PROTEINS: Structure, Function, and Genetics 53:380–388 (2003)

Comparative Modeling in CASP5: Progress Is Evident, but Alignment Errors Remain a Significant Hindrance

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

ABSTRACT **Models for 20 comparative modeling targets were submitted for the fifth round of the "blind" test of protein structure prediction methods (CASP5; http://predictioncenter.llnl.gov/casp5). The modeling approach used in CASP5 was similar to that used 2 years ago in CASP4 (Venclovas, Proteins 2001; Suppl 5:47–54). The main features of this approach include use of multiple templates, initial assessment of alignment reliability in a region-specific manner, and structure-based selection of alignment variants in unreliable regions. The CASP5 modeling results presented here show significant improvement in comparison to CASP4, especially in the area of distant homology. The improvements include more effective use of multiple templates and better alignments. However, a number of structurally conserved regions in submitted distant homology models were misaligned. Analysis of these errors indicates that the absolute majority of them occurred in regions deemed unreliable in the course of model building. Most of these error-prone regions can be characterized by their peripheral location and a lack of conserved sequence patterns. For a few of the error-prone regions, all methods evaluated during CASP5 proved ineffective, pointing to the need for more sensitive energy-based methods. Despite these remaining issues, the applicability of comparative modeling continues to expand into more distant evolutionary relationships, providing a means to structurally characterize a significant number of currently available protein sequences. Proteins 2003;53:380–388.** © 2003 Wiley-Liss, Inc.

Key words: protein structure prediction; sequence-structure alignment; 3D model; model evaluation; distant homology; alignment errors

INTRODUCTION

Thanks to an explosive increase of protein sequence data in recent years and continuing accumulation of experimental three-dimensional (3D) structures, the applicability of comparative or homology-based modeling has expanded enormously. It is not surprising that comparative modeling constitutes an important part of the CASP experiments. Moreover, in CASP4 and even more so in CASP5 we have witnessed a strong tendency of comparative modeling to invade the territory of distant evolutionary relationships among proteins, which just a few years ago was considered to be reserved for the fold recognition methods. These developments suggest that, comparative modeling has already become one of the most effective computational approaches in facilitating structural/functional characterization of many protein-coding sequences across genomes.

I entered CASP5 (as group 425) with a comparative modeling approach having several goals. First, CASP is an ideal setting to test the current capabilities of a modeling method in general as well as various aspects of it. More specifically, I focused on the problem of sequence-structure alignment in cases of distant evolutionary relationship. This includes the ability to maximize correct alignment and to distinguish aligned regions according to their reliability. The latter is especially important if a protein model is to be used for understanding biological function or interpretation of experimental findings. Another goal was to find out whether there has been an improvement in model quality compared with the previous CASPs.[1,2] Finally, CASP provides a very effective reality check, or how good the approach is including its individual components in comparison with the best results, which often are considered to define the current state-of-the-art in protein structure prediction.

In this article I provide an overview of my modeling results and compare them with those obtained in CASP4. I also focus on the causes of alignment errors, paying particular attention to common features of these regions. I Use several examples to illustrate specific comparative modeling problems underscored by the modeling experience of CASP5 prediction targets and discuss two of the more successful predictions.

MATERIALS AND METHODS

The approach that I used in CASP5 was similar to that successfully introduced during CASP4.[2] In short, the main features of the modeling approach are the use of multiple templates to generate a model and differential treatment of the alignment based on the region-specific classification by the alignment reliability.

Č. Venclovas is also affiliated with Institute of Biotechnology, Vilnius, Lithuania.

*Correspondence to: Česlovas Venclovas, Biology and Biotechnology Research Program, L-448, Lawrence Livermore National Laboratory, Livermore, CA 94551. E-mail: venclovas@llnl.gov

Received 24 February 2003; Accepted 7 April 2003

Use of Multiple Templates

When available, multiple templates were used only when the level of structural similarity between each of the templates and the target protein was expected to be approximately the same. The estimation of the expected structural similarity was based on a standard PSI-BLAST sequence search against nonredundant NCBI (http://www.ncbi.nlm.nih.gov/) sequence database (nr) using the target sequence as a probe. Usually, the analysis was performed after an additional two or three iterations following the appearance of the first PDB match with a significant E-value in the results. Additional structural homologues that had comparable E-values with the best matching structure and a similar level of sequence homology were considered to be at approximately the same evolutionary distance from the target and, thus, were expected to share with it a comparable level of structural similarity. However, the evolutionary distance was mostly ignored, if additional templates could contribute structural motifs not present in the closest structural homologue. In all other cases, when the best PDB match was significantly closer to the target, the rest of the templates were not used to build an overall structure. At the same time, suitable fragments of homologous structures were often used to model insertions/deletions.

