PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessment of template-based modeling of protein structure in CASP11

Vivek Modi, Qifang Xu, Sam Adhikari, and Roland L. Dunbrack Jr.*

Fox Chase Cancer Center, Institute for Cancer Research, Philadelphia, Pennsylvania 19111

ABSTRACT

We present the assessment of predictions submitted in the template-based modeling (TBM) category of CASP11 (Critical Assessment of Protein Structure Prediction). Model quality was judged on the basis of global and local measures of accuracy on all atoms including side chains. The top groups on 39 human-server targets based on model 1 predictions were LEER, Zhang, LEE, MULTICOM, and Zhang-Server. The top groups on 81 targets by server groups based on model 1 predictions were Zhang-Server, nns, BAKER-ROSETTASERVER, QUARK, and myprotein-me. In CASP11, the best models for most targets were equal to or better than the best template available in the Protein Data Bank, even for targets with poor templates. The overall performance in CASP11 is similar to the performance of predictors in CASP10 with slightly better performance on the hardest targets. For most targets, assessment measures exhibited bimodal probability density distributions. Multidimensional scaling of an RMSD matrix for each target typically revealed a single cluster with models similar to the target structure, with a mode in the GDT-TS density between 40 and 90, and a wide distribution of models highly divergent from each other and from the experimental structure, with density mode at a GDT-TS value of ~20. The models in this peak in the density were either compact models with entirely the wrong fold, or highly non-compact models. The results argue for a density-driven approach in future CASP TBM assessments that accounts for the bimodal nature of these distributions instead of *Z* scores, which assume a unimodal, Gaussian distribution.

Proteins 2016; 84(Suppl 1):200–220.

Key words: protein structure prediction; CASP; template-based modeling; homology modeling; structural biology.

INTRODUCTION

Template-based modeling is the most basic and widespread task in protein structure prediction and modeling. Papers describing such programs as Modeller,[1] Swiss-Model,[2] I-TASSER,[3] and SCWRL[4] have been cited 1000s of times each, the vast majority of which are applications of these programs in molecular and cellular biology. The basic protocol established in the 1970s and 1980s[5,6] has remained the same: identifying one or more proteins of known structure homologous to the target protein; aligning the sequence of the target to the sequences and structures of these proteins; building backbone and side-chain coordinates according to this alignment, while filling in insertions and repairing deletions due to gaps in the sequence alignment(s); refinement of the coordinates of the model to account for potential differences in orientation or distance between substructures in the target compared with the template(s); and assessment of model quality. The rapidly increasing number of sequences homologous to any target and the development of powerful database search programs based on sequence profiles or HMMs[7] have made the identification and alignment steps robust and relatively straightforward, although ambiguous alignments are sometimes encountered resulting in substantial errors in models. Accurate loop modeling and refinement remain challenging,[8] particularly the building of larger substructures (some including secondary structure) present in the target but not present in any of the available templates.

Template-based modeling has been a component in the CASP experiments since CASP1, and the relevant assessment papers provide a historical progression of the state of the field over the last 20 years.[9–23] In the early years, the main challenge was identifying whether there

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NIH; Grant number: R01 GM084453.

*Correspondence to: Roland L. Dunbrack Jr., Fox Chase Cancer Center, Philadelphia, PA 19111. E-mail: Roland.Dunbrack@fccc.edu

Received 24 January 2016; Revised 4 April 2016; Accepted 11 April 2016
Published online 15 April 2016 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25049

was in fact a template homologous (or even analogous—with the same fold but not related by evolution) to the target protein and producing a reasonable alignment. This challenge was referred to as "fold recognition"[24] and was a distinct category from the targets with obvious homologues in the PDB ("comparative modeling"); targets without any similar structures in the PDB were put in the "new fold" category.[25–28] Since CASP7, there have been two structure prediction categories, "template-based modeling" targets with both easy and hard-to-identify templates in the PDB (mostly homologues)[19] and "free modeling" targets (mostly without homologous templates).[29]

Numerous measures have been developed over the CASP experiments to assess the template-based modeling categories. Alignment accuracy was used as a measure through CASP8[21] (determined by calculating a target/prediction sequence alignment that results from sequence-independent structure alignment). In recent years, however, all of the assessment criteria have been based on either sequence-dependent structure alignment or comparing atom-atom distances within the prediction to the same atom pairs in the target experimental structure.[30] The former category includes the global measures GDT-TS,[31] GDT-HA (a high-resolution version of GDT-TS), GDC-SC (the ends of side chains are compared), and GDC-ALL (all atoms are compared), while the latter category includes the local structure assessment measures such as SphereGrinder (SpG),[32] LDDT,[33] and RPF.[23] The global measures are highly related to each other as are the local measures, while the local and global measures are less well correlated. Thus, the two sets of measures are complementary measures of structure prediction accuracy.

In this article, we present an assessment of template-based modeling in the CASP11 experiment. In CASP11, there were 39 target domains in the TBM category available to all predictors (both servers and human groups) and 42 target domains available to server groups alone (a total of 81 targets for servers). As with earlier CASPs,[34] some targets were broken into two or more assessment units (AUs), usually based on domain boundaries. This generally occurred when few groups were able to predict the relative orientation and distance of the domains with respect to each other.[35] This might occur because no template contained both domains, the arrangement of the domains was not the same in templates that contained them together, or the protein contained both TBM and free-modeling target domains.

In the current assessment, all prediction accuracy measures and *Z* scores were calculated by the Structure Prediction Center at the University of California Davis directed by Krzysztof Fidelis.[30] We experimented with a number of combinations of different measures and settled on a combination of *Z* scores of two global distance measures (GDT-HA and GDC-ALL), *Z* scores of two local distance measures (SpG and LDDT), and a scaled *Z* score of the MolProbity scores[36] (which was also used in CASP10). The MolProbity scores have the affect of urging predictors to attain statistically reasonable Ramachandran map positions, good rotamers and Cα positions, and few steric clashes.[37] Such factors make the models more useful for starting molecular dynamics simulations for instance. The MolProbity *Z* scores were multiplied by 0.2, since they were spread over a much larger range than the global and local prediction accuracy *Z* scores.

We observed a bimodal distribution of each measure for most of the targets. Multidimensional scaling of an RMSD-based distance matrix demonstrated that one peak in each distribution consisted of a cluster of models resembling the native structure (to a greater or lesser degree, depending on the target), while the lower peak was in fact a broad distribution of models not resembling each other or the native structure.

METHODS

Measures of model quality

Root mean square deviation (RMSD)

RMSD is the most commonly used metric for structure comparison. It is calculated as root mean of the sum of squared distances between corresponding atoms of two structures. CASP reports the RMSD calculated by the program LGA[31] which uses a sequence-dependent algorithm to maximize the number of Cα atoms of the model that are within 4 Å of the target structure. We utilized the program Theseus[38,39] to calculate RMSDs. Theseus uses a maximum likelihood, sequence-dependent superposition method that aligns the most similar regions of two (or more) structures and allows greater variance in divergent regions, such as loops. We used these RMSD values for multi-dimensional scaling with the *R* Project program (http://www.r-project.org), using the *R* command *cmdscale*. The top 10 eigenvectors and eigenvalues (representing the greatest variance in the data) were generated.

Global distance test scores

There are four related Global distance test scores routinely used for assessment of models in CASP. Global distance test - total score (GDT-TS) was first used in CASP4 to overcome the limitations of RMSD. It is computed by performing four different superpositions using LGA, each maximizing the number of Cα atom pairs (one from the model and one from the target at the same position in the sequence) within different cutoff distances. GDT-TS is the average percentage of Cα-atoms in the target within 1, 2, 4, and 8 Å after LGA superpositions with these cutoffs:

$$\text{GDT-TS} = \frac{1}{4}(\text{GDT\_}P_{1.0} + \text{GDT\_}P_{2.0} + \text{GDT\_}P_{4.0} + \text{GDT\_}P_{8.0})$$

where GDT_P_d denotes percent of residues under distance cutoff $\leq d$ Å. Global distance test – high accuracy (GDT-HA) is computed in the same manner as GDT-TS but with smaller cut-off values of 0.5, 1, 2, 4 Å:

$$\text{GDT-HA} = \frac{1}{4}(\text{GDT\_}P_{0.5} + \text{GDT\_}P_{1.0} + \text{GDT\_}P_{2.0} + \text{GDT\_}P_{4.0})$$

It is more sensitive to smaller changes in the structure of the model. Global distance calculation for side chains (GDC-SC) is GDT-like metric which uses a characteristic atom near the end of each side-chain type (instead of Cα atoms) for the evaluation of residue-residue distance deviations:

$$\text{GDC-SC} = 2 * (k * \text{GDC-}P_{1.0} + (k-1) * \text{GDC-}P2.0 \ \ldots \ + \ 1 * \text{GDC-}Pk)/(k+1) * k, \quad k=10$$

where GDC-P_k denotes percent of residues under distance cutoff <= 0.5k Å. Similarly Global distance calculation for all atoms (GDC-ALL) is calculated for all atoms of the protein.

Local distance difference test (LDDT)

The LDDT measure evaluates the quality of the model in terms of preserving local interactions in the structure.[33] The atom-atom distances in the model are compared with the corresponding distances in the native structure. An interaction is presumed to be preserved if the difference between the corresponding atom–atom distances of prediction and native is below a specific cutoff. The final score is calculated by averaging the number of correct interactions in the model for four different difference cutoffs of 0.5, 1, 2, and 4 Å.

Sphere Grinder score (SpG)

SpG is an all-atom score evaluating the quality of the model based on its similarity to local substructures of the native structure.[40] For Cα atom of every residue in the experimental structure the residues within a radius of 6 Å are identified. RMSD is calculated only for this set of atoms between the model and native structure. The mean percentage of number of Cα atoms below two different cutoffs for which this RMSD is ≤ 2.0 Å and ≤ 4.0 Å is defined as the SpG score.

Recall, precision, and F-measure score (RPF)

RPF is conceptually similar to LDDT score and was first introduced in CASP10.[23] First, the pairwise atom–atom distances are calculated in the model and native structure. Then a graph is constructed where the atoms of the structure represent the vertices and the interactions (if shorter than a fixed cutoff) between them represent the edges. The distances which are under a specific cutoff in both the structures are considered to be true positives (TP). On the other hand, the atom pairs which are below the cutoff in native but above it in the model are considered to be false negatives (FN). However, for the opposite case the atom pairs are considered to be false positives (FP) when they are below the cutoff in model and above it in the native structure. The cutoff used in CASP is 9 Å. From these numbers, a value of F is calculated from the True-positive rate (TPR = TP/(TP + FN)) and the Positive predictive value (PPV = TP/(TP + FP)):

$$F = \frac{(1+b)^2 \text{PPV} \times \text{TPR}}{b^2 \text{PPV} + \text{TPR}}$$

$$\text{DP} = \frac{F_{\text{model}} - F_{\text{random}}}{1 - F_{\text{random}}}$$

where F_{model} and F_{random} are the F scores comparing the native structure with the model or with a random polypeptide respectively. In essence, it reports the similarity of atom-atom contacts in the model and target.

