PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessment of template-free modeling in CASP10 and ROLL

Chin-Hsien Tai, Hongjun Bai, Todd J. Taylor, and Byungkook Lee*

Laboratory of Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

ABSTRACT

We present the assessment of predictions for Template-Free Modeling in CASP10 and a report on the first ROLL experiment wherein predictions are collected year round for review at the regular CASP season. Models were first clustered so that duplicated or very similar ones were grouped together and represented by one model in the cluster. The representatives were then compared with targets using GDT_TS, QCS, and three additional superposition-independent score functions newly developed for CASP10. For each target, the top 15 representatives by each score were pooled to form the *Top15Union* set. All models in this set were visually inspected by four of us independently using the new plugin, EvalScore, which we developed with the UCSF Chimera group. The best models were selected for each target after extensive debate among the four examiners. Groups were ranked by the number of targets (hits) for which a group's model was selected as one of the best models. The Keasar group had most hits in both categories, with four of 19 FM and eight of 36 ROLL targets. The most successful prediction servers were QUARK from Zhang's group for FM category with three hits and Zhang-server for the ROLL category with seven hits. As observed in CASP9, many successful groups were not true "template-free" modelers but used remote templates and/or server models to obtain their winning models. The results of the first ROLL experiment were broadly similar to those of the CASP10 FM exercise.

Proteins 2014; 82(Suppl 2):57–83.

Key words: CASP10; template-free modeling; ROLL; Top15Union; EvalScore; critical assessment; protein structure prediction; score function; predictor types.

INTRODUCTION

The biennial CASP assessment has provided an objective and independent assessment of protein structure prediction methods for the past twenty years. Although much progress has been made in the prediction of protein structures during this period, structure prediction without a template remains challenging.[1–3] This is the report of the evaluation of models for the Template-Free Modeling (FM) category of targets in CASP10 and in ROLL (see below), for which no sequence findable template structure can be found either because the sequence similarity is low or the structure has a new fold. In CASP10, there were 96 target proteins, which were broken up into 131 evaluation units[4] (EUs, also called domains in this article), of which 112 were classified into the template-based modeling (TBM) category and 20 in the FM category, with one in both TBM and FM categories. There were 9392 models submitted for the 20 FM targets by 147 prediction groups, of which 68 were server groups.

In contrast to more than 100 TBM targets in recent CASPs, the number of FM targets was 13 and 30 in CASP8[2,5] and CASP9,[6] of which only two and four, respectively, were considered as potentially new folds by the respective assessors. In CASP10, only three of the 20 FM targets have potentially new fold.[4] These numbers are not enough to support a statistically meaningful evaluation of the template-free modeling techniques. Therefore, in an effort to increase the number of FM targets,

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

Grant sponsor: National Institutes of Health; Grant numbers: NCRR 2P41RR001081 and NIGMS; 9P41GM103311 for Resource for Biocomputing, Visualization, and Inofrmatics at the University of Caifornia, San Francisco; Grant sponsor: Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Tai, Bai, Taylor contributed equally to this work.

*Correspondence to: Byungkook Lee, Laboratory of Molecular Biology, National Cancer Institute, National Institutes of Health, Bldg. 37, Room 5120, 37 Convent Dr. MSC 4264, Bethesda, MD 20892. E-mail: BKLee@mail.nih.gov

Received 1 July 2013; Revised 23 October 2013; Accepted 29 October 2013
Published online 9 November 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24470

the organizers introduced the ROLL experiment in CASP10. The idea is to collect FM targets year-round rather than just during the several months of CASP season every two years. Except for the continuous collection of targets and predictions, the ROLL experiment is to be run like the regular CASP experiment, including the double-blind feature, where the predictors predict before the structure is made known and the assessors assess without the knowledge of the identity of the predictors.

The first ROLL experiment began in December 2011 with the release of 18 *off-season* target sequences before the CASP10 season. We included only 15 of these because the structures did not become available in time for the other three. During the ensuing regular CASP10 season, 22 target chains were collected from the regular CASP10 targets and added to the ROLL target list (*in-season* target chains; these are, therefore, both ROLL and CASP10 targets). Fifty groups (ROLL predictors) submitted predictions for the off-season targets that we evaluated. More groups participated in the regular CASP10 experiment, but we included only predictions from ROLL predictors for the ROLL assessment. There were 41 ROLL predictors who also participated in CASP10. The assessment was done using the same procedure used in CASP10 FM assessment. The only unexpected complication was on target selection, which we will describe.

Purely score-based, automated ranking of predictions works well for TBM targets, but it is generally acknowledged that manual expert evaluation is required to properly assess FM predictions[1–3] and, as was done by previous FM assessors, we also evaluated these models ultimately by visual inspection. Ideally, one should visually inspect all predictions, but the large number of models submitted and the limited time available prohibited this. Therefore, models were first clustered by pairwise root-mean-square-deviation (RMSD) to reduce redundancy. We then screened the cluster representatives by using five score functions as parallel filters and visually inspected only those that passed any one of the five numerical score-based screening.

Although GDT_TS score has been used as a quantitative measure of prediction quality since CASP3,[7,8] previous assessment on New Fold (NF) and FM predictions had shown a discrepancy between the models ranked top by GDT_TS score and the best ones identified by visual examination.[1,2,9,10] A list of target-model pairs which GDT_TS significantly over- or under-ranked in the judgment of previous CASP NF/FM assessors is in Supporting Information Table S1. To supplement GDT_TS, many other score functions have been developed.[1,2,10] Inspired by these score functions, we developed three new score functions for use as parallel filters: Handedness, Correlation of Distance Matrix (CoDM), and Deformation (DFM) scores. These are all superposition-independent and emphasize only certain aspects of the structure. They are designed to supplement the GDT_TS score rather than to serve as independent structure comparison scores. After some trial with these old and new score functions, we settled on using five functions: GDT_TS, QCS score[3,10] used in CASP9, and the three newly developed ones.

