Assessment of Predictions Submitted for the CASP6 Comparative Modeling Category

Michael Tress,* Iakes Ezkurdia, Osvaldo Graña, Gonzalo López, and Alfonso Valencia
Protein Design Group, CNB–CSIC, Cantoblanco, Madrid, Spain

ABSTRACT **Here we present a full overview of the Critical Assessment of Protein Structure Prediction (CASP6) comparative modeling category. Prediction accuracy for the 43 comparative modeling targets was assessed through detailed numerical comparisons between predicted and experimental structures. Assessments using standard measures for model backbone quality and structural alignment accuracy highlighted a small number of groups with stand out predictions and these findings were backed up by statistical comparisons. We were able to carry out evaluations of side-chain contacts predictions and side-chain rotamer accuracy, for which one group turned out to have statistically better predictions. We also assessed the prediction quality of structurally divergent regions and biologically important sites. Interestingly we were able to show that predictors were not predicting these important functional regions with any greater accuracy than the rest of the structure. In addition we investigated the ability of predictors to build models that improve on the structural template and reached some tentative conclusions from comparisons with the previous CASP experiment. Proteins 2005;Suppl 7:27–45.**

Key words: target structures; structure prediction; 3D models; side chains; rotamers; side-chain contacts; alignments; binding sites; refinement

INTRODUCTION

Comparative modeling is a powerful technique much used by biologists and a few modeling servers in particular are very popular. Homology models are used for studying mutations, for the prediction of binding sites, for the docking of small molecules and for many other biological processes. Given that these methods are perhaps the most obvious point of contact between the prediction community and the outside world, we have a responsibility to provide methods that produce the best possible models.

Biologists and other users are interested above all in accuracy and reliability and in which method they should choose to build their model. Predictors, on the other hand, have invested huge amounts of time and effort in developing methods and are interested an accurate evaluation and what works best in the field. All of them want to know what scientific progress has been made and what future challenges will be. So it is important that these methods are evaluated as thoroughly as possible, while considering both predictors and potential users.

Comparative modeling is a technique capable of generating detailed and accurate 3D structural predictions for proteins of unknown structure. In order to generate a useful model there must be a detectable evolutionary relationship between the target sequence and a protein of known structure (the template) and the alignment between the two must be correct. It is a well known fact that if the evolutionary relationship between query and template is close, the structures of the template and the query are more similar and better 3D models can be produced.[1]

Identifying the best structural template is still a big challenge, but aside from the extent of structural conservation between target and template structures, the quality of models produced by comparative modeling is determined by a range of factors, such as the ability to deduce the correct structural alignment with the template protein and the accuracy of the modeling step. Any evaluation of comparative modeling must also take into account the modeling of structurally divergent regions and the placement of side chains. Finally model refinement, as we will see later, is important and will be increasingly so.

In the previous edition of CASP,[2] the assessors of the Comparative Modeling (CM) section noted that the scope of comparative modeling had been enlarged toward proteins that had very distant relationships with proteins of known structure. While we find that the same is true of the techniques themselves in this edition, the scope of the CM assessment actually shrank in relation to the previous year. In part this was a decision of the organizing committee to include more "easy" comparative modeling targets, in part a decision of the assessors not to define comparative modeling targets beyond those targets for which we were able locate a structural template with PSI-BLAST.[3]

In this report, we present the results of our assessment of the models submitted to the CASP6 comparative modeling section. We assessed predictions for model backbone quality and structural alignment accuracy and were able to compare the predictions in relation to those of the previous CASP. In addition we compared the best predic-

*Correspondence to: Michael Tress, Protein Design Group, CNB–CSIC, Cantoblanco, Madrid, Spain. E-mail: mtress@cnb.uam.es

Received 31 May 2005; Accepted 21 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20720

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

tions to models based on the best possible structural template. We also carried out a thorough analysis of side-chain rotamer prediction, evaluated side-chain contacts and made some assessment of structurally divergent regions and biologically important sites. In addition we investigated the ability of predictors to build models that improve on the structural template and reached some tentative conclusions from comparisons with the previous CASP experiment. The results allow us to propose where comparative modeling might be improved and to make some recommendations for the future of the CASP comparative modeling assessment itself.

RESULTS

The pre-processing of targets and predictions required to allow the smooth running of a CASP assessment is a vital part of the evaluation process. Targets must be processed, categorized and split into domains, and the best structural template must be found for each target before the numerical evaluation of the predictions can begin. This assessment was greatly facilitated by the group at the Lawrence Livermore Prediction Center, who processed the targets and all the model structures, calculated the data for GDT-TS, RMSD, AL0 and more, maintained a web page accessible only to the assessors and provided excellent and timely support at all times.

Targets

There were 66 target proteins evaluated in CASP6. These were split into 95 target domains by the assessors as described in the domain/categorization article.[4] All subsequent comparisons were based on domains rather than whole target proteins because many of the target proteins were split into domains that fell into more than one category and also because of the difficulty of domain orientation prediction.

Of the targets, 47 domains (from 36 different target proteins) were categorized as comparative modeling targets (see Table I). The categorization of domains as comparative modeling targets was based wholly on whether a template structure could be found by BLAST[2] or PSI-BLAST searches of the Protein Databank (PDB[5]). The precise details are outlined in this issue.[4] The one exception to this rule was domain T0237_3, which was found by BLAST but was deemed to be too difficult to model due to the fact that a large part of its 3D structure was determined by interactions with other domains.

Domains in the Comparative Modeling section were further categorized into "easy" and "hard" targets depending on the level of similarity with known structures. "Easy" CM targets were those domains where it was possible to find a structural template with a simple BLAST search, while "Hard" CM targets were the remainder, those for which the structural template could only be found with PSI-BLAST.

Unlike previous CASP experiments where some targets were evaluated as both fold recognition and comparative modeling targets, no target domain overlapped between the comparative modeling and fold recognition categories in this CASP. As a result the CASP6 CM section had fewer difficult, remotely homologous targets. This was a decision of the assessors and did not reflect the reality of the actual prediction strategies employed by the predictors. Although the targets in this section were the only targets evaluated as comparative modeling targets it was clear that many, if not most, of the better prediction groups were using comparative modeling techniques as part of a strategy to build structures for any target with a detectable template structure, including most of the structures in the fold recognition category. While the results for these harder targets were generally less good than those in the CM category, the fact that modeling techniques were used across the board emphasises the overlap between the Comparative Modeling and Fold Recognition (Homology) sections.

Within the 47 target domains in this section there were a number of anomalies that required special attention. Two of the target proteins (T0270 and T0261) were Fold Recognition (Analogy) or New Fold targets at the opening of the prediction season, but in both cases the late deposition of a similar template structure in the PDB meant that they were converted to "easy" CM targets before predictions closed. Many predictors, particularly server-based predictors, were not able to respond in time and this was reflected in the results. Target T0240 was another odd case, a protein that had an identical sequence to another template, but had a different fold. This was another target that fooled the servers. Target T0226 had two domains that were sufficiently similar to cause problems with the sequence-independent evaluation of the alignment accuracy. The Comparative Modeling assessors made the decision to omit four domains from the comparisons between groups, target T0261 (domains 1 and 2), target T0270 and target T0226_2. However, the predictions for targets T0261 and T0270 were used in the side-chain and structurally divergent region (SDR) prediction assessments in order to increase the pool of available targets.

Numerical Evaluations

As in CASP5 two parameters were used to evaluate the quality of the predictions. The structural similarity between the models and the target structures was measured with GDT-TS, a measure of model backbone quality, and structural alignment accuracy was measured with AL0, the percent of correctly aligned residues. The precise details of the AL0 and GDT-TS calculations are detailed elsewhere in this issue.[6] GDT-TS is reliable as a single measure of the quality of protein backbone prediction for comparative modeling predictions where the model structures are not too dissimilar to the target structure. The parameters were chosen for consistency with CASP5 and to allow clear, simple and easily understandable comparisons, both between groups and between CASPs. Predictions were also evaluated by the root-mean-square deviation (RMSD) of the C-α atom, although this was not used in the group comparisons.

Structurally divergent regions and side chains were also analyzed, but only over a subset of sufficiently high quality

TABLE I. Target Domain Breakdown†

Target	Target residues	Category	LGA best template	Sequence identity	LGA structural score	RMSD	Backbone coverage (%)	Target difficulty ranking
T0196	89	HARD	1skqA	29.21	85.5	1.71	89.89	39.67
T0199_1	74	HARD	1fnnB	12.16	85.7	2.21	97.3	46.33
T0200	255	HARD	1hpuD	10.2	50	2.85	82.35	81.5
T0204	297	EASY	1guqC	24.24	83.4	2.18	94.61	43.33
T0205	103	HARD	1h0yA	19.42	76.5	1.64	82.52	62
T0208	344	HARD	1xljB	6.4	42.7	4.08	77.33	90
T0211	136	HARD	1eut	20.59	84.8	1.76	92.65	48.5
T0222_1	264	HARD	1rzmA	12.5	71.2	2.77	90.91	64.33
T0223_1	114	HARD	1kqbB	14.15	82	2.24	99.12	44.5
T0226_1	182	HARD	1mosA	10.44	59	2.45	85.16	76
T0226_2	*95*	*HARD*	*1jxaB*	*16.84*	*61.2*	*2.69*	*98.95*	*47*
T0229_1	24	EASY	1ml8A	29.17	79.8	0.83	100	12.5
T0229_2	102	EASY	1ml8A	34.31	84.9	1.85	93.14	36
T0231	137	EASY	1v6fA	78.83	95.2	1.76	100	4.83
T0232_1	81	HARD	1k0bC	17.28	86.5	2.41	97.53	37
T0232_2	138	HARD	1q4jA	10.14	71.1	2.69	81.16	78.67
T0233_1	66	EASY	1o17C	31.82	97.4	0.98	100	11.5
T0233_2	265	EASY	1v8gA	43.02	91.5	1.69	96.6	15.33
T0234	135	HARD	1g76A	13.33	63.3	2.8	89.63	66.33
T0235_1	309	EASY	1nbfA	24.92	75.8	2.49	91.59	47
T0240	90	HARD	1ihrAB	65.56	57.8	2.14	70	56.33
T0244	296	EASY	1lvwA	16.89	68.7	2.04	79.05	73
T0246	354	EASY	1cnzB	56.78	94.4	1.36	99.72	8.67
T0247_1	150	EASY	1pj7A	24	88.2	1.63	91.33	44.5
T0247_2	135	EASY	1pj6A	24.44	86.9	1.87	94.07	36.33
T0247_3	76	EASY	1pj6A	19.74	86.4	2	94.74	40
T0261_1	*212*	*EASY*	*1vkyB*	*54.25*	*92.2*	*1.76*	*96.23*	*16.33*
T0261_2	*73*	*EASY*	*1vkyB*	*21.92*	*74.7*	*2.4*	*84.93*	*59.67*
T0264_1	116	EASY	1vhvA	37.93	88.1	1.71	94.83	25.67
T0264_2	173	HARD	1vhvB	19.08	57.7	2.34	77.46	71.17
T0265	102	HARD	1ku9B	21.57	67.9	2.84	82.35	65.17
T0266	150	EASY	1dbxB	23.33	87.5	1.78	98.67	31.67
T0267	174	HARD	1j4jB	16.67	76.1	3.23	91.38	57.33
T0268_1	172	EASY	1m6yA	47.09	93.4	1.28	97.67	15
T0268_2	109	EASY	1n2xA	51.38	91	1.31	97.25	15.83
T0269_1	158	EASY	1prxB	37.97	89.3	1.72	95.57	24
T0269_2	61	HARD	1prxB	14.75	55.8	3.69	80.33	73.33
T0270	*247*	*EASY*	*1t0tW*	*53.04*	*92.5*	*1.48*	*95.95*	*16.67*
T0271	161	EASY	1rlhA	35.4	80	1.69	86.96	43.33
T0274	156	EASY	1i0rA	21.79	87.2	1.53	92.95	42.33
T0275	135	EASY	1mjhA	27.41	74.5	2.17	96.3	35.33
T0276	168	EASY	1sbqB	23.81	82.5	1.72	91.07	50
T0277	117	EASY	1jogD	28.21	88.6	1.7	100	19.17
T0279_1	127	HARD	1jr2A	18.11	64.2	2.71	93.7	52
T0279_2	121	HARD	1jr2A	9.92	62.2	2.49	96.69	59
T0280_1	113	EASY	1bd3B	20.35	89.3	1.8	99.12	31.5
T0282	323	EASY	2cevB	18.89	71.4	2.4	81.11	67.67

Group numbers are in italics.
†Breakdown of the 47 domains categorized as comparative modeling targets. Target difficulty ranking is explained elsewhere in the text and is a combination of backbone coverage, sequence identity and template model GDT-TS. Domains in italics were not used in comparisons of model quality.

models as detailed below. Side chains were evaluated by the percentage of correct ($\chi 1$ and $\chi 2$) rotamer angles where correct was defined as within 15° or within 30° of the experimentally determined value. Models were also assessed for the coverage and accuracy of their side-chain contacts. A limited range of SDRs were evaluated by local and global backbone RMSD.

