Progress Over the First Decade of CASP Experiments

Andriy Kryshtafovych,[1] Česlovas Venclovas,[2] Krzysztof Fidelis,[1] and John Moult[3]*
[1]*Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory Livermore, California*
[2]*Institute of Biotechnology, Vilnius, Lithuania*
[3]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*

ABSTRACT **CASP has now completed a decade of monitoring the state of the art in protein structure prediction. The quality of structure models produced in the latest experiment, CASP6, has been compared with that in earlier CASPs. Significant although modest progress has again been made in the fold recognition regime, and cumulatively, progress in this area is impressive. Models of previously unknown folds again appear to have modestly improved, and several mixed α/β structures have been modeled in a topologically correct manner. Progress remains hard to detect in high sequence identity comparative modeling, but server performance in this area has moved forward. Proteins 2005;Suppl 7:225–236.** © **2005 Wiley-Liss, Inc.**

Key words: protein structure prediction, community wide experiment, CASP

INTRODUCTION

The CASP experiments have now spanned a period of a decade, from CASP1 in 1994, to the latest experiment, CASP6, in 2004. The set of results reflect a decade of development in protein structure modeling by a large number of people—over 200 groups from 25 countries participated in the latest CASP. In this paper, we once more provide an overview of progress over the full course of the experiments, with particular emphasis on the last two. The papers by the three assessors in this special issue of PROTEINS focus more on the state of the art now.[1,2,3] The analysis methods have mostly been introduced in the earlier papers.[4,5] New analyses of server performance and improvement of models over single template copying have been added. Details of the CASP6 experiment can be found in the introduction to this special issue.[6]

GENERAL CONSIDERATIONS

Choice of Models to Evaluate

As before, we analyze two aspects of progress—how the quality of the very best models is improving, and to a lesser extent, how the quality of models produced in the field as a whole is advancing. Best performance is evaluated by comparing the most accurate models of targets of comparable difficulty in different CASPs. Progress in the field as a whole is evaluated by comparing the average accuracy of the six best models for a target with the average accuracy of models in other CASPs for targets of similar difficulty.

Relative Target Difficulty

The difficulty of producing a high quality model of a target protein depends on a number of factors. As in the earlier progress assessments, we use a two-dimensional scale to estimate difficulty, incorporating the similarity of the protein sequence to that of a protein with known structure, and the similarity of the structure of the target protein to potential templates. Some other significant factors that affect modeling difficulty are not considered, particularly the number and phylogenetic distribution of related sequences and the number and structural distribution of available templates. The set of related sequences will influence whether or not an evolutionary relationship can be detected, and also the quality of the alignment that can be generated. As discussed later, additional templates may provide models for regions of structure not present in the single best one. These factors add some noise to the relationship between model quality and our difficulty scale.

The difficulty of a target is calculated by comparing it with every structure in the appropriate release of the protein databank, using the LGA structure superposition program.[7] For CASP6, templates were taken from the PDB releases accessible before each target deadline. Templates for the previous CASPs are the same as those used in the earlier analyses.[5] For each target, the most similar structure, as determined by LGA, in the appropriate version of the PDB is chosen as the representative template.

Similarity between a target structure and a potential template is measured as the number of target–template Cα atom pairs that are within 5 Å in the LGA superposition, irrespective of continuity in the sequence, or sequence relatedness. The 5-Å threshold maintains compatibility with earlier target/template comparisons,[4,8,9] which were made using Prosup[10] software. It is a little larger than we now consider most appropriate (3.8 Å), and there

Grant sponsor: the National Institutes of Health; Grant number: LM07085-01; Grant sponsor: Howard Hughes Medical Institute; Grant number: HHMI-55000341; Grant sponsor: the Sixth European Community Framework Programme; Grant number: MIRG-CT-2004-004543.

*Correspondence to: John Moult, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850. E-mail: jmoult@tunc.org

Received 15 May 2005; Accepted 21 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20740

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

is some times significant superimposability between unrelated structures, particularly for small proteins. Sequence identity is defined as the fraction of structurally aligned residues that are identical, maintaining sequence order. Note that basing sequence identity on structurally equivalent regions will usually yield a higher value than obtained by sequence comparison alone. In cases where several templates display comparable structural similarity to the target (coverage differed by less than 3%), but one has clearly higher sequence similarity (around 10% or more) the template with the highest sequence identity was selected. There are a total of 15 of these in previous CASPs, and six in CASP6.

Domains

Many target structures consist of two or more structural domains. Since domains within the same structure may present modeling problems of different difficulty, assessment in CASPs 4–6 has treated each identifiable domain as a separate target. Assessors have the advantage of deriving domains from the experimental structure, whereas predictors only have the sequence. In any case, domain definitions are nearly always subjective. For most of the analysis, we subdivide comparative modeling and fold recognition targets into domains only if these divisions are likely identifiable by a predictor, and require different modeling approaches (i.e., belong to different difficulty categories), or the domains are sequentially related to different templates. There are six such targets in CASP6, seven in CASP5, three in CASP4, and one each in CASPs 2 and 3. For evaluation of nontemplate based models (the FR/A and NF target categories) and some server comparisons, all domains identified by the assessors are treated as separate targets.

TARGET DIFFICULTY ANALYSIS

Figure 1 shows the distribution of target difficulty for all CASPs, as a function of structure and sequence similarity between the experimental structure of each target and the corresponding best available template. Targets span a wide range of structure and sequence similarity in all the CASPs. There are a few very high sequence identity targets (greater than ~ 50% ID), and these all have high superposability with the best template. At lower identities, superposability varies between 80 and 100% for targets with greater than 30% sequence ID to a template, and is some times as low as 55% for those between 20 and 30% sequence ID. Below 20% ID, superposability may fall below 50%, even for targets, which are evolutionarily related to a template. As discussed later, low superposability often places a limit on the quality of a model. In general, the distribution of difficulty is similar for all the CASPs. Figure 1(B) shows the difficulty distribution for only CASPs 5 and 6, with individual CASP6 targets/domains labeled. The distributions are similar, with 71 targets included in CASP6 and 62 in CASP5.

For most analysis purposes, it is more convenient to use a one dimensional scale of target difficulty, though this does result in some loss of resolution. As in the previous analysis, we project the data in Figure 1 into one dimension, using the following relationship:

$$\text{Relative Difficulty} = (\text{RANK\_STR\_ALN} + \text{RANK\_SEQ\_ID})/2,$$

where RANK_STR_ALN is the rank of the target along the horizontal axis of Figure 1 (i.e., ranking by percent of the template structure aligned to the target), and RANK_SEQ_ID is the rank along the vertical axis (ranking by percent sequence identity in the structurally aligned regions). For the CASP6 analysis, we experimented with 15 alternative definitions of difficulty based on both ranking and absolute value schemes. Some placed more weight on sequence identity in high sequence ID cases and more weight on structural similarity for low sequence ID cases. Alternative difficulty scales were assessed by the correlation between difficulty and the quality of the corresponding best model. In spite of considerable effort put into development of alternative scales, the original difficulty scheme proved best, and so was retained.

For assessment, CASP targets are divided into three categories of relative difficulty: comparative modeling (CM), fold recognition (FR), and new folds (NF).[11] Comparative modeling is subdivided into CM easy (those targets where a structural template can be identified by a BLAST search) and CM hard (the rest). Fold recognition is divided into FR/H: "homologous" (those cases where target and template are similar because of a common ancestor) and FR/A: "analogous" (where target and template are similar, but for which there is no evidence of a common ancestor). These regimes approximately map to the one-dimensional difficulty scale, with comparative modeling the easiest, fold recognition in the intermediate difficulty range, and new fold targets the hardest. However, there is some reordering.