MolProbity

MolProbity validates the physical feasibility of the predicted structure. A model with good prediction of overall backbone orientation might still have clashes or bad rotamers in the structure. Molprobity penalizes such instances and hence is an important tool in assessing the geometrical quality of models. This is a knowledge-based metric which is derived from analysis of a large number of high resolution protein structures. The score is composed of four components which include a clash score (clashscore, the number of all-atom steric overlaps > 0.4 Å per 1,000 atom), a rotamer outlier score (rota_out, the percentage of side-chain conformations classified as rotamer outliers, from those side chains that can be evaluated), and a Ramachandran outlier score (rama_out, the percentage of residues with a ϕ,ψ angles outside of the favored regions of the Ramachandran maps as determined by kernel density estimates).

$$\begin{aligned} MPScore = {} & 0.426 \log(1 + clashscore) \\ & + 0.33 \log(1 + \max(0, rota\_out - 1)) \\ & + 0.25 \log(1 + \max(0, (100 - rama\_out - 2))) + 0.5 \end{aligned}$$

Z-scores

Predictors were allowed to submit up to five models and requested to rank them from 1 to 5 in decreasing order of preference. The Z scores were calculated

Table I
Targets Predicted by Human-servers

Target	Domain	UniProt	PDB	Taxon	GDT-HA-Median[a]	Best Model	GDT-HA-Max	Seq identity	PDBBA	PISABA
T0826	D2	Q7DD94_NEIMB	4KAV	Bacteria	90.13	TS008_1	95.11	99.7	A	A
T0759	D1	PEPL_HUMAN	4Q28	Human	77.21	TS044_1	91.91	14.7	A	A
T0773	D1	designed	2N2U	Synthetic	50.37	TS336_2	85.45	14.1	A	A
T0816	D1	Y1502_ARCFU	5A1Q	Archaea	28.68	TS428_1	74.64	19.6	A2	A2
T0759	D2	PEPL_HUMAN	4Q28	Human	28.63	TS328_5	71.37	20.0	A	A
T0769	D1	designed	2MQ8	Synthetic	48.45	TS120_5	69.59	17.1	A	A
T0765	D1	MZRA_KLEP7	4PWU	Bacteria	39.47	TS403_5	67.76	17.9	A2	A4
T0820	D2	(marine phage)	—	Virus	31.25	TS328_1	63.89	67.8	—	—
T0799	D4	Q5DMH0_BPT5	4UW8	Virus	4.90	TS204_2	60.44	22.6	A3	A3
T0828	D2	A4TUL6_9PROT	4Z29	Bacteria	29.76	TS204_5	60.41	20.3	A	A
T0795	D1	F4MI11_9ADEN	—	Virus	52.75	TS340_5	59.93	18.1	—	—
T0848	D1	YCDA_BACSU	4R4Q	Bacteria	43.12	TS184_3	58.88	22.1	A	A
T0783	D1	ISPD_HUMAN	4CVH	Human	51.65	TS346_1	58.64	30.5	A2	A2
T0810	D2	Q6MI90_BDEBA	—	Bacteria	50.11	TS439_1	57.45	34.5	—	—
T0794	D1	VNN1_HUMAN	4CYF	Human	46.44	TS044_3	56.34	24.2	A	A
T0828	D1	A4TUL6_9PROT	4Z29	Bacteria	15.18	TS425_1	50.59	16.4	A	A
T0853	D2	Q99T58_STAAM	2MQB	Bacteria	24.82	TS296_4	49.31	13.8	A	(NMR)
T0808	D1	A5ZGP5_9BACE	4QHW	Bacteria	41.79	TS333_3	48.66	14.8	A	A
T0767	D1	Q931R5_STAAM	4QPV	Bacteria	22.37	TS347_4	48.35	5.9	A	A
T0793	D4	G7ZLR1_STAAU	—	Bacteria	22.06	TS499_5	47.06	7.1	—	—
T0827	D1	B8H2Q8_CAUCN	—	Bacteria	14.12	TS328_2	46.37	10.2	—	—
T0783	D2	ISPD_HUMAN	4CVH	Human	29.64	TS358_1	42.15	12.7	A2	A2
T0803	D1	U3USG4_PEPDI	4OGM	Bacteria	27.61	TS357_1	40.48	25.0	A	A
T0838	D1	R7NY94_9BACE	—	Bacteria	28.77	TS042_1	38.30	11.1	—	—
T0830	D2	Q1LDT6_CUPMC	5EZM	Bacteria	20.05	TS347_1	37.62	9.2	A	A
T0853	D1	Q99T58_STAAM	2MQB	Bacteria	28.94	TS144_2	37.50	20.7	A	(NMR)
T0822	D1	A0A011VZQ5_RUMAL	—	Bacteria	24.34	TS216_4	35.75	14.0	—	—
T0835	D1	A9R6D9_YERPG	—	Bacteria	24.75	TS333_5	34.59	12.0	—	—
T0831	D1	SHPRH_HUMAN	4QN1	Human	23.15	TS153_2	32.58	15.6	A	A
T0814	D3	A7AEG3_9PORP	4R7F	Bacteria	28.91	TS204_3	31.08	23.4	A2	A2
T0830	D1	Q1LDT6_CUMPC	5EZM	Bacteria	23.77	TS044_4	28.12	17.8	A	A
T0774	D1	A6L3B5_BACV8	4QB7	Bacteria	21.32	TS216_3	27.33	19.7	A	A4
T0793	D3	G7ZLR1_STAAU	—	Bacteria	25.13	TS038_5	27.26	16.1	—	—
T0818	D1	B0MNI9_9FIRM	4R1K	Bacteria	21.08	TS281_1	26.68	13.8	A	A
T0800	D1	J7TBR0_CLOSG	4QRK	Bacteria	21.40	TS173_2	26.53	12.1	A	A
T0781	D2	A7B4B4_RUMGN	4QAN	Bacteria	21.71	TS044_4	26.29	17.2	A	A
T0799	D3	Q5DMH0_BPT5	4UW8	Virus	17.16	TS360_1	25.49	18.6	A3	A3
T0848	D2	YCDA_BACSU	4R4Q	Bacteria	18.72	TS333_5	24.32	15.2	A	A
T0812	D1	LAMA2_HUMAN	4YEQ	Human	22.20	TS067_3	23.91	12.2	A	A

[a]GDT-HA median computed from Model 1s.
PDBBA is the first biological assembly given by the PDB (A2 = homodimer, etc.).
PISABA is the biological assembly given by PISA for crystal structures.

separately for model_1s and all models and also for server groups and all groups. When calculated for model 1s, it is a measure of the quality of a model 1 submitted by a group with respect to the entire set of model 1s submitted for that particular target by all the groups. In the first step, Z scores using a certain measure were calculated for each model 1 of every target from the average and standard deviation of all the model 1s submitted by all the groups. In the next step, all the models that were worse than two standard deviations from the mean are removed. The mean and standard deviation were then recalculated on the remaining set, and new Z scores were then calculated on the entire set. All the models with a new Z score less -2.0 were assigned a value of -2.0. The final Z scores summed over targets were then used to rank the performance of participating groups. The same procedure was performed on the set of all models for each target.

RESULTS

A total of 143 groups participated in the template-based modeling (TBM) category of CASP11. This includes 99 human-curated groups and 44 prediction servers. There were a total of 26,028 submissions from all the groups. The Structure Prediction Center allows the groups to submit up to five predictions for every target. We assessed model 1 from each group for each target, and separately the best model from each group for

Table II
Targets Predicted by Servers

Target	Domain	UniProt	PDB	Taxon	GDT-HA-Median[a]	Best Model	GDT-HA-Max	Seq identity	PDBA	PISAA
T0766	D1	A7V9L7_BACUN	4Q53	Bacteria	84.72	TS008_5	87.96	65.1	A	A
T0784	D1	A7M5D7_BACO1	4QEY	Bacteria	72.10	TS008_3	84.80	58.7	A	A
T0854	D1	D7AL49_GEOSK	4RN3	Bacteria	70.83	TS184_3	83.14	29.8	A	A
T0811	D1	TPIS2_RHIME	—	Bacteria	72.46	TS335_4	78.58	37.9	—	—
T0815	D1	A0A077HY11_RHIML	4U13	Bacteria	68.40	TS479_3	76.89	16.5	A2	A2
T0762	D1	I6L927_STRMU	4Q5T	Bacteria	67.31	TS156_5	71.69	38.8	A	A
T0768	D1	A6NQU6_9FIRM	4OJU	Bacteria	47.90	TS216_1	71.68	38.1	A4	A4
T0801	D1	RFFA_ECOLI	4PIW	Bacteria	63.20	TS184_4	70.35	31.6	A2	A2
T0807	D1	V3S9R5_KLEPN	4WGH	Bacteria	58.30	TS335_2	66.52	27.0	A	A
T0819	D1	Q92R63_RHIME	4WBT	Bacteria	55.82	TS184_4	66.42	23.8	A2	A2
T0817	D1	Q92YH7_RHIME	4WED	Bacteria	57.21	TS156_1	66.32	27.1	A	A
T0817	D2	Q92YH7_RHIME	4WED	Bacteria	59.64	TS184_2	66.07	28.3	A	A
T0782	D1	A7V1A8_BACUN	4QRL	Bacteria	36.59	TS184_3	65.23	12.9	A	A
T0776	D1	R6WSE2_9PORP	4Q9A	Bacteria	60.50	TS184_2	64.50	33.7	A2	A2
T0813	D1	Q92MG1_RHIME	4WJI	Bacteria	54.92	TS008_1	63.41	34.7	A2	A2
T0833	D1	A6LGW3_PARD8	4R03	Bacteria	46.87	TS420_3	62.27	20.9	A	A2
T0856	D1	HERC1_HUMAN	4QT6	Human	52.59	TS277_1	62.26	17.7	—	A
T0764	D1	A6LCA7_PARD8	4Q34	Bacteria	56.85	TS011_1	61.37	34.5	A2	A2
T0854	D2	D7AL49_GEOSK	4RN3	Bacteria	50.72	TS184_3	60.36	12.5	A	A
T0805	D1	W6H7D8_MYCTX	—	Bacteria	49.87	TS184_4	59.77	22.5	—	—
T0843	D1	Q0H2X1_9ACTO	4OCA	Bacteria	54.98	TS110_3	59.42	29.6	A2	A2
T0780	D2	Q97PP3_STRPN	4QDY	Bacteria	36.72	TS184_5	59.37	20.3	A2	A2
T0851	D1	Q8KNF6_MICEC	4XRR	Bacteria	46.80	TS479_3	59.30	29.7	A2	A2
T0792	D1	OSKA_DROME	5A49	Drosophila	45.99	TS184_4	58.98	30.4	A2	A2
T0847	D1	FA83A_HUMAN	4URJ	Human	53.10	TS184_3	58.58	20.6	A	A2
T0858	D1	Q8A2J3_BACTN	—	Bacteria	53.22	TS452_4	58.39	26.7	—	—
T0839	D1	SLA2_SCHPO	—	Fungi	38.92	TS420_1	58.10	17.2	—	—
T0760	D1	A5ZGW9_9BACE	4PQX	Bacteria	44.83	TS041_1	57.71	29.2	A	A
T0829	D1	Q1CIG2_YERPN	4RGI	Bacteria	36.20	TS277_4	56.34	11.4	A	A
T0780	D1	Q97PP3_STRPN	4QDY	Bacteria	50.52	TS499_1	54.47	21.4	A2	A2
T0852	D1	C7M590_CAPOD	4W9R	Bacteria	48.82	TS184_1	54.06	19.1	A2	A2
T0845	D1	Q8A7N7_BACTN	4R5O	Bacteria	34.41	TS251_4	52.58	15.5	A	A3
T0772	D1	A6LIT1_PARD8	4QHZ	Bacteria	43.69	TS381_4	49.18	30.6	A4	A4
T0786	D1	Q73EW7_BACC1	4QVU	Bacteria	41.13	TS184_5	49.08	17.2	A4	A4
T0821	D1	R6X2D1_9PORP	4R7S	Bacteria	35.98	TS216_3	49.02	16.4	A	A
T0770	D1	C6IS27_9BACE	4Q69	Bacteria	41.55	TS492_1	48.30	23.5	A2	A2
T0849	D1	D0LNW1_HALO1	4W66	Bacteria	38.45	TS073_2	46.29	22.3	A2	A2
T0823	D1	B8H047_CAUCN	—	Bacteria	37.16	TS184_2	42.10	19.7	—	—
T0852	D2	C7M590_CAPOD	4W9R	Bacteria	22.42	TS277_2	40.08	10.8	A2	A2
T0845	D2	Q8A7N7_BACTN	4R5O	Bacteria	31.16	TS216_2	36.78	18.9	A	A
T0796	D1	A7IZR5_BACTU	4PKM	Bacteria	29.48	TS210_1	36.65	15.9	A2	A2
T0857	D1	A6KYV6_BACV8	2MQC	Bacteria	17.97	TS454_1	34.12	14.5	A	(NMR)

[a]GDT-HA median computed from Model 1s.
PDBBA is the first biological assembly given by the PDB (A2 = homodimer, etc.).
PISABA is the biological assembly given by PISA for crystal structures.

each target. The model 1 results are considered the fairer assessment, since the best model assessment favors groups that provide five models for every target over those that provide fewer than 5.

All the scores for quality assessment of predictions were computed by the Protein Structure Prediction Center.[30] A total of 81 domains from 69 target proteins were available in CASP11 which included 39 human-server (HS) and 42 server only (S) domains, listed in Tables I and II, respectively. The Structure Prediction Center determined which target proteins were made available for servers or for both human and server groups. The tables include the Uniprot code, the PDB code (if available), the likely biological assembly (if available), and the median and maximum GDT-HA scores of the predictions. The maximum GDT-HA (per target) varied from 23.91 to 95.11 for HS targets and 34.12 to 87.96 for S targets. Moreover, the sequence identity of the most closely related template varied from 5.9 to 99.7 for HS targets and from 11.4 to 65.1 for S targets, suggesting inclusion of structures with varying level of difficulty.

