PROTEINS: Structure, Function, and Genetics 53:585–595 (2003)

Assessment of Progress Over the CASP Experiments

Česlovas Venclovas,[1,2] Adam Zemla,[1] Krzysztof Fidelis,[1] and John Moult[3]*
[1]*Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California*
[2]*Institute of Biotechnology, Graičiūno 8, 2028 Vilnius, Lithuania*
[3]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*

ABSTRACT **The quality of structure models produced in the CASP5 experiment has been compared with that in earlier CASPs. The most significant progress is in the fold recognition regime, where the development of meta-servers has allowed more accurate consensus models to be generated. In contrast to this, there is little evidence of progress in producing more accurate comparative models, particularly those based on sequence identities > 30%. For comparative models based on low-sequence identity and for fold recognition models, accuracy depends primarily on the fraction of the target structure that is similar to an available template, and the quality of the alignment. Overall, these results indicate that there are still no effective methods of improving model quality beyond that obtained by successfully copying a template structure. For models of proteins with previously unknown folds, there appears to be a pause in the previous consistent improvement. There is some evidence that more groups are producing top-quality models, however. Although specific progress between successive experiments is sometimes difficulty to identify, over the history of all the CASPs there has been steady, if sometimes slow, progress in all modeling regimes. Proteins 2003;53:585–595.**

Key words: protein structure prediction; communitywide experiment; CASP

INTRODUCTION

Five CASP experiments have now been completed, spanning the period from 1994 through 2002. The results reflect 8 years of work in protein structure modeling by a large number of people. Therefore, it is of considerable interest to ask what progress has been made, and in which areas. Each of the three assessors' articles in this special issue of *Proteins* addresses aspects of the subject. Here we attempt a broader view, looking at all types of three-dimensional prediction and spanning the full set of CASP experiments. Most of the methods we use also provided progress analysis through CASP4.[1] CASP5 results have extended those earlier analyses. Two new evaluation criteria have been added. One uses the popular GDT_TS measure.[2,3] The other relates alignment accuracy to the limits imposed by available template structures.

GENERAL CONSIDERATIONS

Choice of Models to Evaluate

We analyze two aspects of progress: how the quality of the very best models is improving and how the quality of models produced in the field as a whole is advancing. We evaluate progress in best performance 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. We do not take an average over all groups because many participants in CASP are not professional computational biologists, or may be just testing a new idea, without expecting it to produce the most accurate models.

Relative Target Difficulty

The difficulty of producing a high-quality model of a target protein depends on two primary factors: the similarity of the protein sequence to that of a protein or proteins with known structure and the similarity of the structure of the target protein to potential templates. As in the CASP4 progress assessment, we use a two-dimensional scale to estimate difficulty, reflecting these two factors. Potential templates are identified by determining the extent of structure similarity using the LGA[3] software. Each experimental target structure is compared with every structure in the protein databank. For CASP5, templates were taken from the PDB releases accessible before each target deadline. The templates for targets in previous CASPs are taken from PDB releases at the time of each experiment and are the same as in the previous analyses.[1,4] For each target, the most similar structure in the appropriate version of the PDB is chosen as the representative template.

Similarity between a target structure and a potential template is taken to be the number of target-template Cα atom pairs that are within 5 Å in the LGA superposition. Note that this criterion is sequence independent, measuring structure similarity, rather than the align-ability of

Grant sponsor: National Institutes of Health; Grant number: LM07085-01.

*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 20 May 2003; Accepted 23 May 2003

the corresponding sequences. The 5 Å threshold maintains compatibility with earlier target-template comparisons,[1,4,5] which were made by using Prosup[6] software. A slight disadvantage of the relatively large cutoff is that it sometimes leads to substantial superimposability between unrelated structures, particularly for small proteins. Sequence identity is defined as the fraction of structurally aligned residues that are identical. As noted in the earlier study, some templates with maximum structure superimposability are not those with highest sequence identity to a target. As before, in these cases (10 in previous CASPs; 5 in CASP5), the template with the highest sequence identity was selected.

Domains

Many target structures consist of two or more structural domains. Because domains within the same structure may present modeling problems of different difficulty, assessment in CASPs 4 and 5 has treated each identifiable domain as a separate target. From a predictor's perspective, domain parsing is often not possible when only the sequence of a target is known. Here, for evaluation of models based on templates (the CM and FR targets), structures are only parsed into domains if these domain divisions were likely identifiable to a predictor (e.g., a part of the protein has a sequence related to that of another protein). There are a total of seven such cases in CASP5 and three in CASP4, and one each in CASPs 2 and 3. For evaluation of non-template-based models (the FR/NF and NF target categories), all domains identified by the assessors have been 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 best available template and the experimental structure of each target. Targets span a wide range of structure and sequence similarity in all the CASPs, and in general, the distribution of difficulty is also similar for all the CASPs. There are some minor points worth noting: CASPs 1 and 2 had a few targets with exceptionally high-sequence identity to a known structure, and there are no very low superposability targets. Figure 1(B) shows the difficulty distribution for only CASPs 4 and 5, with individual CASP5 targets/domains labeled.

For most analytical purposes, it is more convenient to use a one-dimensional scale of target difficulty, even though this results in some loss of resolution. As in the previous analysis, we have projected 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 and RANK_SEQ_ID is the rank along the vertical axis.

CASP assessment has usually been performed by dividing the targets into three categories of relative difficulty: comparative modeling, fold recognition, and new folds.[7] These regimes approximately map to the one-dimensional difficulty scale used here, with comparative modeling the easiest, fold recognition in the intermediate difficulty range, and new fold targets the hardest. However, there is some reordering.

OVERALL MODEL QUALITY

No single measure completely captures the relative quality of a structure model. As discussed elsewhere,[8] GDT_TS is the best so far devised. 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 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.[3] The procedure is conducted by using thresholds of 1, 2, 4, and 8 Å, and the 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:

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

where Nn is the number of residues superimposed under a distance threshold of "n" Å.

The 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, so that the 1 and 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 and 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 GDT_TS measure useful across a wide range of modeling accuracy.

Although GDT_TS is the best measure so far devised, it is not perfect. This is most noticeable in the new fold regime, where the models are often very approximate. The assessors in both CASP4 and CASP5 found GDT_TS a useful measure for identifying interesting models, but they noted that the highest quality model is not always the one with the highest GDT_TS score. When comparing performance across CASPs, there may also be a limitation for the most accurate comparative models as well. Significant improvement in model quality will result in a fairly small increase in GDT_TS values, and these may sometimes be partly masked by the noise in the target difficulty estimate.