In order to visually examine submitted models quickly and to recognize the best models more easily, we developed a novel tool, EvalScore, with the UCSF Chimera team[11] for displaying superposed target-model structure pairs upon a single mouse click from a scatter plot. With this device, we could visually examine over 1900 models, which represented, with clustering, 25% of the more than 14,600 models submitted for the FM and ROLL target domains.

We will describe these developments in the Methods section, but this report is mainly on the visual inspection of models for each target, the considerations that went into selecting the top models, and the procedure used for ranking the prediction groups.

METHODS

FM target domains

Detailed definition and categorization process of the 20 CASP10 FM target domains can be found in another article in this issue.[4] The categorization was done case by case, but generally a target domain was classified as FM if it was judged that a useable template could not be found by sequence-based searches. Template availability was highly correlated with the fidelity of submitted models: The 90th percentile GDT_TS[8] (90%GDT) score was less than 40 for the FM targets except for T0663-D0 for which it was 41.0. One of these domains, T0653-D1, has the leucine-rich repeat (LRR) fold with a unique bend and was assigned to both FM and TBM. Four others, T0651, T0663, T0690, and T0713, are made of substructures and the evaluation is to be on the relation between the substructures rather than on each individual substructure. These have the D0 designation. One, T0756-D2, is a server-only target for which human groups did not submit models; therefore, it was not considered in ranking the groups, which included both the human and server groups. The size of FM domains ranged from 58 to 535 residues and that of the chain to which they belong ranged from 165 to 770 residues.

ROLL target selection

The Prediction Center selected and announced the off-season ROLL target sequences before the target structure became available, as required for blind predictions of CASP experiment. The Center used the following rule (ROLL chain filter) to select proteins that contain at least one domain without a template.[12]

(1) In BLAST[13] search, bit score < 80, E-value $>$ 1E$-$20, and maximum coverage $< 75\%$. (2) In PSI-BLAST[13] search, bit score < 120, E-value $>$ 1E$-$40 and maximum coverage $< 75\%$. (3) In HHSearch,[14] the best hit had probability $< 60\%$. And (4) at least one good-size sequence stretch was not covered by a template by visual inspection.

The opinions of the experimentalists were also taken into consideration during the selection process. This ROLL chain filter was also applied to select the in-season ROLL targets from the regular CASP10 target sequences.

ROLL target domain selection

Evaluation units, called simply as domains in this article, were defined manually for the off-season ROLL targets in the similar way as for the regular CASP10 targets[4] except that Montelione's group was not involved.

Since the purpose of ROLL experiment was to assess the template-free modeling, we excluded domains for which a good template structure existed. However, the way we decided which domains to include was different from the method used to categorize CASP10 domains into FM, which was manual and done case-by-case[4] (see above). For ROLL domains, we included those for which the 90%GDT score of the models (one best model per ROLL predictor) was less than 50. This cutoff value was chosen after observing that there was a break in the bar graph of the 90th percentile GDT_TS scores at around this value (Fig. 1). This criterion is more generous than CASP10 FM selection method since the 90%GDT scores of all CASP10 FM targets are 41 or below (see above and Fig. 1). TBM-hard target domains are defined in CASP10 as TBM target domains for which predictions' *maximum* GDT_TS score is less than 50.[4] The in-season ROLL target domains selected in the automatic manner includes all FM and TBM-hard targets from chains that passed the ROLL chain filter. We used the 90th percentile of GDT_TS rather than the maximum so that ROLL target selection is not influenced by the presence of a few exceptionally good models.

Evaluation procedure

Evaluation was done separately for the CASP10 FM target models and the ROLL target models, but using the same procedure. All models, not just model 1, were scored by 5 different score functions (see subsequently). Models were first clustered by complete linkage hierarchical clustering using pairwise RMSD with a 3.0 Å cutoff and a representative model, the one with the highest GDT_TS score, was chosen for each cluster. The 15 top-scoring representatives under each of the five score functions were pooled to produce the *Top15Union* set. Four of us independently inspected all the models in the *Top15Union* set and selected the best cluster by consensus. If there were other clusters that were judged to be of equal merit, all of them were selected as bests. Once a cluster was selected, all the models in that cluster were considered selected. Finally, the prediction groups were ranked by the number of targets (Hits) for which their models were among the selected.

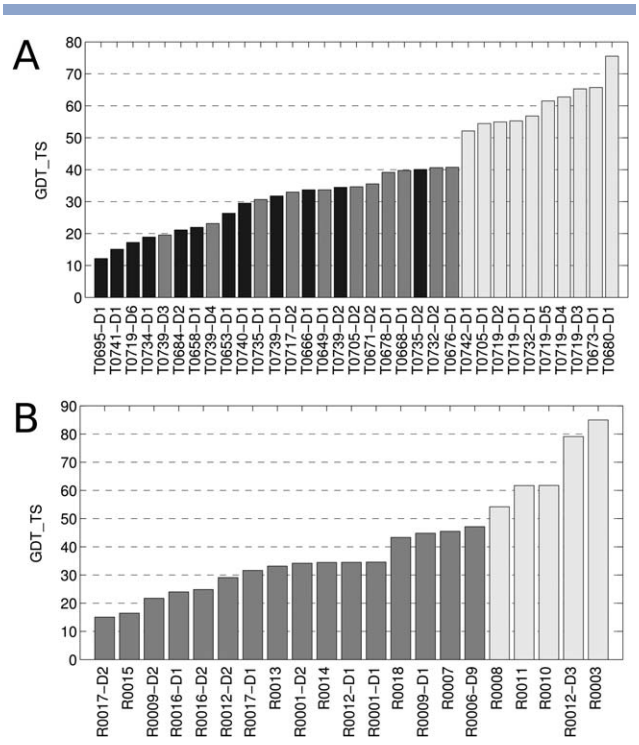

Figure 1

The 90th percentile GDT_TS scores of ROLL target domains. Only the domains from chains that passed the ROLL chain filter are shown. Domains were sorted by the 90th percentile GDT_TS scores. Top panel shows 33 in-season ROLL target domains of which 12 (black bars) belong to FM and 21 (dark and light gray) to TBM category. Bottom panel shows the off-season ROLL target domains. The domains with 90 percentile GDT score greater than 50 are in light gray.