OVER-ALL MODEL QUALITY

Evaluating the quality of approximate models is not simple, and a number of new measures have been introduced in CASP. One of the most useful is GDT_TS.[12] The GDT_TS value of a model is determined as follows. A large sample of possible structure superpositions of the model on the corresponding experimental structure is generated by superposing all sets of three, five, and seven consecutive Cα atoms along the backbone (each peptide segment provides one superposition). Each of these initial superpositions is iteratively extended, including all residue pairs under a specified threshold in the next iteration, and continuing until there is no change in included residues.[7] The procedure is carried out using thresholds of 1, 2, 4, and 8 Å, and the final superposition that includes the maximum number of residues is selected for each threshold. Superimposed residues are not required to be continuous in the sequence, nor is there necessarily any relationship between the sets of residues superimposed at different thresholds. GDT_TS is then obtained by averaging over the four superposition scores for the different thresholds:

$$\text{GDT\_TS} = 1/4[\text{N1} + \text{N2} + \text{N4} + \text{N8}],$$

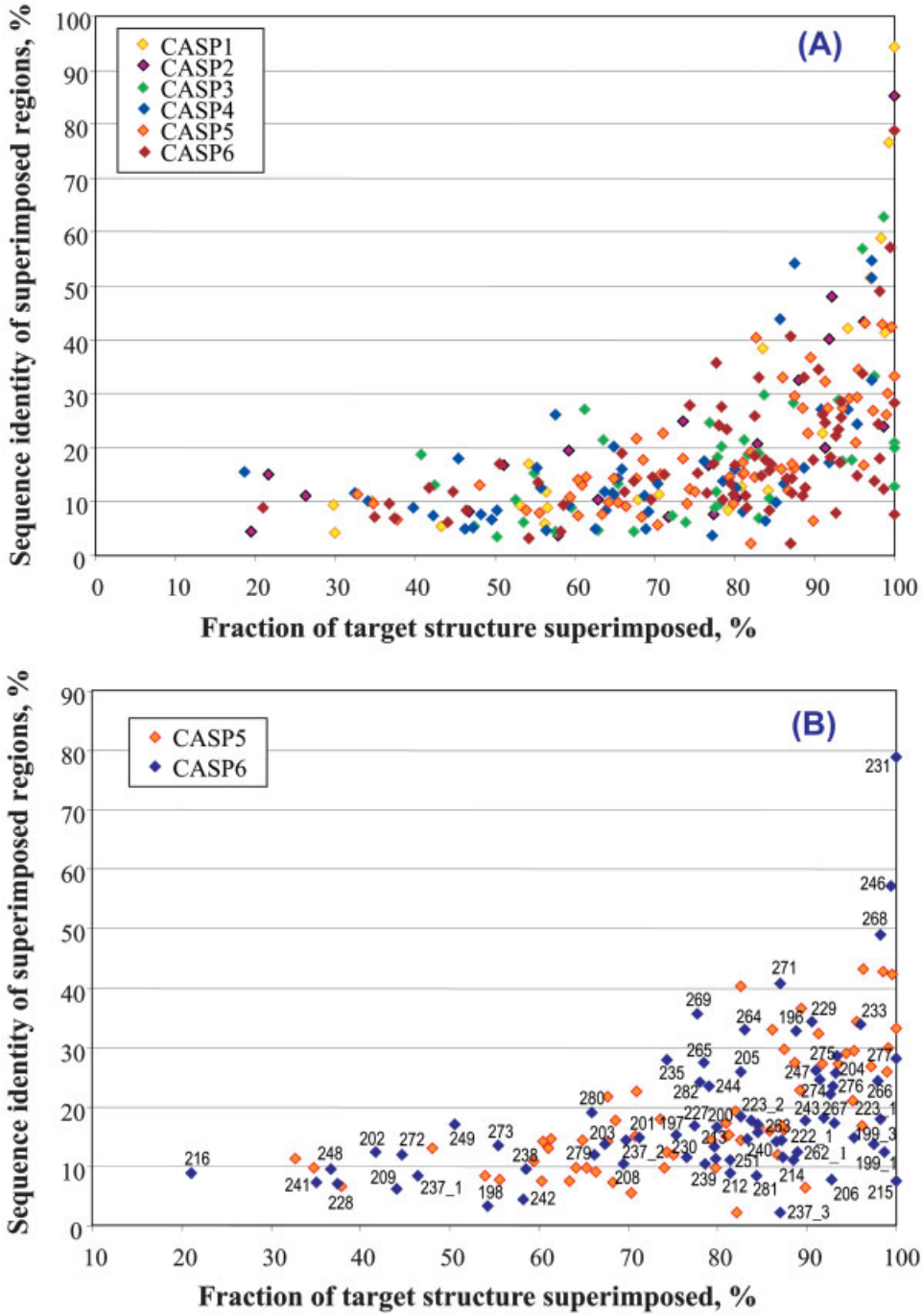

Fig. 1. Distribution of target difficulty. The difficulty of producing an accurate model is shown as function of the fraction of each target that can be superimposed on a known structure (horizontal axis) and the sequence identity between target and template for the superimposed portion (vertical axis). In all CASPs, targets span a wide range of difficulty. **A**: All CASPs. **B**: CASPs 5 and 6 only. CASP 6 targets are labeled.

where Nn is the number of residues superimposed under a distance threshold of "n" Å. GDT_TS may be thought of as an approximation of the area under the curve of accuracy versus the fraction of the structure included. Different thresholds play different roles in different modeling regimes. For relatively accurate comparative models, almost all residues will likely fall under the 8-Å cutoff, and many will be under 4 Å, so that the 1–2 Å thresholds capture most of the variations in model quality. In the new fold regime, on the other hand, few residues fall under the 1–2 Å thresholds, and the larger thresholds capture most of the variation between models. In the intermediate fold recognition regime, all four thresholds will often play a significant role. It is this shift across thresholds that makes the