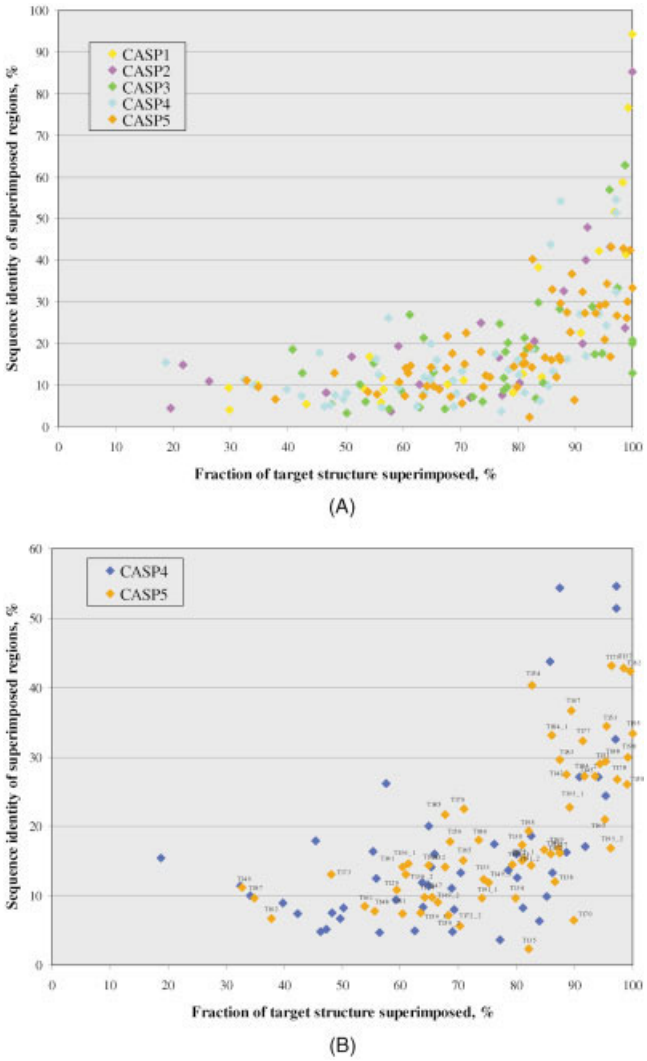


Fig. 1. Distribution of target difficulty. The difficulty of producing an accurate model is shown as a function of the fraction of each target that can be superimposed on a known structure (horizontal axis) and the sequence identity between target and template for the superimposed portion (vertical axis). In both CASPs, targets span a wide range of difficulty. **A:** All CASPs. **B:** CASPs 4 and 5 only. CASP5 targets are labeled.

Figure 2 shows the GDT_TS score for the best model of each target in CASPs 4 and 5. The plot is locally noisy, partly as a consequence of the approximate nature of the difficulty scale. However, for about the easiest two thirds of the targets, there is a trend for CASP5 scores to be higher than CASP4, suggesting progress. Smoothing makes the overall trends clearer. Figure 3(A) shows the same data, with each point an average over five targets. Data for earlier CASPs are also included here. For most of the difficulty range, there is a clear improvement in model