To visually compare models with the target efficiently, we used EvalScore plugin to Chimera, which we developed with the UCSF Chimera group.[11] It takes as input the target and model structures and a matrix of scores. It then produces a scatter plot of the scores and, upon a click of a point in the scatter plot, displays the target-model structure pair side by side with rainbow-coloring according to the target's secondary structure. The EvalScore plugin has a broader utility since the matrix of scores can be replaced with other parameters, such as RMSD and energy.

During visual inspection, the cluster representatives in the Top15Union set were identified only by cluster ID. This was done as an additional measure to avoid potential bias: When model IDs are used, one can frequently assign a character to the predictor IDs from the frequency with

which one sees them among good models, even when the identity of the predictor is hidden. The cluster ID is unique only for each target and cannot be associated with any predictor. For cluster ID, we used the rank in a sorted list of clusters in the Top15Union. The sorting was by the average of five robust *Z*-scores, one from each score function. The robust *Z*-score[15] is defined as:

$$Z_i = \frac{x_i - \text{median}}{\text{MAD}} \quad (1)$$

where x_i is one of the 5 scores used, median is the median score of all models and MAD is the median of the absolute deviations from the median of all models. We used robust *Z*-score rather than conventional *Z*-score because the former uses median. Median represents the population better than the mean because it is less influenced by the presence of outliers (exceptionally good or bad scores).

Score functions

Five structure comparison scores, GDT_TS,[8] QCS,[3,10] deformation score (DFM), Handedness, and Correlation of Distance Matrix (CoDM) score, were used to construct the Top15Union set. GDT_TS scores were provided by the Prediction Center. QCS scores, created for CASP9 FM assessment, were computed locally using the software kindly provided by Lisa Kinch and Qian Cong. Handedness, DFM and CoDM, were developed by us and calculated locally. Brief descriptions of these scores follow. The DFM and CoDM scores are described in more detail in the Supporting Information.

GDT_TS

GDT_TS[8] is the average of GDT_*vi*, with *vi* = 1, 2, 4, and 8 Å. The GDT_*vi* for a given *vi* is the percentage of residues in the largest set of residues that can fit between two structures under the CA distance cutoff of *vi* Å. GDT_TS ranges over [0,100].

GDT_TS has proven to be a very good measure of prediction quality, especially for TBM targets, but has idiosyncrasies and there are cases where it significantly over- or underscores (see Supporting Information Table S1 and Refs. 1–3,9,16,17).

QCS

QCS[3,10] was designed to reproduce human expert assessments when target and prediction do not necessarily align well. The structure is represented as a set of secondary structure element (SSE) vectors. QCS scores are based on comparisons of the SSE lengths, the distances between SSE centers and the center of the protein, angles between SSE pairs, and the distances between the CA atoms in key contacts that reflect the relative packing of SSEs. QCS is a percentage score that ranges over [0,100].

Our own experiments with QCS on FM data from CASP6–CASP9 indicated that this score is better at reproducing human expert assessments than lDDT[18] and Q^2 on FM targets.

Deformation (DFM) score

The target structure is represented by a set of Delaunay tetrahedrons[19] with the CA atoms at their vertices. The model structure is also represented by the same number of tetrahedrons, each made of the same residues that make up each of the Delaunay tetrahedrons in the target. The tetrahedrons in the model are generally distorted from the corresponding tetrahedrons in the target and no longer Delaunay tetrahedrons (Supporting Information Fig. S1). DFM score is a penalty function that depends on the degree of this deformation. For each pair of tetrahedrons, one in the target and the corresponding one in the model, the penalty is computed as

$$f(V, V_0) = \begin{cases} 2.25\left(\dfrac{1}{V/V_0} - 1\right)^2 & \text{if } V/V_0 \geq 1 \\ 9\left(\dfrac{2}{3 - V/V_0} - 1\right)^2 & \text{if } V/V_0 < 1 \end{cases} \quad (2)$$

where V_0 is the signed volume of the Delaunay tetrahedron in the target and V is the signed volume of the corresponding tetrahedron in the model. This function is continuous and smooth, has a value of zero when $V = V_0$ and one when V is zero or 3 times V_0, and reaches 9 or 2.25 when V/V_0 approaches negative or positive infinity, respectively (Supporting Information Fig. S2). The asymmetric definition increases penalty when the handedness of the tetrahedron is changed. The DFM score is a weighted sum of this function over all tetrahedron pairs. The Delaunay tetrahedrons of the target structure were calculated using QHull.[20]

Handedness score

Atoms forming Delaunay tetrahedrons in tessellated proteins tend to be physically close to each other in space and, to a lesser extent, close in sequence. We defined the *handedness score* to compare global conformations of target and prediction. Choose a set of four CA atoms randomly from the target structure and the set of the same four atoms in the prediction. For each set, label the residues in increasing sequence order *i, j, k, l* and compute handedness defined as $\text{sgn}[(\boldsymbol{ij} \times \boldsymbol{ik}) \cdot \boldsymbol{il}]$, where $\boldsymbol{ij}$ is a vector from the CA of residue *i* to the CA of residue *j*, and so forth. Repeat this procedure many times (we did it 50,000 times) and compute the fraction of the pairs that have the same handedness in target and

