods versus sequence-based methods in obtaining sequence-to-structure alignment. Despite low sequence similarity, structures of this porin family have reasonably conserved, 16-strand, antiparallel β-barrels. By using multiple sequences, ten of the barrel strands could be mapped unambiguously, whereas the alignment for the rest of the chain was highly variable and structurally inconsistent. For comparison, pairwise sequence alignment was able to match correctly only 1 of the 16 barrel strands. To align the remaining six strands, a number of manually generated sequence alignments were evaluated for consistency of the corresponding 3D models with an important structural feature specific to the porin family. Porins are located in the membrane, and the outer surface of the β-barrel, which is in immediate contact with the lipid environment, is extremely hydrophobic (Fig. 3b). This feature was exploited successfully to select an alignment that produced the most hydrophobic surface of the β-barrel.

An important problem in identifying correct sequence-to-structure alignment at the 3D level is how to provide adequate sampling of candidate alignments. It seems that, by producing and analyzing series of multiple sequence alignments, one can reduce significantly the number of alternative alignments, at the same time addressing the issue of reliability for each of the considered regions. However, our CASP3 experience suggests that, at very low sequence similarity, sampling beyond series of routinely generated multiple sequence alignments may be necessary.

Selection of Parent Structure (Template)

Although, in general, protein sequence homology strongly correlates with structural similarity,[13] the protein closest by sequence does not always have the most similar 3D structure. An excellent example of such behavior is target T0055. Our choice of structural template was human

selectin (PDB code 1esl), which displays the greatest sequence similarity to the target among related structures (24% sequence identity). However, it turned out that at least three other structures with lower sequence homology were better overall structural templates. Moreover, the one least similar by sequence (1htn; only 16% sequence identity) had the closest structure.

The only other modeled target for which selection of a main parent was an issue was T0070. In that case, sequence similarity with porins of known structure was too weak to be considered a reliable indicator of structure relatedness. Although porin from *R. capsulatus* (2por) had the best pairwise sequence comparison score, *E. coli* porin (2omf) and its close relatives produced more consistent multiple sequence alignments. In addition, *C. acidovorans*, the source organism of the T0070, is evolutionary closer to *E. coli* than to *R. capsulatus*; therefore, we expected the corresponding porin structures to have greater similarity as well. Comparison using Dali[11] revealed that structurally equivalent "core" (residues that deviate within 2.5 Å) is considerably more extensive in *E. coli* porin (198 residues) than in *R. capsulatus* porin (170 residues), providing support for our choice of the parent structure.

Is it possible routinely to select the best template? What other features, apart from sometimes misleading sequence similarity, can be used? Obviously, a number of factors can help identify the best template, including the quality of multiple sequence alignments, functional similarity, nature of ligands, etc. However, although in specific cases such analysis can be effective, it cannot always be applied. It seems that constructing and evaluating models based on all available structural templates can provide a more general solution for selection of the closest parent structure or the specific region of greatest local structure similarity. The latter issue is addressed below in the analysis of parent backbone modification results.

Modification of Parent Back-Bone and Loop Modeling

When the sequence of the target is mapped onto the template structure, regions of protein chain can be classified into two categories: 1) those for which structural correspondence is assigned and backbone structure, in principle, can be inherited directly from the template; and 2) those that must be modeled explicitly due to insertions/deletions respective to the template structure or absence of the corresponding template regions altogether. Analysis of the model quality in regions of the first type can answer the question of whether the modeling was more effective than simple inheritance of template backbone. The quality of regions of the second type is related directly to the loop-building efficiency, assuming that alignment of adjacent regions is correct. Table II summarizes modeling results for both region types. To obtain root mean square deviation (RMSD) values, each modeled structure first was put in the same frame of reference as the parent structure, so that template regions that simply were copied over to the model were superimposed ideally. Next, preserving their mutual orientation, parent and model structures were superimposed with the target using only the "core" (for definition, see Sequence-to-Structure Alignments, above) residues of the parent.

The results presented in Table II indicate that using fragments from alternative parents to model conserved in length stretches of backbone did lead to improvement over template in all but one case, in which, essentially, there was no change (0.96 Å vs. 0.92 Å). At the same time, only one of the fragments taken from nonhomologous structures was able to drive the model closer to the target structure. The quality of explicitly modeled variable regions, in many cases, is determined primarily not by how well local backbone conformation is predicted but by other factors. Significant deviations of backbone in flanking regions and alignment errors are known to be major contributors to the poor quality of modeled loops.[14] This point is illustrated in Table II, where the T0055 loop following the misaligned helix is the worst among all loops for this target.

Accuracy of Side-Chain Prediction

The accuracy of side-chain rotamer prediction, as might be expected, correlates with the level of sequence homology between target and parent. The percentage of correctly predicted χ_1 rotamers (within $\pm$ 30°) ranges from 63% for the model of T0047 to 40% for T0070. If only the structurally conserved part of the target ("core") is considered, then the accuracy gap for χ_1 tends to shrink: The accuracy of side-chain prediction for T0070 increases most, reaching 48%, whereas, for T0047, it remains essentially at the same level. These results confirm previous observations (see, e.g., Chung and Subbiah[15]) that correct selection of side-chain rotamers depends considerably on the backbone accuracy.