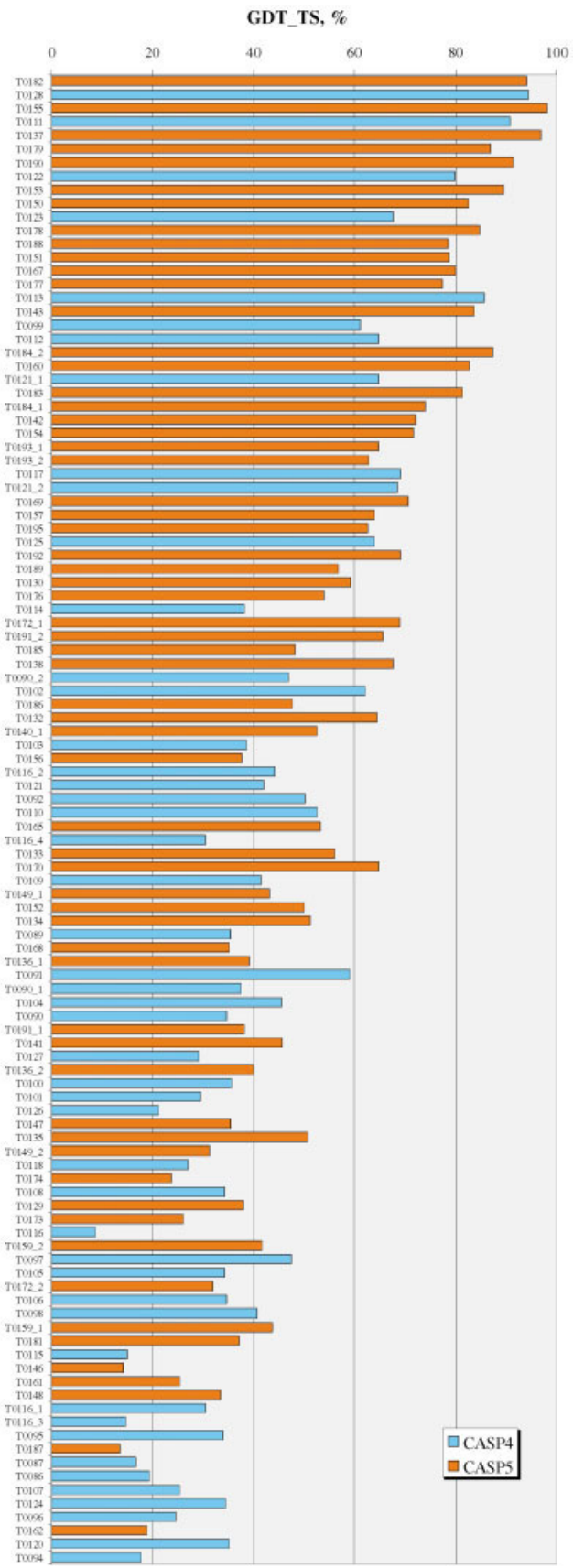
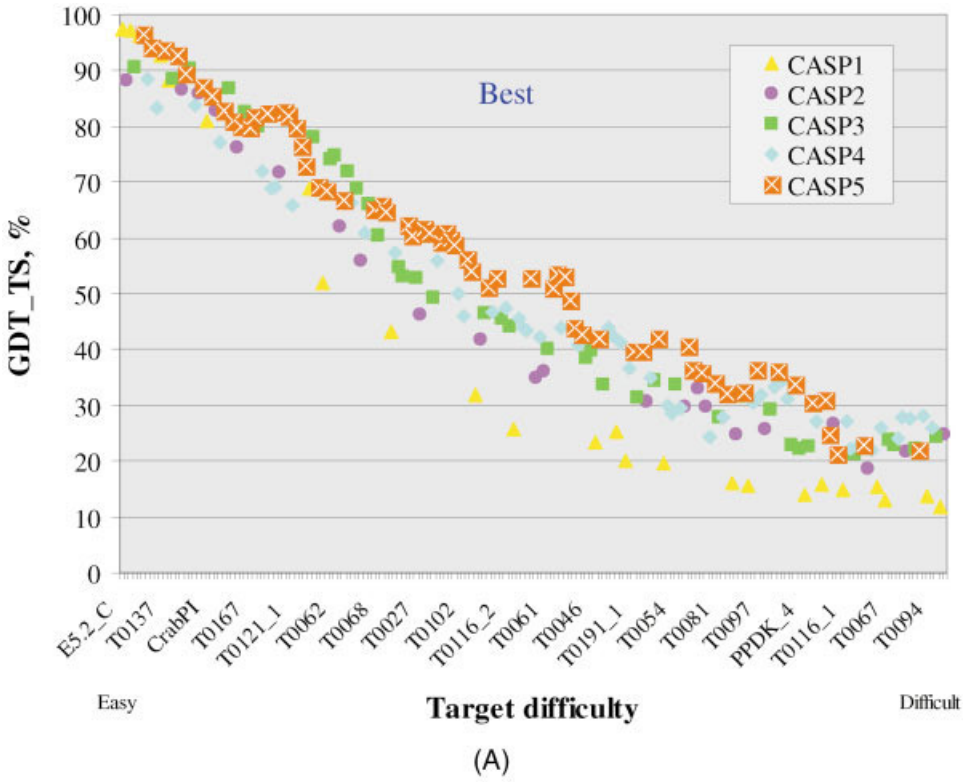
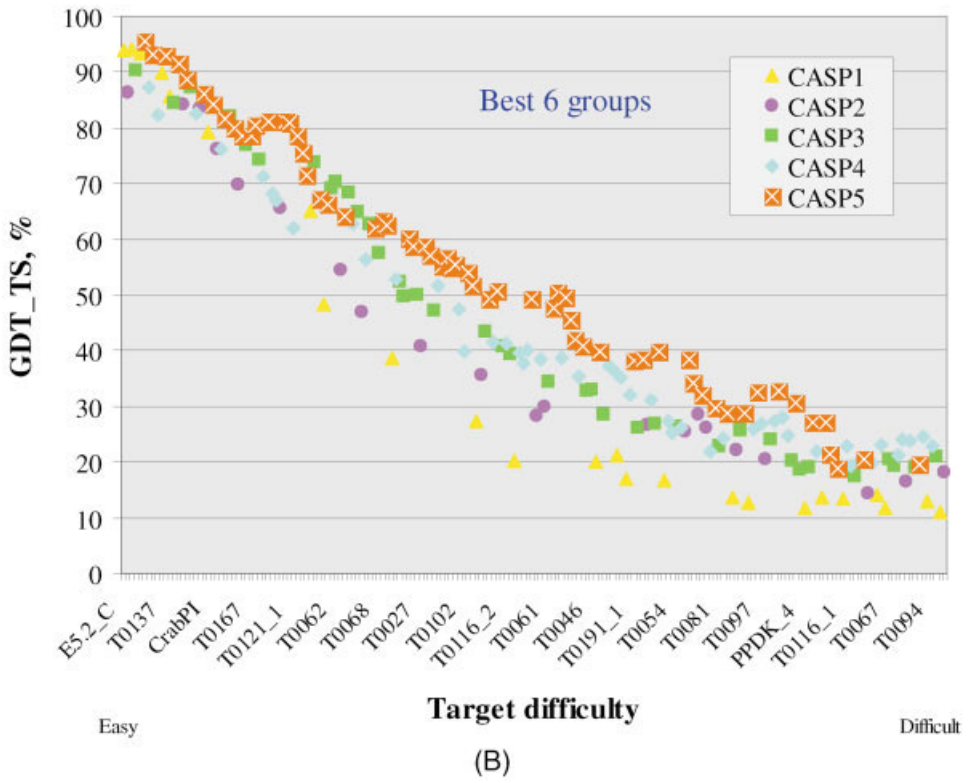


Fig. 2. GDT_TS scores for the best models submitted on each target in CASPs 4 (blue bars) and 5 (red bars). Targets are ordered by modeling difficulty (see text). A GDT_TS score of 100% would correspond to a model in perfect agreement with experiment.



(A)



(B)

Fig. 3. GDT_TS scores for models for targets in all CASPs. Data are smoothed by averaging over sets of five adjacent targets. **A:** Scores for the best models on each target. **B:** Average score over the top six models from different groups. Both plots show a clear improvement from CASP1 to CASP2. There is also an improvement from CASP2 to CASP3 and from CASP4 to CASP5 in the intermediate range of target difficulty. These trends are most apparent in (B) for the average over the six best models.

quality from CASP1 (yellow triangles) to the later CASPs. Performances in CASPs 2, 3, and 4 are harder to distinguish between. For the central section of the difficulty range, though, there is an obvious if modest improvement from earlier CASPs to CASP5 (orange squares). There is no clear improvement in CASP5 for the easiest, comparative modeling targets, or the hardest, new fold, targets.

Figure 3(B) also shows smoothed data, using average GDT_TS values over the six best models from different groups, rather than the single best. There is a similar improvement after CASP1, and in addition, an improvement from CASP2 to CASP3 for the central section. The same improvement in performance in CASP5 for the midrange of difficulty is apparent and stronger. There is also an indication of modest improvement in the comparative modeling regime.

In general, there has been a large improvement between CASP1 and CASP5, with some of that improvement occurring recently. On the other hand, a perfect model would have a GDT_TS value close to 100%. In that sense, except for targets that are very similar to a known structure, there is still a very long way to go before models start to be comparable in quality with experimental structures.

ALIGNMENT ACCURACY

In the comparative modeling and fold recognition regimes, models are generated primarily by mapping the target protein sequence on to one or more template structures. Even though a correct template may have been identified, the mapping presents difficulties, so that the accuracy of this alignment is a critical factor in determining model quality. Positioning a residue one peptide unit away from the correct location causes a main-chain error of 3.8 Å, and four residues away results in an error of about 10 Å. As in previous CASPs, we measure alignment accuracy by counting the number of correctly aligned residues in the LGA superposition of the model 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 experimental atom, and there is no other experimental structure Cα atom nearer.