Quality of predictions based on sequence identity

The sequence identity of the most closely related template to the target (as determined by structure alignment of the target structure to homologues in the PDB[35]) is a

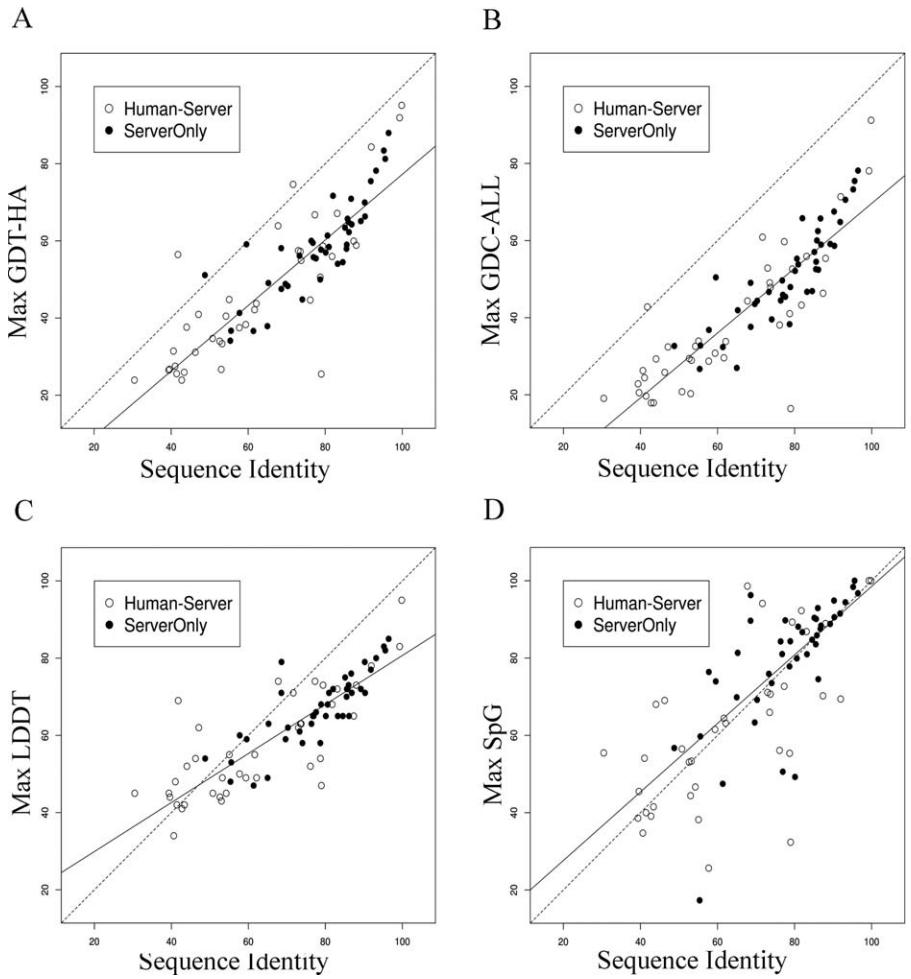

Figure 1

Relationship between sequence identity of the closest template to the target experimental structure and the maximum structure prediction accuracy for each target across all models for different global and local quality assessment measures: (**A**) GDT-HA, (**B**) GDC-ALL, (**C**) LDDT, (**D**) SpG. The line of regression is computed by robust linear regression fitting over all the data points and is shown as a solid line. The dotted line represents x = y.

simple measure of target difficulty, one that can be measured prior to structure determination. We first analyzed the relationship between the sequence identity and the structure prediction accuracy of the best models. Figure 1 shows the maximum value (over all models) of GDT-HA [Fig. 1(A)], GDC-ALL [Fig. 1(B)], LDDT [Fig. 1(C)], and SpG [Fig. 1(D)] versus the sequence identity of the template. The S targets (solid circles in Fig. 1) generally had higher sequence identity to the template than the HS targets (open circles in Fig. 1). The maximum GDT-HA and GDC-ALL values are highly correlated with sequence identity, although there are (not surprisingly) significant outliers since sequence identity is not a perfect measure of target difficulty. Both of GDT-HA and GDC-ALL are global measures of quality assessment suggesting, as expected, that better backbone orientation is predicted if the template is more closely related to the model. However, with the local measures like SpG and LDDT, the relationship with sequence identity is only roughly linear indicating that although good sequence identity leads to better overall fold, in some cases the local substructures can still be poorly defined. The maximum SpG values are the least correlated with sequence identity, suggesting that despite good templates, accurate local packing is difficult to achieve for many targets. Overall, the quality of prediction with respect to sequence identity with the template is strongly correlated with global measures and less strongly with local ones.

Performance of groups based on single and pairwise measures

Z scores for each measure were provided by the structure prediction center along with a means for testing out various combinations of measures and their weights. The performance of different groups on the 39 HS targets

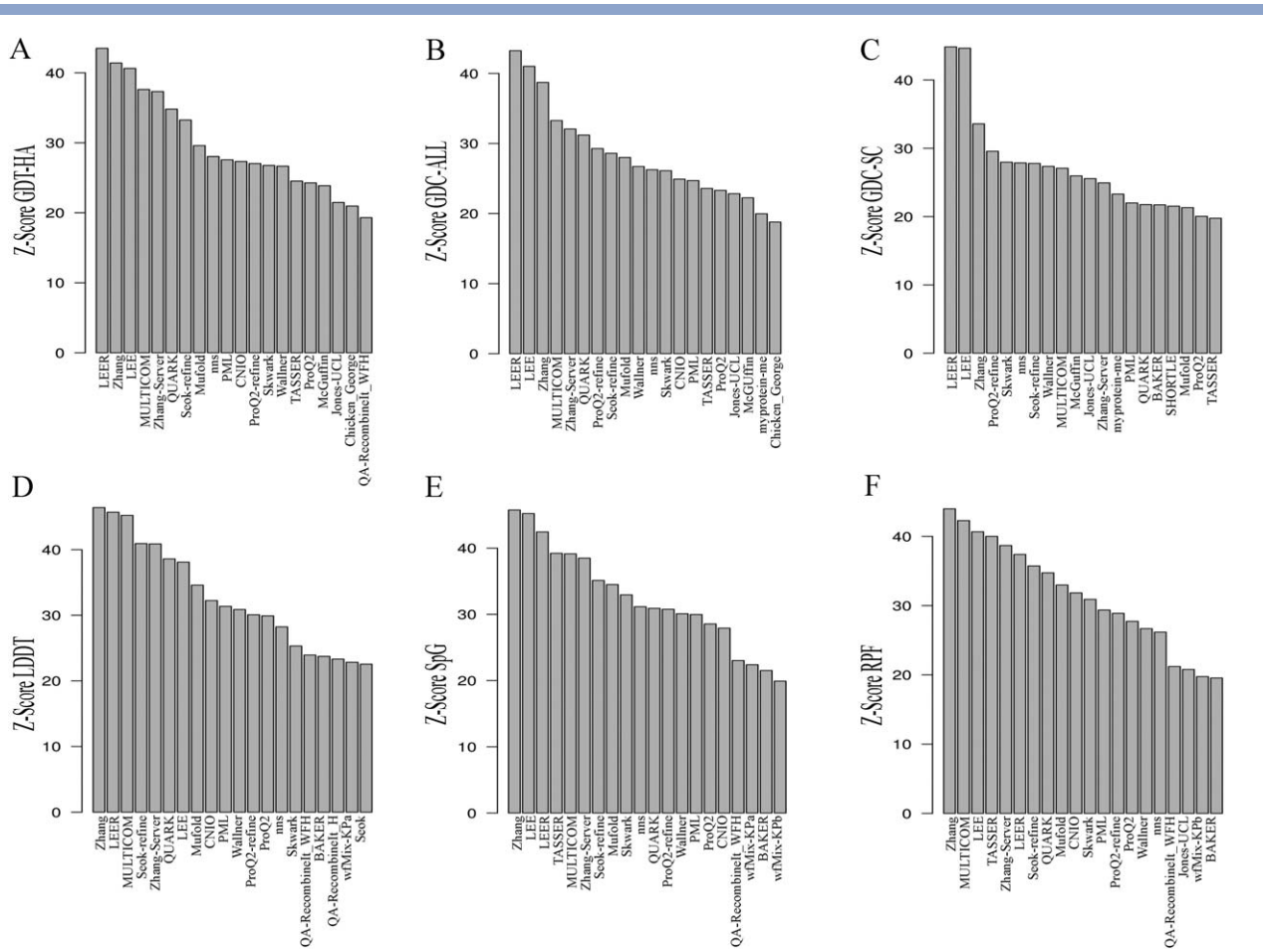

Figure 2

Summary of *Z* scores summed over targets for single metrics of model 1 of top 20 groups for each measure. The *x* axis displays the groups names. The *y* axis is the summed *Z* score for each single metric: (**A**) GDT-HA, (**B**) GDC-ALL, (**C**) GDC-SC, (**D**) LDDT, (**E**) SpG, and (**F**) RPF.

based on the *Z* score of six different single measures, three global and three local, is shown in Figure 2. Based on the global scores (GDT-HA, GDC-ALL, and GDC-SC) LEE, LEER, and Zhang are the top performing groups. The *Z* scores of GDT-HA [Fig. 2(A)] and GDC-ALL [Fig. 2(B)] of the three top groups are comparable but in the case of GDC-SC [Fig. 2(C)], LEER, and LEE outperform Zhang and other groups by a considerable margin. The assessment from *Z* scores of local measure like LDDT, SpG, and RPF also display leading performances by the same three groups, joined by MULTICOM in third place in LDDT and second place in RPF. While the top performing groups exhibit good performance for all the measures, a few other groups display excellent performance in some scores but not others. For instance, TASSER exhibits good performance in local scores like SpG and RPF but performs less well in overall backbone prediction scores like GDT-HA and GDC-ALL. On the other hand, there are groups who perform better with GDT-HA and GDC-ALL but do not do as well in GDC-SC and SpG.

Next, we analyzed the performance of groups by summing up *Z* scores for global and local measures (Fig. 3). In this assessment also LEE, Zhang and LEER outperform the other predictors in most of the combinations, accompanied by MULTICOM and Zhang-Server who also exhibit consistent performance across different combinations of quality metrics. Because every measure exploits a different feature of the model, the ranking of groups change slightly with different combinations of measures used. A fair numerical assessment of groups therefore requires a sum of certain measures which reflects a balanced assessment of both global and local features of the predictions. Therefore to find a suitable combination we have tried to understand the relationship between different metrics.

Relationship between different global and local measures

To evaluate correlations between the different metrics of assessment, we plotted the various metrics against

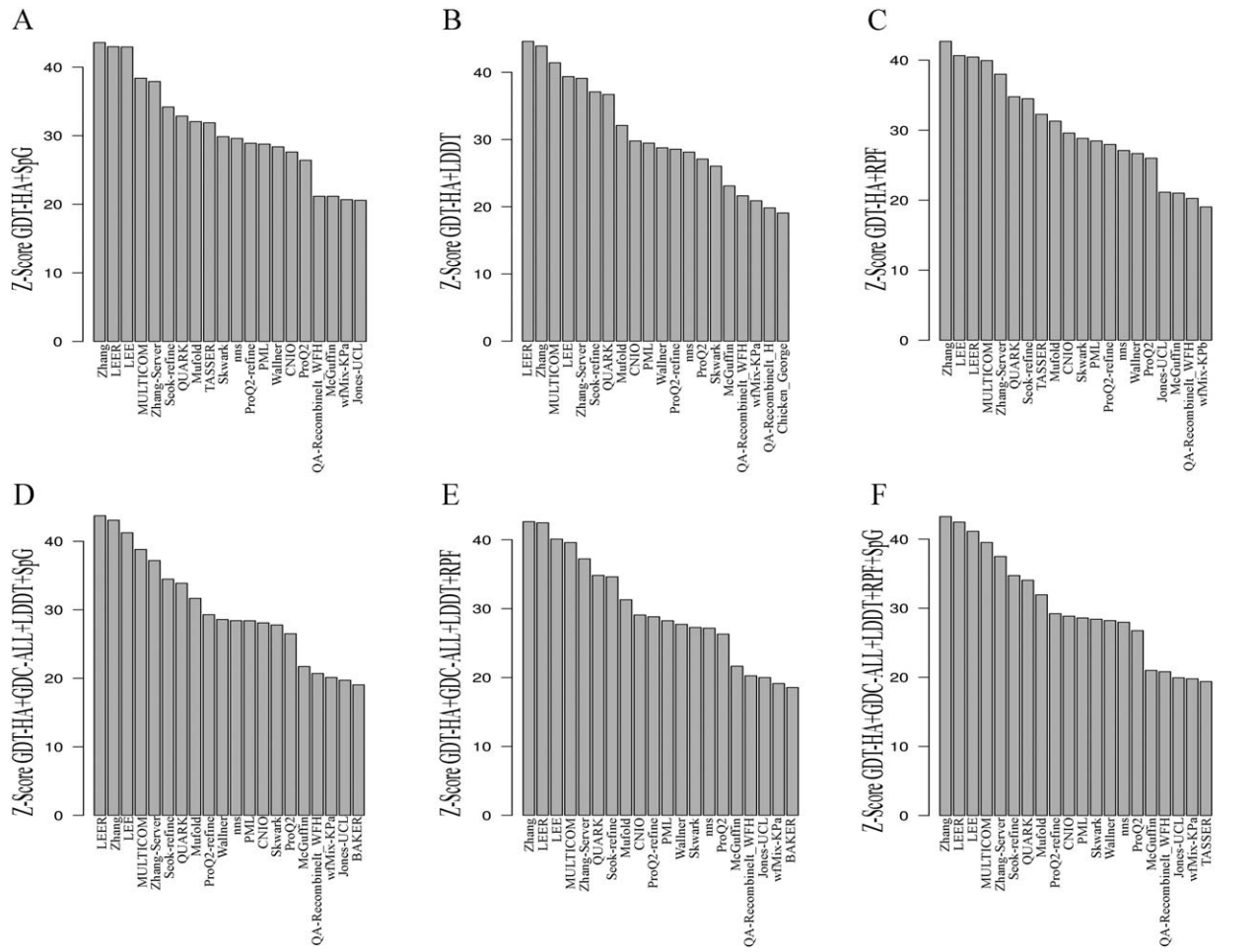

Figure 3

Summary of sum of *Z* scores of metrics of model 1 of top 20 groups for various combinations of global and local metrics. The *x* axis displays the group names. The *y* axis is the *Z* score for sum of metrics: (**A**) GDT-HA + SpG, (**B**) GDT-HA + LDDT, (**C**) GDT-HA + RPF, (**D**) GDT-HA + GDC-ALL + LDDT + SpG, (**E**) GDT-HA + GDC-ALL + LDDT + RPF, and (**F**) GDT-HA + GDC-ALL + LDDT + RPF + SpG.