Sequence-Structure Alignment

For moderate and distant homology targets, sequence-structure alignments were first explored in a region-specific manner using the PSI-BLAST[3] intermediate sequence search (PSI-BLAST-ISS) procedure.[2] In this procedure, a set of sequences (~50–150) that are homologous to both the target and the template are used to generate corresponding PSI-BLAST profiles usually not exceeding five iterations. By using the SEALS package[4] and in-house Perl scripts, the aligned sequences for *the target and the template only* are extracted and compared. The result of this procedure is a multiple sequence alignment where a target sequence is aligned with a number of template sequence copies corresponding to different PSI-BLAST output files. Based on these PSI-BLAST-ISS results, individual regions were then initially classified either as reliably aligned (a single major alignment variant) or as those requiring further assessment (several alternative alignments present). In the latter case, two possibilities were considered: 1) the alignment is not reliable or 2) the local structure of the target is different from that in the template. An assessment of which of these two possibilities is more likely was usually made on examination of the superimposed protein structures related at the superfamily or fold level as defined in the SCOP database.[5] The regions, apparently structurally conserved but lacking reliable alignment, were further explored by using structure-based assessment. First, 3D models were produced on the basis of alternative alignments either taken from PSI-BLAST-ISS results or generated by using the guidance of secondary structure prediction results obtained from the CAFASP metaserver (http://bioinfo.pl/cafasp). The final alignment variant then was usually selected on the basis of consensus of evaluation results for these models. The usual assortment of evaluation procedures included visual inspection for significant structural flaws identified manually or by WHATIF[6] and ProsaII energy Z-scores and profiles [7] for the assessment of the overall quality of the structure and for the region-specific evaluation, respectively.

Modeling Tools

The actual tools used to transform sequence-structure alignments into 3D models included MODELLER[8] and SCWRL.[9] The choice of MODELLER for model building was mainly determined by its ability to incorporate structural information from multiple templates. SCWRL was used to position side-chains in the models for high and moderate homology targets. Notably, a spatial cluster of residues having a large number of residual clashes after rebuilding side-chains with SCWRL was also used as an indicator of significant flaws in a distant homology model, sometimes prompting to reanalyze alignment, the choice of the template, or both. Manual intervention in the positioning of side-chains was minimal.

RESULTS AND DISCUSSION

Overview of the Modeling Results

For the fifth round of the CASP experiment, I submitted models of 20 target proteins. Experimental structures for 18 of them were available in time for the independent assessment before the CASP5 meeting in December 2002. Although the submitted models account for roughly half of the comparative modeling targets, they constitute a representative set, because they do sample an entire range of the target difficulty in the comparative modeling prediction category.

Comparative modeling is essentially a template-based modeling. Thus, the extent of structural similarity between the modeling target and the template seems to be a good criteria in assessing how efficiently the structural information of the template has been used to obtain the model. By assuming the existence of a single template, a simple copying of the template's backbone should produce a model sharing the same number of structurally equivalent residues with the target as does the template. In the ideal case, all structurally equivalent residues also would be correctly aligned. In reality, there often are multiple templates, and in many cases, it is difficult to *a priori* select structurally the closest one. More importantly, as homology goes down, alignment errors become a major factor in determining the model quality.

Following the logic outlined above, a summary of my modeling results is provided in Table I. The data presented in this table were derived from the LGA[10] sequence-independent structural superposition of each target with the closest template and the corresponding model. Both the number of structurally equivalent residues (coverage) and the number of correctly aligned residues were taken directly from the residue correspondencies reported by LGA at 5 Å distance cutoff. For comparison, the data from CASP4, derived in the same way, are also included.

TABLE I. Summary of Modeling Results for CASP5 and Corresponding Data for CASP4

Target	Length (N)	T-P coverage (N)	Parent	Seq id (%)	AlO/T-P coverage (%)	T-M coverage/T-P coverage (%)
CASP5						
T0133	293	218	1hf8_A	12	77.1	**102.3**
T0141	187	113	1aro_L	14	61.9	99.1
T0152	198	134	1kux_A	14	97.0	**108.2**
T0132	147	119	1bvq_A	15	81.5	**102.5**
T0130	100	81	1fa0_B	17	63.0	95.1
T0192	170	144	1qsm_D	17	86.8	99.3
T0169	156	136	1l0c_A	17	90.4	**101.5**
T0160	126	119	1grw_A	21	87.4	99.2
T0150	97	96	1jj2_F	26	99.0	99.0
T0142	280	248	1i9z_A	27	85.5	99.6
T0143	216	202	1agj_A	27	98.5	**100.5**
T0178	219	213	1jcj_A	27	99.1	**100.0**
T0183	247	216	1ktn_A	30	97.7	**102.3**
T0151*	106	97	1eyg_D	33	88.7	**100.0**
T0155	117	117	1dhn	33	**100.0**	**100.0**
T0153	134	128	1euw_A	34	**101.6**	**101.6**
T0182	249	248	3mat_A	42	98.4	99.6
T0137	133	131	1pmp_A	43	**101.5**	**101.5**
CASP4						
T0089	378	242	1dkg_D	14	58.7	99.2
T0092	227	176	1d2c_A	14	71.0	89.2
T0090	199	114	1mut	17	43.0	78.9
T0103	368	239	1ak9	20	53.1	76.6
T0112	348	331	1bxz_D	26	75.5	93.7
T0121	372	218	1b0u	26	85.3	99.5
T0113	255	238	1ahi_B	27	93.7	**100.4**
T0122	241	234	1c29_A	32	84.2	98.3
T0099	56	46	1lck_A	41	93.5	95.7
T0111	430	418	5enl	51	99.3	**100.2**
T0123	160	138	1beb_B	54	75.4	**100.0**
T0128	211	205	1b06_D	55	98.5	99.5

