

## HIGH ACCURACY ASSESSMENT

# Assessment of the protein-structure refinement category in CASP8

Justin L. MacCallum,<sup>1</sup> Lan Hua,<sup>1</sup> Michael J. Schnieders,<sup>2</sup> Vijay S. Pande,<sup>2</sup> Matthew P. Jacobson,<sup>1</sup> and Ken A. Dill<sup>1\*</sup>

<sup>1</sup>Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California

<sup>2</sup>Department of Chemistry, Stanford University, Palo Alto, California

### ABSTRACT

Here, we summarize the assessment of protein structure refinement in CASP8. Twenty-four groups refined a total of 12 target proteins. Averaging over all groups and all proteins, there was no net improvement over the original starting models. However, there are now some individual research groups who consistently do improve protein structures relative to a starting starting model. We compare various measures of quality assessment, including (i) standard backbone-based methods, (ii) new methods from the Richardson group, and (iii) ensemble-based methods for comparing experimental structures, such as NMR NOE violations and the suitability of the predicted models to serve as templates for molecular replacement. On the whole, there is a general correlation among various measures. However, there are interesting differences. Sometimes a structure that is in better agreement with the experimental data is judged to be slightly worse by GDT-TS. This suggests that for comparing protein structures that are already quite close to the native, it may be preferable to use ensemble-based experimentally derived measures of quality, in addition to single-structure-based methods such as GDT-TS.

Proteins 2009; 00:000–000.  
© 2009 Wiley-Liss, Inc.

**Key words:** structure prediction; refinement; CASP.

### INTRODUCTION

The refinement category is a new addition to CASP. In this category, a predictor is given two types of information at the beginning: (1) as usual, the amino acid sequence of the protein, and (2) a starting model that has been pre-generated and judged by CASP organizers to be among the very best predicted by available methods. The task of the predictor is then to attempt to further improve the structure. A similar experiment, called CASPR, was conducted between CASP7 and CASP8. The idea of this category is to see if there are endgame strategies that could add value beyond existing structure-prediction methods. One issue we faced as assessors is what is the “best” metric to know whether or not a structure has been improved. Because the refinement-category structures are often already fairly close to the native structures, we needed metrics that were sensitive to subtle differences, but it is clear that different measures will tend to reflect different aspects of quality. In recent years, CASP has focused on GDT-TS<sup>1</sup> and related metrics and so GDT is an important part of our evaluation. However, here, we augment GDT with a number of other metrics that probe aspects of the structure besides alpha-carbon geometry.

The task of refinement we assess here differs in a subtle but significant way from that commonly found in the structure prediction literature. In the context of structure prediction, refinement typically involves improving a template structure provided by bioinformatics,<sup>2–16</sup> often by modeling changes in loop regions and sidechain packing. Although results have been somewhat mixed, the best current methods are, on average, able to add value to the single best template.<sup>17</sup> However, here, we are given structures

Conflict of interest: The authors state that Matthew P. Jacobson is a consultant to Schrodinger LLC.  
Grant sponsor: NIH; Grant number: GM34993.

\*Correspondence to: Ken A. Dill, Department of Pharmaceutical Chemistry, University of California San Francisco, 600 16th Street, Box 2240, San Francisco, CA 94158. E-mail: dill@maxwell.ucsf.edu  
Received 19 April 2009; Revised 26 June 2009; Accepted 4 July 2009  
Published online 20 July 2009 in Wiley InterScience (www.interscience.wiley.com).  
DOI: 10.1002/prot.22538

**Table I**  
List of Targets Used in CASP8 Refinement Competition

Target	Starting model	Residues	Starting RMSD (Å)	Starting GDT-TS	Method
TR389	407_2-D1	1–134	2.63	81.3	X-Ray
TR429 <sup>a</sup>	057_1	27–55, 75–175	6.72	47.0	X-Ray
TR432	443_5-D1	1–130	1.65	91.5	X-Ray
TR435	453_1-D1	15–58, 73–148	2.15	74.3	X-Ray
TR453	131_1-D1	5–90	1.40	88.8	X-Ray
TR454	178_1	5–196	3.24	64.3	X-Ray
TR461	253_1-D1	20–176	1.63	87.8	X-Ray
TR462 <sup>a</sup>	198_2	1–74, 77–143	2.54	67.1	NMR
TR464	489_5-D1	18–86	2.94	77.5	NMR
TR469	426_3-D1	1–74, 77–143	2.18	88.9	NMR
TR476 <sup>b</sup>	404_2-D1	2–88	6.85	47.1	NMR
TR488	020_5-D1	1–95	1.43	88.2	X-Ray

<sup>a</sup>These targets were two domain proteins. Most groups submitted a combined structure for the two domains. In this case, analysis was performed on the entire structure. Three groups submitted separate domains for these targets that were analyzed individually.

<sup>b</sup>Starting model did not contain side chains. This target was excluded from all analysis that required side chains (GDC\_SC, MolProbity, MCRS, HBsc).

that have already been once-refined. Predictors are given not the best experimental template, but the best refinement of those templates by other predictors. The question here is whether further value can be added beyond what the best predictions have already provided.

## TARGETS USED FOR REFINEMENT IN CASP8

Table I shows the targets that were offered for refinement in CASP8, along with the source of the starting model and measures of the initial quality of the starting model. The starting models were chosen from the top submissions for each target during the normal CASP competition and represent the best efforts of the structure prediction community. One should, therefore, expect that refining these starting models will be a difficult task.

### Metrics used for evaluation

We chose eight metrics for evaluating the refinement predictions. Two of the metrics are calculated by the prediction center<sup>18</sup>: GDT-TS<sup>1</sup> and GDT-HA,<sup>19</sup> which are the standard alpha-carbon-based metrics used extensively in previous CASPs. To supplement these, we have used six additional metrics calculated by Richardson and co-workers<sup>20</sup>: two of them focus on sidechain geometry: GDC-SC and the fraction of correct rotamers (corRot); two metrics focus on hydrogen bonding: the fraction of correct main-chain hydrogen bonds (HBmc) and the fraction of correct sidechain hydrogen bonds (HBsc); and two metrics focus on steric clashes and bond lengths: MolProbity<sup>21</sup> and the main-chain reality score (MCRS). The calculations for these metrics are described briefly below. For a more detailed description of how these metrics were calculated, refer the article by Richardson and co-workers in this issue.<sup>20</sup>

The MolProbity score<sup>21</sup> is calculated as:

$$\begin{aligned} \text{MOLPROBITY} = & 0.42574 \ln(1 + \text{CLASHSCORE}) \\ & + 0.32996 \ln(1 + \max(0, \text{ROTAOUT} - 1)) \\ & + 0.24979 \ln(1 + \max(0, \text{RAMMAIFFY} - 2)) + 0.5, \end{aligned}$$

where CLASHSCORE is the number of unfavorable steric overlaps  $\geq 0.4$  Å as calculated by Probe,<sup>22</sup> ROTAOUT is the percentage of rotamer conformations classified as outliers and RAMMAIFFY is the percentage of backbone conformations classified as Ramachandran allowed or outlier (in other words, not in the favored Ramachandran regions). Lower scores indicate more physically realistic models.

The main chain reality score, MCRS, is calculated as:

$$\begin{aligned} \text{MCRS} = & 100 - 10 \times \text{SPIKE} - 5 \times \text{RAMAOUT} - 2.5 \\ & \times \text{LENGTHOUT} - 2.5 \times \text{ANGLEOUT}, \end{aligned}$$

where SPIKE is the per-residue average of the “spike” lengths between mainchain atoms as calculated by Probe, which indicate the severity of steric clashes; RAMAOUT is the percentage of backbone conformations classified as Ramachandran outliers, and LENGTHOUT and ANGLEOUT are the percentage of bond lengths and angles that are  $> 4\sigma$  from ideal. The MCRS starts at 100 and decreases with each non-ideality in the structure.

There are two metrics that quantify the correctness of hydrogen bonding in the model: HBmc and HBsc, which probe the mainchain and sidechain hydrogen bonding, respectively. The HBmc metric measures the fraction of mainchain–mainchain hydrogen bonds in the native structure that are correctly reproduced in the model, where a hydrogen bond is defined as donor-acceptor pair that are within van der Waals contact. The HBsc metric measures the fraction of sidechain–sidechain and sidechain–mainchain hydrogen bonds that are correctly reproduced in the model; however, the tolerance for HBsc is slightly increased compared to HBmc and donor-acceptor pairs that are within 0.5 Å of van der Waals contact are treated as hydrogen bonds. Note that neither metric is simply counting total hydrogen bonds; rather, they are counting fraction of correct native hydrogen bonds that are reproduced by the model.

The corRot metric measures the fraction of model sidechain conformations that match the native structure. The procedure for determining if the rotamers match involves assigning a letter for each  $\chi$  angle (t = trans, m = around  $-60^\circ$ , p = around  $+60^\circ$ ) and then combining the individual letters to produce a string that serves as the name of the rotamer. The model conformation is considered a match if the strings for the model and experimental structures are identical—meaning that all  $\chi$  angles for that residue must match. For further details, see Reference 20.