each other for the models with the best value for each metric. This exercise helped us to identify a suitable combination of measures for assessment of groups. The relationship of GDT-HA with GDT-TS [Fig. 4(A)] and GDC-ALL [Fig. 4(B)] is nearly linear, which is expected. These metrics are superposition-based global measurements of quality. Figure 4(C) shows the relationship between the best MolProbity score (lower is better) and the best GDT-HA score for each target. As expected, there is no correlation since it is possible to achieve dihedral angle values close to known statistical distributions and few clashes within a model that is not at all similar to the experimental structure of the target. Generally, much better MolProbity scores are achieved when the human groups are included (the HS targets) than with the server groups alone (S targets).

However, the correlations between GDT-HA and the local metrics, LDDT [Fig. 4(D)], RPF [Fig. 4(E)], and especially SpG [Fig. 4(F)] are relatively weak. The comparisons demonstrate that the local and global metrics provide distinctly different pieces of information. Plotting RPF against LDDT indicates that these measures are highly correlated [Fig. 4(G)]. This redundancy is also observed when both of these measures are plotted against SpG [Fig. 4(H,I)]. This high correlation of LDDT and RPF is probably because both are distance matrix-based measures. Both of them in essence represent the similarity of atom–atom contacts between the predicted and native structure.

Performance of groups based on sum of different metrics

It is pertinent that the performance of different groups is assessed by both local and global features while remaining non-redundant. Therefore, we have used the sum of *Z* scores for five measures (GDT-HA + GDC-ALL + LDDT + SpG + 0.2MolP) that are related to each

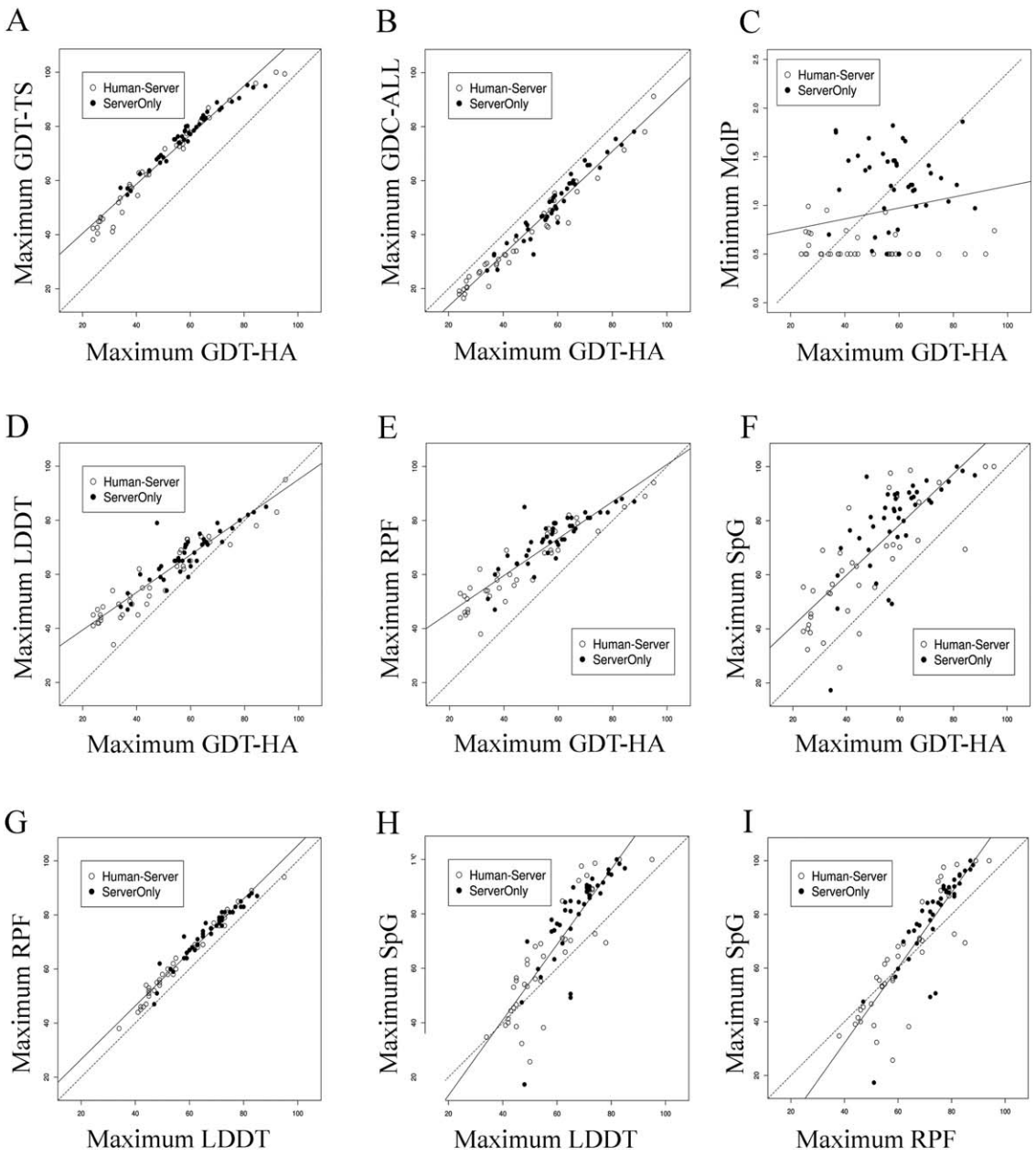

Figure 4

Relationship between different global and local quality assessment measures. Pairwise comparison of maximum GDT-HA for every target is done with (**A**) Maximum GDT-TS, (**B**) Maximum GDC-ALL, (**C**) Minimum (best) MolProbity score, (**D**) Maximum LDDT, (**E**) Maximum RPF, and (**F**) Maximum SpG. Pairwise comparison of maximum LDDT for every target is done with (**G**) Maximum RPF, (**H**) Maximum SpG, and (**I**) Maximum SpG. The line of regression is computed by robust linear regression fitting over all the data points and is shown in a solid line.

other in a useful and non-redundant way. While GDT-HA depends heavily on the accuracy in Cα-atom positions, GDC-ALL shows the confidence in prediction of all atoms. On the other hand LDDT reports the preservation of residue-residue contacts, and SpG examines the packing of local regions. The last quality measure used is MolProbity to assess whether the predicted model is physically feasible or not. The factor 0.2 in front of the MolProbity term is somewhat arbitrary, but the summed Z-scores per group for MolProbity had more than twice the range of the other metrics. Almost any model could be improved in its MolProbity score with some regularization of the structure, so we reduced the factor to 0.2.

The ranking based on sum of *Z* scores (GDT-HA + GDC-ALL + LDDT + SpG + 0.2MolP) for model 1 submitted by all the groups shows that LEER, Zhang, and LEE are the top ranked groups [Fig. 5(A)]. They are followed by MULTICOM, Zhang-Server, and Seok-refine.

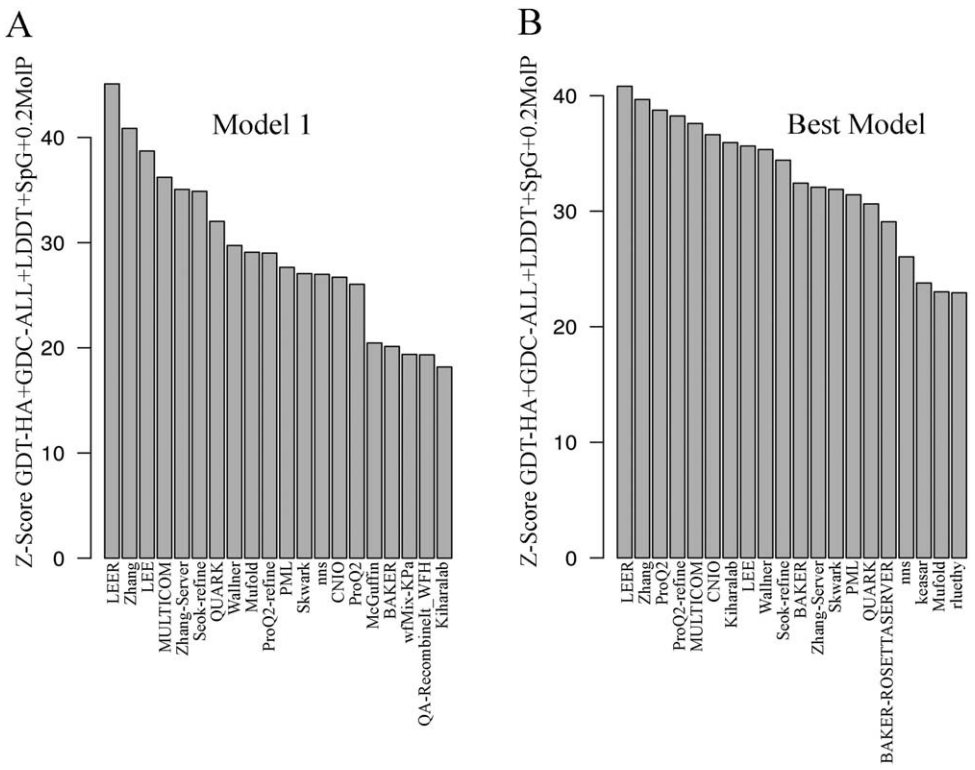

Figure 5

Final ranking of TBM predictor groups using a summed *Z* score consisting of GDT-HA + GDC-ALL + LDDT + SpG + 0.2MolP. The *x* axis displays the group names. The *y* axis is the sum of *Z* scores for computed for (**A**) Model 1 and, (**B**) Best model.

The pairwise bootstrap*p* values for the ranking are provided in Table S1 of the Supporting Information. Bootstrap*p* values were determined by randomly picking (with replacement) *N* target domains from the *N* target domains that two groups have in common, and comparing the total sum of *Z* scores. This procedure was repeated 1,000 times, and the table provides the fraction of times that the higher ranked group had a higher sum than the lower ranked group. LEER has a bootstrap*p* value with Zhang of 0.918, but values >0.99 with all other groups. Zhang has bootstrap*p* values with three of the next four groups below 0.95 (except Zhang-server), and >0.99 with all subsequent groups. Values above 0.95 may be used to show that the relative ranking is statistically significant at a 5% level. Thus LEER and Zhang are quite similar, while their ranks over most of the other groups are statistically significant.

The performance assessed by using the best models is shown in Figure 5(B). LEER and Zhang remain the top two groups, followed by ProQ2 and ProQ2-refine, which exhibit a drastically improved performance over model 1 assessments [Fig. 5(A)]. This is likely because both of these groups selected server models which were made available to predictors prior to the human group deadlines. As a result their best models are the same as the best models submitted by some of the top groups. The reduction in the total score across the top 10 groups is more gradual when calculated for best models than it is for model 1s. This is reflected in the bootstrap*p* values (Supporting Information Table S2), which show that the differences in rank are not significant at the 95% level for groups at least 5 apart in rank. This indicates that the best models submitted by many of the groups are of comparable quality.