The data in the table are sorted by the structure-based sequence similarity (**Seq id**) between **Target** and the template (**Parent**). **Length** is the number of residues in a target; **T-P coverage** is the number of structurally equivalent residue pairs in the target-template superposition; **T-M coverage** is the corresponding number in the target-model superposition, and **AlO** are the part of the residues in the **T-M coverage** that are reported by LGA as correctly aligned. Values in boldface indicate that the model either matches or improves over the best template either only structurally (the last column), or even by the extent of correct alignment (the "AlO/T-P coverage" column).

*The best matching template for T0151 is 1QVC, but its crystal structure has obvious errors (the two last β-strands are misthreaded); therefore, 1EYG is used instead.

In CASP5, my models for 11 of 18 available targets either match or improve on the best template in its structural similarity to the corresponding target (the last column in Table I). The added value in these cases originates from the use of multiple templates and/or loop assignments. These results are an improvement over CASP4, where my models for only 3 of 12 targets matched or exceeded structural similarity between the corresponding target and the best template. However, a good structural match between model and target is merely a prerequisite of a high-quality model: the structurally matching residues must also occupy identical positions in sequence (be correctly aligned). If the correctness of the alignment, which is one of the commonly used measures to assess model quality, is considered, the performance drops down significantly. Only models for three targets (T0137, T0153, and T0155) have an equal or larger number of residues aligned correctly compared with the number of structurally equivalent residues between a target and the corresponding best template (see the "AlO/T-P coverage" data in Table I). Nonetheless, again this is an improvement over CASP4, where none of my models has matched the respective best template with a number of correctly aligned residues. Moreover, the overall values of the "AlO/T-P coverage" ratio in CASP5 are shifted upward (61.9–101.6%) in comparison to CASP4 (43.0–99.3%) despite the opposite shift in the target-template sequence homology. It is important that the largest improvement is seen in distant homology range. For example, in CASP4, the ratio $> 80\%$ is achieved only for targets that are $>25\%$ identical in sequence to the corresponding templates. In CASP5, a similar level of accuracy is achieved for a

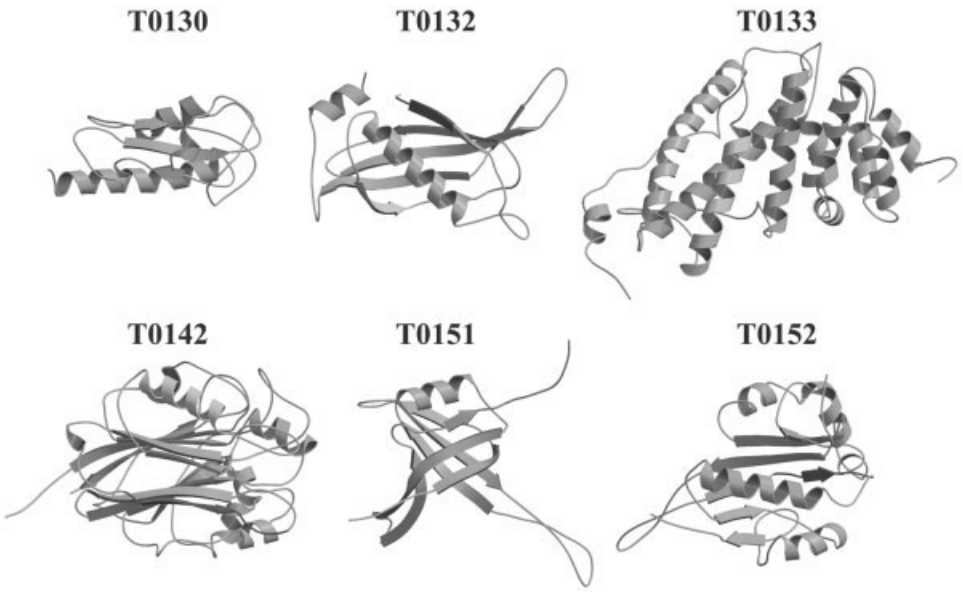

Fig. 1. Alignment errors in my CASP5 models. Misaligned regions are colored according to the difficulty of avoiding errors from the predictor perspective; red, orange, and yellow indicate, respectively, that none of the predictor groups, <5% and >5% of groups were able to produce correct alignment. Here and throughout the article only the highest confidence model ("Model 1") per each group was considered. This and other figures were prepared by using Molscript[23] and Raster3D.[24] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

number of distant homology targets with sequence identity as low as 14%.