The correctness of the sidechain position was evaluated using the GDC-sc metric (global distance calculation for sidechains). This metric is similar to GDT-TS and is also calculated using the LGA program.<sup>1</sup> While GDT usually uses the alpha carbon positions, GDC-sc uses a single reference atom from each sidechain instead (see Reference 20 for the list of atom-residue pairs). First, the optimal backbone superposition between the model and native structures is calculated. Next, the distance between the position of the reference atom in the model and the native structure is calculated for each residue. Each distance is assigned to a bin  $i$ , depending on the native-model distance, with  $i = 1$ , corresponding to distances  $<0.5 \text{ \AA}$  and  $i = 10$  corresponding to distances  $<5.0 \text{ \AA}$ . The reference atom may be assigned to multiple bins, so if the native-model distance is  $<0.5 \text{ \AA}$ , the reference atom would be assigned to all bins. The GDC-sc value is then calculated as:

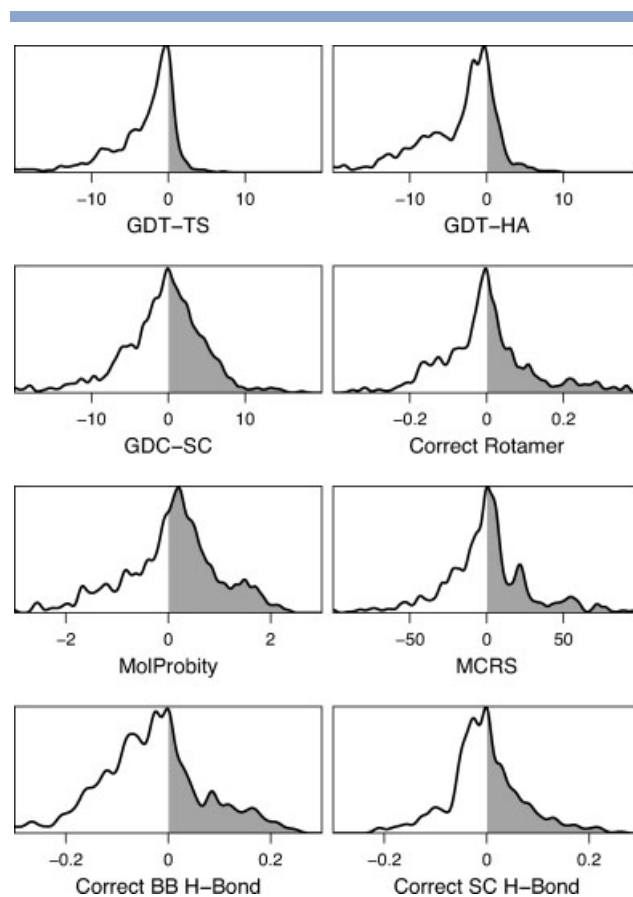
$$\text{GDC-sc} = \frac{200 \sum_{i=1}^{10} (k - i + 1) P_i}{k(k + 1)},$$

where  $k = 10$  is the number of bins and  $P_i$  is the fraction of reference atoms assigned to bin  $i$ . A score of 100 indicates that all reference atoms in the model are within  $0.5 \text{ \AA}$  of their position in the native structure, while a score of zero indicates that all of the reference atoms are  $>5.0 \text{ \AA}$  away from the correct position.

## RESULTS

First, as a baseline reference, we define the null refinement as the procedure that simply returns the starting model without any change. We judge a refinement successful if it scores better than the null refinement. Put another way, methods should *primum non nocere*—first, do no harm.

Figure 1 shows the distribution of improvements relative to the starting model for all predictions from all groups. For most metrics, there are approximately as many improvements as failures and the average values are slightly negative, but near zero. However, for both GDT-TS and GDT-HA, the failures outnumber successes by more than a factor of two. While most of the  $\Delta\text{GDT-TS}$  and  $\Delta\text{GDT-HA}$  values lie within  $\sim\pm 5\%$ , the distribution is skewed to the left – meaning there are more big failures than big improvements. For the other metrics, although the average change is near zero, the range of observed values is quite large; on average the change is small, but there are often big successes and big failures. These results suggest that, on average, one would be better off not refining a structure as the average change in most metrics is close to zero, while GDT-TS and GDT-HA are worse on average. In other words, the entries into the refinement competition, when taken as a whole,



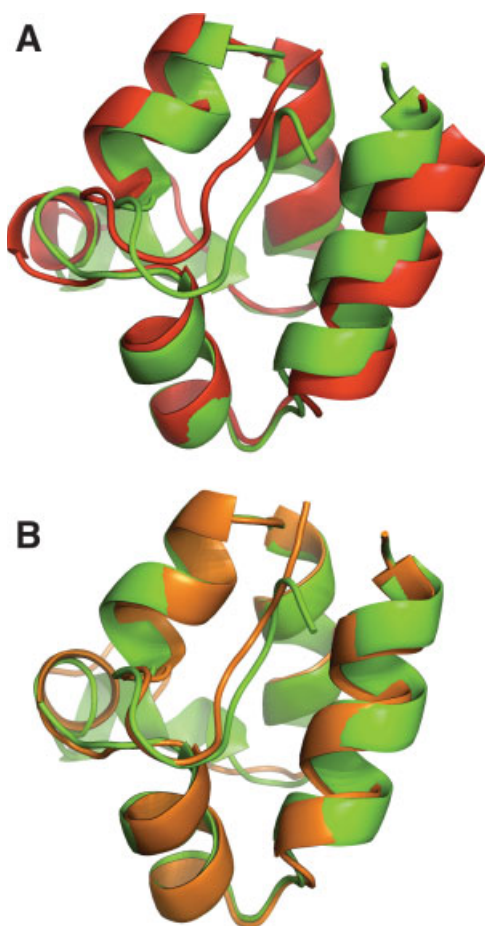
**Figure 1**

Distribution of refinement results across all submitted structures. Distributions are shown for these metrics: GDT-TS, high accuracy GDT (GDT-HA), sidechain GDC (GDC-SC), fraction of correct sidechain rotamer conformations (Correct Rotamer), the additive inverse of the MolProbity score (so that better structures have  $\Delta\text{MolProbity} > 0$ ), main chain reality score (MCRS), fraction of correct backbone hydrogen bonds (Correct BB H-Bonds), and fraction of correct sidechain hydrogen bonds (Correct SC H-Bonds). Values to the right of zero (gray shading) are an improvement over the starting model.

cannot beat the null refinement. Clearly, this analysis ignores the variation between groups and, as we show below, some groups perform much better than others.

There are a number of examples of obviously successful refinement. Figure 2 shows the single largest GDT-TS increase we observed among all groups and targets. Even simple visual examination reveals that the model is much better than the starting model. Aside from obvious visual improvement, this structure also scores higher than the starting model on all eight metrics we consider here. The groups participating in CASP8 were not able to consistently deliver improvements this dramatic, but the presence of results like this suggests that the refinement category is worth pursuing.

To compare the groups with a single score, we have computed Z-scores for each target and metric based on all of the submitted structures. For each metric and tar-



**Figure 2**

The best single improvement in the CASP8 refinement competition was for target TR469. (A) Starting model (red, GDT-TS = 82.94) overlaid on the crystal structure (green). (B) Refined structure (orange, GDT-TS = 90.08) produced by YASARARefine overlaid on crystal structure (green).

get we calculated the mean,  $\mu$ , and standard deviation,  $\sigma$ . Next, we calculated the Z-score for each model. We then repeated this calculation, this time excluding any models with a Z-score of less than  $-2$ ; in other words, we eliminated any negative outliers from the calculation. We did not follow the usual CASP convention of clamping negative Z-scores to zero. For each structure, we then calculated a modified sum of Z-scores where we have multiplied the Z-scores for GDT-TS and GDT-HA by a factor of three. This gives roughly equal weight to the two alpha carbon based metrics and the six other metrics. Although this procedure is arbitrary, we feel this strikes a reasonable balance between the standard CASP metrics and the new ones. For each group we chose the single structure for each target that had the highest modified sum of Z-scores as the best structure.

As shown in Table II, considering only the first model (out of up to five) submitted by each group for each target, there are five groups that are able to perform better

than the null refinement (as judged by the sum of Z-scores): DBAKER, LEE, YASARARefine, FAMSD, and LevittGroup. Of these groups, only LEE is able to improve the GDT-TS and GDT-HA on average, although the declines for the other groups are relatively small. In contrast to the GDT scores, the other metrics are generally improved by the refinement procedures employed by these groups. Clearly, one would judge the average refinement of LEE as a success; the only metric that decreases on average is fraction of correct sidechain hydrogen bonds. However, the situation is less clear for the other groups. We argue that any group that beats the null refinement has been successful; however, this depends heavily on the choice of metrics used and their relative weights. How many GDT units is a correct sidechain rotamer worth? Although there is no correct answer, the relative weights chosen may have a large influence on the relative rankings. For example, if more weight was placed on GDT-TS, then LEE would have scored higher than DBAKER. Thus, not much credence should be placed on the relative ranking of the different top groups.

Each group was allowed to submit up to five models per target. Groups were asked to rank-order their models, with Model 1 being the prediction in which they had the most confidence. We assessed the abilities of groups to rank-order their submissions by calculating the Spearman rank correlation coefficient on both the sum of Z-scores and  $\Delta$ GDT-TS for each model (data not shown). None of the groups was able to rank-order their submissions any better than random (as judged using a permutation test; not shown). As a consequence of the inability to correctly order models, the distribution of results is not changed significantly when considering the entire set of models instead of only Model 1. The inability to rank models is not surprising given the general difficulty in accurately scoring protein models and the fact that most models are generally very similar to one another. However, for some applications it may be possible to use more than one model. For example, during computational drug screening or docking, it could be possible to screen the compounds against different models of the target protein and then take either the lowest or average score for each compound. To that end, we have recalculated the results, but this time examining only the best structure for each group (as judged by sum of Z-scores) for each target (Table III). When judged by their best model for each target, there are eight groups that perform better than the null model. The top two groups, DBAKER and LEE, are able to improve all of the metrics on average.