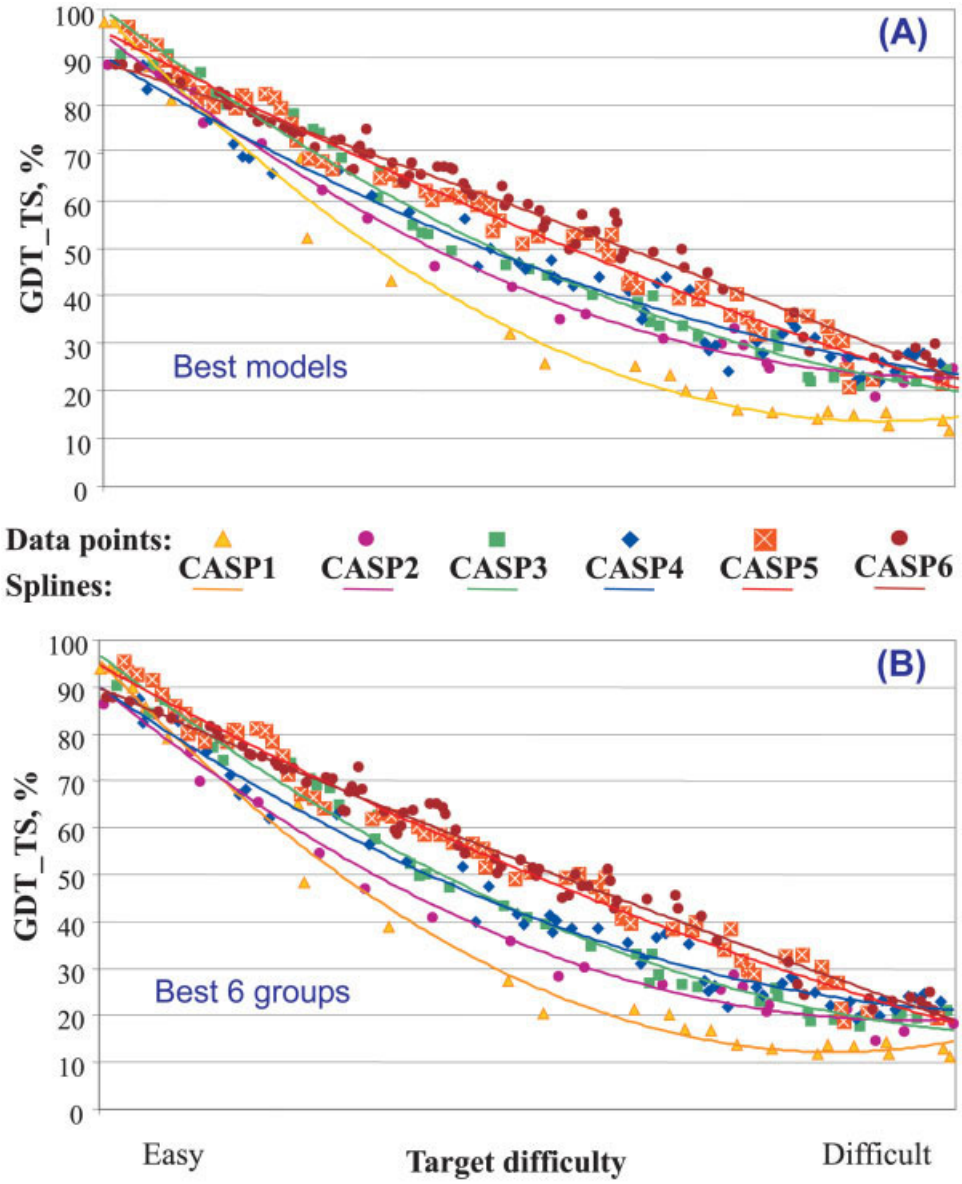

Fig. 2. GDT_TS scores for models for targets in all CASPs. Data are smoothed by averaging over sets of consecutive targets in each CASP. **A**: shows the scores for the best models on each target. **B**: the average score over the top six models from different groups. Trend lines show a clear—though some times modest improvement—from each successive CASP to the next, for both the best models and the best sets of models.

GDT_TS measure useful (though not perfect) across a wide range of modeling accuracy.

In the new fold regime models are often very approximate. In recent CASPs the assessors have found GDT_TS a useful measure for identifying interesting models, but have occasionally visually identified an alternative highest quality model to that found by GDT_TS. In the comparative modeling regime, accuracy improvements are likely to be relatively small-scale. The CASP6 comparative modeling assessors have introduced a finer grained measure, GDT_TL, where the thresholds are 0.25, 0.5, 1, and 2 Å.

Figure 2(A) shows the GDT_TS scores for the best model on each target, for all CASPs, with each point an average over five targets. Quadratic splines have been fitted through the data for each CASP.

A perfect model would not be expected to have a GDT_TS score of 100, since there are errors in the experimental structures. For high sequence ID ($\gtrsim$ 30%) targets, only X-ray crystallographic structures have been used for evaluation of model quality. Errors in the core of X-ray structures are small, typically a few tenths of an Ångstrom,[13] but there may be systematic differences from the solution conformation, caused by the crystal environment. These effects are some times invoked to justify imperfect models. In spite of these factors, the easiest comparative modeling targets (at far left) do consistently score better than 90 on the GDT_TS scale. These have sequence identities of 50%

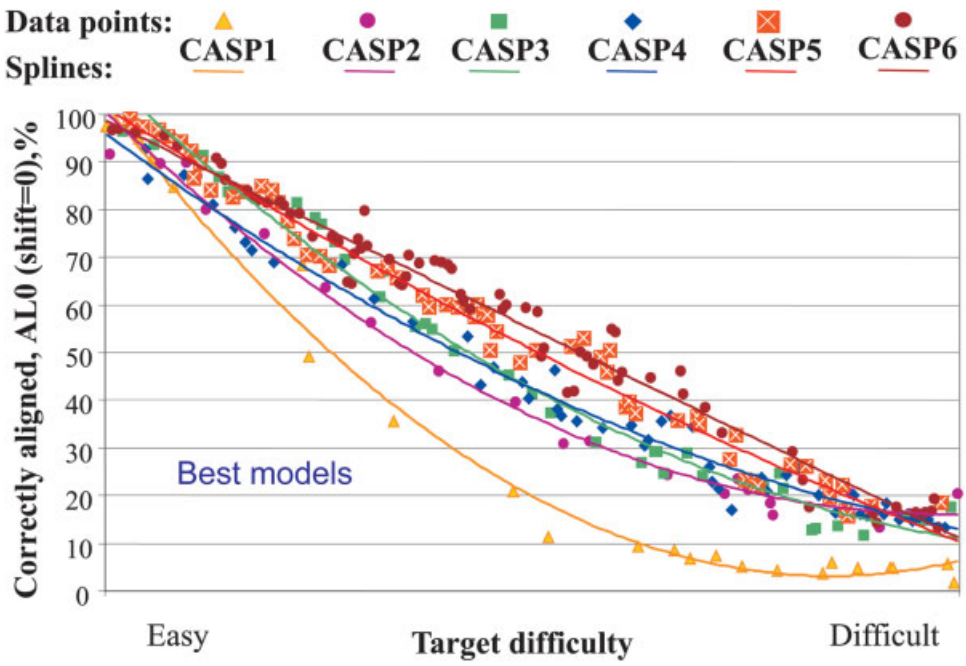

Fig. 3. Percent of residues correctly aligned for the best model of each target in all CASPs, smoothed by averaging over sets of five adjacent targets. Trend lines here follow those in the equivalent GDT_TS plot (Fig. 2) indicating that for many targets, alignment accuracy, together with the fraction of residues that can be aligned to a single template, dominate model quality.

or more to an available template, and usually a high degree of structure superposability (see Fig. 1). The task of modeling structure from sequence will be complete when almost all models score better than 90, irrespective of target difficulty. It is clear that there is still a long way to go to reach that goal.

Nevertheless, there has been very substantial and consistent progress over the CASPs, with the trend line for each experiment higher than the previous one. Progress is clearest in the mid-range of target difficulty. Here, the best GDT_TS scores have doubled from about 30 in CASP1 to 60 in CASP6. According to GDT_TS there is little apparent progress for the closest evolutionary relationships (comparative models at the left hand side of the plot). The more sensitive GDT_TL measure does show a slight improvement between CASP5 and CASP6.[1] GDT_TS is also not optimal for assessing new fold models (far right), and we return to analysis of these later.

Figure 2(B) shows the smoothed average GDT_TS values over the six best models from different groups, rather than the single best. There is a similar improvement trend across the mid-range of target difficulty, although here the change from CASP5 to CASP6 is more modest.

ALIGNMENT ACCURACY

For models based on an evolutionary relationship, correct alignment of the target sequence onto available template structures is a critical and often demanding step. As in previous analyses, we measure alignment accuracy (AL0) by counting the number of correctly aligned residues in the LGA 5 Å superposition of the modeled and experimental structures of a target. A model residue is considered to be correctly aligned if the Cα atom falls within 3.8 Å of the corresponding atom in the experimental structure, and there is no other experimental structure Cα atom nearer.

Figure 3 shows the smoothed alignment accuracy for the best models of each target in all the CASPs. Alignment accuracy is near 100% for the easiest targets, but falls steadily with target difficulty. The spline fits show a steady improvement in alignment accuracy over the CASPs, again most noticeable in the mid-range of target difficulty. It is most dramatic from CASP1 to CASP2, but has continued steadily thereafter. The plots show a similar dependence on target difficulty as for GTD_TS [Fig. 2(A)]. As discussed below, alignment accuracy is only one of two factors contributing to this similarity.