Source of Improvement

I used similar approaches in CASP4 and CASP5, so where is the source of improvement? In the time between the two CASPs, the number of experimental protein structures continued to grow moderately. In contrast, during the same period of time the number of available protein sequences increased dramatically. This greatly expanded sequence space, combined with modern sequence comparison methods such as PSI-BLAST, clearly made a strong impact in detecting distant evolutionary relationships that could be exploited in comparative modeling. I believe that in many cases the better representation of evolutionary changes within protein superfamilies through the increased number of related sequences also contributed to more accurate initial alignments. This resulted in fewer regions defined as unreliable by PSI-BLAST-ISS, effectively decreasing the number of regions for which assessment at the level of 3D structure is required. Last but not least, the experience gained in previous CASPs[1,2] played a significant role in achieving better results. However, it would be difficult to dissect the observed progress into exact contribution of objective factors such as database increase versus human input.

Remaining Hindrances

The data in Table I and the above analysis indicate that the biggest potential for improvement of comparative modeling results is in the area of distant homology. One of the components of improvement rests in the ability to predict structural regions not represented in a structural template, still a daunting task. The other, perhaps more straightforward way toward improvement, is better use of information available in the form of related protein structures. More effective combination of multiple templates, but most importantly, further improvement in sequence-structure alignments are needed to move the comparative modeling field forward. However, the initial step is to understand why alignment errors occurred, do the misaligned regions have some distinct features, and whether these errors could have been avoided.

Alignment Error-Prone Regions: What Is Special About Them?

The results of sequence-independent superposition indicate that my models for several prediction targets have occasional alignment errors in structurally conserved regions (Fig. 1). Most of these alignment errors occurred in cases when target and the structurally closest template share <20% sequence identity. The highest sequence identity at which errors are still present is 33% (T0151).

Are there similar traits with regard to the regions where alignment errors occurred? The analysis of errors in β-strands provides the clearest picture. Several alignment errors were caused either by the appearance (T0132) or the disappearance (T0152) of a β-bulge within a β-strand. Apparently, it is very difficult to detect these very localized structural changes in β-strands unless homologous protein structures provide a structural hint. For example, in T0132 there are two β-bulges in the second β-strand and the adjacent region (V59-F69). Although at least one of the templates had an identical pattern, only a very small

fraction of predictor groups (<10%) were able to correctly identify β-bulges and produce the correct alignment (see summary tables at the CASP5 Web site; http://prediction-center.llnl.gov/casp5/). In the other two cases (the N-terminal strand of T0132 and the 4th strand in T0152) structural templates did not provide any hint about the local change. It is not surprising that not a single group detected these changes, resulting in parts of the affected strands being misaligned.

A prominent feature of the alignment error-prone β-strands is their location within the structure. All but one of the misaligned β-strands are located at the edges of β-sheets. Why are the edge β-strands more prone to being misaligned? These regions are highly exposed; therefore, they are under weaker structural/energy constraints, leading to the accelerated mutation rates. For example, the misaligned region (L100-D110) that includes the 5th β-strand in T0142 and the corresponding region in the closest template (PDB code: 1I9Z) share no identical residues, whereas the overall sequence identity reaches 27%. The same situation is observed in the case of the misaligned C-terminal β-strand (104D-110L) of T0151. There are no identical residues in the corresponding regions of T0151 and the related *E. coli* SSB protein (1EYG), whereas overall these two proteins share ≈33% identical residues. In addition to a lack of sequence similarity, matters were complicated because of the insertions/deletions that had to be introduced in these regions to produce the correct alignment. Although the PSI-BLAST-ISS procedure in both cases did indicate that the alignment is not reliable, structure-based assessment of different alignment variants proved to be ineffective. The reason for this failure was simply the lack of any identifiable structural constraints that could be used to distinguish the correct alignment from an incorrect one. The collective CASP5 modeling results indicate that the alignment for these two particular regions was universally difficult. Thus, <3% of the predictor groups correctly aligned the C-terminal strand of T0151 and none in the case of 5th β-strand of T0142.

The causes for alignment errors within α-helices are less obvious. The analysis is also complicated by frequently observed shifts of helices, introducing ambiguity into the assignment of corresponding residues. Ambiguities aside, there are examples of clearly misaligned α-helices in my models for targets T0130, T0133, and T0142. Targets T0130 and T0142 deserve a special consideration and are discussed in separate sections below. In T0133, the alignment error occurred in the first helix (V160-S192) of the second, triple coiled coil, domain. In this case, the correct alignment was not even present among the initial set of alignment variants produced by PSI-BLAST-ISS. Retrospectively, it appears that secondary structure prediction would have been a much better guide for selecting the correct alignment in this particular case. The predicted boundaries for this helix by PsiPred[11] coincide exactly with the boundaries determined by DSSP[12] in the T0133 experimental structure. This error was clearly preventable because close to 15% of groups were able to identify the correct alignment.

Post-CASP5 analysis of the misaligned regions in my models showed that all of them except two were initially classified as unreliable based on the results of PSI-BLAST-ISS. The two exceptions include either gain (T0132) or loss (T0152) of a β-bulge, and they proved to be impossible to predict by any method in CASP5. Thus, the absolute majority of the alignment errors came as no surprise. This is a positive finding, because for any interpretation involving a 3D model, it is important to know which parts of the model can be trusted. At the same time, I realize that although the knowledge of where errors might occur is valuable, the ability to avoid errors is much more desirable. The analysis indicates that for some (e.g., α-helix in T0133), more emphasis on secondary structure prediction might have been fruitful. For others, (the edge β-strand of T0142), it is obvious that more sensitive energy-based methods are needed to identify the correct mapping of the residues.