Energy Minimization

We used very limited energy minimization to improve the stereochemistry of models for all four targets; however, only for the high homology target T0047, we tested the effect directly. Although it was very small, the effect of energy minimization was positive by several criteria: Cα RMSD went down by 0.01 Å; the contact area difference value,[16] which reflects the similarity of atomic contacts in two structures, improved by 1%; and the number of correctly predicted χ_1 rotamers increased by 2%. However, in general, these values are less than the observed variation in the independently solved X-ray structures for the same protein[17] or between molecules related by noncrystallographic symmetry in the same crystal.[16]

Error Estimates

Considering the simplicity of the approach, estimates of deviation between atoms of predicted and experimental structures in general were reasonable, especially in the structurally conserved regions. The mean values of absolute differences between observed and estimated deviations of "core" Cα values did not exceed 0.8 Å for any of the models. Not unexpectedly, error estimates for variable regions of structure fared worse.

TABLE II. Summary of Structure Regions in Models That Were Assigned Back-Bone Conformation Other Than the Principal Parent†

Modeled region	RMSD T-M	RMSD T-P	Type of the region	Origin
T0048/1dcp-C				
3-ALTQAHCEACRADA-16	20.33	—	None	De novo
17-PH-18	3.33	2.29	P	NH
31-IPD-33	0.95	—	Del(2)	NH
37-EVRDGI-41	7.32	—	Ins(2)	NH
114-AE-115	**3.88**	**4.60**	P	NH
116-GRK-118	8.47	—	None	De novo
T0055/1esl				
2-D-2	2.25	—	None	De novo
24-RGM-26	**0.86**	**2.15**	P	H
29-VSSAM-33	4.83	—	Ins(2)	NH
45-FTEVKGH-51	9.86	4.71	P	NH
59-NLQDGAYN-66	4.34	—	Ins(3)	NH
70-NDGV-73	2.30	—	Del(1)	NH
99-IWSKYNLL-106	4.23	—	Del(4)	H
120-EK-121	0.96	0.92	P	H
T0070/2omf				
82-SGNFG-86	0.56	—	Del(1)	H
144-FGGFN-148	3.68	—	Del(4)	H
180-NGPL-183	**1.69**	**1.83**	P	H
212-YNFG-215	**2.13**	**3.27**	P	H
229-KRDI-GD-235	12.99	—	Del(8)	H
248-PVGGVGE-254	2.98	1.68	P	NH
264-QKAID-268	2.53	—	Del(4)	H

†Numbering of residue regions is according to target structures. T0047 is not included in the table, because none of the parent backbone regions was modified explicitly. RMSD T-M and T-P denote target-model and target-parent values, respectively. Region types indicate whether the template did not have a corresponding region (None), some residues had to be inserted (Ins) or deleted (Del), or it has the same number of residues and could be inherited (P) in the course of model building. Numbers in parentheses indicate how many residues should have been inserted or delted relative to parent structure. Origin indicates how the region was built: De novo, built from scratch; H, fragment taken from a homologous structure; NH, fragment from a nonhomologous structure. Values in boldface indicate improvements over template made in the specified region.

It is unfortunate that often the aspect of assigning some kind of a reliability index to the modeled structure is omitted. However, at the user's end, this factor becomes a necessity when it comes to deciding whether a particular conclusion can be derived from the model.

CONCLUSIONS

1) In comparative modeling, sequence-to-structure alignment and the selection of the optimal template(s) are the most critical factors determining the quality of the model.
2) Multiple sequence alignments are more sensitive than pairwise alignments, and their use may reduce dramatically the number of alternative alignments that should be considered.
3) Evaluation of 3D models is an effective procedure for selecting sequence-to-structure alignment. In many cases, it is sufficient to consider the 3D structure of only conserved regions. However, when target structure deviates considerably in at least some of these regions, evaluating a 3D model based on the preserved template structure can be ineffective and/or misleading. In such cases, including the variable regions either explicitly or implicitly in model evaluation may be necessary to produce correct alignment.
4) Loop construction and side-chain placement depend heavily on the correctness of alignment and the accuracy of the backbone on which they are based.
5) Estimation of position-specific reliability is a prerequisite for the usefulness of the modeled structure.

ACKNOWLEDGMENTS

We thank the crystallographers and nuclear magnetic resonance spectroscopists who provided target proteins. This work was performed in part under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory (contract W-7405-Eng-48).

REFERENCES

1. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.

2. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
3. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 1991;11:635–650.
4. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
6. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
7. Rost B, Sander C, Schneider R. PHD — an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994;10:53–60.
8. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
9. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
10. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
11. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
12. Venclovas Č, Zemla A, Fidelis K, Moult J. Criteria for evaluating protein structures derived from comparative modeling. Proteins Suppl 1997;1:7–13.
13. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
14. Martin AC, MacArthur MW, Thornton JM. Assessment of comparative modeling in CASP2. Proteins Suppl 1997;1:14–28.
15. Chung SY, Subbiah S. How similar must a template protein be for homology modeling by side-chain packing methods? In: Hunter L, Klein TE, editors. Biocomputing: proceedings of the 1996 Pacific Symposium. Singapore: World Scientific; 1996. p 126–141.
16. Abagyan RA, Totrov MM. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J Mol Biol 1997;268:678–685.
17. Flores TP, Orengo CA, Moss DS, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 1993;2:1811–1826.
18. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
19. Nicholls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11:281–296.