Performance of servers on HS + S targets

We have also separately assessed the performance of server groups on the combined Human-Server and Server-only target domains, using the same sum of *Z* scores (GDT-HA + GDC-ALL + LDDT + SpG + 0.2MolP) as displayed in Figure 6. Mean values and standard deviations used to calculate the *Z* scores were based only on the raw scores of the models from the servers, even on the HS targets. Zhang-Server, nns, BAKER-ROSETTASERVER, QUARK, and myprotein-me are the top five groups when assessed using model 1 for all targets [Fig. 6(A)], followed by a significant drop off after the fifth-ranked group This is confirmed by the bootstrap*p* values in Supporting Information Table S3, which show values of ≥0.95 only from the sixth-ranked group and

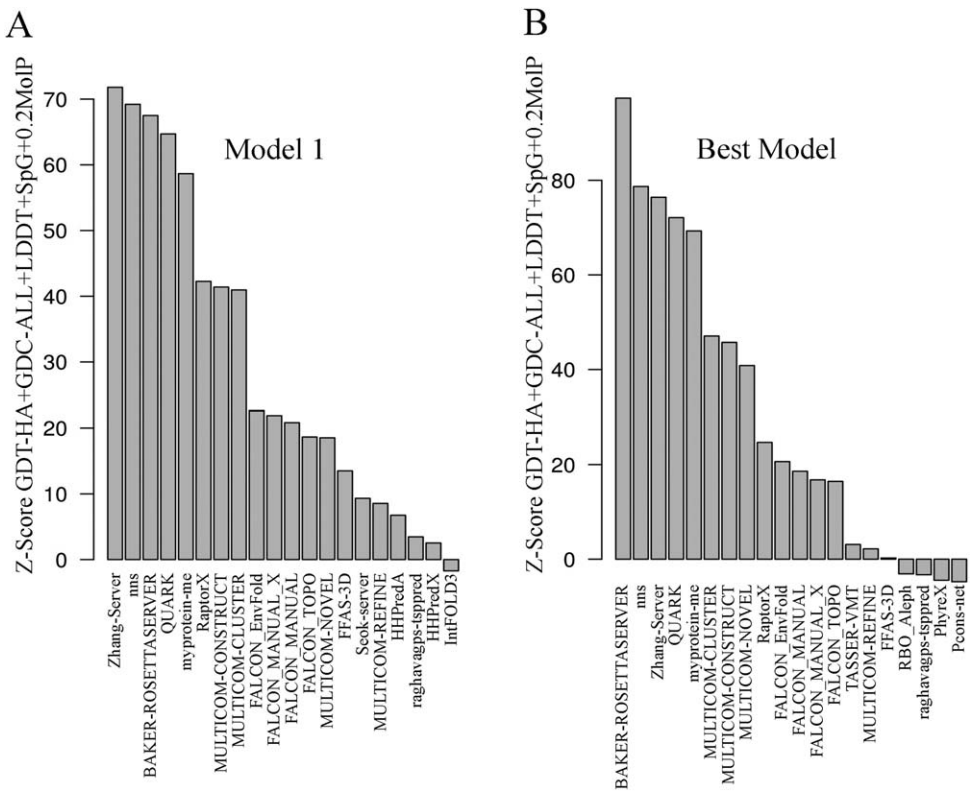

Figure 6

Final ranking of TBM server groups over the HS + S targets. The *x* axis displays the group names. The *y* axis is the sum of *Z* scores for GDT-HA + GDC-ALL + LDDT + SpG + 0.2MolP computed for (**A**) Model 1 and, (**B**) Best model.

onwards. Another significant drop off occurs after the eighth-ranked group.

When the best models are scored for the server groups [Fig. 6(B)], BAKER-ROSETTASERVER is ranked at the top by a significant margin as demonstrated by bootstrap*p* values >0.95 for all subsequently ranked groups (Supporting Information Table S4). The next four groups, nns, Zhang-Server, QUARK, and myprotein-me are all quite similar. We observe that in the case of the best models for servers on the HS + S targets, there is again a steep decline in the total score after the fifth group and again after the eighth group. The distribution of *Z* scores for models 1, 2, 3, 4, and 5 for BAKER-ROSETTASERVER were all quite similar to each other (not shown), indicating that the difference in ranks for model 1s and best models was not due to distinct protocols for each of the five models.

Performance based on molprobity

MolProbity is calculated from statistical analysis of a large number of protein crystal structures.[20,36] It gives negative scores to clashes, unfavorable Ramachandran dihedrals, bad Cβ positions, and bad rotamers—thereby identifying models that are not physically reasonable irrespective of the model's similarity to the native experimental structure. A higher MolProbity score represents a lower quality model. MolProbity *Z* scores were therefore calculated on inverted values. Physical reasonableness is an important component of structure prediction, and therefore we have examined the performance of different groups based on MolProbity separately (Fig. 7). We analyzed the separate components of the MolProbity score for the top-ranked groups (Fig. 5), with the addition of the STAP group, which had the highest sum of MolProbity Z-scores. The results are shown in Figure 8, and the groups are ordered by their total average MolProbity score (smaller values are better) [Fig. 8(A)]. The separate components include percent Cβ outliers [Fig. 8(B)], percent Ramachandran outliers [Fig. 8(C)], percent bad rotamers [Fig. 8(D)], and the number of clashes per 100 residues [Fig. 8(E)]. Several groups have up to 8% Cβ outliers, 10% Ramachandran outliers, nearly 15% bad rotamers, and up to 8 clashes per 100 residues. The groups differ in which components of the MolProbity score they performed well or poorly in. Seok-refine, PML, Zhang, Zhang-Server, QUARK, McGuffin, and TASSER have higher percentages of Cβ outliers than other groups; Zhang, Zhang-Server, QUARK, and TASSER have higher percentage of Ramachandran

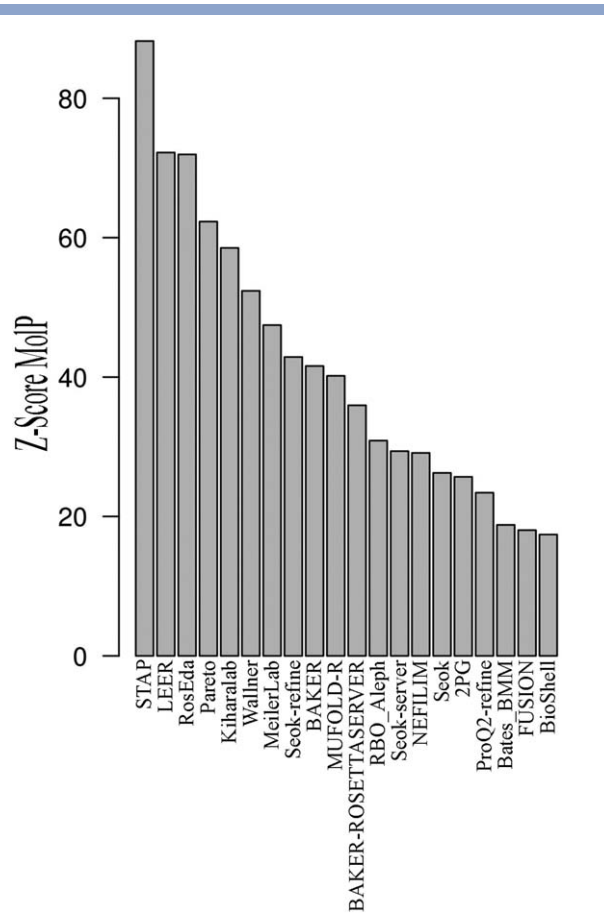

Figure 7

Summed *Z* score of MolProbity for model 1 of top 20 groups for all the targets.

outliers than other groups; PML, Zhang-Server, QUARK, and McGuffin have poor side-chain rotamer scores. A different set of groups performed worse than average on the clashscore [Fig. 8(E)], including Skwark, CNIO, nns, LEE, MULTICOM, Mufold, and TASSER. Because the clashscore [Fig. 8(E)] is weighted more heavily than the other MolProbity components (see Methods), even groups with moderate clashscores do not rank well in overall MolProbity scores [Fig. 8(A)]. The individual MolProbity scores can be utilized by predictors to focus their efforts on improving some aspects of their models in reference to the statistically probable values of bond angles, dihedral angles, and atom-atom distances.

Distribution of scores in predictions of all targets

Model-quality assessment programs have had some success in picking out accurate models from an ensemble of models generated by other programs,[41] and some groups using this method performed well in our assessment of the TBM models in CASP11.[42] We therefore were interested in the analyzing the ensemble of models for each target to determine how the models relate to one another, in addition to how they relate to the target experimental structure. Kernel density estimates of GDT-TS and RMSD (calculated with the program Theseus[38,39]) to native for all 39 human-server targets are shown in Figures 9 and 10, respectively. What is striking is that nearly all of the distributions are bimodal. The same is true of other accuracy measures, including GDC_ALL, LDDT, and SphereGrinder, and it is also true for the scores of most of the server-only target domains (not shown).

Such a phenomenon could occur because there are two (or more) distinct clusters of models, perhaps based on different templates, or it could occur if there is a single cluster of models that are relatively close to the target

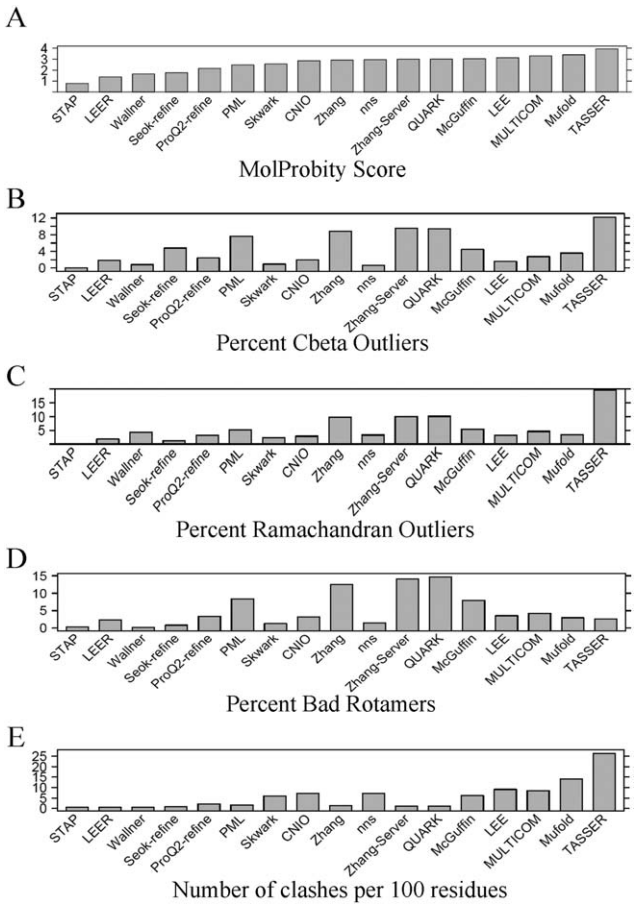

Figure 8

Assessment of different groups on the basis of MolProbity metric. The *x* axis displays the group names and *y* axis shows the score. The groups included are those that ranked highly in the model 1 assessments with the inclusion of the group with top MolProbity score (STAP). (**A**) Total MolProbity score (lower is better); (**B**) Percent Cβ outliers, (**C**) Percent Ramachandran outliers, (**D**) Percent bad rotamers, and (**E**) Number of clashes per 100 residues. In each panel, the groups are ordered by total MolProbity score.