ALIGNMENT ACCURACY RELATIVE TO TEMPLATE IMPOSED LIMITS

The fraction of residues that can be aligned is limited by the fraction of superimposable residues between the target and template structures. Values above that may be obtained by the use of additional templates, where these contribute new information, and by free modeling of additional features, such as loops and secondary structure elements. We define the maximum alignability with respect to the best single template as follows: We first find all target Cα atoms that are within 3.8 Å of any template Cα atom in the 5-Å LGA sequence-independent superposition. Then, we use a dynamic programming procedure that determines the longest alignment between the two structures using these preselected atoms, in such a way that no

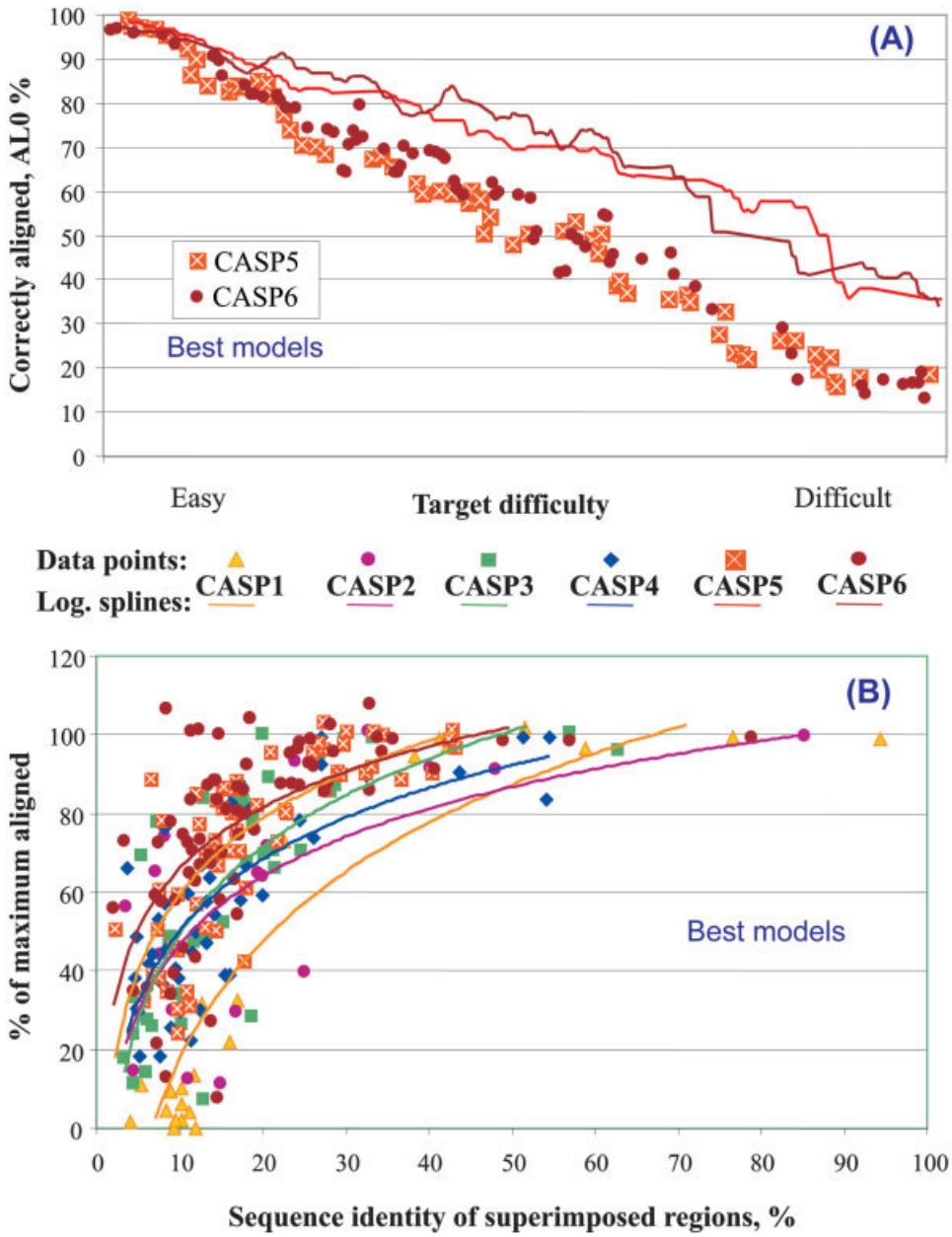

Fig. 4. **A**: Smoothed alignment accuracy and smoothed maximum alignability as a function of target difficulty. Targets for CASPs 5 and 6 are shown. Maximum alignability is defined as the fraction of equivalent residues in a superposition of the target and best template structures. The fraction of this theoretical maximum that is successfully aligned falls steadily with target difficulty. Residues that cannot be aligned to the best template must either be obtained from additional templates, if available, or modeled using template free methods. **B**: Alignment accuracy for the best model of each target in all CASPs, expressed as percent of the maximum number residues that can be aligned by copying from the closest available template structure. Targets are ordered by sequence identity between the target and the closest template. An alignment of 100% indicates that all residues with an equivalent in the template were correctly aligned. A value greater than 100% indicates an improvement in model quality beyond that obtained by copying a template structure. Above 30% sequence identity most, but not all, best models are perfectly aligned to the template. Trend lines show a steady though some times modest improvement over successive CASPs.

atom is taken twice and all the atoms in the alignment are in the order of the sequence. The maximum alignability is then the fraction of target Cα atoms in this alignment.

Figure 4(A) shows the smoothed alignment accuracy for the best models of all targets in CASPs 5 and 6 [a subset of the data in Fig. 3], together with the smoothed maximum alignability. Alignability falls steadily and approximately linearly with increasing target difficulty, but with a smaller slope than that of the fall-off in alignment accuracy. In the mid-range of difficulty, best model alignments are typi-