Note that this definition of alignment accuracy focuses on the mapping of a model structure onto the corresponding experimental structure, not onto a template structure. There are two reasons for this. First, the relationship between model and experiment is most relevant to the accuracy of the model. Second, many methods now use multiple templates, and it is not clear what the most relevant comparison with a template would be. An alternative view of alignment is at the sequence level. In building a model, structure alignment is often deduced from sequence alignment. Although building an accurate sequence alignment is a challenging problem in many cases, it is a step along the way, and not the final result. Furthermore, a "perfect" sequence alignment may not result in a perfect structure alignment for a number of reasons.

Figure 4 shows the alignment accuracy for each target in CASPs 4 and 5. Solid bars show the percent of residues correctly aligned in the models that are most accurate by this criterion, and stripped regions are the additional fraction of residues aligned with an error of no more than four residues. Once again, although the plots are locally noisy, the general trend is that CASP5 alignment accuracy is higher than CASP4, in correctly aligned residues and also residues aligned within plus or minus four residues. Figure 5 shows the corresponding smoothed plots, with all CASP targets included, for both correct alignment (A), and alignment within four residues (B). Alignment accuracy falls approximately linearly with target difficulty, in a manner similar to that seen for GDT_TS in Figure 3. Both plots show trends very similar to those of smoothed GTD_TS (Fig. 3), suggesting that improvements in model quality reflect in improved alignments. The plot of alignment accuracy [Fig. 5(B)] within plus or minus four residues shows a stronger improvement trend over the CASPs than the one for exact alignment [Fig. 5(A)], indicating there is a greater reduction in large alignment errors.

ALIGNMENT ACCURACY RELATIVE TO TEMPLATE-IMPOSED LIMITS

A model built by just copying from a single template has an upper limit of alignment accuracy, equal to the number of residues that may be superimposed between that template and the target structure. In recent CASPs, it is usual for the best models to be based on multiple templates, choosing appropriate regions of structure from different templates. If successful, this procedure should lead to an alignment accuracy above that possible with a single template. Successful modeling of nonalignable regions, such as loops and larger motifs unique to a target structure, should also lead to an alignment accuracy above the single template limit. For these reasons, it is useful to relate alignment accuracy to that achievable by optimum copying from the single best template. We define the maximum alignability of a target as the fraction of Cα pairs that are within 3.8 Å of each other in the LGA superposition of the target and best available template structure. (Because the sequences of template and target are not the same, the additional alignment criterion of the closest Cα pairs corresponding to the same residues cannot be included. In practice, this is a minor factor).

Figure 6(A) shows the smoothed alignment accuracy for the best models of all targets in CASPs 4 and 5 [a subset of the data in Fig. 5(A)], together with the smoothed maximum alignability. For the easiest, comparative modeling targets, CASP5 targets have a slightly higher average alignability than those from CASP4, and this is reflected in the higher alignment accuracy of the models. Alignability falls steadily and approximately linearly with increasing target difficulty, but with a smaller slope than that of the fall off in alignment accuracy. As noted above, alignment accuracy falls with target difficulty in a manner similar to that seen for GDT_TS. Thus, model accuracy, as reflected by GDT_TS, is dominated by two factors. First, more difficult targets have a smaller fraction of residues that can be superimposed on a template, and modeling methods

are not at present successful for the rest of a protein. Second, as target difficulty increases, the fraction of theoretically alignable residues that are successfully positioned also falls substantially.

Figure 6(B) shows the alignment accuracy for all CASPs, as a percent of the maximum alignability. Targets are ordered by the sequence identity between the target and best available template. In all CASPs, most targets with >30% sequence identity to a template have all possible residues 100% correctly aligned. The worst alignment in this zone is ~90% of the maximum possible. It is surprising that there is one target (T0123) with 55% sequence identity to a template with an alignment accuracy only 91% of maximum. This fatty acid binding protein has a conformational difference between target and template,[9] probably associated with ligand binding. The results for all CASPs are similar (i.e., there is no evidence that the incidence of alignment errors has decreased for the relatively high-sequence identity targets). Between 15 and 30% sequence identity, most targets have worse than 90% alignment accuracy, ranging down to 60%. There are some outstandingly poor alignments for some targets from CASPs 1, 2, and 3, suggesting that in CASPs 4 and 5, fewer really poor results are obtained. Most of these poorly aligned targets are remote homologues of the template. Below 15% sequence identity, CASP1 results are clearly worse than in the other CASPs, and CASP5 results appear somewhat better. The worst CASP5 target has an alignment accuracy 27% of maximum, and one CASP5 target at only 5% sequence identity has 94% of maximum. Most of these targets are new fold or analogous fold relationships.

Except for very low sequence identity targets, CASP5 points are intermingled with those from CASPs 2, 3, and 4, suggesting that there has been no significant improvement in alignment quality since CASP2. There are no targets for which there is a significantly higher alignment than provided by correctly copying the best template. Thus, by this measure, the use of multiple templates and modeling of nontemplate regions have yet to have a discernable impact on model quality. Because the definition of maximal alignability is slightly different from the measure of alignment accuracy (see above), it is not possible to rule out some improvement from multiple templates, however, and careful benchmarking[10] as well as hands-on modeling experience suggests that improvement should be expected.

NEW FOLD METHODS

For the most difficult, new fold targets, there are no templates available, and no significant sequence identity to a known structure, so that the difficulty scale used for other targets is not relevant. There are alternative factors that affect modeling difficulty, particularly structure class (in previous CASPs, better results have been obtained for