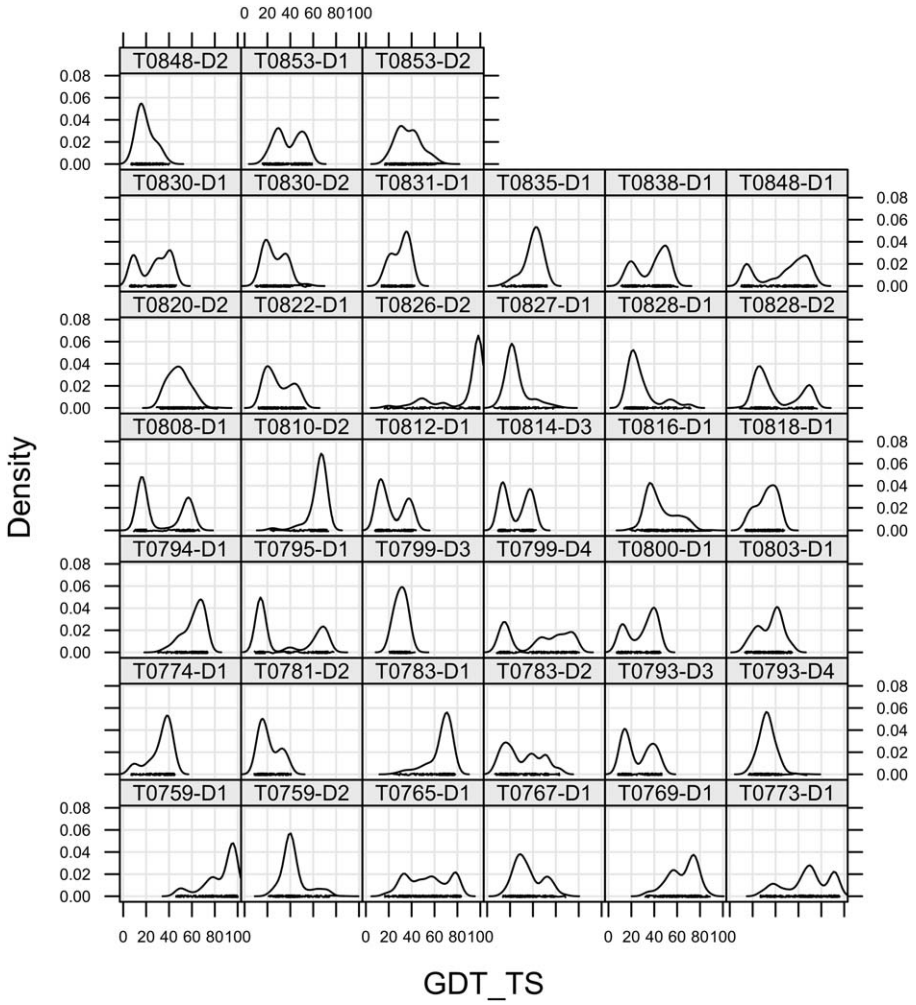

Figure 9

Kernel density estimates of the probability distribution of GDT-TS for 39 human-server targets.

and a distribution of poor models that are not similar to the target nor similar to each other. We performed multi-dimensional scaling (MDS) which represents higher dimensional data in a smaller number of dimensions, while approximately preserving the pairwise distances among all data points. In the current context, the conformational variation of the protein models is a space of $3N_{atoms}-6$. We applied MDS to visualize the structural variation among the models in two dimensions. Structurally distant models are separated from each other and similar models are clustered together on the plot. This requires calculation of pairwise RMSD matrix for all the models of a target. For this purpose we have performed pairwise structure alignments with the program Theseus[38,39] for each of the 39 TBM human-server targets in CASP11. Theseus uses a maximum-likelihood method that includes realistic assumptions that Cα-Cα distances nearby in sequence are correlated and that the variance of Cα-Cα distances along the chain are not uniform.

The results are shown for three representative targets in Figure 11 (T0795-D1), 12 (T0828-D2), and 13 (T0838-D1). In panel A of each figure, a scatterplot of the first two MDS coordinates is colored by RMSD to native (from blue to red) and four structures are marked—the native (panel C), the best model (panel D), a model near the upper mode in the GDT-TS density for the target (panel E), and a model near the lower mode in the GDT-TS density (panel F). The same structures are indicated in a scatterplot of GDT-TS versus RMSD (panel B). The symbols are shown next to the protein structure images.

It is evident that the second scenario explains the bimodal nature of the GDT-TS and RMSD densities in most cases, that is, that there is a single cluster of models relatively similar to the target and to each other, and a

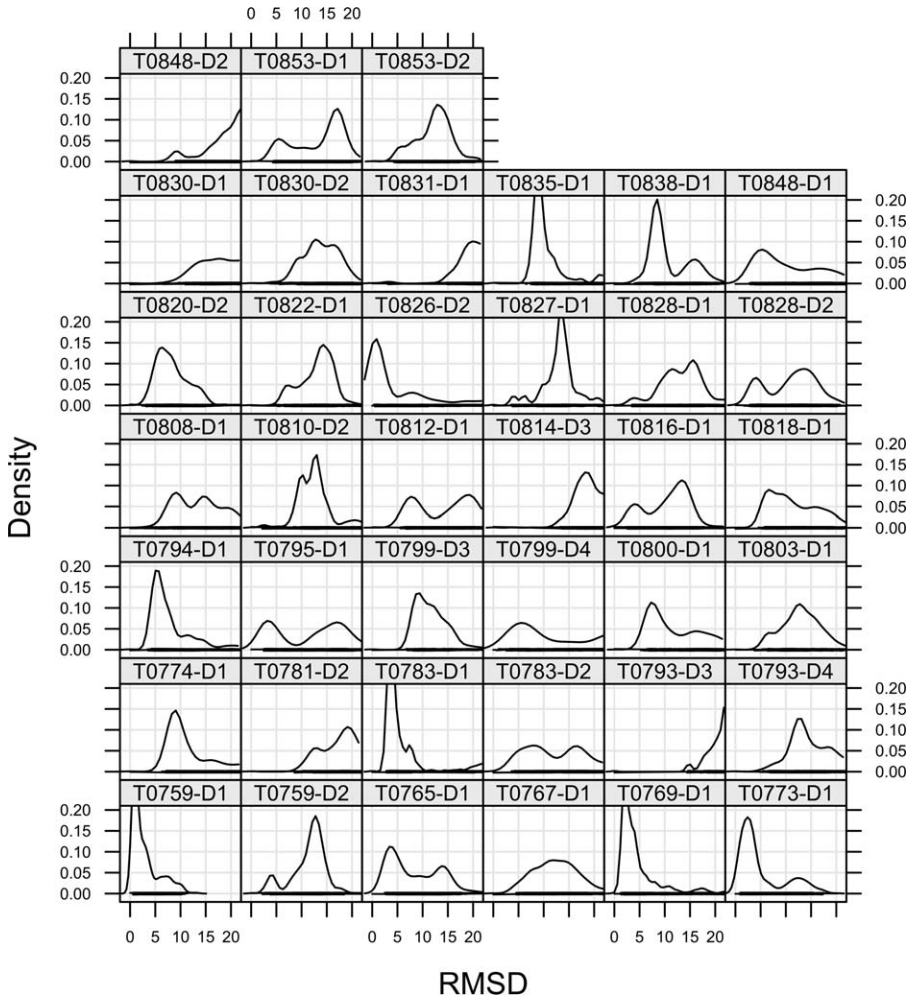

Figure 10

Kernel density plots of backbone RMSD to native for 39 human-server targets. The pairwise RMSD between models and native structure of each target is calculated by doing maximum likelihood superposition using the Theseus program.

broad distribution of other models that are not similar to one another or to the target. The best models for these targets have GDT-TS values of 77.8, 76.2, and 60.5 respectively [Figs. 11(D), 12(D), 13(D)], the models at the mode of the upper peak are at 68.6, 69.0, and 48.8 [Figs. 11(E), 12(E), 13(E)], and those of the lower peak are at 17.5, 26.5, and 19.4 [Figs. 11(F), 12(F), 13(F)]. The structures at the mode of the lower peak have very high RMSD (>14 Å) and are all in a widely scattered group of models in the MDS plots (square symbol, Figs. 11(A), 12(A), 13(A)]. A value of GDT-TS of around 20 appears easy to achieve with a reasonably compact model, even if the fold is entirely incorrect, and explains the presence of the lower peak for most of the targets in Figure 9.

We note that in each case, the proportion of variance in the first two components (eigenvectors) of the multi-dimensional scaling was low—34%, 22%, and 29%. We checked additional dimensions and it is evident that additional clusters do not form. The central cluster spreads out, which correlates with the observation that the native structure appears too close to models than it is in the original distance matrix.

CASP11 compared to previous CASPS

The assessment of progress in CASP in comparison with previous experiments has always been difficult due to two main reasons: a suitable metric for target difficulty is challenging, and the growth of sequence and structure databases means that methods may improve simply because of larger sequence alignments or the existence and effective utilization of multiple templates.

However, to make a broad comparison, we chose as a measure of target difficulty the GDT-TS of a model based on copying the coordinates of the best template

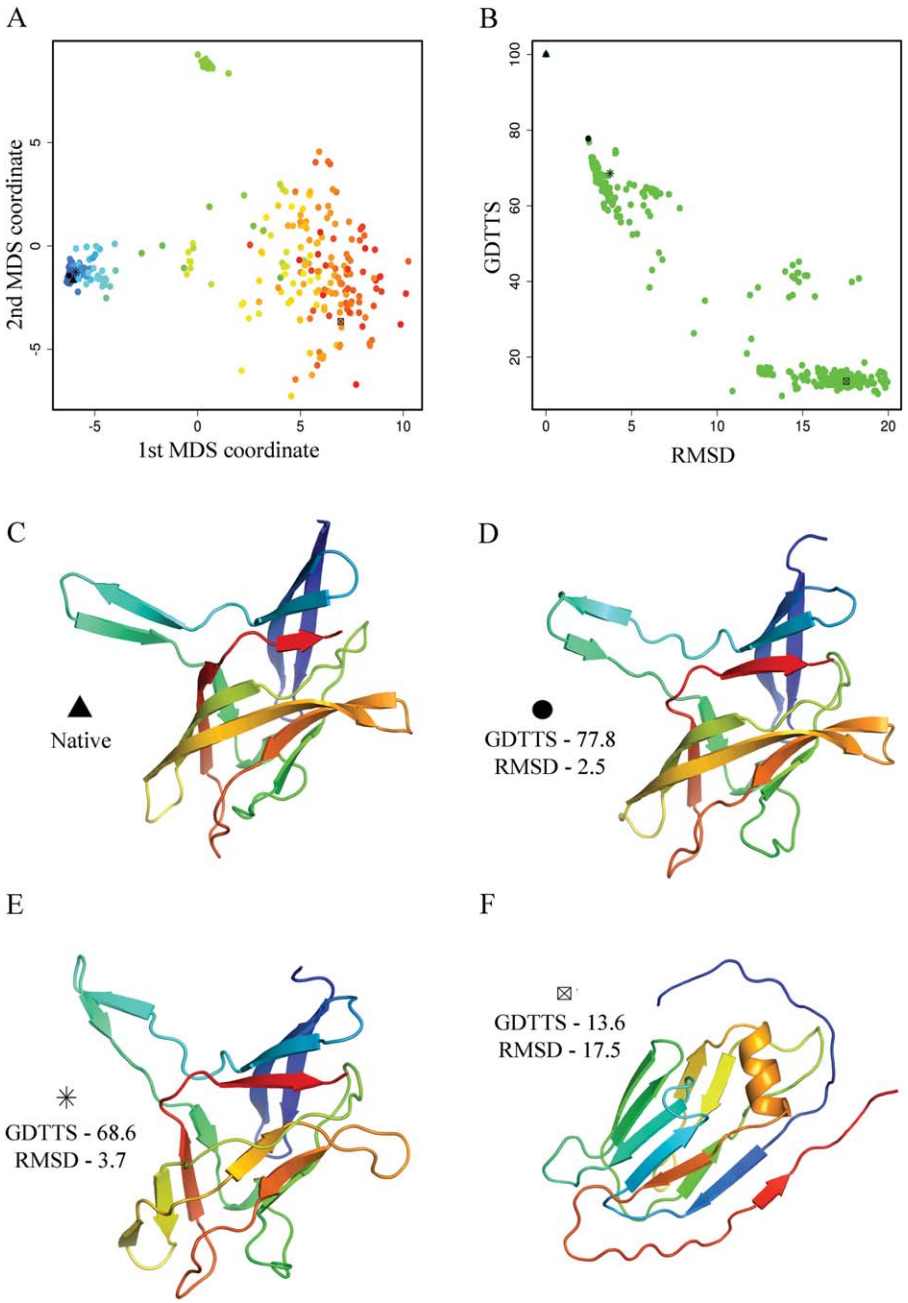

Figure 11

Assessment of predictions for the target T0795-D1. (**A**) A scatter plot of the first two coordinates of multidimensional scaling computed from pairwise RMSD distance matrix including all the models submitted by all the groups. The RMSD to native of every model is represented by the color gradient of blue (lowest RMSD) to red (highest RMSD). (**B**) A scatter plot of the RMSD to native of all the models against GDT-TS. (**C**) The experimentally determined target structure for T0795-D1. (**D**) The model with highest GDT-TS submitted by Handl (TS340_5). (**E**) The model with median GDT-TS of the upper peak (closer to native) in the bimodal distribution of GDT-TS of T0795-D1 in Figure 9. (**F**) The model with median GDT-TS of the lower peak in the bimodal distribution of GDT-TS of T0795-D1 in Figure 9. The structures in C, D, E, and F are highlighted in the graphs in panel A and B with different symbols as shown.

according to a sequence alignment resulting from template-target structure alignment. Predictors may do better than this template in a number of different ways. They may be able to build regions of the structure for which there is no information in the template; even if they do this poorly, GDT-TS may improve because it gives credit for Cα atoms placed with in 1, 2, 4, and 8 Å in a sequence-dependent structure alignment. Predictors may also do better than the best template if they combine multiple templates, and/or they are able to refine