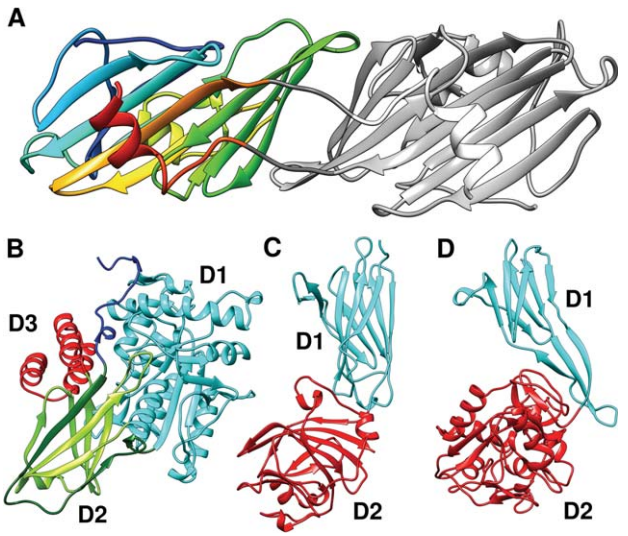

Figure 2

Four of the six multidomain off-season ROLL targets. (**A**) Domain-swapped dimer of R0006. R0006-D9 is the unswapped monomer shown in rainbow colors, blue to red along the sequence. (**B**) R0012. The three domains are colored cyan, green and red, except for a stretch of the N-terminal sequence that spans both D1 (colored blue) and D2 (colored dark green). D1 and D2 are thus segmented domains. (**C**) R0016. (**D**) R0017. The two domains in the last two structures are colored cyan and red.

prediction. This is the Handedness score, which ranges over [0,1].

Only the tetrahedrons with the edge length less than a cutoff value were used. The cutoff value was the diameter of the protein molecule estimated using a formula that one of us (TT) used before,[21] $L < 2 + 7.2n^{1/3}$, where n is the number of residues of the protein. If the fraction of the tetrahedrons that satisfied this length condition in the prediction was less than 0.45 times the fraction in the target, then the prediction was tagged. Such models are much less compact than the target. The tag was, however, inadvertently ignored in this work.

Correlation of Distance Matrix (CoDM) score

CoDM score is a weighted Pearson's correlation of the distance matrices of the target and model structures. It ranges over [−1,1]. Each element k of the distance matrix, where k is the residue pair index that ranges from 1 to n^2, is weighted by w_k given by

$$w_k = w_{k,m} + w_{k,t} \tag{3a}$$

with

$$w_{k,t} = \frac{1}{4\pi d_{k,t}^2} \cdot \frac{1}{1 + \left(d_{k,t}/d_0\right)^2} \tag{3b}$$

for the target structure t and $w_{k,m}$ similarly defined for the model structure m. In (3b), $d_0 = 8.0$ Å and $d_{k,t}$ is the distance between CA atoms of residue pair k in the target structure. The first factor normalizes for the expectation that, for a globular protein, the number of interatomic distances increases with the square of the distance unless the distance becomes large. The second factor gradually switches the weight from 1 to 0 as distance increases, with half weight at the distance of 8.0 Å.

RESULTS

Selection of ROLL target domains

Nine of the 15 off-season ROLL targets were single domain proteins. One (R0006) was in the form of a domain-swapped dimer [Fig. 2(A)]. We "unswapped" it as we did for such domains in regular CASP10.[4] The domain definitions of this and the remaining five multi-domain proteins are given in Table I. Two, R0001 and R0009, form tightly interacting homotrimer and homo-pentamer, respectively (Fig. 3). The three domains of R0012 are not well separated geometrically [Fig. 2(B)]. The N-terminal 54 residues start from D1 [Fig. 2(B), dark blue], then pass through D2 to contribute two β-strands [Fig. 2(B), dark green] before returning to D1. This makes both D1 and D2 segmented domains. The helix from residues 367 to 379 connecting D1 and D2 is included in D1. R0016 and R0017 are clearly made of two domains [Fig. 2(C,D)]. There is a three-stranded β-sheet between the two domains of R0017, made of a hairpin loop excursion from D1 and the C-terminal extension of D1. We included this structure in D1.

The 15 off-season ROLL targets contributed 21 target domains. When the ROLL chain filter was applied to the CASP10 target chains, 22 passed the test, which provided 33 additional domains. Figure 1 shows the 90th percentile GDT_TS (90%GDT) scores of these domains and their names. It also shows that the 90%GDT scores are very high for some domains. Since the purpose of the

Table I

Off-Season Multidomain ROLL Target Domain Definitions

Target	Domain definition
R0001-D1	1–93
R0001-D2	94–183
R0006-D9[a]	A:20–171, B:175–188
R0009-D1	1–50
R0009-D2	54–176
R0012-D1	40–61, 94–379
R0012-D2	63–93, 380–451
R0012-D3	452–499
R0016-D1	32–170
R0016-D2	171–333
R0017-D1	1–110
R0017-D2	111–352

[a]R0006-D9 forms a domain-swapped dimer. The definition given is for an "unswapped" monomer, which is composed of residues 20–171 of chain A and residues 175–188 of chain B.