Below, I present three specific examples that illustrate problems and successes in modeling the CASP5 targets.

Example of the Difficulty (and the Value) of Multiple-Template Combination (T0130, protein HI0073, *H. influenzae;* PDB code 1NO5)

T0130, a small α/β protein (114 amino acids), is a member of a protein family thus far found only in archaea and bacteria.[13] Although the cellular function of this protein family is not yet known, on the basis of sequence analysis it was suggested that the proteins in this family represent the minimal domain of the polβ nucleotidyltransferase superfamily. Therefore, the proteins in this family including T0130 were designated as "minimal" nucleotidyltransferases, or MNT.[13] Indeed, the structure of T0130 has a characteristic glycine-rich loop as well as the three aspartates corresponding to the active site in other nucleotidyltransferases of the polβ superfamily (Fig. 2). The closest structural template, yeast poly(A) polymerase (1FA0), is approximately 17% identical in amino acid sequence. Although small by itself, T0130 has only three strands and a single helix that are also universally present in other members of the polβ superfamily (Fig. 2). Outside this conserved core, there is a varying repertoire of structural motifs and their respective arrangements. The main challenge in this case was to identify the additional building blocks needed to assemble the complete 3D structure of the target. It is most interesting that none of the template structures had a complete set of these building blocks. One of the best choices for representing the second α-helix could be found in ERA GTPase (1EGA). It is interesting that in the SCOP classification, ERA belongs to a different fold, but the chain topology is clearly related to that of other T0130 templates. The corresponding helix is also present in the closest template, poly(A) polymerase, but it is preceded by β-hairpin, which is absent in T0130. In addition, the orientation of the helix is somewhat different. The C-terminal helix of T0130 does not have a very good structural match in any of the

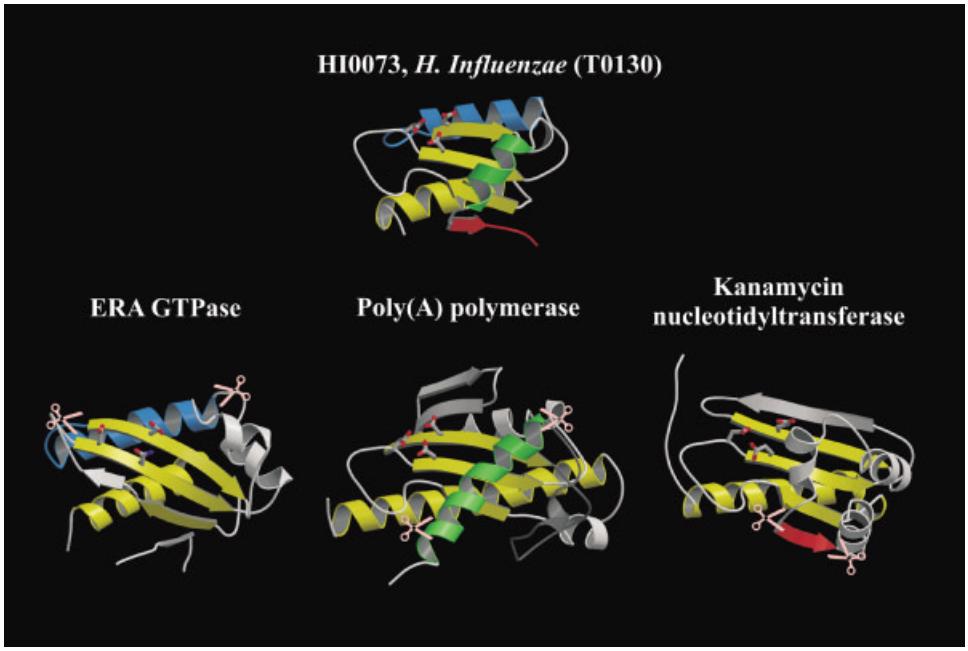

Fig. 2. "Protein lego" construction of the T0130 structure. Yellow denotes the conserved structural core. Three different proteins provide some of the better matches for modeling the remaining T0130 structural motifs colored in blue, green, and red. Approximate boundaries for these motifs within the template structures are indicated with scissors. No single template that had all three motifs together was available.