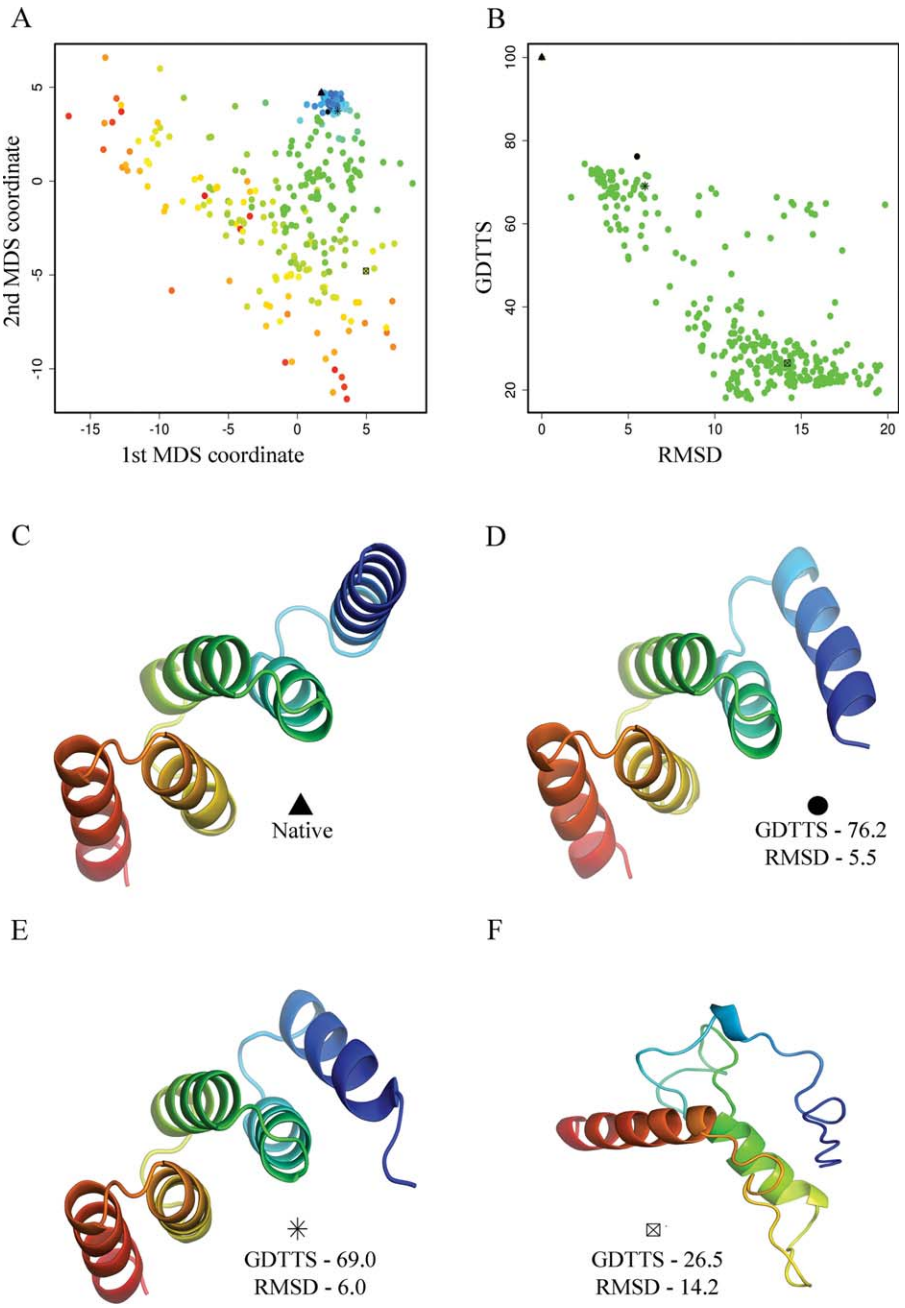

Figure 12

Assessment of predictions for the target T0828-D2. (**A**) A scatter plot of the first two coordinates of multidimensional scaling computed from pairwise RMSD distance matrix including all the models submitted by all the groups. The RMSD to native of every model is represented by the color gradient of blue (lowest RMSD) to red (highest RMSD). (**B**) A scatter plot of the RMSD to native of all the models against GDT-TS. (**C**) The experimentally determined target structure for T0828-D2. (**D**) The model with highest GDT-TS submitted by Zhang (TS204_5). (**E**) The model with median GDT-TS of the upper peak (closer to native) in the bimodal distribution of GDT-TS of T0828-D2 in Figure 9. (**F**) The model with median GDT-TS of the lower peak in the bimodal distribution of GDT-TS of T0828-D2 in Figure 9. The structures in C, D, E, and F are highlighted in the graphs with different symbols as shown.

the model successfully from the initial alignment-based model toward the template. Predictors may do worse than the best template if they choose a less suitable template and/or they align the sequence of the target to the template incorrectly. Because about two thirds of predictors do not list templates they use in their submitted predictions, it is difficult to tell the difference.

We have plotted the GDT-TS of models built directly from the best template and a structure-based sequence alignment versus the GDT-TS of the best predictor

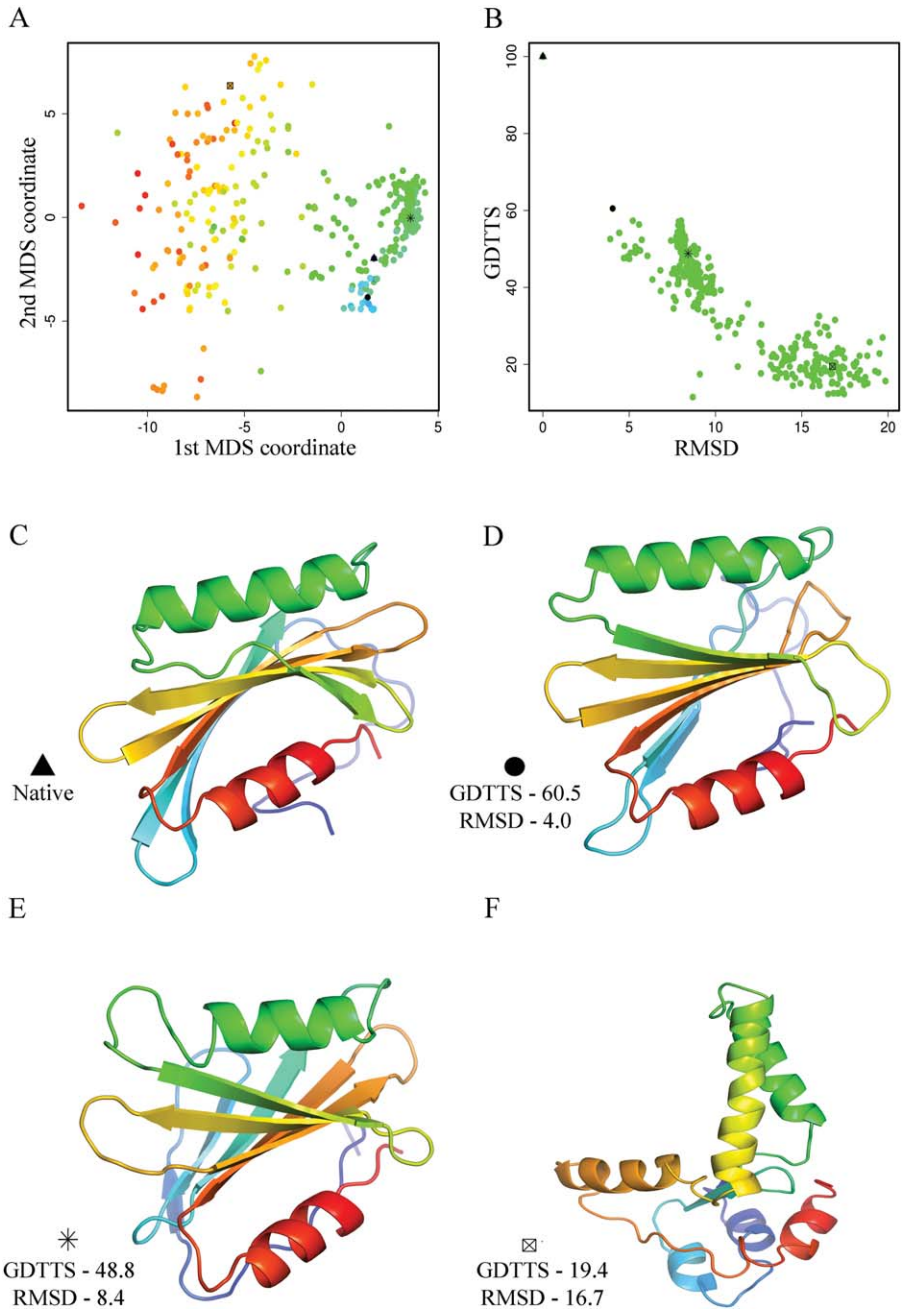

Figure 13

Assessment of predictions for the target T0838-D1. (**A**) A scatter plot of the first two coordinates of multidimensional scaling computed from pairwise RMSD distance matrix including all the models submitted by all the groups. The RMSD to native of every model is represented by the color gradient of blue (lowest RMSD) to red (highest RMSD). (**B**) A scatter plot of the RMSD to native of all the models against GDT-TS. (**C**) The experimentally determined target structure for T0838-D1. (**D**) The model with highest GDT-TS submitted by wfHHPred-PTIGRESS (TS034_5). (**E**) The model with median GDT-TS of the upper peak (closer to native) in the bimodal distribution of GDT-TS of T0838-D1 in Figure 9. (**F**) The model with median GDT-TS of the lower peak in the bimodal distribution of GDT-TS of T0838-D1 in Figure 9. The structures in C, D, E, and F are highlighted in the graphs with different symbols as shown.

models of CASP11 with CASP5, CASP8, CASP9 and CASP10 (Fig. 14). Density estimates of the best template GDT-TS and the differences between model and template GDT-TS are plotted in Figures 15(A,B) respectively (where positive values mean the model is better than the single best template). Note that the point with best-template value of GDT-TS of 69 and best-model GDT-TS of 43 is a small target (T0799-D3) of three β−sheet strands from a viral spike protein with numerous structural repeats of β−strands. It is very difficult to find the

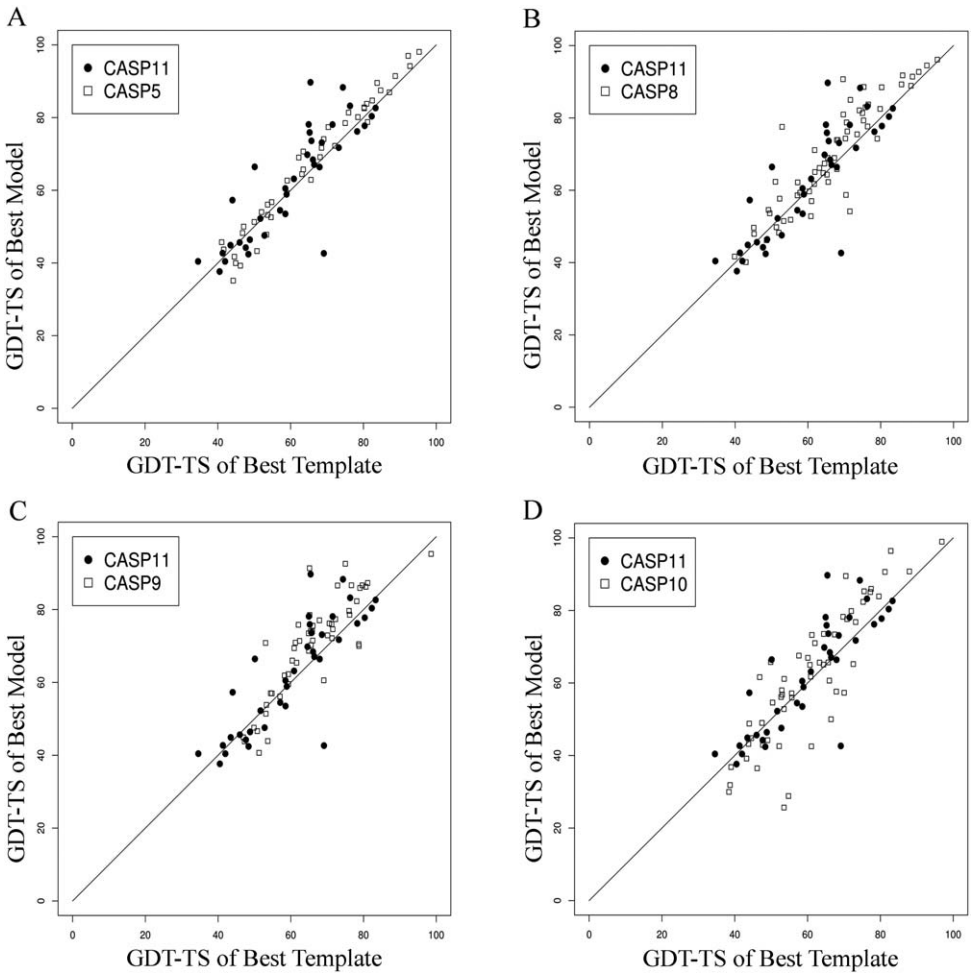

Figure 14

Comparison of predictions in CASP11 with previous CASP experiments. The GDT-TS of the best structural template is plotted against GDT-TS of the best predicted model (highest GDT-TS) for (**A**) CASP11 versus CASP5. (**B**) CASP11 versus CASP8. (**C**) CASP11 versus CASP9. (**D**) CASP11 versus CASP10. The GDT-TS of the model built from the best template is derived by copying coordinates of the template according to a sequence alignment obtained from a structure alignment of the template and the experimental structure of the target.

best template given the large sequence divergence of the target and the templates.