Evaluation Details

While many groups made predictions for all models, some groups built models for just some of the targets, and while some groups predicted up to five models, other groups only predicted a single model for each domain. Here all comparisons between methods were adjusted to take

account of the differences in the number of predictions submitted and where possible were based on predictions for common domains.

The scoring scheme used in this evaluation was the same as the one used in CASP5. The scoring calculation was twofold to take account of the fact that not every group predicted every domain. Groups were compared using first models only. For groups that submitted refined and unrefined models only the refined first model was evaluated and where groups submitted predictions for domains in the form of several fragments only the highest scoring fragment was evaluated. There were 4533 first models assessed in the CM section.

z-Scores of each of the predictions were calculated for AL0 and for GDT-TS in order to smooth over as much as possible differences in the scores due to the varying difficulty of the targets. z-Scores were calculated as described below.

Models that were worse than two standard deviations below the average for each target were removed from the calculations. The mean and standard deviation were then recalculated for the distribution of the remainder of the predictions and z-scores for each prediction were calculated from these parameters. All models that were worse than average (including those models eliminated from the calculation) were reassigned a z-score of 0. This meant that models that were of average or worse quality would not over-penalize the overall mean of the groups z-scores—with the idea of not over-penalizing less tested and riskier methods.

This comparison is valid for the majority of groups that predicted all targets. But to make the comparison valid for groups that predicted less targets, groups were also compared head-to-head over common subsets of predicted targets. The results of these head-to-head comparisons for GDT-TS and AL0 are reported as statistical tables.

To complement the results for each group we also calculated the average and standard deviation of the GDT and AL0 z-scores for each target (figure not shown). The results show that the z-scores for targets T0226_2, T0261_1, T0261_2, and T0270 have by far the highest standard deviation and that target T0226_2 in particular would have biased the final rankings, essentially backing up our decision to remove these targets from the group comparisons.

Cα–Cα Clashes

Ideally the assessors would like to be able to evaluate as many models as possible by visual inspection. However, this is no longer feasible with so many predictions. Fortunately this year the CASP support group at the Lawrence Livermore provided basic information on Cα–Cα clashes, so it was fairly easy to spot many physically impossible models. The quantity of poor models submitted to the CM section was alarming. Clearly unfeasible models were widespread and submitted not just by automatic servers but also "human" groups, even though a visual inspection would have sufficed to have detected the problem. Not all of the impossible models were low scoring. Indeed three of the top models for target T0235_1 were penalized for clashes. In total 73 groups had at least one model that qualified for penalization, though in some cases the errors may have originated from groups cherry-picking the "best" predictions from servers that were generating physically impossible models.

We made a decision to penalize those models that were clearly unacceptable. The decision was made because we felt that models in the CM assessment should be of use to the biological community, and that models with obvious knots and clashes would be rejected out of hand and might actually discredit the server. Models produced by CM techniques should provide users with models of sufficient quality to act as a guide to experiments and to help to interpret results. If end users doubt the quality of the some models produced by a server or human group they are not likely to trust more reliable results from the same source in the future.Those models with greater than 50 bumps (where the Cα–Cα distances were between 1.9 Å and 3.6 Å) or that had more than four severe clashes (Cα–Cα distances of less than 1.9 Å) were penalized. The choice of cut-offs was rather arbitrary, but also fairly generous. We checked a selection of 1000 chains from the PDB and found just one chain with more than 16 minor clashes. Penalized models were inspected manually and those that contained visible backbone–backbone clashes or were that were otherwise clearly unfeasible [Fig. 1(a,b)] had both their AL0 and GDT-TS z-scores set to 0. In total, 55 first models were penalized in this way. The remainder had their GDT-TS z-score reduced to the mean z-score for that target, while their AL0 score was unaffected. In total, 155 first models were penalized.

One group in particular (TS176) produced models that almost always had a very high number of clashes. One model had 325 Cα–Cα clashes, a number of which were severe. Despite this they were among the highest scoring groups. It was clear that the models resembled the correct folds but that they were compressed in certain regions of the backbone. Calculations showed that compression was most common in loop regions and at the N- and C-terminals, and that this also had a detrimental effect on their prediction of side-chain rotamers. The compression clearly gave the group an advantage when the GDT-TS scores were calculated and as such we felt we were quite correct in penalizing the models. Since this is a new method we hope that the problem of compression is only temporary.

Assessment of the Overall Model Quality

The final AL0 and GDT-TS scores for each group were the mean of the z-scores over all the submitted models. Table II shows the mean overall GDT-TS, AL0 and RMS measures for each group along with the mean of the z-scores and the mean of the combined GDT-TS and AL0 z-scores. The final results are very similar to the previous CASP. Four groups (450 – Ginalski, 591 – Venclovas, 21 – Bujnicki-Kolinski and 176 – Skolnick) produced, on average, models of higher overall quality based on the overall

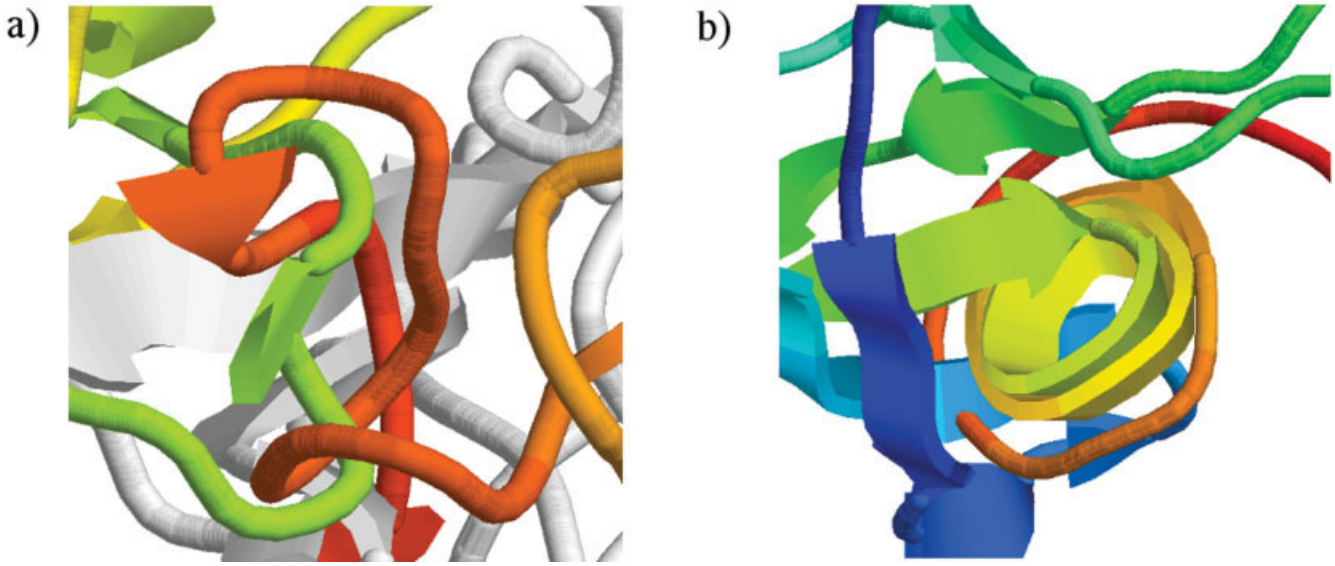

Fig. 1. Impossible models. Close up details of two models submitted in the comparative modeling section: (**a**) the common tangle, (**b**) a β whorl. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

z-scores. Three of these four groups also stood out in CASP5.

Three of the four groups submitted models for all 43 targets. Group 591 modeled just 28 of the domains, but the head-to-head comparisons over these 28 domains do not affect the order of any of the top four groups. In addition the same four groups stand out when the targets are divided into Easy targets and Hard targets. There are minor differences in the order of the top-scoring groups if you plot the separate GDT-TS (Fig. 2) and AL0 (Fig. 3) z-scores. In particular, group 176 is not among the top groups in GDT-TS because so many models were penalized for compression.

We also compared the GDT-TS scores of each pair of groups directly against each other by performing a paired t-test on the GDT-TS and AL0 values for the common subset of predicted targets. The results for the comparison between the 25 highest scoring groups (combined AL0 and GDT-TS z-scores) are shown in Table III. Note that this head-to-head comparison was made with the actual AL0 and GDT-TS values and prior to penalizing models with clashes. Again the results are remarkably similar to those of CASP5. The top three groups can be distinguished from the rest, while a fourth-placed group (021) is distinguishable from all but 035, 367, 319, and 588.

The best-performing groups used much the same techniques with similar success as in previous years. The top three groups all considered a range of templates and alignments from a range of sources before proceeding with the modeling step, though all did develop some new techniques to improve their predictions. Ginalski used Meta-BASIC,[7] a novel meta profile alignment method, Venclovas employed ISS-BLAST with the aim of improving alignments and Bujnicki-Kolinski introduced a high resolution CABS lattice. Only Skolnick developed an entirely new procedure based on the PROSPECTOR server,[8] tertiary structure assembly with fragments and lattice Monte Carlo simulation. As some of these methods were developed with the Fold Recognition/New Fold categories in mind, their success in Comparative Modeling was perhaps a bit of a surprise.

Comparisons With CASP5

In order to make comparisons between predictions made for the CASP5 and CASP6 Comparative Modeling sections we needed to rank the targets by difficulty. As noted by previous assessors pairwise sequence identity between target and template is not an effective parameter for describing the difficulty of a target. One of the most important factors influencing the ability to predict accurate models is the extent of structural conservation between target and template.[9]

The target difficulty rankings we used were based on the standard target difficulty scales used in CASP.[6] The difficulty score used here was a combination of the rankings of three separate scores. The first two, sequence identity between target and parent and the percentage of the target that was covered by the template structure, were taken from the LGA[10] sequence-independent alignment between the experimental structure and its closest parent. The third score was a measure of structural similarity between target and template, the GDT-TS of the template model. The template model was constructed from the LGA structural alignment by mapping the target residues onto the structure of the template. This extra ranking applies well to CM and FR targets, but would not be suitable for new fold targets due to the difference in template and target structures.