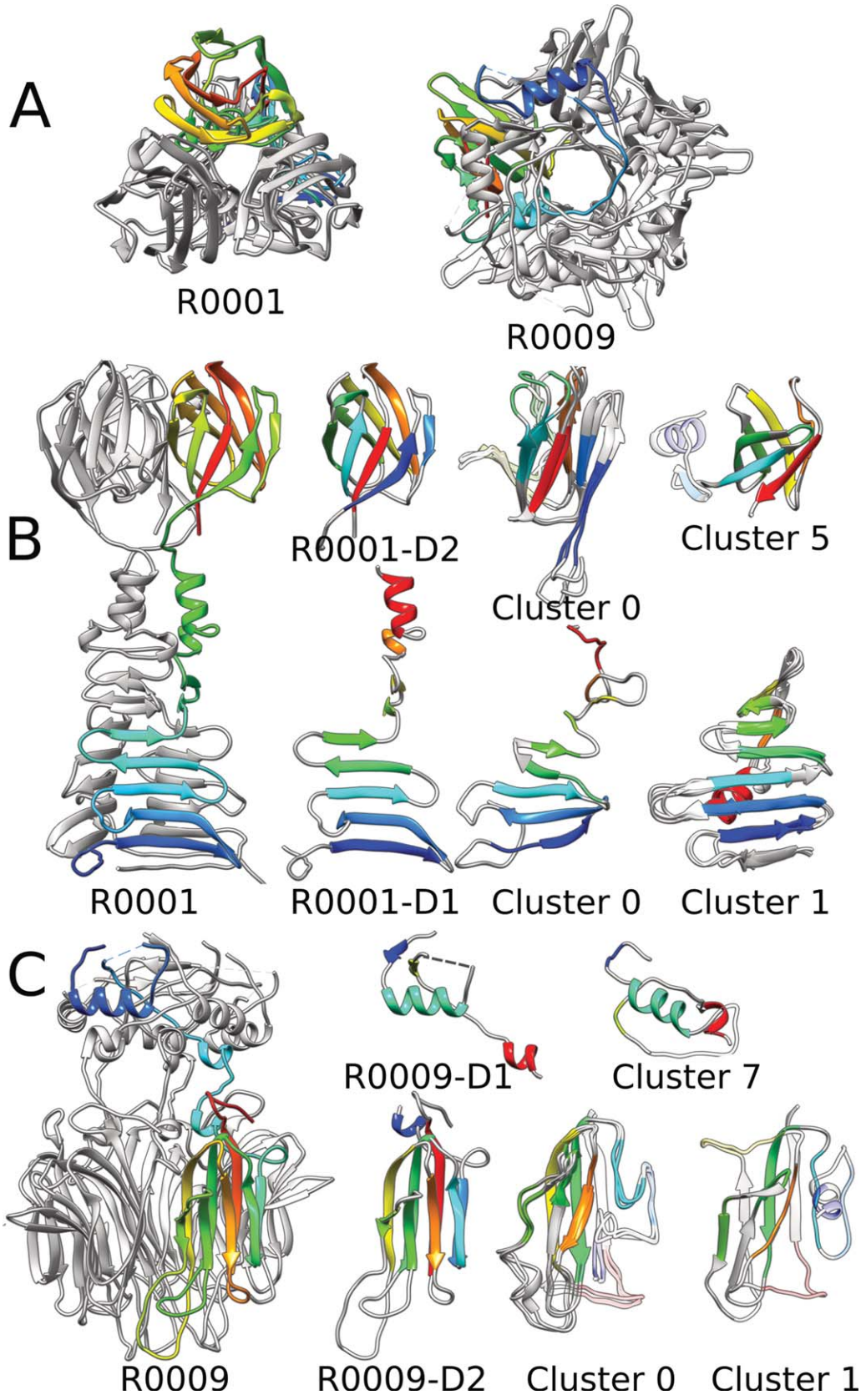

Figure 3

Oligomeric off-season ROLL targets, R0001 and R0009, and selected top models. Secondary structure elements of one monomer of the target are rainbow-colored, from blue to red along the sequence. (**A**) Views of the targets down the symmetry axis. (**B**) From left to right, Target R0001, the two domains of target R0001, and two sets of models for each of the two domains. (**C**) From left to right, Target R0009, the two domains of target R0009, and one and two sets of models for D1 and D2, respectively. Residues in models are colored the same as corresponding residues in the target regardless of the structure of the former. The unmatched residues of the models are colored semitransparent to highlight the matched part. Multiple models in a cluster are superimposed together.

Table II
Selected Best Models for FM Targets

Target[a]	Cluster ID[b]	GDT score[c]	Rank[d] GDT	QCS	Hand	CoDM	DFM	Minimum rank[e]	Models in the cluster[f]
T0651-D0	0	35.63	5	7	13	1	2	4	TS301_3
	1	34.15	22	16	8	3	1		TS405_4, **TS114_4,** TS267_3, TS475_2
	34	36.12	4	112	47	92	78		**TS292_5, TS330_3,** TS164_1, TS164_2, TS350_1, TS350_2
T0653-D1	3	30.22	2	14	109	99	72	2	TS201_2
*T0658-D1**	1	21.39	6	4	8	13	3	3	TS473_4[g]
	4	23.49	2	7	11	8	7		**TS463_1**
	5	22.14	4	2	13	12	2		TS473_3[g]
	6	21.08	10	3	12	11	5		TS473_5[g]
	7	22.89	3	5	23	15	8		TS473_1[g], TS473_2[g]
T0663-D0	36	41.61	10	172	5	150	120	7	TS043_2
	45	23.19	264	223	4	129	178		TS152_1
	50	22.04	270	251	7	184	257		TS152_4
*T0666-D1**	11	31.11	7	6	5	30	48	5	TS079_1[g]
*T0684-D2**	0	22.17	5	1	1	4	1	1	TS045_4[g]
	1	24.85	1	2	6	5	2		TS045_3[g]
T0690-D0	0	29.11	3	2	3	44	48	2	TS344_4
T0693-D1	2	31.00	8	7	4	4	2	2	TS437_3
	3	35.25	2	1	72	1	4		TS315_4_1, **TS335_1, TS488_3,** TS130_4, TS267_5, TS280_1, TS285_3, TS405_2, TS079_4, TS428_5
*T0695-D1**	6	13.22	7	1	96	2	4	2	**TS330_4[g],** TS029_3, TS164_2, TS475_2
	8	10.73	44	2	99	5	7		TS479_4
	9	10.44	51	19	58	1	13		TS201_1
T0713-D0	0	28.74	10	55	3	152	95	13	TS317_4_1
	2	28.34	17	110	5	147	42		TS317_1_1
	18	28.61	13	130	15	137	85		TS317_5_1
*T0719-D6**	0	27.15	1	19	2	52	14	7	TS315_2_6[g], **TS411_2**
	1	23.77	4	26	4	57	5		**TS456_2**
	2	24.23	2	47	9	67	2		**TS456_1**
	4	23.93	3	32	11	83	7		**TS148_4**
	5	23.62	7	38	8	75	17		**TS411_1**
T0726-D3	1	35.17	2	6	55	5	1	4	TS172_1_2
	3	29.24	15	3	95	1	7		TS315_2, **TS114_2, TS035_2,** TS315_4, TS388_3
	4	31.36	7	9	69	4	5		TS315_5, **TS108_3,** TS350_4
T0734-D1	27	20.28	3	237	66	183	54	4	**TS358_1_2**
	44	19.81	4	288	58	205	294		**TS358_2_1**
*T0735-D2**	0	41.48	2	8	1	40	1	9	**TS035_2[g]**
	2	39.49	7	1	3	19	4		TS315_1[g], TS388_1[g], TS130_2, TS267_2[g], **TS114_5[g],** TS197_3, TS197_5, TS428_2[g], TS489_2[g],
	3	38.35	12	4	4	14	3		TS237_3[g]
	8	35.23	35	43	11	13	9		**TS333_5_2**
T0737-D1	0	40.60	1	1	2	12	2	1	TS045_5
T0739-D1	0	34.12	2	12	1	25	1	1	**TS498_4**
T0739-D2	7	32.76	9	6	162	9	20	6	TS111_5
	1	32.33	11	20	3	11	39		**TS258_2**
T0740-D1	0	32.1	3	1	1	1	19	1	TS287_4
*T0741-D1**	19	16.8	2	1	236	245	243	2	**TS413_5[g]**
	16	13.6	37	2	279	42	25		TS344_2[g]
T0756-D2	0	43.9	1	1	2	14	18	1	**TS114_5**