The later CASPs (9, 10, and 11) appear to have a few more points substantially above the diagonal (Fig. 14) with ΔGDT-TS >15 [Fig. 15(B)] than CASP8 and particularly CASP5. Most of the relatively easy targets (with GDT-TS of the best structural template >55) were consistently improved in each of the CASP experiments, probably because alignments were accurate and missing loops in the template were filled in, which automatically improves GDT-TS if the loops are within 8 Å of the native. Similarly, CASP11 has lower density than the other CASPs at values of ΔGDT-TS worse than -15 even though CASP10 and CASP11 had a larger number of targets where the GDT-TS of the best structural template was below 55 (relatively hard targets, Fig. 15(A)], compared to CASP8 and CASP9. The number of targets is small in both categories and we do not want to make too much of the differences or caterogically say there is progress in modeling in recent CASPs. The differences in performance over the CASPs have been analyzed in significantly greater detail in another article in this issue (Moult *et al.*).[43]

DISCUSSION

We have presented the assessment results for the template-based modeling category of CASP11. After examining the relationships among the different scoring metrics provided by the Structure Prediction Center, we chose a set of four metrics equally balanced between global (GDT-HA and GDC-ALL) and local (LDDT and SphereGrinder) measures with the addition of the

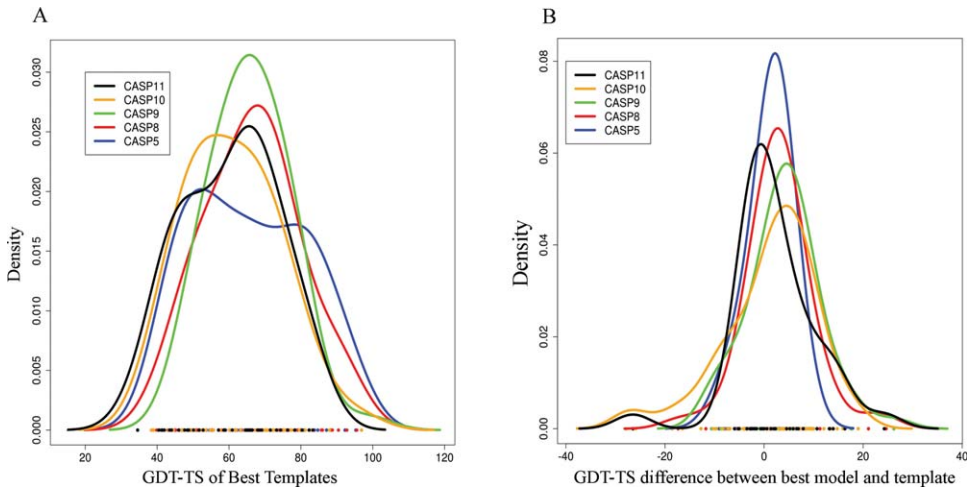

Figure 15

Kernel density plots of (**A**) GDT-TS of best templates and (**B**) Difference between GDT-TS of best model and template for all human-server targets in CASP5, CASP8, CASP9, CASP10, and CASP11.

MolProbity score to favor or penalize models with realistic or unrealistic features of bond and dihedral angles and atom-atom clashes. Because the MolProbity score is relatively easy to improve with energy minimization with a molecular mechanics or knowledge-based scoring function without significantly altering the global and local accuracy measures, we scaled the MolProbity *Z* scores by 0.2.

By analyzing the resulting *Z* score sums of individual metrics and pairs of metrics, we observed that the top five groups differ somewhat with individual metrics since some groups perform better with local or global metrics or vice versa. When local and global metrics are combined pairwise, the top five groups are always the same with some reordering. In the final scoring function, the top five groups are LEER, Zhang, LEE, MULTICOM, and Zhang-Server. The LEER and LEE methods[44] are based on a selection of models from the CASP11 server models (including their own server, nns), which were made publicly available 48 h after the release of the targets, followed by a conformational space annealing refinement with a backbone-dependent rotamer library for side-chain conformations[45] supplemented with a "residue-specific rotamer library."[44] The LEER method supplemented the LEE method by a final refinement step with molecular dynamics simulations. Zhang and Zhang-Server combine threading methods and a free-modeling approach to assemble the three-dimensional structure,[46] side chains built with the program REMO,[47] which utilizes SCWRL3,[48] followed by "fragment-dependent molecular dynamics simulations." MULTICOM ranked server models with model-quality assessment scores and combined the models with a "model combination approach" and side-chain modeling with SCWRL, and further refinement with an energy-minimization approach implemented in 3Drefine.[49] When taking the best model submitted by each group, according to a comparison with the native structure, the top 5 groups include LEER, Zhang, ProQ2, ProQ2-refine, and MULTICOM. ProQ2 selects the best model from the CASP11 server models with a model-quality assessment approach, while ProQ2-refine repacks the side chains of these models with Rosetta.[50]

With the same assessment scoring function, we evaluated the CASP11 servers on 81 targets. The top 5 groups were separated by the remaining groups by a significant drop in score. The top 5 groups when only model 1 was considered were Zhang-Server, nns (from the LEE group), BAKER-ROSETTASERVER, QUARK (from the Zhang group), and myprotein-me. The top 5 groups when the best model was considered were BAKER-ROSETTASERVER, nns, Zhang-Server, QUARK (from the Zhang group), and myprotein-me.

By comparing the GDT-TS of the best predicted model for each target and a model for each target built from the best template in the PDB (by copying coordinates according to a sequence alignment from a structure alignment of the template with the experimental structure of the target), we showed that the performances of the best predictors in recent CASPs (8, 9, 10, and 11) has been relatively steady with a small increase in the number of models with ΔGDT-TS > 15 (i.e., models improved from the best template) in CASP10 and CASP11 compared to the earlier CASPs.

We analyzed the distribution of evaluation metrics for each target, and identified a bimodal probability density for most targets and measures. Multidimensional scaling of the RMSD distance matrix revealed that this was due

to a single cluster of models reasonably close to the experimental structure and a broad distribution of models with very high RMSD to each other and to the native structure. The latter group appears in a separate mode in the probability distribution of each measure.

The data illustrate an important difference between evaluation parameters such as GDT-TS and LDDT, which measure accuracy as a percentage of the total structure, and thus bounded in value on both the left (low accuracy, a value of 0) and right (high accuracy, a value of 100) and a distance metric such as RMSD, which is bounded on the left (high accuracy, RMSD = 0.0) but unbounded on the right. The boundedness of low accuracy models in GDT-TS results in a piling up of low-accuracy structures in a Gaussian-shaped peak at GDT-TS of about 20 in most of the targets. This peak is spread out in RMSD density and in the MDS plots of the RMSD values, since the RMSD value is unbounded and the models do not resemble each other.

We conclude that perhaps assessment methods should take account of the bimodal nature of these distributions, rather than assessment methods that depend on *Z* scores, which presume a single, normal distribution of scores.

ACKNOWLEDGMENTS

The authors thank Peter Huwe for comments on the manuscript.

REFERENCES

1. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
2. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 2003; 31:3381–3385.
3. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinform 2008;9:40.
4. Krivov GG, Shapovalov MV, Dunbrack RL, Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009; 77:778–795.
5. Browne WJ, North A, Phillips D, Brew K, Vanaman TC, Hill RL. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 1969;42:65–86.
6. Greer J. Model for haptoglobin heavy chain based upon structural homology. Proc Natl Acad Sci USA 1980;77:3393–3397.
7. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
8. Li Y. Conformational sampling in template-free protein loop structure modeling: an overview. Computat Struct Biotechnol J 2013;5: e201302003.
9. Mosimann S, Meleshko R, James MN. A critical assessment of comparative molecular modeling of tertiary structures of proteins*. Proteins Struct Funct Bioinform 1995;23:301–317.
10. Lemer CMR, Rooman MJ, Wodak SJ. Protein structure prediction by threading methods: evaluation of current techniques. Proteins Struct Funct Bioinform 1995;23:337–355.
11. Martin AC, MacArthur MW, Thornton JM. Assessment of comparative modeling in CASP2. Proteins Struct Funct Genet 1997;29 (Suppl 1):14–28.
12. Jones TA, Kleywegt GJ. CASP3 comparative modeling evaluation. Proteins Struct Funct Genet 1999;37:30–46.
13. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins Struct Funct Genet 2001;45 (Suppl 5):22–38.
14. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. Proteins Struct Funct Genet 2003;53(Suppl 6): 395–409.
15. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins Struct Funct Genet 2003;53 (Suppl 6):352–368.
16. Wang G, Jin Y, Dunbrack RL, Jr. Assessment of fold recognition predictions in CASP6. Proteins Struct Funct Genet 2005; 61 (Suppl 7):46–66.
17. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins Struct Funct Genet 2005;61 (Suppl 7):27–45.
18. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 2007;69 (Suppl 8):27–37.
19. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 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, Block JN, Zemla A, Richardson DC, Richardson JS. The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins 2009;77(Suppl 9):29–49.
21. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins 2009;77 (Suppl 9):18–28.
22. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. Assessment of template based protein structure predictions in CASP9. Proteins 2011;79 (Suppl 10):37–58.
23. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. Proteins 2014;82 (Suppl 2):43–56.
24. Sippl MJ, Lackner P, Domingues FS, Koppensteiner WA. An attempt to analyse progress in fold recognition from CASP1 to CASP3. Proteins Struct Funct Genet 1999;37 (Suppl 3):226–230.
25. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins Struct Funct Genet 1999;37:149–170.
26. Lesk AM, Lo Conte L, Hubbard TJ. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins 2001;45 (Suppl 5):98–118.
27. Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins Struct Funct Genet 2003;53 (Suppl 6):436–456.
28. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins Struct Funct Genet 2005;61 (Suppl 7):67–83.
29. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69 (Suppl 8):57–67.
30. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins Struct Funct Bioinform 2014;82 (Suppl. 2):7–13.
31. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
32. Antczak PLM, Ratajczak T, Blazewicz J, Lukasiak P. SphereGrinder-reference structure-based tool for quality assessment of protein structural models. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015. IEEE. p 665–668.
33. Mariani V, Biasini M, Barbato A, Schwede T. LDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29:2722–2728.

34. Taylor TJ, Bai H, Tai CH, Lee B. Assessment of CASP10 contact-assisted predictions. Proteins 2014;82 (Suppl 2):84–97.
35. Kinch LN, Li W, Schaeffer RD, Dunbrack RL, Jr, Monastyrskyy B, Kryshtafovych A, Grishin NV. CASP11 target classification. Proteins, 2016 in press.
36. Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 2004;32:W615–W619.
37. Lovell SC, Davis IW, Arendall WB, III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins Struct Funct Genet 2003;50:437–450.
38. Theobald DL, Wuttke DS. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics 2006;22:2171–2172.
39. Theobald DL, Steindel PA. Optimal simultaneous superpositioning of multiple structures with missing data. Bioinformatics 2012;28: 1972–1979.
40. Lukasiak P, Antczak M, Ratajczak T, Szachniuk M, Blazewicz J. Quality assessment methodologies in analysis of structural models. Proceedings of the 25th European conference on operational research, Vilnius, Lithuania, 2012. pp 8–11.
41. Larsson P, Skwark MJ, Wallner B, Elofsson A. Assessment of global and local model quality in CASP8 using Pcons and ProQ. Proteins 2009;77 (Suppl 9):167–172.
42. Cao R, Bhattacharya D, Adhikari B, Li J, Cheng J. Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11. Proteins 2015.
43. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical Assessment of methods of protein structure prediction (CASP)—progress and new directions in Round XI. Proteins, in press.
44. Joo K, Joung I, Lee SY, Kim JY, Cheng Q, Manavalan B, Joung JY, Heo S, Lee J, Nam M, Lee IH, Lee SJ, Lee J. Template based protein structure modeling by global optimization in CASP11. Proteins 2015.
45. Shapovalov MV, Dunbrack RL, Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011;19:844–858.
46. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins Struct Funct Bioinform 2012;80:1715–1735.
47. Li Y, Zhang Y. REMO: a new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins 2009;76:665–676.
48. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003; 12:2001–2014.
49. Bhattacharya D, Cheng J. 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. Proteins Struct Funct Bioinform 2013;81: 119–131.
50. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 2011;487:545–574.