These three scores were converted into ranks and the difficulty rankings for the 97 Comparative Modeling targets in CASP5 and CASP6 was calculated as follows:

$$[RANK(SeqId) + RANK(\%\ target\ backbone\ covered) + RANK(template\ model\ GDT\text{-}TS)] \,/\, 3$$

TABLE II. Detailed Results by Group†

Group	Number of predictions	Mean GDT-TS	Mean AL0	Mean RMS	Mean AL0 z-score	Mean GDT-TS z-score	Combined AL0 and GDT-TS z-score
450	43	73.54	73.71	4.09	1.02	1.12	2.14
591	28	73.24	*76.16*	3.86	1.03	1.03	2.06
021	43	70.71	*70.89*	4.28	0.78	0.84	1.62
176	43	71.91	*72.78*	4.02	0.93	0.54	1.47
035	43	69.57	*68.72*	4.35	0.67	0.76	1.42
319	40	69.27	*70.09*	4.28	0.70	0.67	1.37
100	*43*	67.46	67.33	6.13	0.66	*0.66*	*1.32*
367	42	69.26	*68.58*	5.08	0.64	0.68	1.32
454	43	68.48	68.92	4.98	0.65	0.66	1.31
272	43	67.72	66.03	5.02	0.59	0.68	1.27
561	43	68	*70.04*	4.65	0.72	0.54	1.25
003	43	67.95	68.34	4.98	0.61	0.60	1.21
506	43	68.04	69.35	4.21	0.58	0.61	1.19
242	43	68.03	67.45	5.88	0.57	0.61	1.18
588	16	71.77	*72.48*	3.13	0.57	0.59	1.16
109	43	67.9	68.82	4.26	0.55	0.59	1.14
166	43	67.14	67.13	5.26	0.58	0.55	1.13
680	37	67.1	*68.3*	4.24	0.56	0.56	1.12
501	42	68.29	*69.28*	4.87	0.57	0.55	1.12
210	43	66.5	67.98	4.36	0.61	0.51	1.12
007	38	70.4	*71.27*	4.28	0.56	0.55	1.11
207	43	67.44	67.47	6.05	0.55	0.55	1.10
042	41	67.28	*67.7*	4.77	0.57	0.53	1.10
237	43	67.77	67.84	3.96	0.54	0.54	1.08
579	42	66.47	*66.41*	6.26	0.54	0.54	1.07
096	43	66.15	66.31	5.53	0.51	0.55	1.06
026	42	66.39	66.16	5.5	0.53	0.54	1.06
400	43	67.07	67.25	6.33	0.50	0.56	1.06
390	9	76.97	79.3	2.31	0.52	0.53	1.05
067	18	68.54	68.99	4.81	0.50	0.51	1.01
604	43	65.62	64.38	11.38	0.51	0.48	0.99
504	43	66.51	67.27	4.29	0.52	0.46	0.98
217	29	72.13	73.91	3.32	0.48	0.50	0.98
101	43	65.36	63.99	6.3	0.51	0.46	0.97
394	43	65.6	66.57	5.29	0.52	0.44	0.96
164	41	63.47	65.59	3.88	0.51	0.43	0.95
314	28	62.56	63.65	4.83	0.51	0.43	0.95
223	43	68.2	69.07	4.88	0.52	0.42	0.94
344	39	64.81	64.21	6	0.46	0.48	0.94
157	43	66.69	65.93	5.03	0.45	0.48	0.94
232	33	63.33	65.35	4.51	0.48	0.43	0.91
152	5	58.68	58.39	6.87	0.48	0.42	0.90
079	43	64.3	63.51	6.47	0.46	0.44	0.89
418	31	61.92	61.89	5.16	0.49	0.40	0.89
612	39	64.64	63.31	5.43	0.44	0.45	0.89
513	42	63.11	64.76	5.02	0.47	0.42	0.89
011	43	63.21	61.63	4.93	0.46	0.41	0.86
573	43	67.56	68.68	5	0.47	0.39	0.86
009	43	65.28	63.83	5.22	0.40	0.45	0.85
030	43	58.98	55.06	6.4	0.43	0.43	0.85
376	43	64.11	64	5.63	0.41	0.42	0.84
375	43	62.84	61.48	6.35	0.41	0.42	0.84
033	43	61.54	62.99	5.73	0.44	0.39	0.83
113	43	63.41	63.22	5.23	0.38	0.44	0.83
607	31	64.5	63.63	5.92	0.40	0.42	0.82
137	32	64.53	62.36	5.9	0.39	0.43	0.82
163	36	64.09	65.3	4.52	0.43	0.38	0.81
308	36	57.59	57.36	6.02	0.41	0.40	0.81
156	29	70.21	70.99	3.88	0.41	0.40	0.81
199	40	65.39	65.4	4.42	0.42	0.39	0.80
451	42	63.86	61.99	5.89	0.37	0.44	0.80
213	43	66.57	66.71	4.7	0.44	0.36	0.79
197	43	63.55	61.29	5.86	0.35	0.44	0.79
185	43	62.75	64	6.08	0.38	0.40	0.78
139	38	64.23	65.4	4.63	0.43	0.35	0.78

TABLE II. (Continued)

Group	Number of predictions	Mean GDT-TS	Mean AL0	Mean RMS	Mean AL0 z-score	Mean GDT-TS z-score	Combined AL0 and GDT-TS z-score
267	20	67.33	67.07	5.43	0.40	0.37	0.77
060	42	58.37	53.98	7.09	0.39	0.38	0.77
126	43	61.71	60.29	7.53	0.37	0.39	0.76
046	20	61.62	60.61	5.73	0.38	0.37	0.75
331	41	62.32	61.17	6.08	0.38	0.36	0.74
290	39	63	62.64	5.15	0.40	0.34	0.74
231	43	61.86	61.4	6.25	0.36	0.36	0.72
160	40	63.18	63.91	5.74	0.39	0.32	0.71
398	42	62.75	61.88	5.37	0.36	0.35	0.71
352	43	60	59.86	5.73	0.40	0.30	0.70
530	43	61.06	59.25	6.63	0.36	0.34	0.70
112	43	62.07	64.64	4.79	0.35	0.32	0.67
111	41	57.39	58.67	5.98	0.35	0.31	0.66
014	23	68.13	68.98	3.68	0.35	0.31	0.66
110	40	57.39	60.27	6.23	0.34	0.32	0.66
051	43	61.65	60.68	6.42	0.34	0.32	0.66
381	42	57.16	54.7	6	0.34	0.31	0.66
105	17	59.9	54.67	5.21	0.34	0.31	0.65
243	43	60.01	61.37	6.21	0.35	0.30	0.65
229	43	60.62	60.73	6.88	0.34	0.30	0.64
287	41	59.96	62.42	6.27	0.36	0.27	0.63
550	43	62.98	61.98	6.55	0.29	0.34	0.63
631	42	58.56	57.06	7.92	0.30	0.32	0.62
186	42	58.81	59.56	6.58	0.31	0.31	0.62
596	41	63.14	62.15	5.21	0.32	0.29	0.61
289	41	61.2	60.72	5.14	0.30	0.29	0.59
439	9	66.68	69.79	4.29	0.27	0.31	0.58
032	16	57.14	56.71	3.64	0.30	0.25	0.55
629	43	60.81	59.67	5.83	0.27	0.27	0.53
338	*43*	57.12	60.08	6.97	0.26	*0.27*	*0.52*
475	34	52.04	55.31	3.46	0.28	0.23	0.51
052	43	58.58	56.78	7.59	0.26	0.24	0.50
384	41	55.38	56.11	6.09	0.25	0.24	0.49
230	43	55.08	56.01	7.75	0.25	0.21	0.47
244	40	54.87	52.88	7.44	0.23	0.23	0.46
324	42	56.94	53.96	6.98	0.21	0.24	0.45
261	41	61.29	60.53	4.98	0.26	0.18	0.44
263	33	56.15	55.52	7.94	0.26	0.17	0.42
283	40	63.7	63.53	5.68	0.26	0.15	0.42
482	38	60.58	62.6	5.88	0.27	0.14	0.41
122	43	53.47	54.63	8.73	0.22	0.17	0.40
081	41	54.79	54.37	6.54	0.18	0.22	0.39
383	38	55.66	53.91	7.46	0.20	0.19	0.39
304	41	57.43	56.66	6.59	0.18	0.20	0.38
393	26	56.36	53.07	5.81	0.21	0.16	0.37
514	39	51.26	49.46	7.57	0.19	0.18	0.36
420	27	57.88	56.59	7.06	0.18	0.18	0.36
532	29	56.26	53.88	6.7	0.16	0.19	0.35
428	19	49.77	46.45	7.45	0.19	0.15	0.34
305	43	56.14	49.13	6.51	0.14	0.18	0.31
013	27	48.14	49.47	9.39	0.15	0.15	0.31
406	42	39.55	36.22	14.59	0.16	0.14	0.31
307	7	61.39	57.89	5.83	0.14	0.16	0.30
472	42	50.57	46.82	8.74	0.15	0.14	0.29
114	43	40.39	38.59	29.5	0.13	0.16	0.29
092	35	50.89	52.65	8.06	0.13	0.14	0.27
478	40	58.73	58.49	4.98	0.15	0.12	0.27
061	41	50.86	46.72	7.69	0.14	0.11	0.25
094	36	51.84	49.61	8.41	0.14	0.10	0.24
512	36	45.36	36.82	8.26	0.11	0.11	0.22
106	33	48.81	51.91	5.78	0.12	0.07	0.19
245	39	51.11	44.13	7.96	0.08	0.11	0.19
291	34	35.16	41.83	6.26	0.12	0.06	0.18
519	35	39.2	37.08	10.04	0.08	0.07	0.15
247	40	51.99	46.88	7.52	0.06	0.08	0.14

TABLE II. (Continued)

Group	Number of predictions	Mean GDT-TS	Mean AL0	Mean RMS	Mean AL0 z-score	Mean GDT-TS z-score	Combined AL0 and GDT-TS z-score
431	26	51.2	46.61	7.83	0.09	0.04	0.12
552	42	35.79	29.74	11.37	0.07	0.04	0.11
089	42	39.7	28.9	9.89	0.06	0.03	0.09
656	12	40.77	25.53	6.03	0.03	0.06	0.09
606	43	50.28	45.09	6.87	0.05	0.01	0.06
321	32	19.48	4.21	15.5	0.00	0.01	0.01
320	33	28.13	19.61	10.63	0.00	0.01	0.01
356	16	44.58	43.53	13.23	0.00	0.00	0.00
444	5	17.62	0.25	21.6	0.00	0.00	0.00
104	5	17.55	0	15.72	0.00	0.00	0.00
057	5	18.87	7.66	11.38	0.00	0.00	0.00
031	7	23.78	12.17	10.12	0.00	0.00	0.00
402	7	23.56	4.15	14.08	0.00	0.00	0.00
276	9	17.06	6.19	13.3	0.00	0.00	0.00
599	42	18.29	2.48	15.39	0.00	0.00	0.00
018	43	20.46	2.59	15.4	0.00	0.00	0.00
348	43	21.84	3.18	13.63	0.00	0.00	0.00
019	43	20.13	2.33	15.23	0.00	0.00	0.00
545	10	20.67	0	19.27	0.00	0.00	0.00
064	11	20.63	1.59	15.42	0.00	0.00	0.00
027	12	24.33	8.62	10.19	0.00	0.00	0.00
049	41	15.92	0.27	31.27	0.00	0.00	0

†The calculation of z-scores for AL0 and GDT-TS is explained in detail in the text. Groups are sorted by combined AL0 and GDT-TS z-Score.