Figure 3 shows the improvement in various metrics considering only the best structure from each target for the top five groups. When compared to Figure 1, all of the distributions are shifted to the right. Clearly, the best structures from top five groups perform much better than the average of all groups. However, even the top groups cannot accurately rank-order their submissions,

**Table II**

Average Changes Relative to Starting Model Over All Models with a Model ID of One

Group	GDT-TS	GDT-HA	GDC-SC	Correct rotamers	Correct MC H-bonds	Correct SC H-bonds	MolProbity <sup>a</sup>	MCRS	Sum of Z-scores
DBAKER	-0.19	-0.69	2.88	0.07	0.04	0.09	1.18	-2.11	9.34
LEE	0.09	0.27	2.42	0.05	0.01	-0.01	0.22	16.95	7.54
YASARARefine	-1.64	-3.10	1.02	-0.00	0.03	0.05	1.77	19.22	7.46
FAMSD	-0.73	-1.16	1.47	0.03	-0.04	0.03	0.11	-3.07	5.33
LevittGroup	-0.42	-0.99	1.38	-0.00	0.01	0.01	0.04	5.13	5.08
Null	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.81
TASSER	-1.48	-1.54	0.78	0.00	-0.06	-0.02	0.09	1.14	4.78
FEIG_REFINE	-0.39	-0.67	1.19	-0.00	-0.07	0.03	1.05	-10.88	3.78
SAM-T08-HUMAN	-2.31	-2.88	0.73	0.02	0.02	0.02	0.07	11.55	3.38
PS2-server	0.25	0.18	-0.05	-0.05	-0.01	-0.03	-0.20	12.70	3.23
Tripos_08	-1.58	-1.92	-0.01	0.05	-0.05	0.00	0.05	8.92	2.68
Xianmingpan	-3.56	-4.89	-2.25	-0.03	-0.04	0.09	1.13	6.54	2.04
Jacobson	-2.98	-4.03	-2.13	-0.04	-0.07	0.04	0.11	0.57	1.21
Kolinski	-4.10	-6.71	-2.36	-0.05	-0.04	0.08	0.44	0.22	-0.34
Abagyan	-1.19	-2.31	-3.32	-0.07	-0.07	-0.01	-0.38	-4.09	-1.15
BATES_BMM	-3.68	-5.48	-1.32	0.02	-0.04	-0.02	0.07	4.53	-1.58
A-TASSER	-3.12	-4.23	-3.05	-0.04	0.00	-0.06	0.29	-0.20	-2.37
SAMUDRALA	-4.53	-3.40	-2.63	0.03	-0.08	-0.00	0.51	9.41	-4.55
FAMS-multi	-5.93	-6.82	-2.87	0.01	-0.11	0.05	-0.74	-17.08	-4.70
Kearar	-5.24	-7.84	-4.07	-0.01	0.03	0.03	0.19	-48.40	-7.05
POISE	-8.76	-11.38	-6.14	-0.02	-0.07	0.01	1.07	8.17	-8.02
EB_AMU_Physics	-6.19	-8.54	-7.32	-0.14	-0.06	0.02	0.00	-4.82	-8.43
MidwayFolding	-5.45	-6.66	-6.46	-0.06	-0.10	-0.03	-0.83	-14.53	-9.19
Jones-UCL	-12.69	-16.11	-11.93	-0.07	-0.12	-0.04	-0.67	-10.91	-20.68
Elofsson	-20.82	-19.25	-12.98	-0.00	-0.24	-0.01	0.09	-6.19	-30.92
POEM	-15.29	-20.15	-14.78	-0.22	-0.41	-0.13	-1.82	-31.34	-37.36

Groups are sorted by average sum of Z-scores (refer text). In some cases, groups submitted incomplete models or models without sidechains, which precluded us from calculating one or more metrics. In these cases, all results for that particular structure were omitted from this table. Target 476 was excluded from all calculations as the starting model did not contain sidechains.

<sup>a</sup>We use the additive inverse of MolProbity, such that improved structures have  $\Delta$ MolProbity  $> 0$ .

which suggests, at least for refinement, that scoring may be harder than search. That is, groups are able to generate structures that are an improvement over the starting model, but they have more difficulty ranking various structures relative to one another. Essentially, we are giving participants up to five guesses per target. If groups were able to accurately rank order their structures, then only a single model would be necessary. It would be desirable if the metrics we have examined here could be used to help improve predictions. Unfortunately, of the eight metrics we use here, all but MolProbity and MCRS are native-centric—meaning that you need to know the native structure to calculate them—and are thus not useful for structure prediction.

The analysis we have presented so far either examines all structures from each group or chooses one structure per target that is defined as the best structure over a combination of metrics. To account for the fact that different metrics may be more important for different end uses, we have also performed calculations where we select the best structure for each metric individually (Table IV). For example, the GDT-TS column indicates the average change in GDT-TS for each group, where the average is over the structure with the highest GDT-TS for each target. On the basis of these results, GDT-HA, GDT-TS, and fraction of correct main chain hydrogen bonds appear to be the most

difficult metrics to improve consistently, while nearly all of the groups are able to improve MCRS, MolProbity, and fraction of correct sidechain hydrogen bonds.

Overall, the top-ranking groups were able to improve most metrics, usually with only a small cost to GDT-TS and GDT-HA. The situation is even more promising if we set aside the difficulty that groups had in rank ordering their structures and instead focus on the best structures from each group.

As we discuss below, choosing metrics is difficult, and partly arbitrary. This is further compounded by the relatively small sample size; there were only 12 targets and so a single bad (or good) result could easily shift the score of a group by a relatively large amount. Therefore, while we feel that our analysis gives a general impression of the groups that are doing well and of the overall performance of the field, a comparison between closely ranked groups is unlikely to yield meaningful results. More meaningful statistics may be possible in future years if there are more groups or more targets.

#### Different metrics measure different aspects of quality

What is the best metric for measuring the quality of protein structures predicted? In principle, the most physically meaningful metric is nature's own: comparison

**Table III**

Average Improvement Relative to the Starting Model for the Best Models from Each Group

Group	GDT-TS	GDT-HA	GDC-SC	Correct rotamers	Correct MC H-bonds	Correct SC H-bonds	MolProbity <sup>a</sup>	MCRS	Sum of Z-scores
DBAKER	2.22	3.11	5.29	0.08	0.05	0.09	1.24	1.28	14.39
LEE	0.38	0.48	2.72	0.05	0.01	0.01	0.12	14.31	8.35
YASARARefine	-1.64	-3.10	1.02	-0.00	0.03	0.05	1.77	19.22	7.46
FEIG_REFINE	0.42	0.52	1.97	0.02	-0.05	0.05	1.04	-12.97	6.95
LevittGroup	-0.18	-0.36	1.16	-0.00	0.01	0.02	0.29	6.28	6.52
TASSER	-0.41	-0.42	1.01	0.00	-0.05	-0.01	-0.01	1.18	6.28
SAM-T08-HUMAN	-0.71	-0.44	2.93	0.02	0.03	0.03	0.11	9.87	6.28
FAMSD	-0.67	-1.05	1.84	0.03	-0.04	0.04	0.19	3.70	6.06
Null	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.81
A-TASSER	-0.36	-0.91	0.16	-0.01	0.01	-0.04	0.43	2.58	4.81
tripos_08	-0.72	-0.90	1.10	0.06	-0.04	0.01	0.07	6.77	4.74
BATES_BMM	-1.32	-2.33	0.04	0.04	-0.03	-0.02	0.19	6.35	3.64
PS2-server	0.25	0.18	-0.05	-0.05	-0.01	-0.03	-0.20	12.70	3.23
Kolinski	-0.93	-0.93	0.57	-0.07	0.01	0.00	-0.06	7.51	3.06
Jacobson	-1.81	-2.74	-1.55	-0.04	-0.07	0.04	0.15	1.12	2.73
Xianmingpan	-3.56	-4.89	-2.25	-0.03	-0.04	0.09	1.13	6.54	2.04
POISE	-3.99	-5.42	-1.69	-0.02	-0.02	0.02	1.29	10.43	1.24
Abagyan	-1.07	-1.63	-2.90	-0.06	-0.06	0.01	-0.39	0.13	0.76
SAMUDRALA	-3.59	-2.36	-0.86	0.06	-0.03	0.00	0.69	11.90	-1.12
FAMS-multi	-5.83	-6.81	-2.48	-0.00	-0.10	0.05	-0.65	-12.65	-4.00
Keasar	-3.43	-5.87	-3.00	-0.01	0.04	0.03	0.19	-49.31	-4.55
MidwayFolding	-4.94	-6.20	-6.15	-0.06	-0.12	-0.03	-0.83	-11.49	-8.23
EB_AMU_Physics	-6.19	-8.54	-7.32	-0.14	-0.06	0.02	0.00	-4.82	-8.43
Jones-UCL	-7.14	-9.48	-7.87	-0.06	-0.07	-0.03	-0.59	-8.61	-10.30
Elofsson	-20.82	-19.25	-12.98	-0.00	-0.24	-0.01	0.09	-6.19	-30.92
POEM	-12.60	-17.00	-11.77	-0.21	-0.41	-0.12	-1.79	-27.60	-32.00

For each group, a single model was chosen for each target based on the sum of Z-scores (refer text). Groups are sorted according to the sum of Z-scores of their best models. In some cases, groups submitted incomplete models or models without sidechains, which precluded us from calculating one or more metrics. In these cases, all results for that particular structure were omitted from this table. Target 476 was excluded from all calculations as the starting model did not contain sidechains.