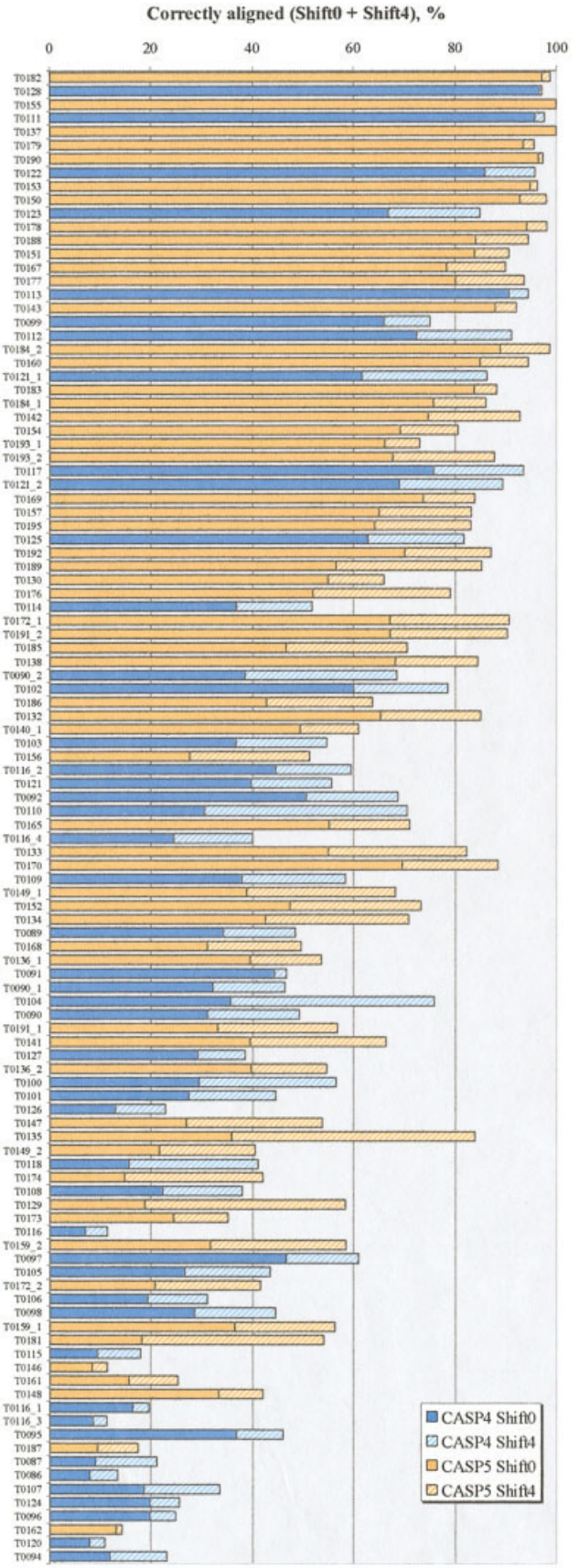


Fig. 4. Alignment accuracy for the best models for all targets in CASP4 (blue) and CASP5 (red). Targets are ordered by difficulty. Solid bars show the fraction of residues correctly aligned, and the hashed regions show the additional residues aligned to within four residues of the correct position.

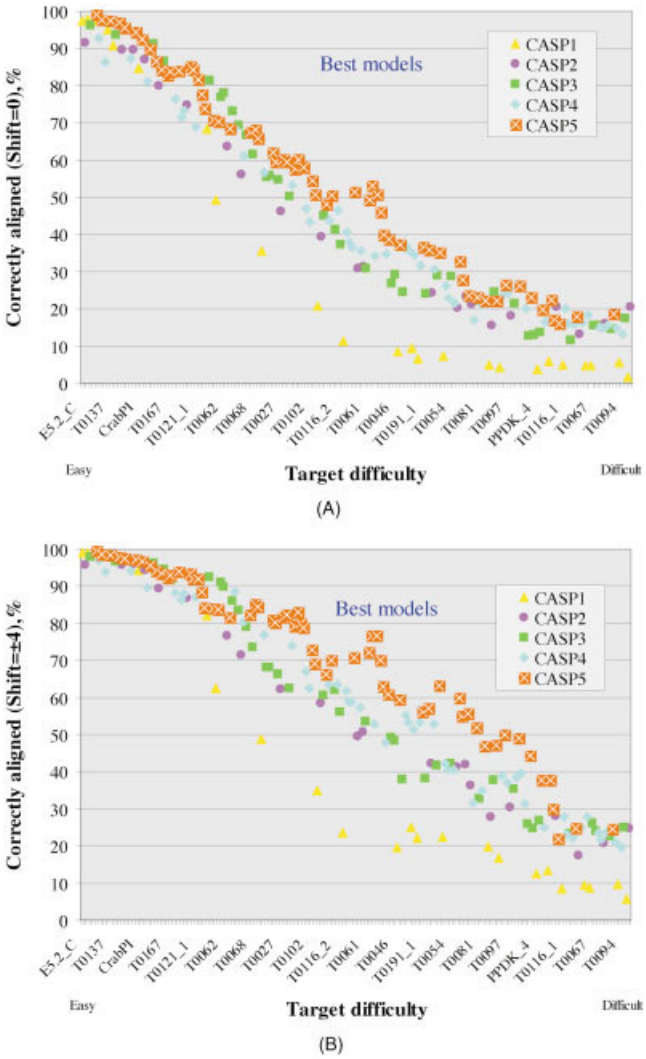


Fig. 5. Alignment accuracy for the best models for each target in all CASPs, smoothed by averaging over sets of five adjacent targets. **A:** Percent of residues correctly aligned. **B:** Percent of residues aligned to within four residues of the correct position. Trends here follow those in the equivalent GDT_TS plots (Fig. 3) indicating that for many targets, alignment accuracy dominates model quality.

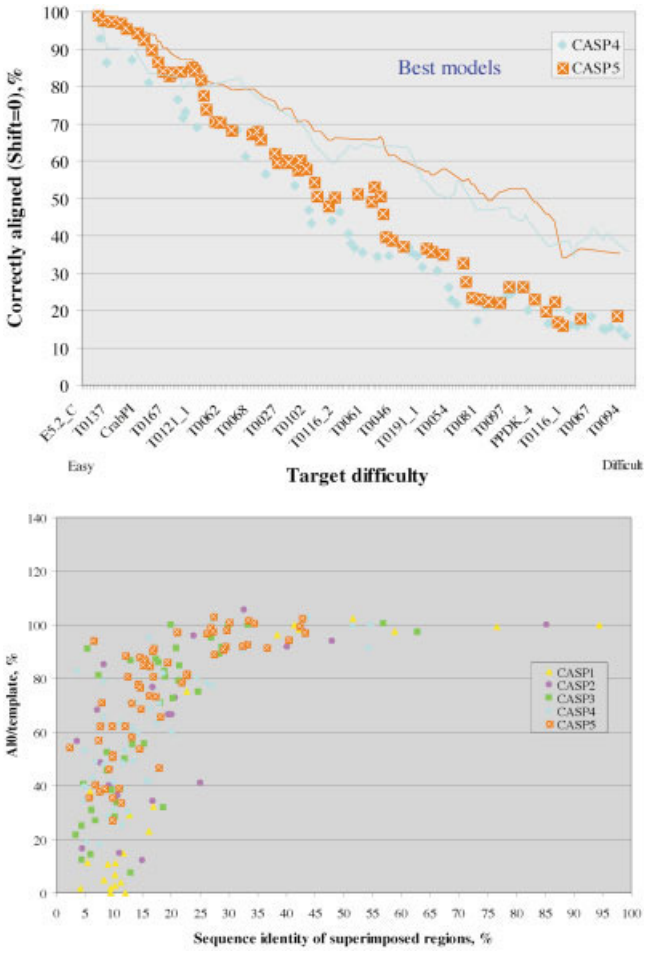


Fig. 6. **A:** Smoothed alignment accuracy and smoothed maximum alignability as a function of target difficulty. Targets for CASPs 4 and 5 are shown. Maximum alignability (continuous line) is defined as the fraction of equivalent residues in a superposition of the target and best template structures. The fraction of this theoretical maximum that is successfully aligned falls steadily with target difficulty. **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 the fraction of 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 >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, but there is little evidence of an improvement over template copying.