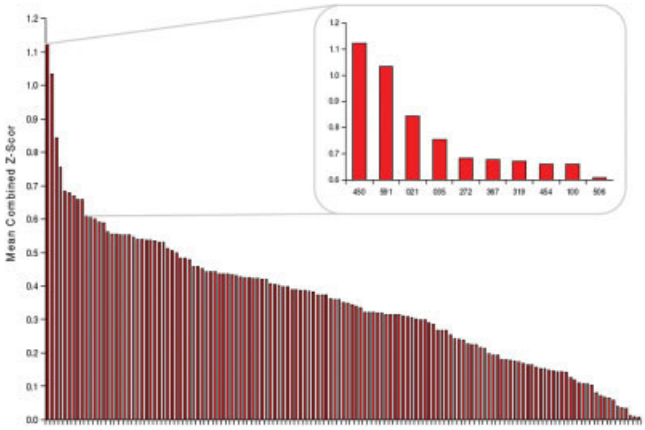

Fig. 2. Rankings by GDT-TS z-scores. The mean GDT-TS z-scores for all groups with the 10 best-scoring groups as amplified in the inset. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

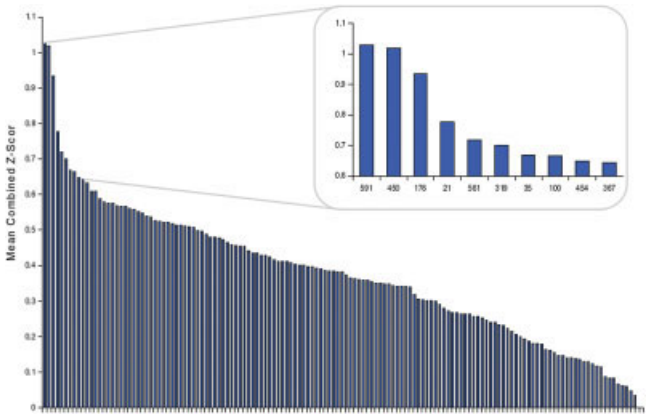

Fig. 3. Rankings by AL0 z-scores. The mean AL0 z-scores for all groups with the 10 best-scoring groups as amplified in the inset. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

One thing that was clear from the difficulty ranking is that since there was no overlap between Comparative Modeling and Fold Recognition targets this year, CASP6 has many fewer difficult targets than CASP5 (figure not shown).

For the comparison between CASP5 and CASP6 we plotted the best-scoring predictions and the best 20 predictions for each target against the difficulty rankings. The mean of the best 20 was used rather than the mean of all models since many groups produced models for all targets regardless of categorization. This results in a number of very poor models that would bias the mean. The CASP5–CASP6 comparison was made for both GDT-TS and AL0 (Fig. 4).

The plots show that there has been little or no improvement in prediction in the last two years. While Figure 4 suggests that there may have been some slight improvement in AL0, there is no hint of an improvement in GDT-TS between CASP5 and CASP6. In fact GDT-TS can also be plotted solely against template model GDT-TS (Fig. 5) and here the results confirm that there is no difference in model quality GDT-TS between CASP5 and CASP6. The comparison also shows clearly that the quality of the best models is most dependent on the similarity between target and template structures, something that has been suggested in previous CASP experiments.

There was a suggestion at the CASP meeting in Gaeta that the cut-offs used to calculate GDT-TS (1 Å, 2 Å, 4 Å,

TABLE III. Statistical Significance[†]

	X591	X450	X176	X021	X319	X035	X007	X588	X561	X454	X367	X506	X100	X109	X501	X003	X680	X242	X210	X166	X042	X207	X272	X237	X579
X591	NA	0.96	0.15	**0.01**	**0**	**0.01**	**0**	**0.01**	**0.01**	0	**0.01**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
X450	0.6	NA	0.56	0.09	**0**	0.06	**0.02**	**0**	0.17	**0.01**	**0**	**0**	**0**	**0**	**0**	0.01	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
X176	0.26	0.11	NA	0.13	**0.02**	**0.01**	**0**	**0.02**	0.11	**0.01**	**0**	**0**	**0.01**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
X021	**0.01**	**0**	0.16	NA	0.16	0.3	0.13	0.87	0.93	0.17	0.14	0.09	0.1	**0.01**	0.17	0.06	0.05	**0.03**	0.07	0.08	**0.03**	**0.04**	**0.04**	0.1	**0.03**
X319	**0**	**0**	**0**	0.08	NA	0.84	0.74	0.25	0.36	0.84	0.61	0.98	0.62	0.69	0.98	0.79	0.38	0.62	0.68	0.33	0.8	0.44	0.3	0.27	0.19
X035	**0.02**	**0**	**0.01**	0.31	0.43	NA	0.94	0.6	0.35	0.93	0.4	0.88	0.52	0.91	0.53	0.78	0.45	0.48	0.72	0.47	0.1	0.47	0.36	0.68	0.39
X007	**0**	**0**	**0**	**0**	0.65	0.28	NA	0.57	0.19	0.39	0.78	0.53	0.78	0.85	0.45	0.57	0.33	0.51	0.65	0.94	0.44	0.64	0.61	0.83	0.36
X588	**0.01**	**0**	**0.05**	0.41	0.21	0.13	0.49	NA	0.67	0.73	0.24	0.75	0.74	0.98	0.91	0.38	0.78	0.69	0.23	0.86	0.97	0.93	0.87	0.87	0.56
X561	**0**	**0**	**0**	**0.01**	0.77	0.21	0.61	0.34	NA	0.23	0.24	0.26	0.12	0.07	0.45	0.07	0.16	**0.03**	0.12	0.08	0.05	**0.03**	**0.04**	0.11	0.06
X454	**0**	**0**	**0**	**0.04**	0.84	0.42	0.17	1	0.64	NA	0.77	0.85	0.48	0.48	0.77	0.69	0.17	0.25	0.43	0.38	0.51	0.22	0.11	0.53	0.19
X367	**0.01**	**0**	**0**	0.15	0.22	0.22	0.18	0.06	0.31	0.37	NA	0.53	0.69	0.41	0.81	0.81	0.19	0.62	0.73	0.53	0.58	0.66	0.18	0.69	0.38
X506	**0**	**0**	**0**	**0.03**	0.87	0.25	0.72	0.93	0.96	0.66	0.19	NA	0.54	0.62	0.83	0.77	0.18	0.61	0.62	0.31	0.94	0.46	0.28	0.42	0.21
X100	**0**	**0**	**0**	**0.03**	0.66	0.16	0.99	0.85	0.71	0.53	0.32	0.65	NA	0.74	0.64	0.58	0.31	0.96	0.76	0.89	0.63	0.95	0.63	0.76	0.83
X109	**0**	**0**	**0**	**0**	0.77	0.12	0.94	0.54	0.9	0.5	0.2	0.86	0.71	NA	0.42	0.94	0.33	0.83	0.88	0.53	0.83	0.6	0.36	0.62	0.31
X501	**0**	**0**	**0**	**0.03**	0.79	0.22	0.81	0.65	0.82	0.88	0.38	0.95	0.92	0.81	NA	0.41	0.24	0.22	0.38	0.24	0.2	0.2	0.16	0.38	0.1
X003	**0**	**0**	**0**	**0.01**	0.77	**0.02**	0.67	0.53	0.97	0.64	0.5	0.94	0.69	0.94	0.8	NA	0.42	0.45	0.78	0.5	0.33	0.49	0.33	0.77	0.39
X680	**0**	**0**	**0**	**0.01**	0.35	0.05	0.16	0.65	0.39	0.12	**0.05**	0.12	0.27	0.17	0.34	0.27	NA	0.47	0.32	0.57	0.49	0.75	1	0.74	0.96
X242	**0**	**0**	**0**	**0.02**	0.87	0.13	0.97	0.71	0.97	0.6	0.31	0.99	0.69	0.87	0.72	0.92	0.21	NA	0.7	0.87	0.73	0.98	0.52	0.84	0.65
X210	**0**	**0**	**0**	**0**	0.22	**0.03**	0.17	**0.05**	0.24	**0.04**	**0.01**	0.11	0.56	0.12	0.12	0.2	0.45	0.1	NA	0.67	0.85	0.67	0.35	0.94	0.44
X166	**0**	**0**	**0**	**0.01**	0.27	0.1	0.98	0.69	0.52	0.35	0.13	0.38	0.75	0.52	0.62	0.55	0.54	0.51	0.68	NA	0.84	0.87	0.65	0.65	0.79
X042	**0**	**0**	**0**	**0**	0.44	**0**	0.3	0.44	0.35	0.29	0.09	0.47	0.46	0.34	0.12	0.14	0.63	0.19	0.53	0.82	NA	0.97	0.48	0.87	0.7
X207	**0**	**0**	**0**	**0.01**	0.51	**0.03**	0.7	0.73	0.59	0.24	0.07	0.57	0.99	0.59	0.33	0.52	0.47	0.25	0.29	0.83	0.59	NA	0.49	0.83	0.62
X272	**0**	**0**	**0**	**0.03**	0.73	0.26	0.89	0.68	0.83	0.47	0.2	0.8	0.87	0.87	0.64	0.86	0.38	0.81	0.36	0.71	0.66	0.82	NA	0.4	0.69
X237	**0**	**0**	**0**	**0.01**	0.31	0.17	0.83	0.71	0.84	0.55	0.19	0.8	0.77	0.89	0.99	0.88	0.42	0.84	0.34	0.54	0.62	0.77	0.97	NA	0.47
X579	**0**	**0**	**0**	**0**	0.24	0.05	0.43	0.52	0.39	0.18	**0.04**	0.25	0.7	0.32	0.11	0.41	0.74	0.31	0.93	0.74	0.81	0.56	0.56	0.43	NA

[†]Statistical comparisons between the top-scoring 25 groups. Comparisons in the top right of the table were for AL0, comparisons on the bottom left were for GDT-TS. Comparisons were made over common targets and reported as *P*-values from the *t*-tests. Significant differences between groups (*P*-values of less than 0.05) are highlighted.

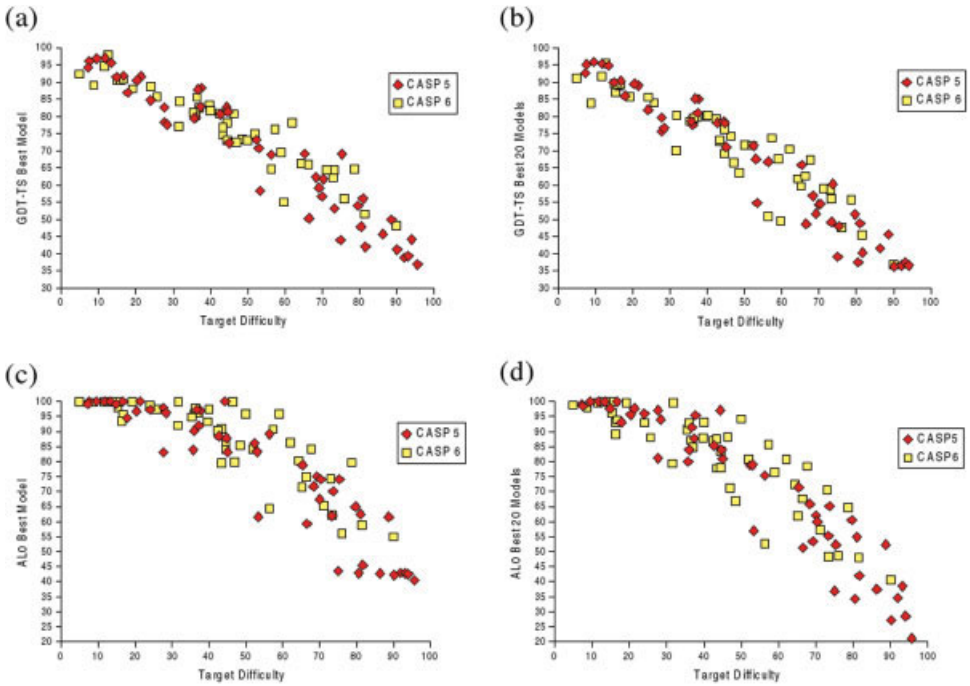

Fig. 4. CASP5 and CASP6 comparisons. Raw GDT-TS and AL0 scores are plotted against the difficulty ranking of each target as explained in the text: (**a**) the GDT-TS of the best predictor for each target, (**b**) the mean GDT-TS of the best 20 predictors for each target, (**c**) the AL0 of the best predictor for each target, (**d**) the mean AL0 of the best 20 predictors for each target. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