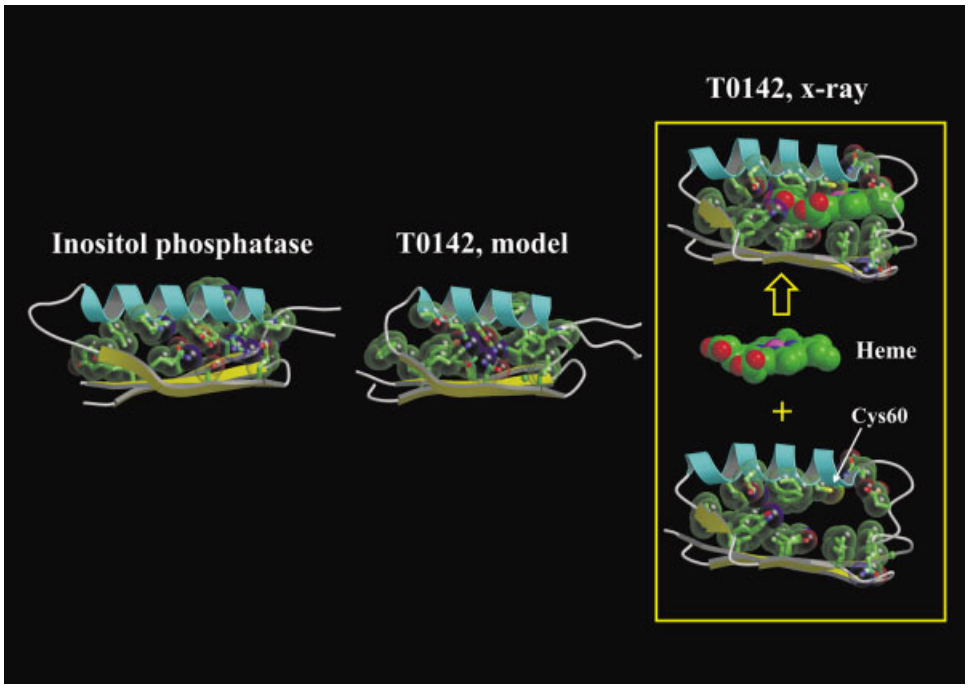

Fig. 3. Unexpected gain of novel function complicated modeling of heme-binding region in T0142. The residue side-chains contributing to heme-binding in T0142 and corresponding side-chains in both the model and the template are represented as sticks with semitransparent spheres indicating their physical (Van der Waals) volume. The heme is shown in a solid space-filling representation. Bonds and surfaces are colored according to the atom type: carbon: green; oxygen: red; nitrogen: blue; sulfur: yellow, and heme iron: magenta. Cys60 provides the proximal ligand for the heme iron.

templates. The orientation of the corresponding helix perhaps is the closest in the best template, poly(A) polymerase, but the preceding long loop needs to be excised. The counterpart of T0130 C-terminal β-strand is present only in kanamycin nucleotidyltransferase (1KNY), yet another structural template. Thus, an effective template-based

modeling of T0130 requires not only the use of divergent structural templates but also a robust selection of individual structural elements from these templates. In this respect, the assembly of the T0130 structure can be regarded as "protein lego" construction. In the course of modeling, identification of the structural core was straightforward, but the assembly of variable building blocks deteriorated as it progressed toward the C-terminus. Thus, both helices outside the conserved core were modeled, but the alignment is correct only for the first helix (colored blue in Fig. 2). The C-terminal β-strand (red) is missing entirely, because I failed to recognize that the long loop preceding the last α-helix is absent in T0130. As a result, a purely structural match of the model is similar to that of the best template, but the correctly aligned regions are significantly smaller (see Table I).

The T0130 modeling clearly shows the usefulness of multiple templates. It also illustrates the challenges for automatic template-based modeling methods. It is not surprising that the best model for this target generated by a human group (020) outperforms the best automatic server (029) prediction by 10.5 in the GDT_TS score (GDT_TS is an average of GDT values[14] at 1, 2, 4, and 8 Å distance thresholds). T0172 is the only other comparative modeling target, for which human groups outperformed servers by a larger margin. However, in that case the human superiority was based not on the template combination but on the recognition that the conserved structure of a methyltransferase-like domain is interrupted by the insertion of another domain.

Biologically Important Regions and Protein Modeling (T0142, salivary nitrophorin from the bed bug *C. lectularius;* PDB code: 1NTF)

In most cases protein models are built to gain insight about their biologically important regions and, thus, to advance the understanding of their function. However, there is also another side to the relation of functional regions and modeling. Perhaps T0142 is the best example of the knowledge of the functional regions and the protein functional state being critical for the ability to produce a correct model. T0142, a heme protein, named nitrophorin, is used by the blood-sucking bed bug to store and transport nitric oxide (NO) from the salivary glands to the skin of the host.[15] Once saliva is injected, because of a pH change, NO is released from the nitrophorin and serves as a main vasodilator.

There would not be anything unusual about this protein if not for the fact that none of the related structures are heme-binding proteins. T0142 itself and related structures all have a DNAseI-like fold including inositol phosphatase, *E. coli* exonuclease III, human apurinic/apyrimidinic endonuclease, and DNAseI. All these related proteins, as their names indicate, have entirely different functions. T0142 and the closest available structure—inositol phosphatase[16]—are quite similar in the sequence (27% sequence identity) and the structure [1.95 Å root-mean-square deviation (RMSD) for 248 corresponding Cα atoms], yet there is this sharp contrast between their respective biological functions.