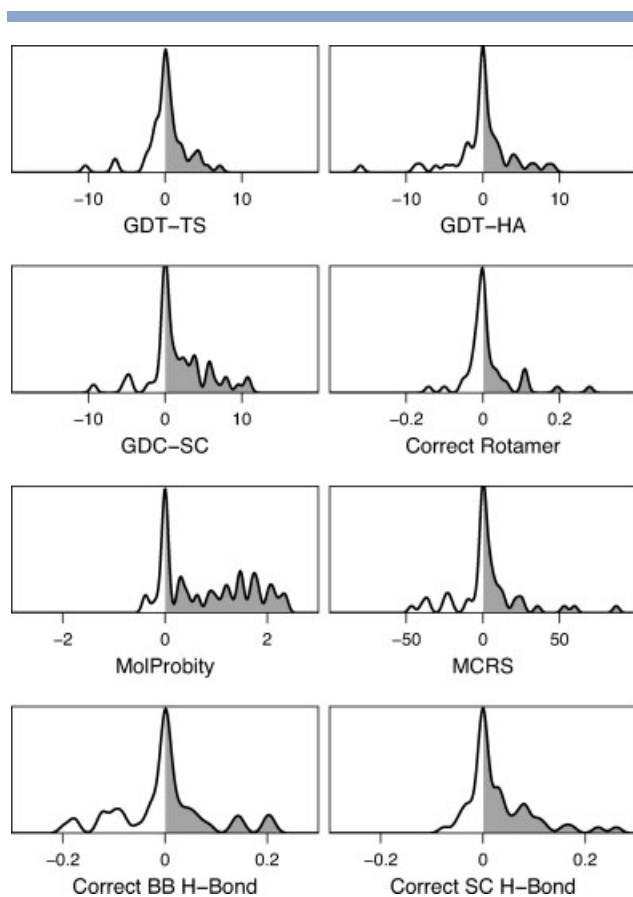
<sup>a</sup>We use the additive inverse of MolProbity, such that improved structures have  $\Delta$ MolProbity > 0.

against the true free energy. Unfortunately, nature's true free energy function is not known, so we must rely instead on structure-based surrogate measures, which will always be a compromise and arbitrary to some degree.

In recent years, CASP has focused primarily on GDT-TS<sup>1</sup> and related metrics. While GDT-TS has some advantages over RMSD, it is still fundamentally geometric, taking no account of energetics, and is focused on the backbone, taking no account of sidechains. Richardson and her colleagues in this issue<sup>20</sup> note that focusing on the alpha-carbon atoms ignores 90% of the protein. For example, imagine a structure that has all of the alpha carbons in exactly the correct locations, i.e., it has a GDT-TS of 100 and an RMSD of zero. Now, move some of the sidechains around such that there are overlapping atoms. This would lead to structures with very high energy, therefore, unphysical and unnatural, and yet GDT-TS would imply the errors are small. Richardson and co-workers point out numerous examples of this throughout CASP8—there are models that score well by GDT-TS and related metrics that are sometimes very unphysical. This was also seen in CASP7.<sup>19</sup> So, we followed previous<sup>19,23</sup> and current assessors,<sup>20</sup> who have applied multiple metrics. There is no fundamentally correct single metric; choosing a metric is partly arbitrary and will depend on the structure-prediction objective.

### Other metrics of quality are based on comparing to experimental structural data

All of the analysis presented thus far has compared a predicted model with a model derived from the experimental data. However, it should be noted that for both x-ray crystallography and NMR, the models generated are exactly that—models. Almost all of the metrics we have employed in our analysis, including GDT-TS, require the definition of a single experimental structure. Of course, proteins are flexible and dynamic and adopt a range of different conformations, which give rise to the native ensemble. Both crystallography and NMR attempt to model conformational heterogeneity, but in different ways. Typically, NMR structures are reported as an explicit ensemble containing 20–50 different conformations. For CASP8, the average NMR structure was calculated and then the single structure from the ensemble that was closest to the average was taken as the experimental structure. Structures determined by crystallography generally consist of a set of average atomic coordinates with the heterogeneity of each atomic position modeled using an isotropic Gaussian distribution. However, it has been shown that using an isotropic model of heterogeneity leads to the estimates of model accuracy that are too high.<sup>24,25</sup> After the selection of a single experimental structure, the standard CASP protocol is then



**Figure 3**

Distribution of refinement results across the best structures of the top five groups (489, 407, 298, 335, 442) as judged by the highest average sum of Z-scores. Distributions are shown for these metrics: GDT-TS, high accuracy GDT (GDT-HA), sidechain GDC (GDC-SC), fraction of correct sidechain rotamer conformations (Correct Rotamer), the additive inverse of the MolProbity score (so that better structures have  $\Delta\text{MolProbity} > 0$ ), main chain reality score (MCRS), fraction of correct backbone hydrogen bonds (Correct BB H-Bonds), and fraction of correct sidechain hydrogen bonds (Correct SC H-Bonds). Values to the right of zero (gray shading) are an improvement over the starting model.

to excise any regions that appear to be unstructured or otherwise highly flexible. Any remaining heterogeneity, as encoded by crystallographic B-factors or NMR ensembles, is then ignored. This protocol is practical and indeed necessary to use most common structural quality metrics. However, this practice also has the undesirable side effect of reducing an ensemble of many possible conformations down to a single structure. The ideal way to evaluate structure predictions would be to compare an ensemble of structures from each group with ensemble averaged experimental data. Here, we attempt to go half way and compare single structural models with ensemble averaged data. We do not treat the five structures submitted by the predictors as an ensemble, as the groups were not told to treat their predictions in this way. In future, it may be interesting to allow groups to submit an ensemble

of structures, each of which they believe is approximately equally likely and lies within the native basin.

### There are structures that agree better with NMR data that have poorer GDT-TS scores

The dynamical properties or flexibility of proteins are crucial to account for their stability and biological function in solution.<sup>26–28</sup> In this respect, nuclear magnetic resonance (NMR) spectroscopy can provide information for both structure and dynamics of proteins in a physiologically relevant environment.<sup>29</sup> The intensities of nuclear overhauser effect (NOE) are among the most important NMR parameters for structure determination because they provide both local and non-local (in sequence) inter-proton distance information. For well structured large molecules such as proteins, NOE intensities are directly related to interatomic distances with reasonable neglect of the contribution from intramolecular motion.<sup>30</sup> Protein structural ensembles derived from molecular simulations are often compared to experiment by back-calculating NOE distances.<sup>31,32</sup> This procedure is attractive because of its relative simplicity.

For each CASP8 refinement target whose native structure was measured by NMR spectroscopy, the experimental distance restraints derived from experimentally observed NOEs were obtained from protein data bank and processed by the AQUA program suite (version 3.2).<sup>33–35</sup> For non-stereospecifically assigned hydrogen groups such as methyl and methylene in experiment, pseudo atoms were constructed by AQUA based on geometry average of corresponding proton positions. In general, the average distance between hydrogen atom  $i$  and  $j$  should be calculated as an ensemble average using  $\langle r_{i,j}^{-p} \rangle^{-1/p}$  (where  $p = 3$  or  $6$  based on the size of protein<sup>36</sup>). In this case, we have only used single structures, rather than ensembles, so we do not average the distances. A hydrogen pair ( $i,j$ ) is considered violating the NOE upper bound distance  $R_{ij}$  when  $v_{i,j} = r_{i,j} - R_{ij}$  is positive.<sup>32</sup> The upper bound distances are estimated based on the experimental NOE data and are typically used as one of the primary forms of data during NMR structural determination. Thus, rather than comparing to a single experimental structure, we are instead comparing to a set of upper bound distances determined from the experimental data. The violation  $v_{i,j}$  is considered zero when its value is negative. The average upper bound violation was calculated as  $v = \frac{1}{N} \sum_{i,j} v_{i,j}$  by averaging over the  $N$  distances derived from the experimentally determined NOEs.

As an alternative way to evaluate the success of prediction, the NOE upper bound violation analysis is compared with the more traditional metric GDT-TS (Fig. 4). The improvement in average upper bound violation for all refined models ( $-\Delta v$ , note the minus sign) was plotted versus the improvement in GDT-TS ( $\Delta\text{GDT}$ ) for all NMR targets by comparing with the corresponding unre-