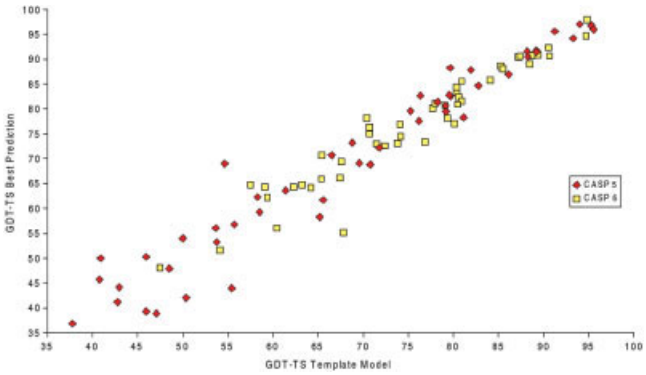

Fig. 5. CASP5 and CASP6 GDT-TS comparison. A comparison of the best-scoring GDT-TS predictions in CASPs 5 and 6. This time for each CM target in the CASP5 and CASP6 experiments, the best-scoring prediction was plotted against a ranking generated from the raw GDT-TS of the model that was built from the parent structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

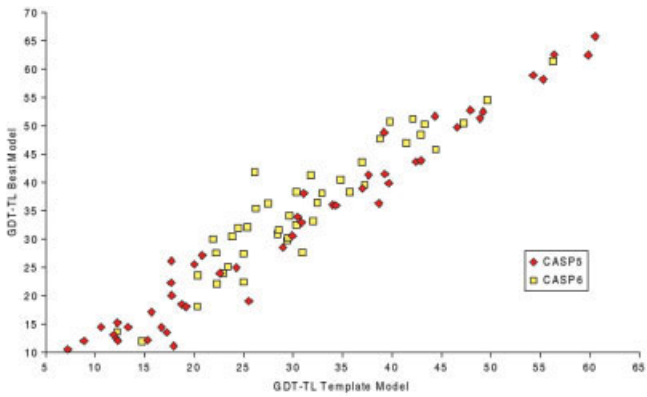

Fig. 6. CASP5 and CASP6 GDT-TL versus template model GDT-TL. Comparisons between the best-scoring CASP5 and six predictions for GDT-TL. For each CM target in the CASP5 and CASP6 experiments the GDT-TL for the highest scoring prediction is plotted against the GDT-TL of the model built from the parent structure. GDT-TL calculation is explained in the text. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

and 8 Å) might not be able to detect the very small improvements in backbone quality that may have occurred between CASP5 and CASP6. We experimented with new GDT combinations to see if we could find higher resolution improvements in model quality since the previous CASP and came up with cut-offs of 0.25 Å, 0.5 Å, 1 Å, and 2 Å. The comparison between CASP5 and CASP6 using the new GDT combination (GDT-TL) is shown in Figure 6. The GDT-TL scores of the best predictions for each target are plotted against the GDT-TL scores of the template model. The figure suggests that there has at least been some small improvement in fine-grained model quality between CASP5 and CASP6.

Template Model Comparisons

In previous CASP experiments it has been demonstrated that predictors are rarely able to predict models that are closer to the target structure than the structure of the closest template. Previous comparisons between tem-

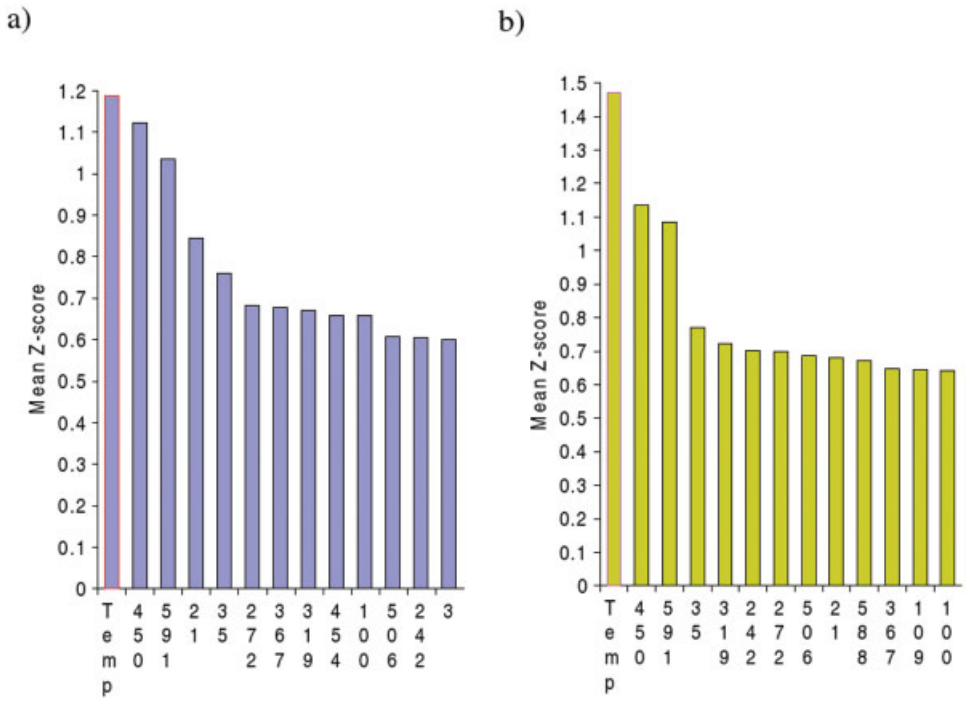

Fig. 7. Predictors versus template moel quality. **a**: The mean of the GDT-TS z-scores for the best groups in comparison with the mean GDT-TS of the template model (*Temp*). **b**: The mean of the GDT-TL z-scores for the best groups in comparison with the mean GDT-TL of the template model (*Temp*). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

plate models and predictors were based uniquely on RMSD. Here we have used GDT-TS for the comparisons. While comparing predictions to templates by RMSD gives a slight advantage to the template models over the predictors because templates do not rebuild gapped regions, comparisons with GDT-TS give the predictors a slight advantage since predictors predict the full model and can often improve on template model GDT-TS simply by closing small loops. Despite the advantage that the predictors have over the template, Figure 7(a) shows that if a group had submitted the perfectly aligned best template without any refinement, they would have had the best overall GDT-TS z-score. Interestingly, while the differences between the best-scoring groups and the template are not so high when comparing with GDT-TS, they are much more significant when comparisons are made with GDT-TL [Fig. 7(b)].

While no group scored better overall than the template model, for 33 out of the 43 targets at least one group had a better GDT-TS than the model built from the template and in head-to-head decisions group 450 had a better GDT-TS than the template model for 23 out of 42 target domains (one was a tie). In Figure 8(a) the GDT-TS for the template model, the best prediction and the best ten predictions for each target are plotted against target difficulty ranking for each of the original 47 target domains. It is clear that not only does the best predictor improve on the template model on most occasions, but also there are a number of targets for which the best 10 predictors also had better GDT-TS than the template model. The effect is less noticeable with AL0. Figure 8(b) shows that the predictor rarely beat the AL0 of the template model—the best predictor has a better AL0 for 11 of 47 domains, while the best 10 predictors better the template model only four occasions.

Figure 8(a) also demonstrates that the best predictors tended to improve on the template models for the easier targets. This trend can also be seen in Figure 9, in which the targets were grouped into three similar-sized sets based on their difficulty rankings. Here the mean template model GDT-TS is compared with the mean GDT-TS for the best 12 groups. It is clear that when the targets are easier more predictors are closer to the template model. Many groups have mean scores close to the template model GDT-TS when the targets are easier, while only two groups (450 and 591) are able to get close to the template model GDT-TS for the harder targets.

With the easier targets the templates are usually clear, the alignments are simpler and the biggest differences between target and template are usually contained in short divergent regions with fixed end points that are close in sequence and relatively easy to close. The goal for the easier targets should be to improve on the structural template and this seems to be quite possible for many targets, at least under the GDT-TS scoring scheme.

Making a prediction that is better than the best template is a much more difficult task for the harder domains. Given current methodologies the ideal goal would be to predict models that approach the accuracy of the template model, but for the harder targets errors in both template and alignment are hampering efforts. There were several

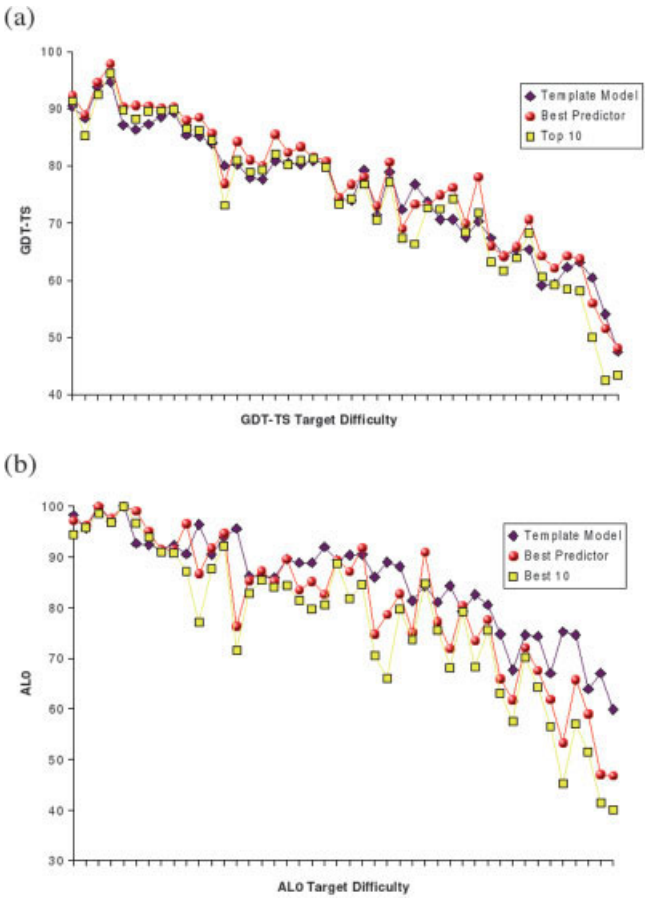

Fig. 8. Target by target comparison of predictors and template models. This plot shows the results for the template model, the best predictor and the mean of the best 10 predictors for each CM target. Results are plotted against target difficulty ranking, easiest targets first. **a**: the results for GDT-TS, **b**: the same plot for AL0, this time target difficulty is calculated using template model AL0 ranks instead of GDT-TS ranks. In both graphs the results for domains T0240, T0226_2 and T0261_2 have been left out for the sake of clarity.

harder targets where the template model GDT-TS was much better than any of the predictors.

For the harder targets the evolutionary distance between target and template is much greater and this makes the alignment more difficult. However, while it would seem that choosing the template should be the easier task, there are pitfalls even here. We have been able to check the list of templates used by the predictors in the "PARENT" line of the models and the results seem to bear out the fact that template choice is not always correct, even in the easier cases.

There are several examples where only one or very few suitable templates exist yet half the templates used by the predictors are not suitable. One extreme example is target T0276, where the best structural templates (by some distance) come from PDB structure 1sbq, yet only 26 of the approximately 350 templates used by the predictors come from 1sbq. While some of the less suitable templates may have been used in fragment-based assembly, the sheer quantity of less suitable templates suggests that, at least for some targets, alignment errors are not the only thing holding back the prediction of models.