targets with a predominantly α secondary structure, worst for predominantly β), contact order (how local the contacts in the experimental structure are),[11] domain structure, and size. The extent of sequence-dependent structure superposition is the most relevant measure of model quality. To improve the resolution of the analysis, we consider the four terms that contribute to the GDT_TS measure, rather than just that single value (i.e., the number of residues that can be superimposed under 1, 2, 4, and 8 Å.

All domains identified by the assessors are treated as separate targets in this evaluation. Domains that are unambiguously new folds (NF targets) and domains that have a faint or partial relationship to known folds (FR/NF targets are included, providing a total of 15 targets in CASP5. The FR/NF targets are considered because any relationship to a known fold is too weak for template-based modeling to be very effective.

Figure 7 shows the results. In each CASP, targets are ordered by size. Fold type is indicated by the usual Greek letter classification. The number of residues superimposed under the distance thresholds of 1, 2, 4, and 8 Å are depicted by the components of each target's bar. Recall that this is the number of residues for which the largest error is less than or equal to each threshold. The root-mean-square (RMS) error on such a set is typically about half the threshold. Thus, substructures meeting the 8 Å threshold are those that visual inspection would usually rate as correct. A convenient way of

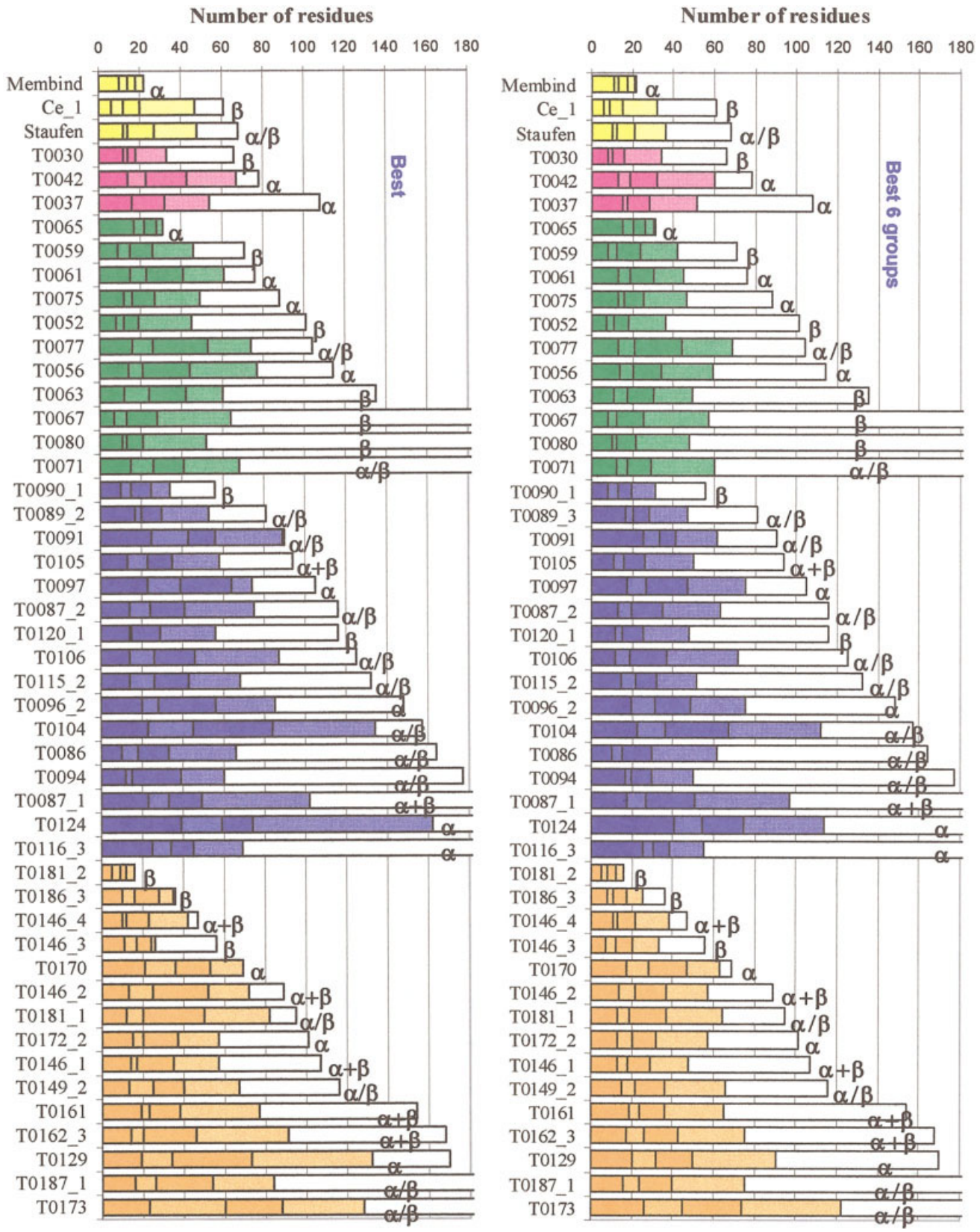


Fig. 7. Model quality for the best (**A**) and averaged over the six best (**B**) new fold category targets, for all CASPs. CASPs are distinguished by color: yellow: CASP1; pink, CASP2; green, CASP3; blue, CASP4; orange, CASP5. For each target, the lowest bars show the number of residues superimposed between model and target to closer than 1 Å, the next bar, the number superimposed to 2 Å, then 4 Å, and then 8 Å. The open bars show number of residues superimposed to >8 Å. Greek letters indicate the fold type. A trend for improved accuracy can be seen for CASPs 1–4. It less clear whether there has been any improvement from CASP4 to CASP5.

comparing model quality in each CASP is to examine the number of targets for which >40 residues are closer than 4 Å and the number for which 60 residues are closer than 8 Å. 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. Results for CASPs 1 through 4 were also presented in the previous analysis.[1] As noted then, there has been a steady improvement over the first four CASPs, apparent from the number of residues superimposed under the 4 and 8 Å thresholds. The best models plot shows no clear overall progress between CASPs 4 and 5 by these measures, although the small number of targets and the variation in performance because of the peculiarities of each target could be masking some advance. There are some notable performances in CASP5 on individual targets, particularly the two smallest ones: domain 3 of T0186, a predominately β structure; and T0170, an α structure. For both these targets, almost all residues of the best model are superposable on the target structure with a threshold of <8 Å. (In CASP4, there was one, slightly longer target, completely correct by this criterion.) Figure 7(B) does provide evidence of some improvement in average performance between CASPs 4 and 5. In CASP5, all but three of the targets longer than 60 residues have at least 60 residues under the 8 Å threshold, whereas in CASP4, only 6 of the 14 reached this threshold. Performance is also markedly superior to the CASPs before that by this criterion, with only one target in CASP3 having >60 residues under 8 Å.