**Table IV**  
Best Results for Each Metric Averaged Over Targets

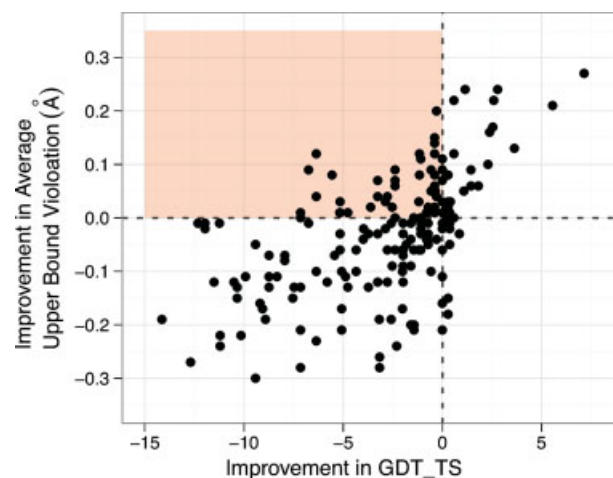
Group	GDT-TS	GDT-HA	GDC-SC	Correct rotamers	Correct MC H-bonds	Correct SC H-bonds	MolProbity <sup>a</sup>	MCRS
DBAKER	2.38	3.38	6.24	0.10	0.07	0.13	1.39	2.41
FEIG_REFINE	0.89	0.93	2.83	0.03	-0.03	0.07	1.25	-3.06
LEE	0.59	0.76	3.27	0.07	0.03	0.03	0.26	20.33
PS2-server	0.25	0.18	-0.05	-0.05	-0.01	-0.03	-0.20	12.70
LevittGroup	0.19	-0.01	2.03	0.01	0.04	0.05	1.37	16.16
Null	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-TASSER	-0.01	-0.61	0.69	0.02	0.04	-0.02	0.56	5.32
SAM-T08-HUMAN	-0.12	-0.28	3.22	0.03	0.05	0.05	0.43	14.55
FAMSD	-0.38	-0.57	2.06	0.04	-0.01	0.05	0.29	6.89
TASSER	-0.41	-0.42	1.04	0.01	-0.05	-0.01	0.15	4.26
Abagyan	-0.51	-1.39	-2.12	-0.05	-0.06	0.03	-0.33	1.16
tripos_08	-0.67	-0.85	2.23	0.06	-0.04	0.02	0.07	9.66
Kolinski	-0.93	-0.93	0.57	0.00	0.01	0.14	0.52	7.51
BATES_BMM	-1.22	-2.20	0.68	0.08	-0.01	0.02	0.33	10.80
YASARARefine	-1.64	-3.10	1.02	-0.00	0.03	0.05	1.77	19.22
Jacobson	-1.76	-2.64	-1.15	-0.01	-0.06	0.08	0.33	3.63
Keasar	-3.27	-5.87	-2.04	-0.00	0.06	0.05	0.34	-42.41
SAMUDRALA	-3.36	-2.05	0.08	0.06	-0.02	0.02	0.72	14.02
Xianmingpan	-3.56	-4.89	-2.25	-0.03	-0.04	0.09	1.13	6.54
POISE	-3.83	-5.27	-1.63	-0.01	-0.01	0.03	1.44	14.27
MidwayFolding	-4.94	-6.20	-5.83	-0.04	-0.10	-0.02	-0.81	-10.93
FAMS-multi	-5.50	-6.44	-1.18	0.03	-0.09	0.07	-0.60	-8.49
EB_AMU_Physics	-6.19	-8.54	-7.32	-0.14	-0.06	0.02	0.00	-4.82
Jones-UCL	-6.99	-9.22	-6.41	-0.03	-0.05	0.00	-0.45	-0.24
POEM	-12.60	-17.00	-10.38	-0.20	-0.40	-0.10	-1.79	-21.71
Elofsson	-20.82	-19.25	-12.98	-0.00	-0.24	-0.00	0.10	-5.69

In contrast to Table III where we chose a single best structure for each target, here, we choose the structure that maximizes each metric separately. The values in each column represent the average improvement in that metric, taken over the the models (one for each target) that give the biggest improvement for that metric. Groups are sorted according to the average change in GDT-TS. In some cases, groups submitted incomplete models or models without sidechains, which precluded us from calculating one or more metrics. In these cases, all results for that particular structure were omitted from this table. Target 476 was excluded from all calculations as the starting model did not contain sidechains.

<sup>a</sup>We use the additive inverse of MolProbity, such that improved structures have  $\Delta\text{MolProbity} > 0$ .

finer models. The agreement with NMR data is roughly correlated with GDT, i.e., the larger the improvement in  $\nu$  for refined models, the larger the improvement in GDT score. However, the correlation is not perfect; there are models that are worse as measured by GDT-TS ( $\Delta\text{GDT} - \text{TS} < 0$ ) that actually have better agreement with NMR data than the unrefined model and can be identified by using average upper bound violation ( $-\Delta\nu > 0$ ) (shaded region, Figure 4). On the other hand, most of the better models selected by GDT-TS are seldom regarded as worse models by average upper bound violation. Therefore, based on current analysis of NMR targets, it appears that direct comparison with NOE data can provide additional information beyond that available when comparing only with a single model. In future, comparison with NMR measurements might be useful in identifying more successful predictions in addition to those selected by traditional evaluation methods such as GDT-TS.

Table V shows the average improvement in NOE upper-bound violation by group. When scored according to their best model, more than half of the groups are able to do better than the null refinement. This performance is clearly better than for GDT-TS where only five



**Figure 4**

There is a correlation between improvement in GDT-TS and agreement with NMR data. The agreement with NMR data is quantified by calculating the average upper bound violation distance for the set of distance restraints derived from experimentally observed NOEs. However, there are many structures which are in better agreement with the NMR data, but that give poorer agreement as judged by GDT-TS (shaded region).

**Table V**  
Average Improvement in NOE Upper-Bound Violations by Group

Group	Number of targets <sup>a</sup>	$\langle \text{Improvement} \rangle_{\text{Bestmodel}}$	$\langle \text{Improvement} \rangle_{\text{Model1}}$
DBAKER	5	0.19	0.07
Jacobson	2	0.15	0.09
Lee	5	0.13	0.10
SAM-T08-HUMAN	5	0.12	0.08
SAMUDRALA	5	0.06	0.02
YASARARefine	5	0.04	0.04
FEIG-REFINE	4	0.03	-0.05
Tripos_08	1	0.03	-0.12
Jones-UCL	5	0.03	-0.07
LevittGroup	5	0.01	-0.01
FAMSD	5	0.00	-0.03
Null	5	0.00	0.00
BATES_BMM	5	-0.02	-0.02
POISE	5	-0.05	-0.14
Abagyan	4	-0.06	-0.07
A-TASSER	5	-0.08	-0.20
Xianmingpan	4	-0.08	-0.08
MidwayFolding	5	-0.08	-0.14
FAMS-multi	5	-0.13	-0.17
Elofsson	5	-1.93	-1.93

<sup>a</sup>There were four NMR targets, but TR429 contained two domains that were analyzed separately.

groups are able to improve on average. There are a number of possible factors at work here. First, GDT-TS uses a single set of fixed alpha carbon coordinates, while the NMR analysis uses a set of upper bounds that is inherently more flexible (all distances less than the upper bound are considered equally “good”). In other words, the upper bound distances allow for structural flexibility. Second, this flexibility is evident from the fact that the model of the NMR data is an ensemble rather than a single structure. To run tools like GDT, this ensemble of structures must be collapsed to a single structure. For CASP8, the structure closest to the average of the ensemble was chosen, which we feel is a reasonable choice. However, there is no way that a single structure can encode the conformational heterogeneity that is present in the experimental data. Third, the analysis presented here is extremely simplistic. Calculating upper-bound distance violations is probably the simplest way to compare structures with NMR data. This simplicity is why we chose the method, however, there may be more rigorous ways to approach this problem. At the very least, it would be desirable to compare ensembles of predicted structures with the experimental data.

There were only four NMR targets in the refinement competition, and therefore, the statistics are extremely limited. We would urge extreme caution when comparing the relative ranks of different groups due to the small data set. At most each group was only judged on five models.

### Comparison with crystallography data

The primary goal of X-ray crystallography is to determine the electron density within the unit cell of a crystal,

which is used as a map to construct an atomic resolution model. The electron density  $\rho(x,y,z)$  is given by a summation of Fourier transforms over complex structure factors  $F(h,k,l) = |F(h,k,l)| \exp[i\alpha(h,k,l)]$ , where  $(x,y,z)$  are fractional coordinates ( $0 \leq x < 1$ ) and  $(h,k,l)$  are the Miller indices of a Bragg reflection. The fundamental relationship is:

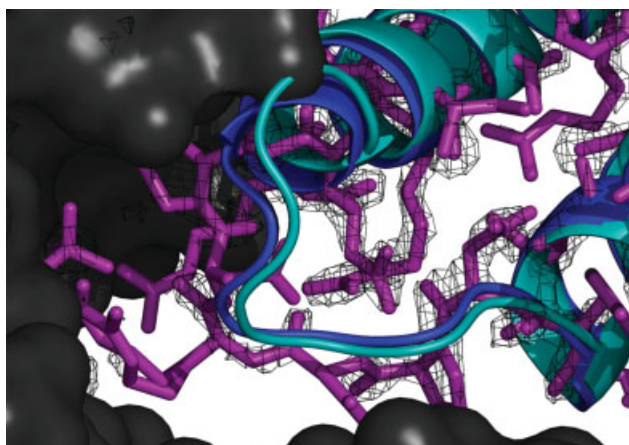
$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F(h, k, l)| \times \exp[-2\pi i(hx + ky + lz) + i\alpha(h, k, l)] \quad (1)$$

where  $V$  is the volume of the unit cell. Although structure factor magnitudes are directly derived from observed diffraction intensities  $I_o(h,k,l) \sim |F_o(h,k,l)|^2$ , phase angles  $\alpha_o(h,k,l)$  are not.<sup>37</sup> This is commonly referred to as the phase problem.