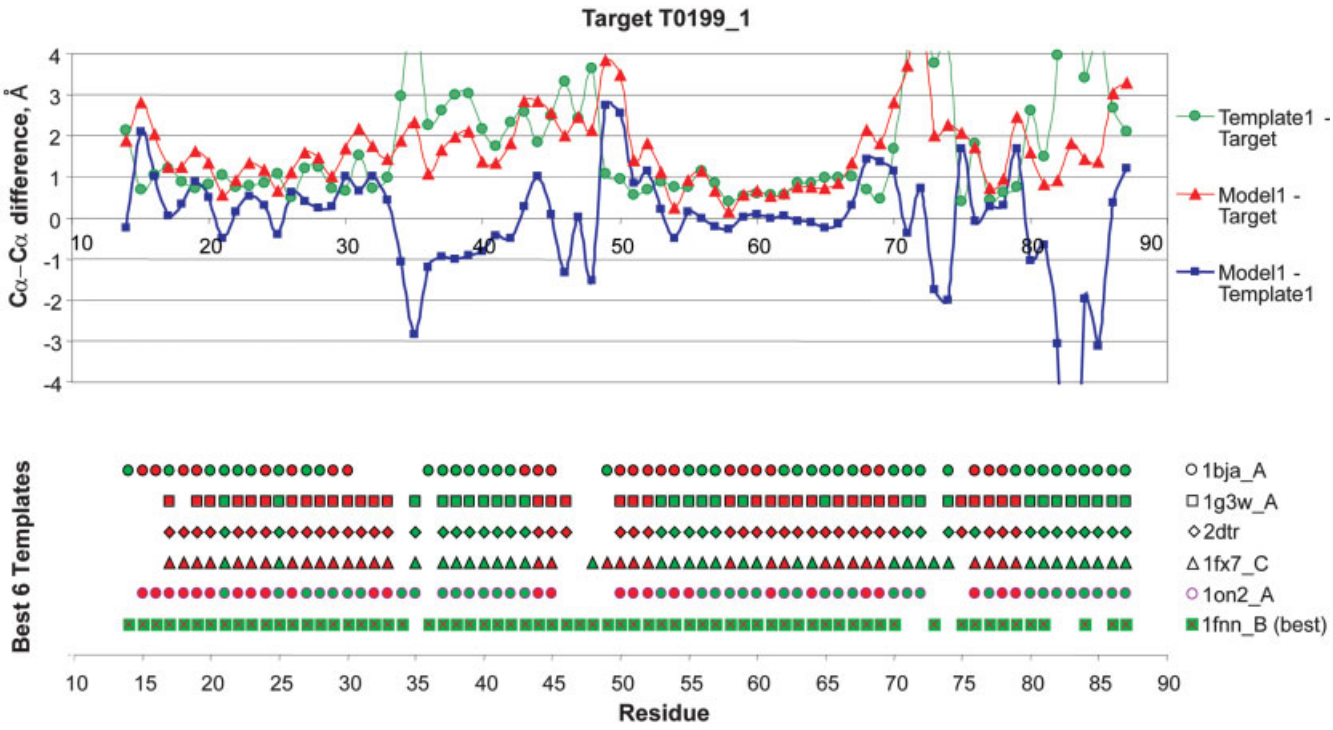

Fig. 5. An example of Improvements in a model over copying from a single best template, CASP6 comparative modeling target 199_1. The red curve shows the error in the model, ϵ^i(model), and the green curve, the "error" in the template, ϵ^i(template). The blue curve shows the difference, $\Delta\epsilon^i$. In most regions, the two error curves are similar, and the model has the accuracy obtained by simply copying the template ($\Delta\epsilon^i$ close to zero). Two regions, residues 34 to 50 and 80 to 86, are significantly more accurate than can be obtained from the best template. The lower panel shows where the six best templates (best one postitioned lowest) provide a model for which residues. It can be seen that extra information is provided by several templates in these regions.

cally within 20% of optimum, but up to 40% of the structure cannot be aligned at all.

Figure 4(B) shows the alignment accuracy for all CASPs, as a percent of the maximum alignability. Log fits to the data for each CASP are also shown. Targets are ordered by the sequence identity between the target and best available template. In all CASPs, the majority of targets with greater than 30% sequence identity to a template have all possible residues correctly aligned. The curves show steady progress over the CASPs, with the smallest increment between CASP5 and CASP6. Closer inspection shows that for the first time in CASP6 there are a few targets at low sequence identity that have a higher fraction of the residues aligned than is possible from a single template.

FEATURES OF MODELS NOT AVAILABLE FROM A SINGLE BEST TEMPLATE

As noted earlier, not all residues can be modeled by copying from a single evolutionary related structure. It is of interest to ask whether models contain features not available from a single template. Additional features may be added in three ways: by refining aligned regions away from a template structure towards the experimental one (requiring adjustments of up to about 4 Å), by the use of template-free modeling methods, and by the identification of features that are present in other available template structures. We searched for these "added value" features using an error difference function, $\Delta\epsilon$:

$$\Delta\epsilon^i = \epsilon^i(\text{model}) - \epsilon^i(\text{template}),$$

where ϵ^i(model) is the error in the model for the Cα atom of residue "i," and ϵ^i(template) is the distance between the template and target Cα atoms of residue "i" (the error in a correctly aligned template-based model at that position). Inter-Cα distances are taken from the LGA sequence-independent superposition of target and template. Negative $\Delta\epsilon$ values represent regions of a model that are more accurate than could be obtained by simply correctly aligning the single best template.

In general, there is no sign of model improvement by refinement, but occasional utilization of additional templates and limited free modeling is evident.

Figure 5 shows results for the CASP6 comparative modeling target 199_1. The red curve shows the error in the model, ϵ^i(model), and the green curve, the "error" in the template, ϵ^i(template). The blue curve shows the difference, $\Delta\epsilon^i$. In most regions, the two error curves are similar, and the model has the accuracy obtained by simply copying the template. Two regions, residues 34 to 50 and 80 to 86, are significantly more accurate than can be obtained from the best template. The lower panel shows where the six best templates provide a model for which residues. In this case, multiple templates are available for much of this winged helix structure, and in particular, provide additional information in the two improved regions. The best model likely incorporates information from

one or more additional templates. There are other examples, indicating that current methods are at least some times able to successfully combine multiple templates.

Some small features, particularly loops between secondary structure elements in comparative models, are built moderately accurately by template-free methods. An example is two loops in CASP6 target 266. As noted earlier, there are also a few examples of larger features modeled in this way, such as a helix in target 205.

SERVER PERFORMANCE

Human prediction teams often start with models generated by publicly available servers. Thus, overall progress in server performance has an immediate effect on the art of protein structure prediction in general. In addition, servers are the only option for high-throughput modeling, so their performance is independently important. We have examined server predictions in CASPs 5 and 6, comparing them to the corresponding human predictions. Approximately the same number of structure prediction servers participated in CASP5 and CASP6 (53 and 50 respectively) making the two sets of data comparable.

Figure 6(A) shows how many server models were among the "best six" for three categories of target difficulty. For "easy" CM targets (CMe, those where a template can be detected with BLAST), the percent of server models has risen by about a factor of two, to the same level as in the other categories ($\sim$ 20%). In contrast to this, CASP6 server performance in the other two categories has slipped slightly compared to humans, relative to CASP5. Figure 6(B) shows the number of targets with at least one server among the best six models and the number of server groups with at least one model in a top six set, for the easy comparative modeling targets. By these measures too, it is clear that relative server performance has improved for this class of targets. These analyses are made on all assessor-defined domains.

Figure 6(C) shows the ratio of the quality of best server models to best human models as a function of target difficulty, for CASPs 5 and 6, as measured by GDT_TS. By this measure, it is again clear that there has been an improvement in the relative performance of servers for comparative models, while in other areas, they are less competitive. "Predictor" domains were used for this figure.

Overall, servers provided the best models (or tied with humans) for 7% of targets in both CASP5 and CASP6.

TEMPLATE FREE METHODS

For "new fold" targets, and most targets judged to be similar to a known fold because of convergence rather than an evolutionary relationship ("FR/A" targets), template free modeling methods are required. Different factors affect the quality of model in this category. In early CASPs, structures with a high fraction of β structure were more difficult to model than those that were mostly α-helical, though this problem has now largely been solved. Other factors are contact order (how local the contacts in the experimental structure are, see Bonneau et al.[14]), domain structure, and size.

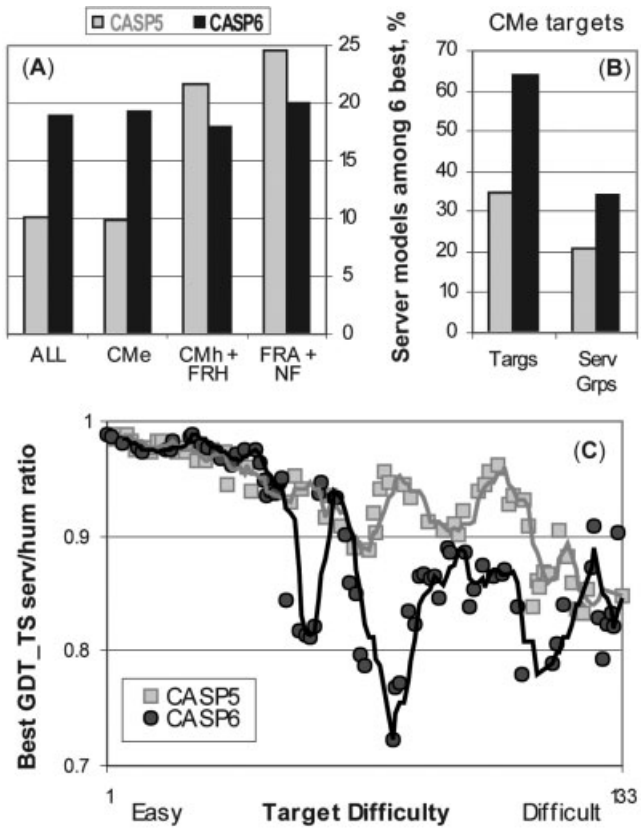

Fig. 6. **A**: Percent of server predictions among the "best six models" for three categories of target difficulty in CASPs 5 and 6. Relative server performance improved for easy comparative models. (CM: comparative models, e: easy, h: hard; FR/H: homologous fold recognition; FR/A: analogous fold recognition; NF: new folds). **B**: Percent of easy comparative modeling targets for which at least one server is among the top six best models, and percent of server groups having at least one model in a top six set. By these measures too it is clear that there has been an improvement in relative server performance for this class of target. **C**: Ratio of the quality of best server models to best human models as a function of target difficulty, for CASPs 5 and 6, as measured by GDT_TS, averaging over five adjacent points. By this measure, there is a small improvement in the relative performance of servers for comparative models, while in other areas, server models are less competitive.