In any case these results in no way disagree with the comparisons that were carried out in the CASP5 CM assessment, the apparent differences stem from the scoring scheme used to compare templates and models. When we used GDT-TS to compare template models and predictors for the CASP5 targets we found similar results.

We also compared predictor and template model RMSD for each target domain in CASP6. In CASP6 predictors were only able to improve on the RMSD of the template models for 11 of the targets. Eight of these 11 targets were the easiest eight targets in the target difficulty rankings, showing that predictors can improve on the best template, but generally only with the easier models.

In the few harder targets where predictors were able to improve on the RMSD of the best template, the improvements were not systematic. For example, the biggest improvement over the template model was 0.6 Å by two groups (176 and 530) for target T0267 (one of the harder targets). In this target there is a 22-residue loop that is angled differently in the experimental structure and the template and these two groups placed the loop closer to the target than to the template. As a comparison the biggest improvement in RMSD by any group over the template model in CASP5 was 0.4 Å.

CM Issues—Structurally Divergent Regions and Side Chains

Previous CASP experiments have suggested that the prediction of SDR and side chains can only be successful when the model backbone is accurately predicted. In the past this quality threshold has meant that there were too few side chains and SDR regions for meaningful comparisons. However, this year the CASP organizers provided easier targets in order to get around this problem. In addition two targets (T0261_1 and T0270) turned into easy CM targets during the competition and increased the pool of targets available for SDR and side-chain evaluation.

SDR regions and side chains had to fit the following criteria. The structure of the target must have been determined by X-ray crystallography. If the original structure had more than one chain, the chains should not differ by more than 0.8 Å RMSD. Target residues must have a B-factor of less than 40, and not be involved in crystal packing. The models analyzed had to have a GDT-TS value of greater than 80.

In addition SDR regions had to be at least three residues long and the target Cα residues had to diverge from the template chain by at least 2.5 Å. In addition, a loop was rejected if it could be shown that another close template could be used to recreate the SDR.

The SDR regions satisfying the above criteria included 133 residues from just 20 loops, whereas 1878 nonglycine and nonalanine residues from 20 domains were available for the side-chain analysis.

Structurally Divergent Regions

There were few conclusions that could be drawn from the SDR regions, mostly because there were so few regions

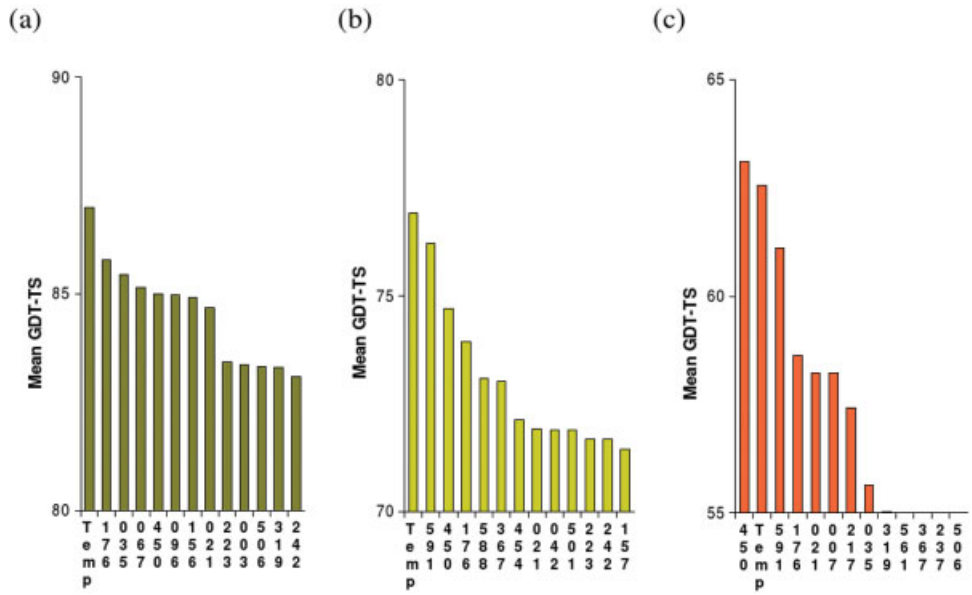

Fig. 9. Results for easier, medium, and harder targets. The mean GDT-TS for the template model (*Temp*) and for the top predictors with the targets split into three sub-groups, easier (**a**), middle (**b**), and harder (**c**) targets. Targets were categorized according to the target difficulty ranking. Minimun of six targets predicted in each subset. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

to compare. In comparison to CASP5 there were actually less SDR regions, despite the increase in easier targets, and on average the SDR regions were smaller this time. In general, as in CASP5, the larger the SDR—the greater the local and global RMS of the SDR. However, the low number of SDRs included in the comparison meant that there was little significant difference between most pairs of groups after t-tests were conducted on common sets of models (data not shown).

While in CASP5 there were no loops of greater than five residues with RMSD of less than 1.0 Å, groups 367, 067, 450, and 272 all managed this task and most groups produced very low local RMSDs for the SDR in T0231 because the loop retained the same form despite being angled differently. However, these are just anecdotes, there were too few loops in CASP5 and CASP6 to draw any firm conclusions.

Side-Chain Evaluation Details

The standard criteria for measuring the correctness of side-chain placement is the accuracy of rotamers. A rotamer is usually deemed to be accurate if it is within 30° or 40° of the same rotamer in the native structure. We also used rotamer accuracy to make comparisons between the groups, not only at 30° and 40° but also at 15°.

In addition we wanted to introduce a measure of global correctness of side chains, so we also evaluated the residue–residue contacts made by the subset of good side chains. For the side-chain contact assessment we calculated all the inter-atomic distances in the native structures and in the models. These inter-atomic distances were measured as the distances between the van der Waals radii of the heavy atoms in the protein, with the van der Waals distances defined as follows: C = 1.548 Å, O = 1.348 Å, N = 1.4 Å, S = 1.808 Å. A side chain was assumed to be in contact if any of the heavy atoms in the side chain were within 1 Å of another side chain or the backbone of a residue that was separated by at least six residues in sequence. Side-chain–side-chain contacts were counted only once and backbone to side-chain contacts were ignored.

We evaluated side-chain contacts quality using a GDT-like scoring system. For each of the target domains we recorded all native contacts between heavy atoms below 1 Å. We scored side chains that made the same residue–residue contacts. Side-chain contacts scored one point for being between 0.125 Å further apart and 0.0675 Å closer than the equivalent contact in the native structure, another point if the equivalent contact in the model was between 0.25 Å further apart and 0.125 Å closer than the native contact, a third point if the equivalent contact in the model was between 0.5 Å further apart and 0.25 Å closer than the native contact and another point if the equivalent contact in the model was between 1 Å further apart and 0.5 Å closer than the native contact. The cut-offs were tighter when the predicted contact was closer than the native contact because we observed that heavy atom contacts closer than 0.5 Å were minimized in the native proteins. All points gained for each model were divided by the number of native contacts for that target and multiplied by 25 to give a score between 0 and 100.

For both rotamers and contacts we report the overall percentage accuracy. We also compared groups head-to-head over common targets. From this we calculated the number of head-to-head wins between groups that had at least five targets in common and calculated the significance of each head-to-head comparison using *t*-tests. We

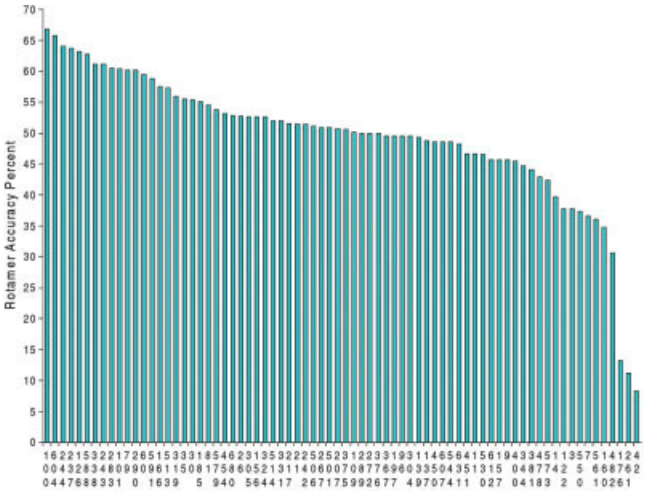

Fig. 10. Mean accuracy of $\chi 1$ predictions. Comparison of the mean accuracy of the predicted rotmers for each group. Group must have predicted rotamers for at least five targets. Here the standard of correctness is 15° and comparison is only made for $\chi 1$ rotamers. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

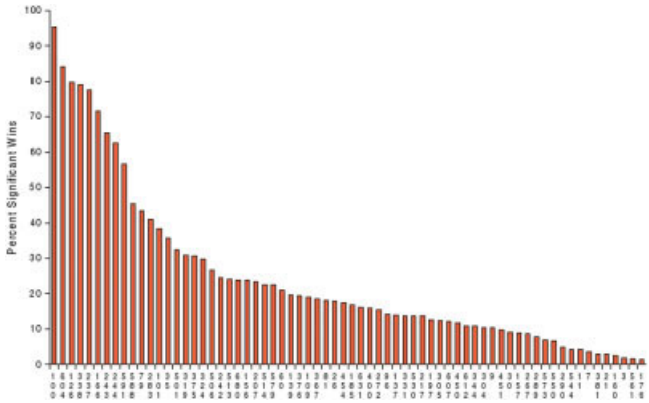

Fig. 11. Percentage of signifiant wins against other groups over common targets. The percentage of statistically significant wins for each group from head-to-head comparisons of rotamer predictions over common tagets. Here the standard of correctness is 15° and comparison is only made for $\chi 1$ rotomers. The significance of the wins in head-to-head comparisons is calculated from standard paired *t*-tests and over a minimum of five common targets. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

were able to report the head-to-head comparisons as win percentages and also as percentage of significant wins.

Rotamer Accuracy

We compared the accuracy of $\chi 1$ and of $\chi 1$ and $\chi 2$ rotamers combined at standards of correctness of 15°, 30°, and 40°. The results were similar no matter which parameters were chosen, though not surprisingly as the number of allowed degrees increased the differences between the groups became blurred. Although the number of side chains involved in the side comparison was still rather small, there were enough differences between groups for the assessors to draw some tentative conclusions.

For standards of correctness of 15° and 30° we calculated overall percentage accuracy and the win percentages from head-to-head comparisons between pairs of groups with common targets. The mean accuracy at 15° with $\chi 1$ rotamers only is shown in Figure 10. One group tops virtually all these comparisons, group 100 (Baker). Group 100 also tops significant wins when rotamer accuracy is measured with both $\chi 1$ and $\chi 2$, and at 30 degrees, though the results here are slightly less conclusive.

We also performed t-tests on the results from the head-to-head comparisons over common targets in order to calculate the number of statistically significant wins. The percentage of statistically significant wins for each groups against all other groups with at least five common targets are shown in Figure 11. At 15°, group 100 is significantly better at rotamer prediction than all but four groups: 604, 591, 290, and 338—of these groups 290 and 591 only have five targets in common with group 100. Group 604 comes from the same source as 100.

We split the results at 15° into subsets based on the categorization of the residue side chains as surface (greater than 40% accessible surface area) or core (less than 20% accessible surface area) and as "changed" or "conserved" in relation to the target–template structural alignment. Here group 100 has a big lead on the rest of the groups except for conserved surface residues (Fig. 12) where it lags behind group 243. Group 243 conserves the rotamers of identical side chains, showing the importance of conserving identical rotamers, even on the surface. It is noticeable that the prediction of surface rotamers is worse than the prediction of buried rotamers.