Molecular replacement (MR) is a solution to the phase problem that avoids the need for additional experiments. The idea is simply to calculate approximate phases  $\alpha_c(h,k,l)$  from an atomic model, which is typically derived from a previously solved protein structure with high homology to the unknown structure. By reversing Eq. (1), structure factors  $F_c(h,k,l)$  can be calculated from the Fourier transform of the homologous electron density. The key steps in MR are choosing the surrogate atomic model, optimally positioning the model within the unit cell via six rigid body coordinates and finally, if possible, using the observed amplitudes and calculated phases,  $\alpha_c(h,k,l)$ , to produce an electron density map via Eq. (1) that allows the structure to be solved. These three steps will be discussed in the context of evaluating CASP

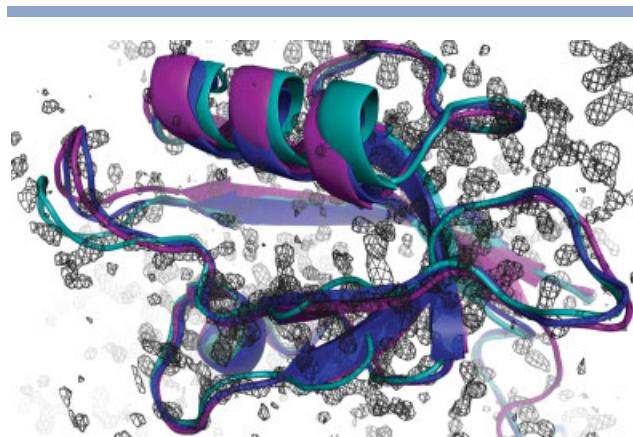
entries using MR and the important role for computational methods developed by CASP participants toward further enhancement of MR. Prediction of biomolecular structures to a level of quality that facilitates successful MR is a significant litmus test for computational methods.<sup>38–40</sup>

To evaluate CASP refinement entries against the experimental diffraction data, MR was performed on the deposited structure (a positive control), the starting template, and the entries. Optimal positioning of each structure within the unit cell used the Phaser software package version 2.1.3<sup>41</sup> in MR\_AUTO mode. In brief, a maximum likelihood rotation function (MLRF) was searched to find optimal sets of Euler angles. Second, a maximum likelihood translation function was searched using Euler angle sets with MLRF values of at least 75% of the top peak found. An important detail is the estimate of how the accuracy of each entry falls off as a function of resolution, which is specified by RMS coordinate error. For consistency, a value of 1.78 Å was used for all targets, which is the default Phaser value for homology models with 20% sequence identity to the target. For comparison, in the limit of 0% sequence identity, the default Phaser RMS is 2.60 Å. Other inputs to Phaser included the experimental diffraction data, unit cell dimensions, space group, and the composition of the unit cell. Four of the eight X-ray crystallographic targets contained multiple copies of the entry within the unit cell, which is called non-crystallographic symmetry



**Figure 5**

Shown is a loop connecting two alpha-helices from Target 432, with the starting model (teal), a top prediction (blue), and the deposited PDB structure (3DAI, purple). Solid molecular surfaces of nearby symmetry mates of the PDB structure are colored dark gray, which clash with both the starting model and prediction. The  $2F_o - F_c$  iso-contour at  $3\sigma$ , shown in wireframe, emphasizes that the loop and side chains are well resolved by the experimental data. The loop conformation may well be influenced by interactions with symmetry mates. Although it is a challenge to take into account crystal-packing effects, this may be a limiting factor in the agreement of CASP entries with experiment as measured by MR statistics.

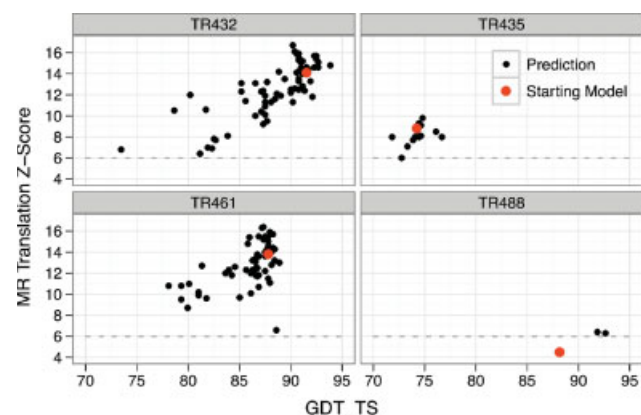


**Figure 6**

Two entries for Target 488 produced Z-scores consistent with possible molecular replacement success (6.4, 6.3), whereas the starting model did not (4.5). Shown in blue is model 3 from DBAKER, with the starting model and the deposited PDB structure (3DAI) colored teal and purple, respectively. The  $2F_o - F_c$  iso-contour at  $3\sigma$ , shown in wireframe, demonstrates that the loop on the left side of the image is located within density for both the entry and PDB structure, but not for the starting model.

(NCS). In these cases, Phaser was instructed to find rigid body coordinates for each NCS copy in turn.

The key metric that will be compared between the deposited structures, templates, and CASP entries is the Z-score assigned to the best rigid body solution after application of the translation function (abbreviated herein as TFZ). This is calculated based on the mean and standard deviation of the log likelihood gains for a set of



**Figure 7**

The GDT-TS score is plotted against molecular replacement translation function Z-score for all entries with a Z-score above 6.0, which indicates a possible MR solution. Four of the eight targets evaluated using experimental X-ray diffraction data had entries of this quality. Red circles represent the starting model; red symbols represent refinement predictions. In three cases, the starting models were already of sufficient quality to facilitate molecular replacement. However, for target TR488, the starting model fails molecular replacement, but two refinement entries succeed.

**Table VI**  
Average Improvement in Translation Function Z-Score (TFZ) by Group

Group	Number of targets <sup>a</sup>	$\langle \text{Improvement TFZ} \rangle_{\text{Bestmodel}}$	$\langle \text{Improvement TFZ} \rangle_{\text{Model1}}$
Tripos_08	2	2.15	0.10
Jacobson	2	1.60	1.40
LEE	3	1.33	-0.13
FAMSD	3	1.20	0.97
FEIG_REFINE	3	1.13	0.23
DBAKER	3	0.60	-1.98
PS2-Server	1	0.50	0.50
BATES_BMM	3	0.43	0.43
A-TASSER	2	0.30	-2.10
SAMUDRALA	3	0.03	-1.43
LevittGroup	3	0.03	0.03
Null	4	0.00	0.00
POISE	2	-0.60	-1.35
SAM-T08-HUMAN	3	-0.77	-2.73
FAMS-multi	1	-1.70	-4.90
Keasar	2	-1.80	-2.40
Elofsson	1	-2.10	-2.10
xianmingpan	1	-2.10	-2.10
Abagyan	3	-2.40	-2.40
MidwayFolding	2	-2.70	-4.45
YASARARefine	3	-3.20	-3.20
Jones-UCL	2	-3.55	-3.95
EB_AMU_Physics	1	-4.20	-4.20

<sup>a</sup>Only targets TR432, TR435, TR461, and TR488 and models with TFZ  $\geq 6$  are included.

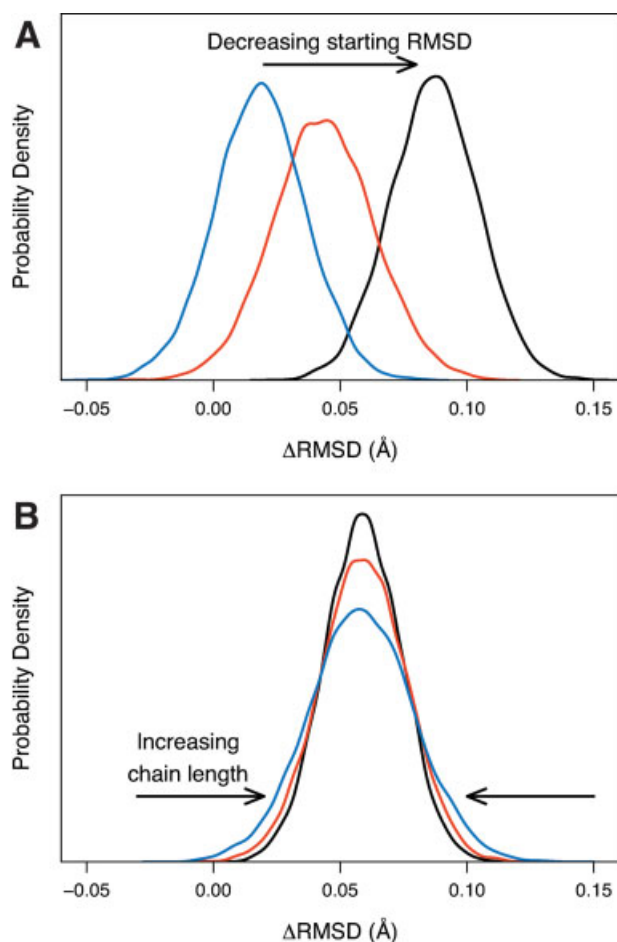
random orientations and translations. The TFZ measures how much better than random a given orientation and translation of the model within the unit cell is. However, TFZ is also very sensitive to the quality of the model itself; it is not possible to do much better than random if the model itself is very wrong. As an approximate guide, a TFZ score above 8 usually indicates successful phasing, while scores in the range of 6–8 are promising, but require further inspection. It is unlikely, although possible, that a structure with a score below 6 is a useful starting point for further refinement. In addition, the reliability of rankings based on low TFZ scores is diminished. Therefore, we focus on entries with TFZ scores of 6 or higher, although this is somewhat arbitrary.

Presentation of a few representative examples provides insight into the assessment of CASP refinement entries via X-ray diffraction data. The template with the highest translation function Z-score (TFZ) was for target 432, which achieved a value of 14.1. Of the entries for this target, 21 achieved higher TFZ scores than the template. In Figure 5, we compare a top entry to the template and target structure. Two salient features are obvious. First, the  $2F_o - F_c$  density map, shown as a wireframe isocontour at  $3\sigma$ , shows that the backbone and side chain conformations are well defined in the experimental data. In other words, the backbone and sidechains can be positioned unambiguously from the experimental data. Second, it is clear that both the template and top entry clash with crystallographic symmetry mates of the fully refined structure. Without taking into account crystal packing it

may be unreasonable to expect refinement methods to achieve greater agreement with the experimental structure.