[a]Names of targets that are also ROLL targets are italicized; those for which a ROLL predictor's model was one of the best are indicted by *.
[b]Clusters which would have been missed without the Handedness, CoDM, and DFM scores are gray shaded.
[c]GDT score of the representative of the cluster.
[d]The rank is among the representatives after clustering, not among all models submitted. Ranks greater than 15 are shaded in gray.
[e]The minimum rank cutoff needed to cover all selected models for a target.
[f]The first model listed is the representative of the cluster. Models from server groups are in bold face.
[g]Models from ROLL predictors.

ROLL experiment is to collect difficult targets, we excluded domains that have the 90%GDT scores above 50 (see Discussion). This resulted in 23 in-season ROLL target domains (12 FM and 11 TBM) and 16 off-season target domains.

Many models were identical or nearly identical to another model

To reduce the number of models to be visually inspected, models were first clustered using pairwise RMSD with a 3.0 Å cutoff and each cluster was represented by a single model, the member with the highest GDT_TS score. For FM, clustering reduced the number of models by 41%, from 9392 to 5573 (Supporting Information Table S2a). Such a large redundancy presumably indicates that many models were copies of a common template or a server model with little or no modification.

We did not re-examine the models for seven FM ROLL targets, indicated in Table II and Supporting Information Table S2a with a superscript *, because these were examined as FM target models and the best models submitted by ROLL predictors were among the bests by all CASP10 predictors. The models from the ROLL predictors for the five remaining FM ROLL targets, the 11 TBM ROLL targets, and the 16 off-season targets were clustered. The number of models decreased by 31% in this case, from 5243 to 3598 (Supporting Information Table S2b). This is somewhat less than the 41% reduction for the CASP10 FM models, presumably reflecting a difference in the composition of predictor types (see Discussion) between the ROLL and FM predictors.

Union of top 15 by 5 score functions produces a manageable number of models per target

The cluster representatives were prescreened by retaining only the top 15 models by each of the five score functions and pooling them to form the Top15Union set for each target (see Methods). The top 15 was a conservative choice because a preliminary testing using previous CASP NF/FM models indicated that all assessor-selected models were in the union of top 10 models by these score functions.

The prescreen reduced the number of clusters from 279 to 43 per target on average for FM and 112 to 34 for ROLL (Supporting Information Tables S2a and S2b). The numbers after the prescreen were small enough that we could visually inspect all representatives using Chimera with the EvalScore plugin (see Methods). The total numbers of models visually inspected were 857 and 1075 for FM and ROLL respectively, representing 1736 (18% of total) and 1905 (36% of total) models. The larger fraction covered for ROLL is the consequence of fewer predictors participating in the ROLL experiment—50 ROLL predictors versus 147 CASP10 FM predictors—which resulted in fewer total ROLL models per target (164 models per ROLL target vs. 470 models per FM target on average).

Selected best models

Tables II–IV show the visually selected best model clusters and the models in them by their ID. The structures of these models and the rationale for selecting them as the best are described later and in the Supporting Information.

One TBM (T0671-D2) and two FM (T0653-D1, and T0739-D2) in-season ROLL targets were excluded because we could not identify models from ROLL predictors that were distinctly better than others. These are not included in Table IV. T0653-D1 is a LRR protein smoothly bent with the β-sheet side convex [see subsequently and Fig. 4(C)]. No model from ROLL predictors had the correct bend. T0671-D2 is another LRR, which is almost straight or slightly bent to make the β-sheet side slightly convex [Fig. 4(B)]. A few ROLL predictions had it broken into two nearly straight segments, arranged perpendicular to one another as in the template 3sb4 [Fig. 4(G)]. Twenty-five other models from ROLL predictors had similar LRR smoothly bent the wrong way. T0739-D2 is a small β-barrel of six strands with jelly roll topology [Supporting Information Fig. SF6(D)]. All models submitted by the ROLL predictors were quite poor for this target.