Because of the lower accuracy of these models, different evaluation methods are required. In earlier analyses, we used the extent of sequence-dependent structure superposition to measure model quality, considering the four terms that contribute to the GDT_TS measure, rather than just that single value. That is, the number of residues which can be superimposed under 1, 2, 4 and 8 Å. Although most observers felt there was progress between CASP4 and CASP5, it was difficult to detect by this measure. We have used the same measure again, and added a second: the fraction of superposed residues between the model and the target, in a structure independent superposition. This measure still requires that superposed residues be in sequence order, and so will not capture approximate topology features that may be visually pleasing. For example, the β-sheet in the best GDT_TS model of CASP6 target 201 has all strands superposed onto target structure strands, but there is

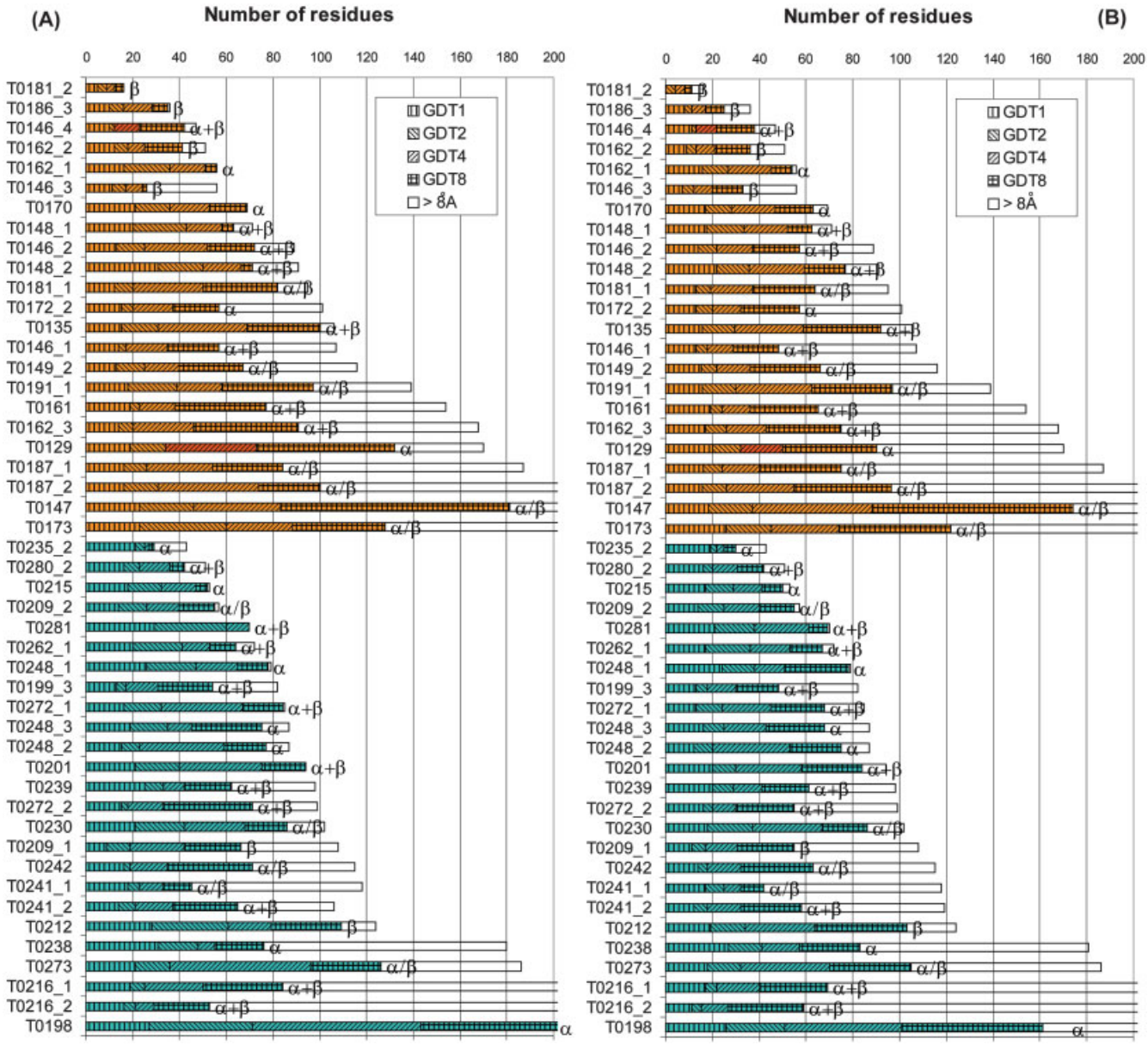

Fig. 7. Model quality for the best (**A**) and averaged over the six best (**B**) "new fold" category targets, for CASPs 5 and 6 for each target. The lowest bars show the number of residues superimposed between model and target to closer than 1 Å, the next bar, the number superimposed to 2 Å, then 4 Å, then 8 Å. The open bars show number of residues superimposed to greater than 8 Å. Greek letters indicate the fold type. Targets in each CASP are ordered by size. Bars for residues under 8Å are colored orange for CASP6 and green for CASP5.

an error in the strand order. We have also added best fit "trend lines" to the analysis, allowing smaller differences in performance to be seen. As in the earlier analyses, all domains identified by the assessors are treated as separate targets. Domains which are unambiguously new folds (NF targets) and domains in the analogous fold recognition category (FR/A) are included, providing a total of 23 targets in CASP5 and 25 in CASP6. The FR/A targets are considered since any relationship to a known fold is usually too weak for template-based modeling to be very effective.

Figure 7 shows the results of the GDT threshold analysis for CASPs 5 and 6. Targets are ordered by size. Fold type is indicated by the usual Greek letter classification. The stacked bars show the number of residues superimposed under the distance thresholds of 1, 2, 4, and 8 Å, i.e., the number of residues for which the largest error is less than or equal to each threshold. The RMS error on such a set is typically about half the threshold, thus substructures meeting the 8-Å threshold would usually be judged excellent by visual inspection.