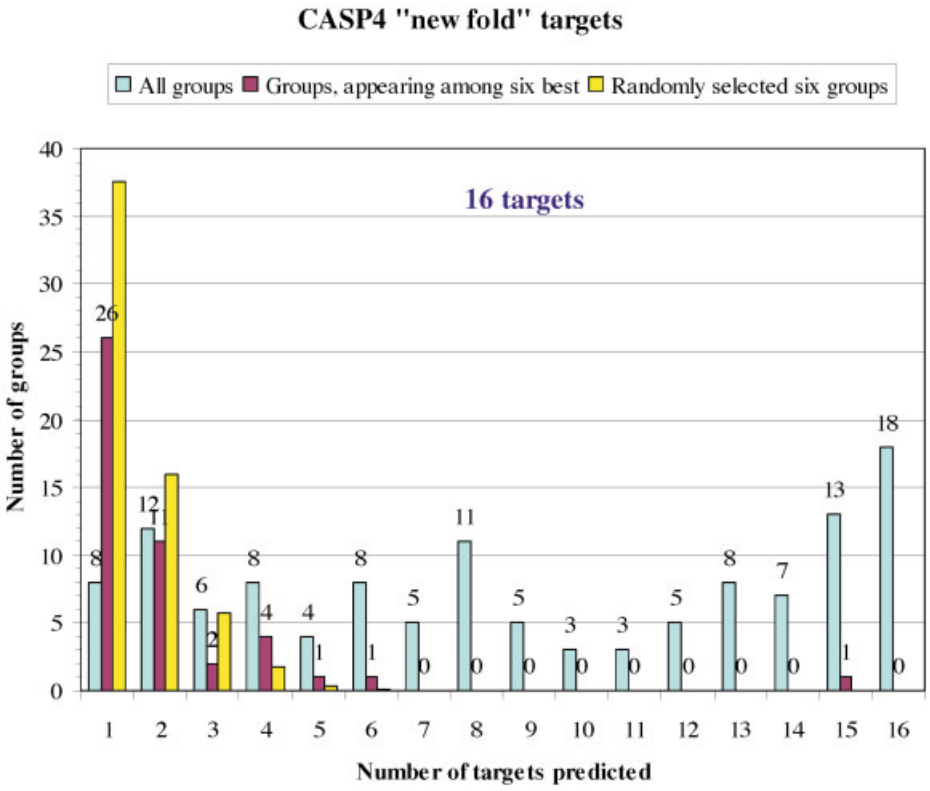
ANALYSIS OF SUSTAINED PERFORMANCE FOR NEW FOLD TARGETS

The data in Figure 7(B) suggest that although the very best models are not improved from CASP4, more groups are producing good models. We now ask whether this represents an improved sustained high-quality performance by particular groups. A measure of that is provided 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 (i.e., a group submitting four models was 4 times as likely to be selected as one submitting a single model for a particular target). Figure 8 shows these data for the 16 CASP4 targets and 15 CASP5 targets. Also shown is information on how many groups submitted models for different number of targets. Blue bars show the number of groups submitting predictions on 1, 2, 3, . . . up to the maximum number of targets in CASP4 (A) and CASP5 (B). There were substantially more groups making new fold predictions in CASP5 than in CASP4 (167 vs 124), and in CASP5, each group submitted more targets on average, most strikingly, 80 CASP5 groups submitted models for all 15 targets, whereas in CASP4, only 18 groups submitted on all 16 targets. 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 distribution is sharper in CASP5 than in CASP4, with a larger probability of selection for a single target, because of the larger number of predictions and predictors. That is, the chances of randomly achieving top six status for more than one target were significantly lower in CASP5 than in CASP4. 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 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. In CASP4, only one group is well separated from the random distribution, ranking in the top six for 15 of 16 targets, and a further six groups ranked for ≥4 targets. For CASP5, there are two groups well separated from random, one ranking for 10 targets, and the other for 6. There are a further two groups ranking for four targets. By this measure, somewhat contrary to the impression given by Figure 7(B), there is no strong evidence of an improvement in sustained performance between CASPs 4 and 5.