Some best models would have been undetected without the new score functions and Top15Union strategy

Tables II–IV also show the ranks of the selected clusters by the score functions used for the prescreen. The rank is highlighted in gray if it is not in the top 15, in which case the model would have been "missed" by the score. The newly introduced Handedness, CoDM, and Deformation scores cannot replace the GDT_TS and QCS scores because many best models ranked poorly by these new scores. However, they complemented GDT_TS

Table III
Selected Best Models for Off-Season ROLL Targets

Target	Cluster ID	GDT score[a]	Rank[b] GDT	QCS	Hand	CoDM	DFM	Minimum rank[c]	Models in the cluster[d]
R0001-D1	0	33.59	3	3	2	2	1	1	TS292_4
	1	33.85	2	1	4	27	15		TS079_3, **TS292_5**, TS315_3, **TS330_1**, TS453_3
R0001-D2	0	33.62	11	4	1	1	1	2	TS341_1, TS068_2, TS388_5, **TS330_3**
	5	38.51	2	12	29	9	4		TS045_5
R0006-D9	0	47.14	1	1	1	1	1	1	TS489_1, **TS035_1**, **TS114_3**, TS237_3, TS267_2, TS267_3, TS388_1, TS388_3, TS428_2, TS428_3, TS453_1, TS079_4, TS489_2
R0007	5	44.72	6	1	8	5	8	1	**TS035_2**
R0009-D1	7	42.44	13	29	69	2	3	2	TS165_2
R0009-D2	0	32.11	1	1	1	1	1	2	TS315_2, TS315_5, **TS125_3**, **TS125_4**, **TS222_2**
	1	25.20	2	2	2	29	2		**TS247_2**
R0012-D1	19	35.51	1	8	12	78	28	11	**TS308_4**
	20	32.40	11	10	19	77	26		**TS308_3**
	21	35.35	2	20	14	74	34		TS301_4
	23	32.40	11	11	18	79	30		**TS308_5**
	24	32.40	11	16	28	71	13		TS301_5
R0012-D2	0	30.41	4	2	5	3	1	1	TS237_5
	2	33.25	1	14	15	16	30		TS477_5, **TS330_5**
R0013	0	41.98	1	2	2	1	2	1	**TS113_1**
R0014	1	37.50	2	3	2	1	2	1	TS413_2
R0015	0	19.27	4	1	1	1	1	1	**TS222_2**
R0016-D1	0	28.80	1	1	4	2	1	2	**TS222_1**
	1	24.64	3	2	3	29	3		**TS125_2**
	2	24.82	2	4	2	32	5		**TS424_2**
R0016-D2	0	25.15	1	1	2	20	2	1	TS315_4, TS267_3, TS267_5, **TS292_2**, **TS330_1**, TS428_3, TS428_5, TS079_5
R0017-D1	0	35.91	3	1	1	4	2	1	**TS222_5**
	1	41.82	1	2	3	49	12		TS477_2
R0017-D2	35	17.05	4	82	18	69	87	4	**TS113_4**
	37	21.49	1	2	74	68	21		TS045_3
R0018	0	46.43	1	5	1	1	3	2	**TS114_5**, TS489_4
	1	45.41	2	7	2	3	4		TS045_3
	3	43.88	3	10	6	13	2		TS344_2

[a]GDT score of the representative of the cluster.
[b]The rank is among the representatives after clustering, not among all models submitted. Ranks greater than 15 are shaded in gray.
[c]The minimum rank cutoff needed to cover all selected models for a target.
[d]The first model listed is the representative of the cluster. Models from server groups are in bold face.

and QCS scores and rescued some best models that would otherwise have been overlooked. As highlighted in gray in Tables II and IV, seven clusters of 10 best models for six targets would not have been examined visually if only GDT_TS and QCS scores were used to construct the Top15Union set. Some chosen models ranked lower than 200 among the cluster representatives by GDT_TS or QCS, and would be even lower if all models were counted individually. In particular, the selected best models for the D0 targets, for which the relative orientation between the substructures is the focus of evaluation, often scored poorly by the GDT and QCS scores, but all ranked 15 or better by the Handedness score. Figure 5 shows some model structures of T0663-D0 as an example.

Minimum rank required to retain all best models

Minimum rank of a cluster is the minimum of the five ranks from the five score functions for the cluster.

Table IV
Selected Best Models for In-Season ROLL Targets

Target[a]	Cluster ID[b]	GDT score[c]	Rank[d]					Minimum rank[e]	Models in the cluster[f]
			GDT	QCS	Hand	CoDM	DFM		
T0649	0	33.83	3	2	4	1	1	1	TS315_4
	1	33.70	5	1	3	2	2		**TS114_4**, TS428_4, TS267_5, **TS035_3**
T0668-D1	1	39.74	3	1	2	7	3	5	TS317_4, TS317_1, TS317_2
	9	35.58	12	39	5	25	13		TS388_5
T0676	3	35.55	3	2	4	7	9	4	TS435_1
	4	33.38	4	4	6	13	4		TS435_2
T0678	0	42.37	2	1	2	2	2	15	TS237_5
	1	39.61	3	2	5	1	1		**TS087_1**, **TS087_2**, **TS087_3**, **TS087_4**, **TS087_5**
	2	38.15	5	3	1	3	5		TS489_2, TS428_5
	6	36.20	13	10	7	4	12		**TS413_3**
	19	30.52	31	17	17	15	28		**TS222_3**
T0705-D2	0	36.48	1	3	1	1	1	11	TS344_1_2, **TS035_1**, TS237_1
	1	34.08	4	1	4	3	2		TS079_3, TS428_5, TS388_5, TS267_5, **TS114_1**, TS079_4, TS315_1_2
	2	34.81	2	2	2	4	4		TS477_4
	3	31.03	8	4	3	2	3		TS477_5
	10	29.14	11	23	12	13	19		TS045_2
T0717-D2	0	48.80	1	1	1	2	1	1	TS045_3
T0732-D2	0	39.29	4	1	12	1	5	2	TS301_1
	2	40.93	3	2	3	8	3		TS301_2
T0735-D1	0	39.16	1	1	1	1	1	3	TS237_5
	1	30.90	3	4	2	7	4		TS298_1
	2	31.65	2	9	6	2	2		TS315_1, TS489_2, TS315_2, TS428_2, TS428_1, TS388_2, TS388_1, TS267_2, TS267_1, **TS114_5**, **TS035_5**, TS489_1
	3	30.04	4	2	4	3	6		TS477_4
	4	28.86	7	3	5	5	5		TS477_1
	11	29.29	6	11	12	6	3		**TS113_4**, **TS113_5**
T0739-D3	0	25.65	1	1	1	11	15	1	TS477_4
T0739-D4	3	23.15	4	2	4	15	3	2	TS489_5, **TS488_1**
T0734-D1	20	20.99	1	1	115	32	32	13	**TS488_5**
	33	17.57	13	18	100	78	14		TS045_1
T0739-D1	6	31.77	4	11	8	34	6	4	**TS413_1**
T0740-D1	3	29.03	6	4	26	3	11	3	**TS035_4**