The best methods at rotamer prediction are Baker (100), Baker-Robetta-04 (604), CASPITA-Fox (338), Sternberg (237), Karplus (166), nanoModel (243), nanoFold (244), Pan (126), and Venclovas (591) because these are the groups that come up the most often in the top five groups no matter how side-chain rotamers are evaluated (at 15°, 30°, or 40°, with just $\chi 1$ angles, or with $\chi 1$ and $\chi 2$ angles together) or what measurement you use to compare the groups (mean accuracy, wins against other groups, statistically significant wins). It is not so easy to make conclusions about the methods here, although apart from the two Baker groups we know that many of the groups (such as groups 338, 237, 591) that used SCWRL[11] did well.

Side-Chain Contacts Evaluation

Side-chain contacts at a distance of 1 Å were evaluated as described above. The results (Fig. 13) show that the groups that have the best rotamers are not always those that have the best contacts. While many groups that did well in rotamer accuracy were using SCWRL, many of the best side-chain contacts groups are using methods based on packing and energy minimization. Although groups 243 and 244 are among the best-scoring groups in raw scores for both rotamers and contacts, this is partly because their predictions were for easier targets.

Groups 156 (CHEN-WENDY), 319 (honiglab), and 026 (SAMUDRALA) come out on top in significant win percentage when groups are compared over common targets. The proportion of significant wins for the top groups is notice-

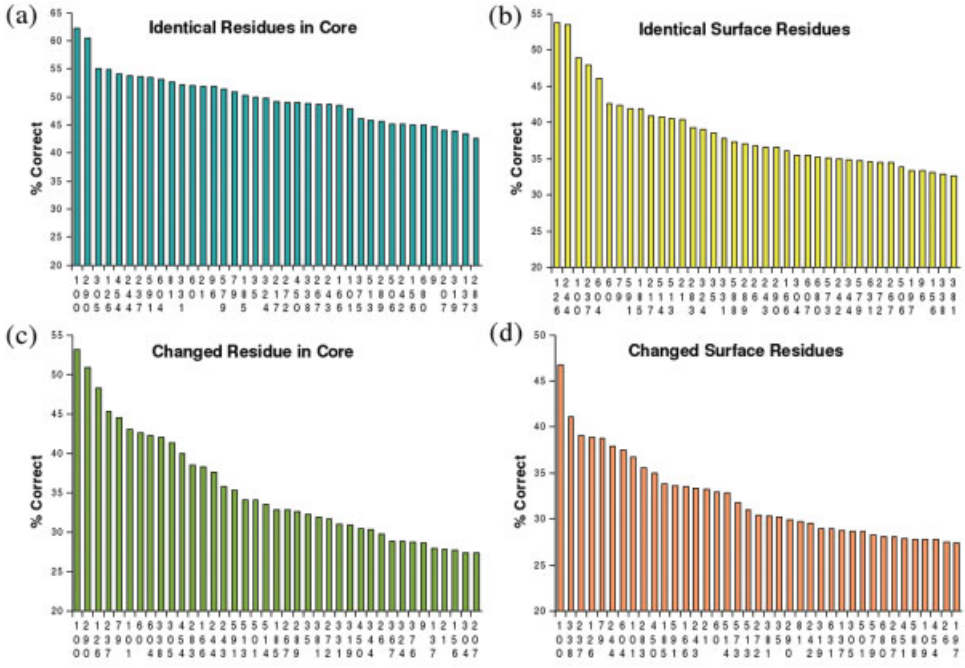

Fig. 12. Breakdown of side-chain rotamer prediction asccuracy. The mean side-chain rotamer accuracy for rotamers from (**a**) conserved side chains in the core of the protein, (**b**) conserved side chains on the surface of the protein, (**c**) nonconserved side chains in the core of the protein, and (**d**) nonconserved side chains on the surface of the protein. Side chains were categorized as surface or core and as identical or changed as in the main text. The standard of correctness is 15° and comparison is only made for $\chi 1$ rotamers. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ably smaller than with rotamers, these three groups are the only groups with significant win percentages above 50%.

Investigations by eye into the models produced by various groups showed that those groups with the best side-chain contact coverage produced models with close side-chain packing distributions that were much more similar to those of the native structures than the methods that had the best rotamer predictions. This demonstrates that good rotamer prediction does not necessarily lead to native-like packing of side chains and that side-chain prediction cannot be unlinked from main-chain prediction.

There was no obvious relationship between model GDT-TS and either rotamer accuracy or side-chain contact values. In fact the two side-chain measures were more related to each other than either was to backbone quality. However, as only models with GDT-TS scores of greater than 80 were included in the comparison, differences in GDT-TS were quite small.

CM Issues—Refinement

There were few attempts to submit both refined and unrefined models. In fact there was only one group that submitted refined and unrefined models for more than just a few targets. For that reason the assessors decided not to assess refinement in this CASP. Refinement is an issue that needs to be rethought for the next CASP competition.

Automatic Versus Human Predictions

The results for the predictions submitted by server groups are shown in Figure 14. Four servers stood out for their predictions over the comparative modeling targets. The server of group 242 (ZHOUSPARKS2), a stand-alone server, performed slightly better than the rest. In the head-to-head t-tests between the best servers, group 242 was significantly better than all but three servers 210 (Eidogen-SFST), 207 (zhousp3), and 400 (ACE) in GDT-TS model quality. Differences in AL0 between the top groups were less obvious.

Inspection of the results shows that 242 was better for the easier CM targets, while group 400 performed slightly better for the harder CM targets. Z-score were also calculated for the server groups alone and here group 242 had a slightly bigger lead. It should be noted that a number of the better performing servers were penalized for having poor-quality models.

Overall the top server had the fourteenth-best combined AL0 and GDT-TS score, showing that the top servers can still beat many of their human counterparts. Indeed, comparisons between performance in CASP5 and CASP6[6] show that server predictions in the CM section have actually improved relative to predictions submitted by human groups, particularly over the easy targets.

Biologically Important Sites

We also assessed the effect of biologically important sites on the quality of the predictions. Previous assessors[12]

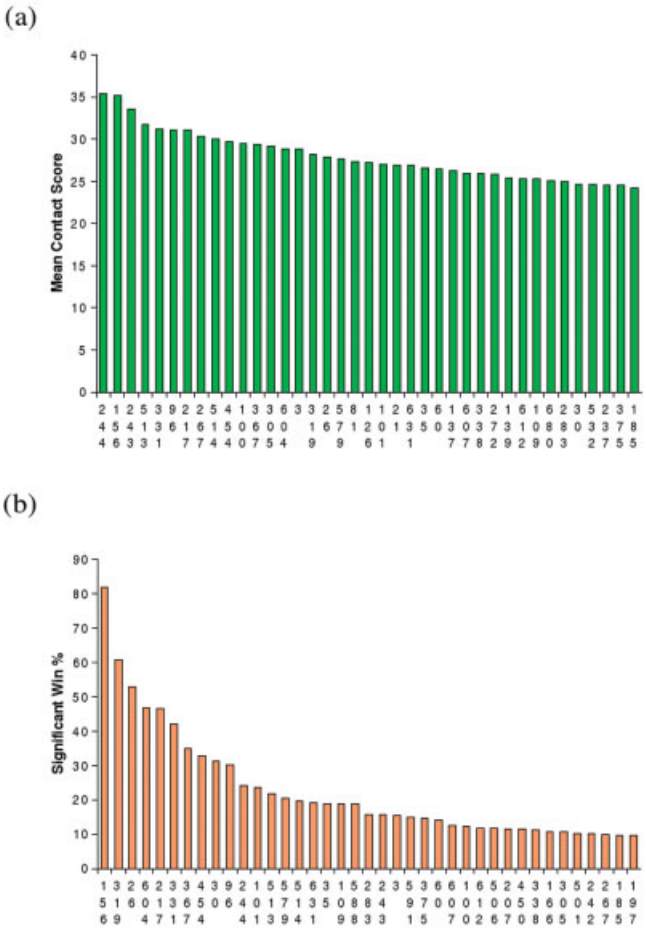

Fig. 13. Comparison of side-chain contact results. Comparison of the mean accuracy and significant win percentage for each group. Groups must have had at least five targets. **a**: mean side-chain contact score as calculated in the text. **b**: statistically significant win percentage between groups based on side-chain contact score. The percentage of significant wins from each group is from head-to-head comparisons of side-chain contacts over at least five common targets and calculated from standard paired t-tests. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

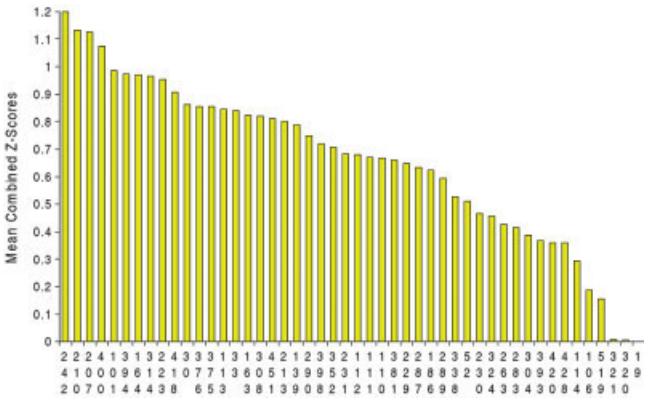

Fig. 14. Server results. Rankings are by combined AL0 and GDT-TS z-scores with the 10 best groups amplified in the inset. Note that target T0240 was removed from the comparison because of the very poor performance of the servers on this target. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

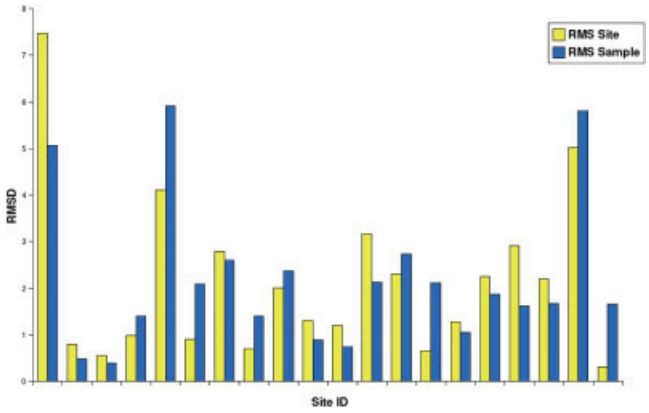

Fig. 15. Prediction of biologically important sites. RMSD of biologically important sites for predicted models and the mean RMSD of samples of the same diameter and number of residues for each of the 20 biologically important sites located. RMSD values are the mean of the best 20 predictors. Biolgically important sites were deduced from information from web pages, from papers, and by homology. Sites deduced by homology were confirmed using the SQUARE alignment reliability server.[13] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

had noted an improvement in prediction at functionally important sites and initial results suggested the same had occurred in CASP6. However, when we adjusted for the difference in size and number of residues of sites and non-sites these differences disappeared.

First we located the known and probable sites. This was not an easy task, since many of the targets came from structural genomics projects and had not been studied in detail. Some functional-site information was found on web pages relating to the structure, some came from papers and some information was deduced by homology.

Once we had found the site information we calculated the diameter of the active-site cluster and the number of residues in the cluster. We used this information to sample populations of residues that fitted the same criteria, taking 40 samples per binding site. We used LGA to calculate RMSD for the functional site residues for the best 20 predictions and did the same for the samples. We compared the local RMSD of the site residues for the best 20 predictions with the mean local RMSD of the sampled residues. The results are shown in Figure 15.

Predictions for the biologically important residues are better than the samples in just 9 of 20 cases. We found that local structural factors (such as the size of the functional site, the quantity and size of SDRs in the structural superposition of template and target alignment) played a much bigger role in the ease of functional site prediction.