In Figure 7(A), the performance in terms of the best models for each target is shown, and in Figure 7(B), the performance averaging over the six best models for each target. A convenient way of comparing model quality in each CASP is to examine the number of targets for which

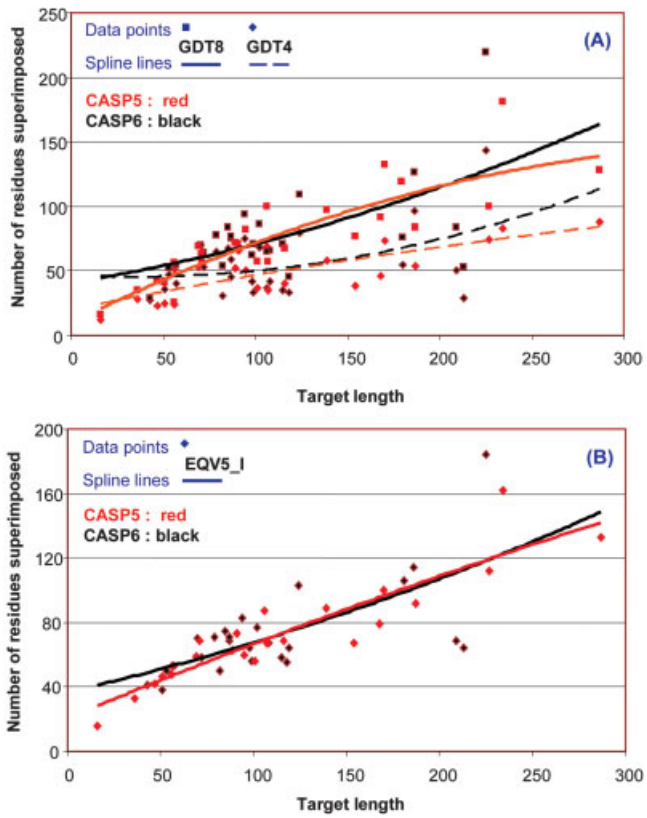

Fig. 8. **A**: Number of residues falling under the GDT_4 and GDT_8 thresholds for the best models of each "new fold" category target in CASPs 5 and 6, as a function of target size. The 8-Å trend line shows no apparent progress, but there is a very slight improvement for the 4-Å one. **B**: Number of model and target residues superposed in a sequence independent superposition for the best models of each "new fold" category target in CASPs 5 and 6, as a function of target size (EQV5_I: all target residues within 3.8 Å of a model residue in the LGA alignment, using a 5-Å sequence-independent superposition). No progress is apparent.

more than 40 residues are closer than 4 Å and the number for which 60 residues are closer than 8 Å. For the best models, 17 out of 25 CASP6 targets meet the 40 residues under 4-Å criterion, versus 15 out of 23 in CASP5. For 60 residues under 8 Å the numbers are: 18 out of 21 in CASP6 versus 15 out of 17 in CASP5. The numbers are similarly close for the average over the six best models. Overall, as between CASP4 and CASP5,[5] there is no clear difference between CASPs 5 and 6 by this measure.

Figure 8(A) shows the number of residues falling under the GDT4 and GDT8 thresholds for CASPs 5 and 6, as a function of target size, together with quadratic spline fits. There is no apparent progress for the 8-Å threshold, and very slight improvement for the 4 Å one. Figure 8(B) is a similar plot, using the sequence independent EQV5_I measure (all target residues within 3.8 Å of a model residue, in sequence order), instead of a GDT threshold. No progress is apparent between CASP5 and CASP6 by this measure.

Over-all, both measures show at most only very marginal progress. Yet, visual inspection suggests there is more progress. Targets 198, 201, 215, 248 domain 2, and 281 have very impressive looking models, and they include structures with a substantial portion of β structure, often problematic in previous CASPs. Target 281 is particularly noteworthy. There is an analogous fold, and the best model (apparently obtained by template-free methods) has a substantially more accurate structure than the template would provide.

As noted earlier, all assessors in this modeling regime have occasionally overridden GDT_TS in choosing a best model. It appears we are still lacking an adequate measure of quality for these more approximate models.

ANALYSIS OF SUSTAINED PERFORMANCE FOR NEW FOLD TARGETS

With so few new fold targets having reasonable models, it is appropriate to ask to what extent particular groups are performing consistently, as opposed to groups occasionally getting lucky, and happening to produce the best score for a target. As before, we address that by comparing the distribution of success of individual groups with the distribution of success expected by chance. Success is measured as the number of targets for which a group had a model ranking among the top six. The chance distribution was generated by randomly choosing six groups as the best scoring for each target. The chance distribution was constrained so that only groups predicting on that target were included, and the draw was weighted by the number of models submitted. For example, a group submitting four models was four times as likely to be selected as one submitting a single model for a particular target.

Figure 9 shows these data for the 23 CASP5 targets and 25 CASP6 targets. Also shown is information on how many groups submitted models for different number of targets. The blue bars show the number of prediction groups submitting predictions on at least 1, 2, 3,. . . up to the maximum number of targets in each CASP. The yellow bars show the probability of a group scoring among the top six for one target, two targets, three targets, and so on, if the results were random. The red bars show the number of groups actually falling among the top six for one target, two targets, and so on. The more different this distribution is from random, the more significant the results. In both CASPs, ranking in the top six for a single target has no significance, and ranking among the top six for two or three targets, little significance. For CASP5, there are three groups well separated from random, ranking for 8, 11, and 16 targets. There are also two groups ranking for six targets. In CASP6, there are three groups with six top-ranked predictions, already very significantly more than random, and also single groups top-ranked for 7, 8, 9, 10, and 13 targets. Thus, by this measure, there has been an improvement in sustained performance between CASP5 and CASP6.

CONCLUSIONS

The general picture that emerges from this analysis is very similar to that for the previous two CASPs: There is steady but modest progress in difficult comparative modeling and homologous fold recognition, in terms of the extent of sequence-dependent superposition between model and