It was exactly the failure to identify the region responsible for the functional change that led to an error in the alignment of the T0142 helix involved in heme binding. Although CASP5 predictors were informed that this prediction target is a heme-binding protein, neither the location of heme iron-coordinating residue nor its type (His or Cys) were known. As part of the modeling process, PSI-BLAST-ISS indicated that the T0142 α-helix, which turned out to be the main heme-binding structural motif after the experimental structure was revealed, was one of the unreliable alignment regions. Therefore, a number of variants, including the one corresponding to the experimental structure, were assessed further to identify the one that is the most compatible with the 3D structure. The final choice produced a reasonably good structure in this region of the model evidenced by the residue packing, pairwise interactions, and the nature of the exposed surface. However, post-CASP5 analysis showed that the alignment variant chosen for the submitted model is shifted by a helical turn in comparison to the one observed in the crystal structure. In retrospect, it is obvious that to choose the correct alignment variant, instead of an optimal one, the "worst" residue packing had to be selected. This point is illustrated in Figure 3. If the heme is removed, there is a big cavity in the T0142 crystal structure, much of it due to the type of residue side-chains in the helix. Once heme is bound, the structural integrity is completely restored. On the other hand, it is obvious that in the unbound state, T0142 structure should undergo significant rearrangement at least in the vicinity of the heme-binding site. It would be not surprising to observe, for example, a shift of the heme-binding helix in such a way that the currently "wrong" alignment in the submitted model would become the "correct" one. Unfortunately, the structure of T0142 in the unbound state currently is unavailable, precluding the possibility to compare the two functional states.

This example suggests that in some cases it may be impossible to correctly generate models without knowing the details about the exact biological function of the protein and its functional state. It is especially important to keep this in mind while modeling proteins for which only sequence information is available, such as proteins predicted directly from genome sequences. Evolution sometimes does play tricks on proteins, and protein structure prediction methods are not immune to these tricks either.

Distant Homology Modeling: Approaching the Limits Set by the Templates (T0152, protein Rv1347c, *M. tuberculosis;* T0169, protein yqjY, *B. subtilis;* PDB code 1MK4)

These two prediction targets are analyzed together because they both have a number of things in common. Both proteins belong to the same superfamily of GCN5-related N-acetyltransferases (GNAT), enzymes that use acetyl coenzyme A to acetylate amino groups of a wide variety of substrates (reviewed in Ref. 17). These two

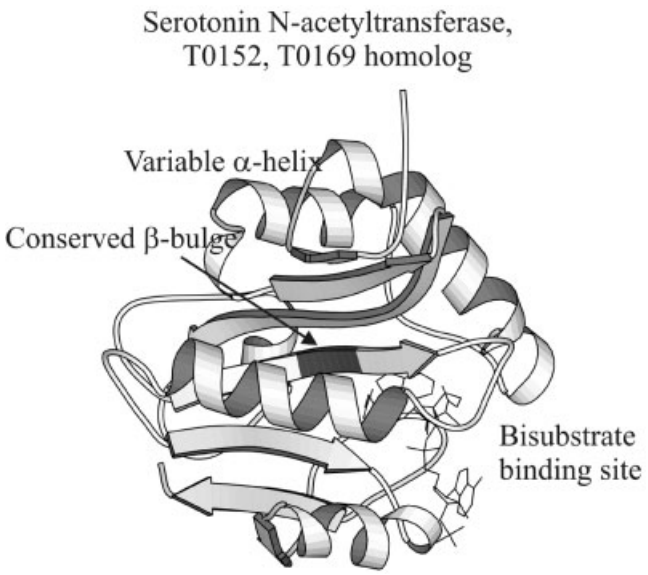

Fig. 4. The structure of serotonin N-acetyltransferase (1CJW), one of the best templates for modeling both T0152 and T0169. The conserved β-bulge, absent in T0152, is indicated by the darker shading.

proteins also happen to have the same structure, a sheep serotonin N-acetyltransferase[18] shown in Figure 4, among the most similar templates. In addition, at least by the sequence identity to serotonin N-acetyltransferase (14% for T0152 and 17% for T0169), they both are in a similar category of prediction difficulty. Overall, there is a large number of known protein structures belonging to this superfamily. Therefore, one of the important problems in modeling these prediction targets was the optimal selection of structural templates. This is especially challenging in distant homology, because sequence identity below ~20% becomes an increasingly poor indicator of structural similarity (e.g., see Fig. 1 in Ref. 19). Therefore, for modeling both T0152 and T0169, I selected multiple templates, six and five, respectively, that would adequately represent the conformational richness of the GNAT superfamily. In neither case, the template presumed to be the structurally closest one turned out to be the closest one. However, the best structural matches determined by structural superposition in the CASP5 aftermath for both T0152 and T0169 were among the templates that I selected for modeling these targets.