CONCLUSIONS

Although there appears to have been very little improvement with respect to CASP5, a small number of groups are approaching the quality of the best templates, something that is true even for the more difficult targets. Still, there are big differences between these groups and many of the rest of the predictors and these differences are statistically significant. In part these differences stem from the meth-

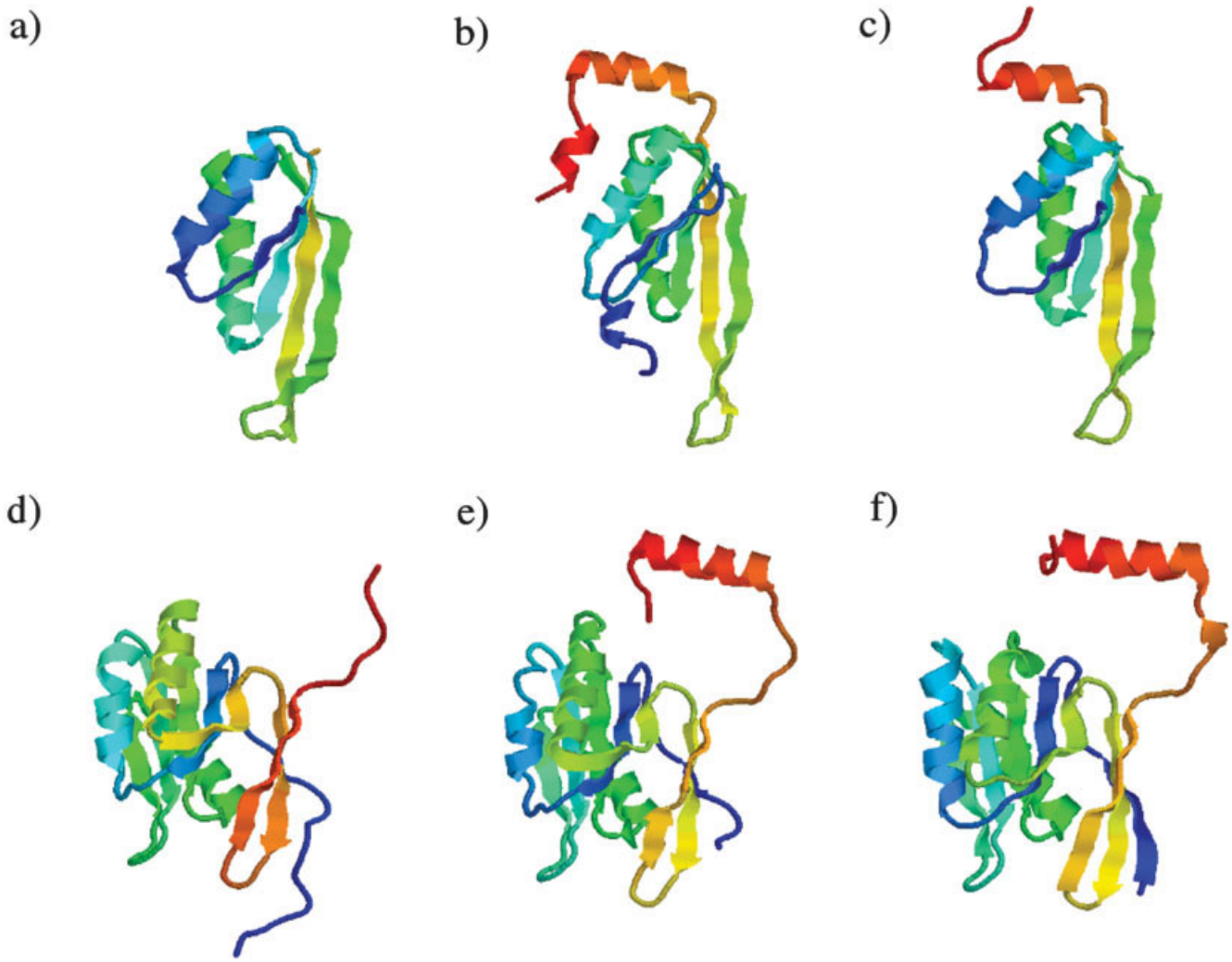

Fig. 16. In both cases, a C-terminal helix was present in the target but not in any of the close templates. The predicting group was able to predict not just the secondary structural type, but also the orientation of the helix with respect to the rest of the protein. **a**: the best template for target T0205, 1hy0A. **b**: the structure of the target T0205 with C-terminal helix. **c**: the prediction from group 450. **d**: the best template for tsrget T0271, 1rlhA. **e**: the structure of the target T0271 with C-terminal helix. **f**: the prediction from group 035.

ods used by these groups to select and align templates, in part the differences are due to expert knowledge.

For most of the harder comparative modeling, targets predictors are doing as well as could be expected, given the distance between the experimental structure and its parent (see Fig. 16). Meanwhile for those targets that have parents that are structurally very similar, predictions are as good as or almost as good as the structural templates used to build them. Despite this, as the plot of GDT-TS versus template model GDT-TS (Fig. 5) shows clearly, predictions are very close to the parent structures and still far from the target structure.

The techniques used by many groups to predict comparative modeling targets were often same as those used for many of the fold recognition targets. The consensus method of the better human predictors for most targets in the comparative modeling and fold recognition categories seemed to be the detection of templates through 3D-Jury,[14] followed by some form of alignment improvement before the alignments were modeled directly, most often with MODELLER.[15] This overlap between the two sections was reflected in the overall results where the same groups came out top in both the Fold Recognition (Homology) and Comparative Modeling categories.

Given that the same methods are being used in both categories we would suggest that the next CASP experiment might merge the Comparative Modeling and Fold Recognition categories into a single "template-based modeling" section, leaving the easiest targets with clear structural parents and trivial alignments for an "easy" comparative modeling category. Comparative Modeling assessors will be able to concentrate on evaluating SDRs, side chains, and refinement and in this way the field may be pushed towards developing new and more effective refinement techniques.

Two of the most interesting developments to come out of this year's CASP were the results of the side-chain evaluation and the assessment of the prediction of biologically

important sites. There were statistically measurable differences between groups in rotamer selection with one group in particular standing out. However, the evaluation of side-chain contacts shows that rotamer accuracy is not the only way side chains can be evaluated. The side-chain rotamer and contact evaluation results do not complement each other—some of the better groups in side-chain contacts have relatively poor rotamer accuracy and some of the models with the best rotamers are not so well packed. The suggestion that side-chain packing improvements can only come at the expense of rotamer accuracy shows that side-chain prediction is closely tied to overall structure prediction and that improvements will require advances in refinement techniques. While it was possible to compare side-chain predictions between groups, it was not at all clear whether there had been any improvement in side-chain prediction in relation to CASP5.

The results for prediction of functional sites is somewhat surprising. While predictions are better for the functional sites in nine of 20 cases, there are just as many cases where the predictions for the functional sites are comparatively worse than the sampled residues. Visual inspection of the sites suggested that local structural factors relevant to each of the target structures, the similarity to the templates and the sites themselves play a much bigger role than conservation in the ease of prediction of biologically important regions.

Comparisons with CASP 5

Comparisons with the comparative modeling section of CASP5 show that there has been little change in the quality of comparative models in the last two years. The improvements we have noted in model quality and alignment accuracy are almost imperceptible and almost certainly not statistically significant. The analysis of prediction performance by Venclovas et al.[9] after CASP5 noted that there was no evidence of alignment improvement in comparative modeling since CASP2. We found disappointingly little evidence that the field had moved forward by any discernible amount in CASP6 either. Despite this there are signs that there may have been some small improvements in model quality at a much finer level than is currently measured.

While the CASP5 assessors found that at least the performance difference between methods was levelling out, we could find little evidence for this in CASP6. As in CASP5 the results of the top groups are still statistically different from the remaining groups (Table III). Comparisons of the top 20 predictors for each target in CASP5 and CASP6 confirmed that there was little improvement among the top twenty predictors either (Fig. 4).

Results from this CASP leave no doubt that there is still room for improvement in all areas of modeling. Previous assessors have noted that although there has been progress in the field since the inception of CASP, many major bottlenecks still exist in comparative modeling including template selection, alignment, modeling of the core (such as the prediction of β-bulges), as well as refinement and the modeling of side chains and SDRs. These bottlenecks still exist and it is not clear that they will be solved by the increasing size of sequence and structural databases.

The quantity of structural information available in the databases is a natural limit for template-based approaches and the best predictors seem to have already reached that limit in CASP5, at least for the easier targets.[16] The increase in the rate of structure deposition to the PDB from structural genomics initiatives therefore is welcome and should lead to more and higher quality structural parents for modeling because of the improved sampling of structural space. However, there was little evidence in CASP6 that this increase in the available structures has actually improved the quality of predictions.

The growth of protein sequence databases parallels the growth of the structural databases. More sequences should mean that evolutionary relationships are easy to detect and that alignments improve,[17] although research in this laboratory suggest that this assumption does not always hold true (results not shown). In fact, particularly for the easiest targets evaluated here, the key difficulty is not template detection and alignment but high-quality refinement.

In the long run improved refinement techniques will also be necessary if predictors are to predict models that are closer to the target structures than to the parent structures. Structures with a high sequence similarity often have differences in main-chain conformation of up to several Ångstroms. These changes are brought about by the sort of subtle differences in atomic contacts and packing that were highlighted in the side-chain contact evaluation. These differences will require the introduction of all atom refinement methods.

While refinement was an issue for this CASP, the fact that only one group consistently entered both refined and unrefined structures suggests that refinement is not being taken so seriously under the current CASP format. The current CASP format may not be sufficiently adequate to attract predictors using refinement techniques and answer the question as to whether additional structural improvement is possible. The format may need to be changed for future CASP experiments.

Increased computer power should lead to the development of more sophisticated algorithms for alignment, side-chain packing and refinement and that large-scale automatic benchmarking procedures such as EVA and LiveBench[18,19] should help in the incremental improvement of methods. However, in the long term, nothing short of a revolution in methodology is needed to improve the field beyond its current position.

ACKNOWLEDGMENTS

Thanks to all those at the Livermore, especially Andriy Krysthafovych. We would also like to recognize the assistance of Anna Tramontano and assessors Roland Dunbrack and B.K. Lee in making this assessment process as smooth as possible. This work was funded by grants from BioSapiens (LSHC-CT-2003-505265) and GENEFUN (LSHG-CT-2004-503567).

REFERENCES

1. Chothia C, Lesk A. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;4:823–826.
2. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;Suppl 6:352–368.
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. Tress ML, Chin-Hsien T, Wang L, Ezkurdia I, López G, Valencia A, Lee BK, Dunbrack RL Jr. Domain definition and target classification for CASP6. Proteins 2005;Suppl 7:8–18.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
6. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;Suppl 7;225–236.
7. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L. Detecting distant homology with Meta–BASIC. Nucleic Acids Res 2004; 32:W576–W581.
8. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR: a new approach to threading. Proteins 2001;42:319–331.
9. Venclovas C, Zemla A, Fidelis K, Moult J. Progress in prediction performance over successive CASP experiments. Proteins 2003; Suppl 6:585–595.
10. Zemla A. LGA—a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
11. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph–theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.
12. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;Suppl 5:22–38.
13. Tress ML, Graña O, Valencia A. SQUARE–determining reliable regions in sequence alignments. Bioinformatics 2004;20:974–975.
14. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D–jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
15. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 1993;234:779–815.
16. Contreras-Moreira B, Ezkurdia I, Tress ML, Valencia A. Empirical limits for template-based protein structure prediction: the CASP5 example, FEBS Lett 2005;579:1203–1207.
17. Cozzetto D, Tramontano, A. Relationship between multiple sequence alignments and quality of protein comparative models, Proteins 2005;58:151–157.
18. Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Graña O, Pazos F, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res. 2003;31:3311–3315.
19. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench–1: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.