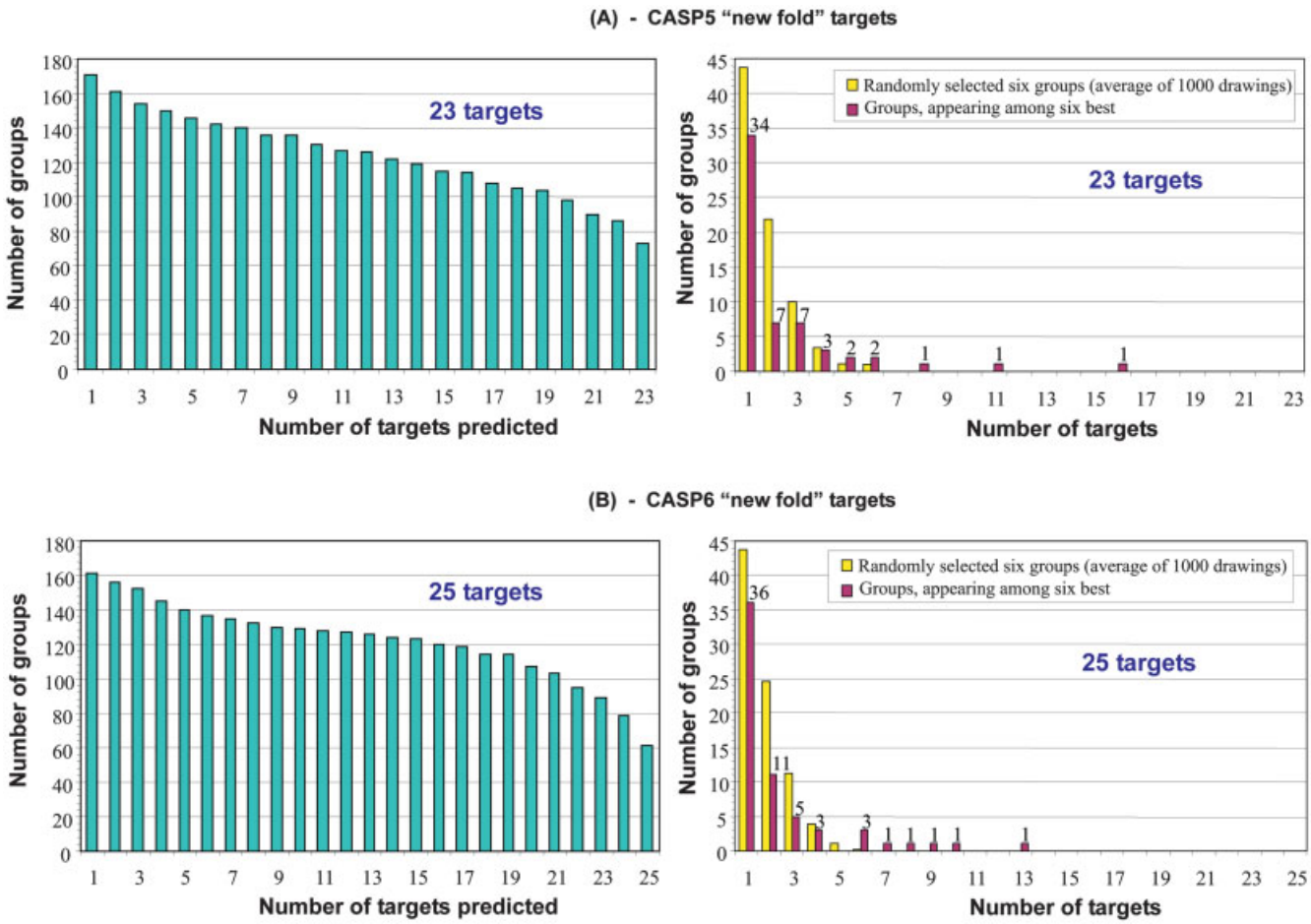

Fig. 9. Distribution of success in predicting "new fold" targets for individual groups in CASP5 (**A**) and CASP6 (**B**), compared with that expected by chance. Left-hand panels (green bars) show the number of groups submitting predictions for at least 1, 2,. . . up to the maximum number of targets in each of these CASPs. In the right panels, red bars show the number of groups ranked in the top six for one target, two targets, and so on. Yellow bars show the distribution of ranking expected by chance. In CASP5, three groups did very significantly better than chance, and a further four groups are in the tail of the chance distribution. In CASP6, there are three groups with six top-ranked predictions, already very significantly more than random, and also single groups top-ranked for 7, 8, 9, and 13 targets. Thus, by this measure, there has been significant improvement in sustained performance between CASP5 and CASP6.

target (as measured by GTD_TS), and in alignment accuracy. Scores for both measures have approximately doubled from CASP1 to CASP6, and another decade of this level of progress would result in excellent models.

For models such as these, based on an evolutionary relationship, accuracy is dominated by three factors. First, parts of the structure correctly aligned with a template will have errors of up to a few Ångstroms, because of differences in main-chain conformation.[15] These errors are the smallest, but their reduction will require the introduction of all-atom refinement methods. It is also likely that this problem must be solved before fully accurate alignments can be obtained, since the information that determines how segments of a protein interact is contained in the detailed atomic interactions and packing, not represented in the current approximate models. This problem is common to the other areas of modeling, and is discussed further below.

Second, for targets in the mid-range of difficulty, a significant fraction of residues will usually be misaligned, introducing larger errors. Alignment accuracy is strongly influenced by the amount of relevant structure and sequence information available for a target, and it is not clear to what extent the improvements over the CASPs have been due to greater quantities of data, and to what extent from real methodological improvements. As mentioned earlier, the impact of these factors is not considered in our difficulty scale. Even if most progress can be ascribed to the availability of more data, it should not be regarded as a minor achievement. Utilization of multiple sequences has required the development of sophisticated new methods.[16–20] Similarly, making effective use of structural information to detect remote homologs and to improve alignments has been possible because of the implementation of a range of methods.[21–28]

Third, the more remote the evolutionary relationship, the smaller the fraction of residues that can be superimposed on a single template. For mid-range difficulty targets, a lot of the model deficiencies are due to this factor, so that a different sort of progress will be required from now on. There are two

ways in which the remaining residues may be modeled. First, other templates may contain some of the missing features. It is encouraging that the analysis shows clear examples of secondary template information being successfully captured in CASP6. Second, template-free modeling methods may be used to add structural features beyond those present in templates. This is occasionally detectable, for example, in CASP6 target 205, where a helix not present in any template was modeled, although in a different orientation to that found in the target. Given the steady improvement in template-free methods, we may expect to see progress on this basis in future.

Visual inspection suggests progress in template-free modeling for small targets, but the analysis only very weakly supports that conclusion. There is also evidence of sustained performance by more prediction groups. The quality of models for large targets remains generally very poor. One of the difficulties in this area is identifying domains in a structure. Assessment of domain identification in CASP6[29] shows that this is far from a solved problem. Modelers also report that for small targets, accurate models are often generated, but cannot be distinguished from less accurate ones. The absence of a reliable scoring function for discriminating between more and less accurate models is an obstacle to progress. As with obtaining accurate alignments, it is likely that successful discrimination will require refinement of each model at atomic resolution, since only then will the interactions that determine the relative free energy of each conformation be included.

Once more, by the measures used here, there is disappointingly little change in high-sequence-identity comparative-model quality. The basic difficulty is the absence of effective refinement techniques—the analysis shows that at 30% or higher sequence identity, most of the best models already have accurate alignments. Side-chain accuracy is closely coupled to main-chain accuracy,[30] and so is unlikely to improve further without full refinement. Similarly, one of the major limitations on free modeling of loops is the error in the surrounding structure.[31] Much attention is now focused on the problem of refinement of comparative models, and there are some encouraging signs outside of CASP, so maybe things will be better in the next experiment.

ACKNOWLEDGMENTS

This work was performed in part under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under contract W-7405-Eng-48. This work was also partly supported by NIH grant LM07085-01 (to K.F.) and by grants from Howard Hughes Medical Institute (HHMI-55000341) and the 6th European Community Framework Programme (MIRG-CT-2004-004543) (to Č.V.).

REFERENCES

1. CASP6 comparative modeling assessment.
2. CASP6 FR assessment.
3. CASP6 NF assessment.
4. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;Suppl 5:163–170.
5. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins 2003;Suppl 6:585–595.
6. CASP6 introduction paper.
7. Zemla A. LGA–a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
8. Venclovas C, Zemla A, Fidelis K, Moult J. Some measures of comparative performance in the three CASPs. Proteins 1999; Suppl 3:231–237.
9. Lackner P, Koppensteiner WA, Domingues FS, Sippl MJ. Automated large scale evaluation of protein structure predictions. Proteins 1999;Suppl 3:7–14.
10. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. ProSup: a refined tool for protein structure alignment. Protein Eng 2000;13: 745–752.
11. Moult J, Melamud E. From fold to function. Curr Opin Struct Biol 2000;10:384–389.
12. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
13. Zhang H, Huang K, Li Z, Banerjei L, Fisher KE, Grishin NV, Eisenstein E, Herzberg O. Crystal structure of YbaK protein from Haemophilus influenzae (HI1434) at 1.8 Å resolution: functional implications. Proteins 2000;40:86–97.
14. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.
15. Liao DI, Herzberg O. Refined structures of the active Ser83→Cys and impaired Ser46→Asp histidine-containing phosphocarrier proteins. Structure 1994;2:1203–1216.
16. 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.
17. Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 2001;17:713–720.
18. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
19. Kahsay RY, Wang G, Gao G, Liao L, Dunbrack R. Quasi-consensus based comparison of profile hidden Markov models for protein sequences. Bioinformatics 2005;21:2287–2293.
20. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;13:1071–1087.
21. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2001; Suppl 5:39–46.
22. Wrabl JO, Grishin NV. Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. Proteins 2004;54:71–87.
23. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19:874–881.
24. Przybylski D, Rost B. Improving fold recognition without folds. J Mol Biol 2004;341:255–269.
25. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 2003;Suppl 6:491–496.
26. von Grotthuss M, Pas J, Wyrwicz L, Ginalski K, Rychlewski L. Application of 3D-Jury, GRDB, and Verify3D in fold recognition. Proteins 2003;Suppl 6:418–423.
27. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
28. Venclovas C. Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. Proteins 2003;Suppl 6:380–388.
29. CASP6 domain assessment.
30. Chung SY, Subbiah S. The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. Protein Sci 1995;4:2300–2309.
31. Samudrala R, Moult J. Handling context-sensitivity in protein structures using graph theory: bona fide prediction. Proteins 1997;Suppl 1:43–49.