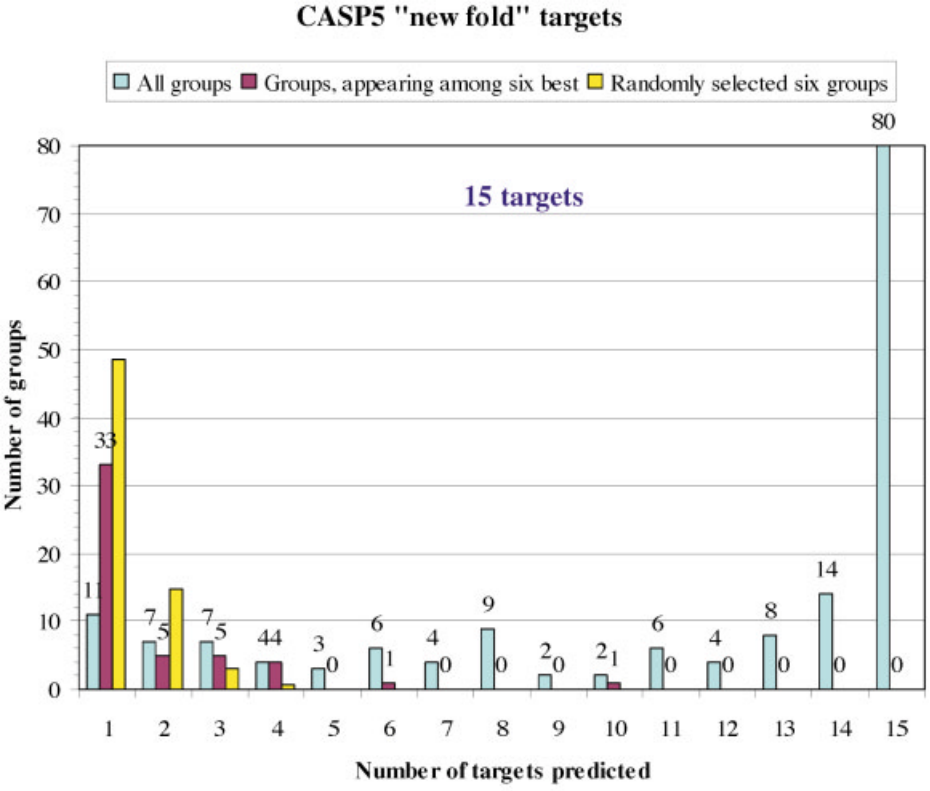
CONCLUSIONS

As the number of CASP experiments increases, evaluation of progress and conversely, identification of bottlenecks, becomes increasingly important. We have extended the earlier analyses to include the CASP5 results and added two more analysis tools: a new alignment-related measure and the popular GDT_TS score. Comparison of performance remains an imperfect art. The primary problems are establishing a reliable scale of the relative difficulty for modeling different targets, unscrambling different contributions to total error, and the absence of exact measures of model quality. Despite these difficulties, the results do show some clear trends.

It is encouraging that the smoothed GDT_TS measure shows evidence of improvement in the accuracy of the best models in CASP5 for all regimes of difficulty except in some regions of comparative modeling and prediction of structures with new folds. The improvement is more pronounced when the quality of the six best models on each target is considered, and then extends to more of the comparative modeling regime. On the other hand, the improvement is generally modest, and values of GDT_TS are still low except for high-sequence identity comparative modeling, showing that although the field is moving forward, there is still a very long way to go before models competitive with experiment are produced. Smoothed alignment accuracy for template-based modeling shows a similar improvement to GDT_TS, suggesting that for most types of modeling, alignment accuracy is the most significant cause of error.


CASP4 "new fold" targets
All groups
Groups, appearing among six best
Randomly selected six groups
16 targets
Number of groups
Number of targets predicted
0
5
10
15
20
25
30
35
40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
8 26 12 11 6 2 8 4 4 1 8 1 5 0 11 0 5 0 3 0 3 0 5 0 8 0 7 0 13 1 18 0

(A)


CASP5 "new fold" targets
All groups
Groups, appearing among six best
Randomly selected six groups
15 targets
Number of groups
Number of targets predicted
0
10
20
30
40
50
60
70
80
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
1 33 7 5 7 5 44 3 0 6 1 4 0 9 0 2 0 2 1 6 0 4 0 8 0 14 0 80 0

(B)

Figure 8.

Closer examination of alignment accuracy as a fraction of that achievable by correctly copying a single best template shows that alignments are often close to the maximum obtainable for cases based on >30% sequence identity. There is no indication that the use of multiple templates, loop building, or other refinement techniques improves the quality of models beyond that obtained from copying a single best template. Improvements of that sort may be masked by remaining alignment issues. Nevertheless, substantial changes would be apparent. Comparison with earlier CASPs shows no evidence of alignment improvement in the rest of the comparative modeling regime by this measure. These two bottlenecks (no improvement in alignment in the comparative modeling regime and no refinement methods that significantly improve template built targets) have persisted since CASP2.

For new fold targets, there may be a pause in the steady progress seen since CASP1, with no clear improvement at CASP5.

Bottlenecks and pauses notwithstanding, there is clear evidence of an overall steady improvement over the history of the CASP experiments. This is most evident in Figure 3, which shows the GDT_TS scores for the best models in all CASPs. CASP5 scores are not only way above CASP1, they are also well differentiated from CASP2 and that is true across the whole range of modeling difficulty. In this sense, although progress in any particular regime between any particular successive pair of CASPs is often hard to identify, the overall trend is encouraging and suggests the field is moving steadily, if sometimes slowly, forward.

What factors are contributing to progress? Methods in all modeling regimes now rely heavily on the knowledge base of known sequences and structures. Is progress just the result of increasing size of these data sets, or are the methods really improving? In comparative modeling, the increased number of available templates for a typical target and the increased number of sequence relatives to use in determining an alignment should lead to improved backbone accuracy. Most successful fold recognition methods now depend heavily on sequence data, both in detecting remote relationships and to produce accurate secondary structure predictions. The most dramatic impact of increased data availability has been in the new folds regime. The more successful methods are based directly on using sequence and structure information, rather than the older physics-based approaches. In one sense, then, most progress may be attributed to increased data availability. Nevertheless, in all modeling regimes, making use of the available data has required very substantial algorithm development, so perhaps a more balanced answer is that it is a bit of both.

Fig. 8. Distribution of success in predicting new fold targets for individual groups in CASP4 (**A**) and CASP5 (**B**), compared with that expected by chance. Blue bars show the number of groups submitting predictions for 1, 2, . . . up to the maximum number of targets in each of these CASPs. More groups submitted in this category in CASP5 than in CASP4, and more of those groups submitted on all targets. In CASP5, 80 groups submitted models for all targets (the corresponding bar is truncated in the figure). 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 (i.e., randomly drawing from the submitted models on each target). In CASP4, there was one group who ranked in the top six very much greater than chance (on 15 of 16 targets), and a further six groups in the tail of the chance distribution. In CASP5, two groups did very significantly better than chance (ranking in 10 of 15 and 6 of 15), and a further four groups are in the tail of the chance distribution.

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 (to KF).

REFERENCES

1. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;Suppl 5:163–170.
2. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13–21.
3. Zemla A. LGA–a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003. Forthcoming.
4. Venclovas C, Zemla A, Fidelis K, Moult J. Some measures of comparative performance in the three CASPs. Proteins 1999; Suppl 3:231–237.
5. Lackner P, Koppensteiner WA, Domingues FS, Sippl MJ. Automated large scale evaluation of protein structure predictions. Proteins 1999;37:7–14.
6. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. ProSup: a refined tool for protein structure alignment. Protein Eng 2000;13: 745–752.
7. Moult J, Melamud E. From fold to function. Curr Opin Struct Biol 2000;10:384–389.
8. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;Suppl 5:2–7.
9. Hoedemaeker FJ, Visschers RW, Alting AC, de Kruif KG, Kuil ME, Abrahams JP. A novel pH-dependent dimerization motif in beta-lactoglobulin from pig (Sus scrofa). Acta Crystallogr D Biol Crystallogr 2002;58:480–486.
10. Venclovas C. Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. Proteins 2001;Suppl 5:47–54.
11. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.