The extent of structural conservation in each of the two targets compared to serotonin acetyltransferase is similar but not the same. T0152 is somewhat more divergent than T0169. It has a large number of nonconserved loops and shifted secondary structure elements. A prominent structural difference of T0152 is the absence of a β-bulge, characteristic of other members of GNAT superfamily (Fig. 4), leading to a different curvature of the β-sheet. The disappearance of the β-bulge in T0152 was not detected; as a result, the β-strand after the bulge was misaligned. However, as discussed above, such local structural changes are difficult to detect without a hint from the structures of related proteins. More extensive structural variations, such as the shift of the second α-helix (indicated in Fig. 4), were easy to detect, because the corresponding helix varies greatly in the length and orientation in the GNAT proteins. By using multiple templates, the position of this helix was modeled to match quite closely that of the target structure.

The resulting models for both T0152 and T0169 were the best in CASP5 according to the GDT_TS score: 50 and 70.7, respectively. How do these models compare to the closest structural templates? Sequence-independent superposition between the experimental target structures and corresponding models in both cases produces more equivalent residue pairs than the target-template superposition (Table I). In that sense, the use of multiple templates resulted in an improvement over the single closest-matching template. Moreover, in T0152, the number of correctly aligned residues approaches the extent of structural conservation between the target and the best template. Although such a result may be routinely obtained when the template sequence is 40% or more identical to the target, at sequence identity below 20% this level of accuracy is not a trivial result.

CONCLUSIONS

The analysis of my CASP5 results indicates that there is a visible improvement over CASP4 in model quality, both in producing a greater number of models that structurally outperform the corresponding best templates and in the overall alignment quality. Although in this article the improvement in comparative modeling is tied to a single predictor group, it also suggests overall progress in the field. The argument for this is that both in CASP5 and CASP4, the results I achieved were considered by the independent assessment to be among the best.[20,21] The overall improvement in comparative modeling results perhaps is more difficult to see when they are pooled together with the results of significantly lesser quality such as those for predictions of analogous or new folds.[22]

Although the largest improvement is seen in distant homology modeling, this is also the most problematic area of comparative modeling. One of the important issues is the effective use of multiple templates, enabling the extension of models beyond the consensus structure of related proteins (T0130 being a good example). Occasional, even though often times anticipated, alignment errors present an even larger hindrance. Analysis of these alignment errors suggests that some of them could be prevented by applying existing techniques. However, there are cases where all methods tested at CASP5 failed (e.g., β-strand in T0142). Peripheral location and the absence of detectable sequence conservation patterns seem to be a common denominator of these regions. Therefore, more sensitive energy-based methods rather than those relying on pattern conservation are needed for the successful modeling in such regions.

Despite a number of remaining problems, comparative modeling is continuously moving deeper into the realm of distant protein evolutionary relationships. Although methods are improving, arguably the most significant factor

contributing to this is the explosive increase in the number of available protein sequences. On one hand, larger sequence databases allow easier detection of very distant homologues; on the other hand, the enrichment of protein superfamilies leads to more reliable alignments. Thus, the applicability even of exclusively template-based comparative modeling is poised to spread even further, perhaps close to the borderline of analogous fold relationships.

ACKNOWLEDGMENTS

I thank the structural biologists who provided their proteins as prediction targets for CASP5 and Dr. Dorota Sawicka for critically reading the manuscript. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

REFERENCES

1. Venclovas Č, Ginalski K, Fidelis K. Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins. Proteins 1999;Suppl 3:73–80.
2. Venclovas Č. Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. Proteins 2001;Suppl 5:47–54.
3. 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.
4. Walker DR, Koonin EV. SEALS: a system for easy analysis of lots of sequences. Proc Int Conf Intell Syst Mol Biol 1997;5:333–339.
5. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
6. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
7. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
8. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
9. 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.
10. Zemla A. LGA program—a method for finding 3-D similarities in protein structures. 2000. Accessed at http://PredictionCenter.llnl.gov/local/lga/lga.html.
11. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
12. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
13. Aravind L, Koonin EV. DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. Nucleic Acids Res 1999;27:1609–1618.
14. Zemla A, Venclovas Č, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
15. Valenzuela JG, Ribeiro JM. Purification and cloning of the salivary nitrophorin from the hemipteran Cimex lectularius. J Exp Biol 1998;201:2659–2664.
16. Tsujishita Y, Guo S, Stolz LE, York JD, Hurley JH. Specificity determinants in phosphoinositide dephosphorylation: crystal structure of an archetypal inositol polyphosphate 5-phosphatase. Cell 2001;105:379–389.
17. Dyda F, Klein DC, Hickman AB. GCN5-related N-acetyltransferases: a structural overview. Annu Rev Biophys Biomol Struct 2000;29:81–103.
18. Hickman AB, Namboodiri MA, Klein DC, Dyda F. The structural basis of ordered substrate binding by serotonin N-acetyltransferase: enzyme complex at 1.8 Å resolution with a bisubstrate analog. Cell 1999;97:361–369.
19. Venclovas Č, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;Suppl 5:163–170.
20. Tramontano A, Morea V. Assessment of homology based predictions in CASP5. Proteins 2003;Suppl 6:352–368.
21. Tramontano A, Morea V, Leplae R. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;Suppl 5:22–38.
22. Venclovas Č, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins 2003;Suppl 6:585–595.
23. Kraulis PJ. Molscript—a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24: 946–950.
24. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.