The template for target 488 fails at MR, as judged by its low TFZ score. However, two entries achieved TFZ scores (6.4 and 6.3) above 6.0, which indicates that phasing may have been successful. Refinement of a structure that fails MR into one that succeeds is of great practical significance to field of X-ray crystallography. One easily visible improvement of both top entries compared to the template is presented in Figure 6. Specifically, a template loop on the left side of the image is displaced from the loops of the entry and target structures.

A summary of the assessment via X-ray diffraction data of the CASP refinement category is presented in Figure 7. The GDT-TS scores are plotted versus molecular replacement TFZ scores for four targets that had entries of sufficient quality that molecular replacement was successful, where success was defined as the target having entries with TFZ scores greater than 6. It is clear by eye that the GDT-TS score and TFZ score are correlated, which is not surprising. One striking feature of the summary plot is that a greater number of entries achieved TFZ scores that were superior to the template for the two targets with the best templates (432 and 461), compared to the two targets with poorer templates (435 and 488). Although TFZ correlates roughly with GDT-TS, there are numerous examples where TFZ is improved but GDT-TS is worse, and vice versa. As pointed out during the assessment of CASP7,<sup>23</sup> this is likely due to the fact



**Figure 8**

The predicted difficulty of refinement depends on starting RMSD and chain length. (A) Refinement is predicted to be more difficult close to the crystal structure. Distribution of  $\Delta$ RMSD values after randomly perturbing structures with approximately the same chain length and increasing starting RMSD: TR453 (black, 91 residues, starting RMSD: 1.40 Å), TR464 (red, 89 residues, starting RMSD: 2.94 Å), and TR476 (blue, 116 residues, starting RMSD: 6.85 Å). (B) Refinement is predicted to be more difficult for longer chains. Distribution of  $\Delta$ RMSD values of randomly perturbing structures with decreasing chain length and approximately the same starting RMSD: TR435 (black, 135 residues, starting RMSD: 2.15 Å), TR488 (red, 95 residues, starting RMSD: 2.18), and TR469 (blue, 65 residues, starting RMSD: 2.15 Å). Structures with negative  $\Delta$ RMSD values are an improvement over the starting model. Structures were perturbed by displacing the alpha carbon positions according to a Gaussian distribution with a mean of zero and a standard deviation of 0.5 Å.

that GDT-TS and TFZ have very different functional forms and are sensitive to different types of errors, which is precisely what makes TFZ a potentially useful metric. Although a large number of entries improved upon the template in terms of TFZ score, we expect this number could be higher if the crystallographic environment is accounted for by prediction methods.

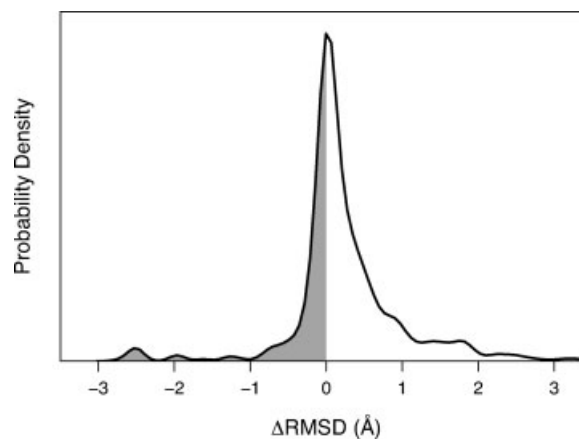
For half of the targets, even models that were significant improvements over the template still had RMSDs

from the refined structure that are larger than ideal to expect successful MR. Crystallography metrics are best suited to judge refinement of templates that are at most 2–3 Å RMSD from the true structure. With respect to the four targets with models that achieved MR success, comparison directly to the crystallography data and visualization of the electron density gives insight into why models improved upon the template. Participants may gain considerable insights by applying MR to their entries post facto to appreciate the degree of importance of precise bond lengths, bond angles, and crystal packing to MR success and lower *R* values.

Table VI shows the average improvement in TFZ by group. The data set is very small because only four of the templates were of good enough quality for MR to succeed. Therefore, we urge caution in comparing the relative ranks of different groups. At most, each group was evaluated using only three structures. It appears that half of the groups were able to beat the null refinement. However, this result is heavily skewed by the fact that we only included models with a TFZ > 6. Any models with TFZ < 6 were excluded and did not count against the group that produced them.

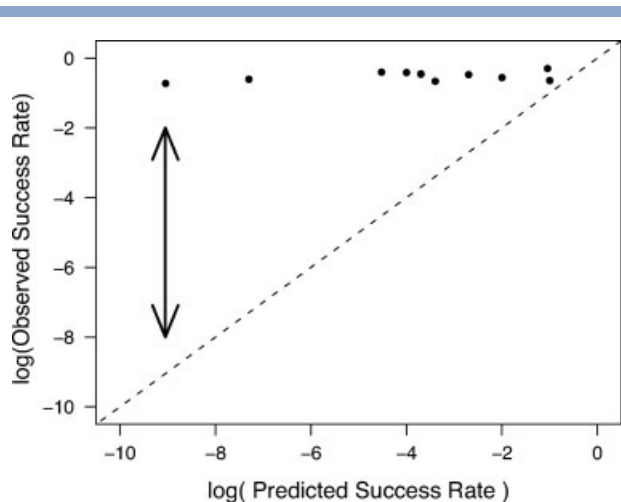
#### Assessing target difficulty using a random-perturbation model

How difficult is a given target structure to predict? Sullivan and Kuntz noted that there are far more conformations of the protein far away from the native structure than are close to it.<sup>42,43</sup> So, refinement should get harder as the starting model gets closer to the crystal structure. This argument is essentially one of conformational entropy; there is only one crystal structure, while



**Figure 9**

Observed distribution of  $\Delta$ RMSD values for all CASP8 refinement submissions. The actual distribution extends on the right to a maximum  $\Delta$ RMSD of 24.64 Å. Structures with negative  $\Delta$ RMSD values (gray shading) are an improvement over the starting model.



**Figure 10**

Refinement is successful much more frequently than predicted by a simple random model (refer text). The observed success rate (as judged by  $\Delta$ RMSD) is essentially constant over a huge range of predicted target difficulty, leading to success rates that are eight orders of magnitude more successful than predicted for target TR461 (left most point, 189 residues, 1.63 Å RMSD). The dotted line indicates perfect agreement between observed and predicted success rates. Target TR453 is not included in this plot as the observed success rate was zero.

there are many, many more possible structures as we move away from the crystal structure. The dimensionality of the search space also increases with the chain length, so larger target proteins should be more difficult.

Here, we use a simple model to estimate the target difficulty. Starting from the starting model, we randomly jiggle its coordinates. We perturb each alpha carbon in a random direction by a distance chosen from a Gaussian distribution with zero mean and a standard deviation of 0.5 Å. This perturbed model is then aligned to the experimental structure and the alpha carbon RMSD calculated in the standard way. The random models that have improved the structure are those where the RMSD of the perturbed model is lower than the RMSD of the starting model, or in other words, where  $\Delta$ RMSD < 0. This model is highly simplified; it ignores chain connectivity, excluded volume, and energetics. It would be interesting to apply energy minimization to the “jiggled” structures to compare the effect on the size of conformational space when energetic terms are included, although we have not done that analysis here. For a more rigorous analysis of the distribution of protein conformation space refer (42,43). However, despite these shortcomings, our simple model captures the qualitative features described earlier for indicating the entropic difficulty of the search for comparing different target proteins.

We apply our simple random model using  $\Delta$ RMSD as a metric of success. We choose  $\Delta$ RMSD rather than  $\Delta$ GDT-TS primarily for computational convenience; we need to calculate scores for millions of structures gener-

ated by our model and performing a single RMSD calculation is much faster than the many alignments required by GDT. As GDT-TS and RMSD are related measures, we thus expect that the insight gained by examining  $\Delta$ RMSD should be roughly transferrable to GDT-TS. Figure 8 shows the results of applying this simple model to several CASP8 refinement targets. The mean of the  $\Delta$ RMSD distribution shifts to the right as the starting RMSD decreases [Fig. 8(A)], while the  $\Delta$ RMSD distribution narrows with increasing chain length [Fig. 8(B)]. The structures of interest for refinement are those with negative  $\Delta$ RMSD values, i.e., structures that are closer to the crystal structure than the starting model. Because we are looking at the tail of a Gaussian distribution, we expect that the success rate of refinement should drop off rapidly as the initial RMSD is decreased and chain length is increased. This simple random model also predicts a wide range of target difficulties, with target TR476 (116 residues, 6.85 Å starting RMSD) predicted to be the easiest target, where  $\sim 1$  in 10 random perturbations lead to improvement in RMSD. In contrast, target TR461 (189 residues, 1.63 Å starting RMSD) is predicted to be the most difficult with  $\sim 1$  in  $10^9$  random perturbations leading to successful refinement.

### Refinement in CASP8 was much more successful than random chance

Figure 9 shows the  $\Delta$ RMSD values for the predictions submitted during the CASP8 refinement competition. The distribution is non-Gaussian due, presumably, to the small number of groups and targets. The range of  $\Delta$ RMSD values is from  $-2.73$  Å at best to 24.64 Å at worst. Most structures fall within the  $\Delta$ RMSD range of  $-1.0$  to 2.0 Å. The overall observed success rate (predictions with an RMSD to native that is lower than the starting model) across all groups and all targets is 0.29, which is higher than predicted for even the easiest target (TR476), which has a predicted success rate of 0.1.

Interestingly, we find no correlation between the observed refinement success of a given prediction and our measure of target difficulty (Fig. 10). This has a simple interpretation: For starting models that are far away from native, the entropic part of the search is not the hard part; a large fraction of perturbations can lead in the direction toward the native structure, but the predictors’ scoring functions are unable to provide guidance from that far away. In contrast, for starting models close to native, which are the most difficult targets, the actual CASP8 predictions are  $\sim 8$  orders of magnitude more successful than the random model predicts. This implies that the energy or scoring functions used provide a great deal of information that guides the search in the region of configuration space around the native structure. This may be the mirror result of the Baker group with Rosetta, where within a certain “radius of convergence”

of around 3 Å they are able to minimize towards the native structure.<sup>8,44</sup> Beyond the radius of convergence, the potential does not appear to offer much guidance towards the native state.

The actual success of the CASP8 predictions is roughly constant across a wide range of predicted difficulty and there is no correlation between success and predicted difficulty, starting GDT-TS, or chain length (data not shown). We have also examined several different measures of refinement success: the fraction of structures with improved GDT-TS, the average change in GDT-TS, the GDT-TS of best model from each group, and the GDT-TS of best model from all groups. We did not find any significant correlation between any of the measures of target difficulty and any of the measures of refinement success (data not shown). At this point, we cannot propose any simple metric that accurately predicts the difficulty of refining a given target.

### Suggestions for future refinement competitions

Here are some proposals for future CASP refinement tests.

- It would be valuable to have more targets, to increase the statistical significance of the evaluations.
- In some cases, knowing the crystal packing is essential for accurate refinement. For example, refinement of the loop in TR432 (Fig. 5) is likely impossible without knowing the unit-cell dimensions as the conformation of the loop appears to be heavily influenced by neighboring symmetry mates. In future, it may help to give information (unit cell, non-crystallographic symmetry) of this type, although it is not clear that predictors would be able to effectively use this information. Alternatively, targets where refinement would be influenced by crystal packing could be excluded.
- We believe that it is desirable to incorporate direct comparison with experimental data into the assessment of future CASP experiments whenever possible.
- We believe it is useful to apply a variety of physical metrics in addition to GDT-TS.

### SUMMARY

We assessed the refinement category in CASP8. Some groups were able to consistently improve from the given starting models. We assessed not only backbone errors, using GDT-TS and GDT-HA scores, but also several new metrics developed by Jane Richardson and colleagues.<sup>20</sup> The combination of these measures appears to give a good overview of performance. We also compared the refinement results directly with experimental data, either NMR or X-ray diffraction. We found several examples of

structures that are in better agreement (compared to the starting model) with the experimentally derived NOE distances that have negative  $\Delta$ GDT-TS. That is, some structures agree better with experimental data, even though they perform worse by the dominant metric used in recent CASP competitions. We also tested the models against X-ray diffraction data using the molecular replacement Z-score as our metric. We find that the Z-score is highly sensitive to being extremely close to the native structure. Most of the models we examined are not good enough for a meaningful comparison with experimental data to be made.

### ACKNOWLEDGMENTS

The authors thank Andriy Kryshchak and the prediction center for the automated analysis. Several new metrics were calculated by Jane Richardson and colleagues (Daniel A. Keedy, Christopher J. Williams, Jeffrey J. Headd, W. Bryan Arendall III, Vincent B. Chen, Gary J. Kapral, Robert A. Gillespie, Adam Zemla, David C. Richardson, and Jane S. Richardson) and the authors acknowledge that the analysis would not have been possible without their work. J.L.M. and K.A.D. thank the CASP organizers, John Moult, Krzysztof Fidelis, Andriy Kryshchak, Burkhard Rost, and Anna Tramontano, for the invitation to participate as CASP assessors. J.L.M. is supported by fellowships from the Alberta Heritage Foundation for Medical Research and the Natural Sciences and Engineering Research Council of Canada. The authors thank Tim Fenn for helpful discussions. M.P.J. is a consultant to Schrodinger Inc. M.P.J. and L.H. were participants in this category during CASP8 (group 470). Their main contribution to the article was the comparison to NMR data and their participation in the evaluation in no way skews the results to be more favorable to their group.

### REFERENCES

1. Zemla A. LGA: a method for finding 3d similarities in protein structures. *Nucleic Acids Res* 2003;31:3370–3374.
2. Chopra G, Summa CM, Levitt M. Solvent dramatically affects protein structure refinement. *Proc Natl Acad Sci USA* 2008;105:20239–20244.
3. Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci* 2004;13:211–220.
4. Ishitani R, Terada T, Shimizu K. Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations. *Mol Simulation* 2008;34:327–336.
5. Jagielska A, Wroblewska L, Skolnick J. Protein model refinement using an optimized physics-based all-atom force field. *Proc Natl Acad Sci USA* 2008;105:8268–8273.
6. Lee MR, Tsai J, Baker D, Kollman PA. Molecular dynamics in the endgame of protein structure prediction. *J Mol Biol* 2001;313:417–430.
7. Lu H, Skolnick J. Application of statistical potentials to protein structure refinement from low resolution ab initio models. *Biopolymers* 2003;70:575–584.

8. Misura KMS, Baker D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 2005;59:15–29.
9. Misura KMS, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with rosetta can be more accurate than their templates. *Proc Natl Acad Sci USA* 2006;103:5361–5366.
10. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: predicting loops in inexact environments. *Proteins* 2008;72:959–971.
11. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? i. Large scale amber benchmarking. *J Comput Chem* 2007;28:2059–2066.
12. Wroblewska L, Jagielska A, Skolnick J. Development of a physics-based force field for the scoring and refinement of protein models? *Biophys J* 2008;94:3227–3240.
13. Zhu J, Fan H, Periole X, Honig B, Mark AE. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins* 2008;72:1171–1188.
14. Chen J, Brooks CL, III. Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins* 2007;67:922–930.
15. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. *Protein Sci* 2000;9:1753–1773.
16. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc Natl Acad Sci USA* 2004;101:15346–15351.
17. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. *Proteins* 2007;69(Suppl 8):194–207.
18. Kryshtafovych A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T, Fidelis K. New tools and expanded data analysis capabilities at the protein structure prediction center. *Proteins* 2007;69(Suppl 8):19–26.
19. Kopp J, Bordoli L, Battey JND, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Predictions for template-based modeling targets. *Proteins* 2007;69(Suppl 8):38–56.
20. Keedy DA, Williams CJ, Headd JJ, Arendall WB, III, Chen VB, Kapral GJ, Gillespie RA, Zemla A, Richardson DC, Richardson JS. The other 90 percent of the protein: assessment beyond the alphas for CASP8 template-based models. *Proteins*, in press.
21. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, III, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 2007;35:W375–W383.
22. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 1999;285:1711–1733.
23. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins* 2007;69(Suppl 8):27–37.
24. DePristo MA, de Bakker PIW, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by x-ray crystallography. *Structure* 2004;12:831–838.
25. Furnham N, Blundell TL, DePristo MA, Terwilliger TC. Is one solution good enough? *Nat Struct Mol Biol* 2006;13:184–185.
26. Benkovic SJ, Hammes-Schiffer S. A perspective on enzyme catalysis. *Science* 2003;301:1196–1202.
27. Eisenmesser EZ, Bosco DA, Akke M, Kern D. Enzyme dynamics during catalysis. *Science* 2002;295:1520–1523.
28. Rasmussen BF, Stock AM, Ringe D, Petsko GA. Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature* 1992;357:423–424.
29. Teng Q. *Structural biology: practical NMR applications*. New York: Springer; 2005.
30. Lipari G, Szabo A. Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules. 1. Theory and range of validity. *J Am Chem Soc* 1982;104:4546–4559.
31. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. *Nature* 2005;433:128–132.
32. Zagrovic B, van Gunsteren WF. Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us? *Proteins* 2006;63:210–218.
33. Laskowski RA, Rullmann J, MacArthur M, Kaptein R, Thornton J. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 1996;8:477–486.
34. Doreleijers JF, Rullmann JA, Kaptein R. Quality assessment of NMR structures: a statistical survey. *J Mol Biol* 1998;281:149–164.
35. Doreleijers JF, Ravest ML, Rullmann T, Kaptein R. Completeness of NOEs in protein structure: a statistical analysis of NMR. *J Biomol NMR* 1999;14:123–132.
36. Tropp J. Dipolar relaxation and nuclear overhauser effects in non-rigid molecules: the effect of fluctuating internuclear distances. *J Chem Phys* 1980;72:6035–6043.
37. Drenth J, Mesters J. *Principles of protein x-ray crystallography*, 3rd ed. New York: Springer; 2007.
38. Raimondo D, Giorgetti A, Bosi S, Tramontano A. Automatic procedure for using models of proteins in molecular replacement. *Proteins* 2007;66:689–696.
39. Giorgetti A, Raimondo D, Miele AE, Tramontano A. Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics* 2005;21(Suppl 2):72–76.
40. Das R, Baker D. Prospects for de novo phasing with de novo protein models. *Acta Crystallogr D Biol Crystallogr* 2009;65:169–175.
41. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Crystallogr* 2007;40:658–674.
42. Sullivan DC, Kuntz I. Conformation spaces of proteins. *Proteins* 2001;42:495–511.
43. Sullivan DC, Kuntz I. Distributions in protein conformation space: implications for structure prediction and entropy. *Biophys J* 2004;87:113–120.
44. Bowman GR, Pande VS. Simulated tempering yields insight into the low-resolution rosetta scoring functions. *Proteins* 2009;74:777–788.