[a]Names of the three FM targets are italicized. Others are TBM targets. For the seven other FM ROLL targets with a ROLL predictor model as one of the selected best, see Table II.
[b]Clusters which would have been missed without Handedness, CoDM, and DFM are gray shaded.
[c]GDT score of the representative of the cluster.
[d]The rank is among the representatives after clustering only the models from ROLL predictors. Ranks greater than 15 are shaded in gray.
[e]The minimum rank cutoff needed to cover all selected models for a target.
[f]The first model listed is the representative of the cluster. Models from server groups are in bold face.

Minimum rank of a target, which generally has more than one selected best cluster, is the largest rank (worst cluster) among the minimum ranks of the selected clusters.

The minimum ranks for the FM targets (Table II) show that, even the worst of the selected models for the FM targets emerge within top 10 union set except for T0713-D0, which has the minimum rank of 13. For the off-season ROLL target models (Table III), all the selected best models are in the union of top 4 by each score function, except one (R0012-D1) for which top 11 need to be used [Fig. 2(B) and Supporting Information Fig. SR2].

For the in-season ROLL target models that were newly screened (Table IV), the situation is similar except for two targets for which the cutoff values of 13 (T0734-D1) and 15 (T0678) are needed. The structure of the target and models for T0734-D1 will be described later, but

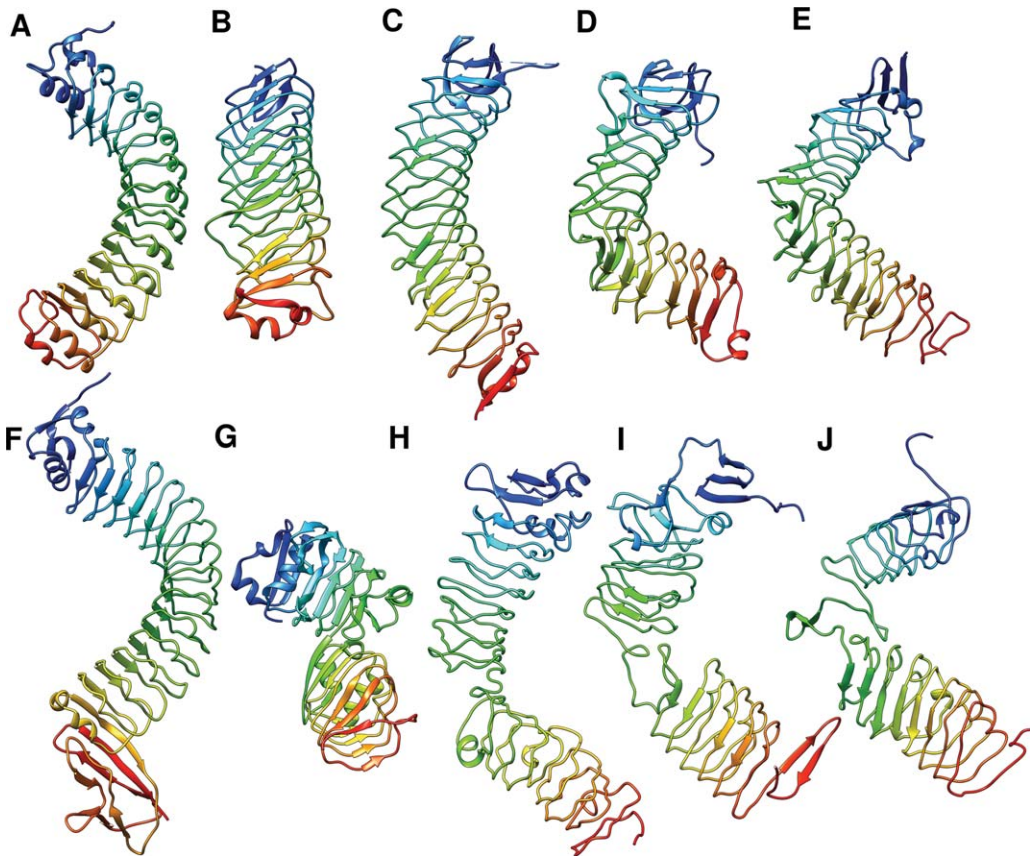

Figure 4

LRR targets, templates and models. All structures are rainbow-colored from N- to C-termini. Top row shows the five targets with a LRR fold. (**A**) T0650, a regular LRR and a TBM target. (**B**) T0671-D2, a TBM target with a small curvature. (**C**) T0653-D1, LRR with the β-strands on the convex side, a TBM/FM target. (**D**) T0690-D0 and (**E**) T0713-D0, two kinked LRRs, FM targets. Bottom row shows two templates and three selected models for the three FM targets T0653-D1, T0690-D0, and T0713-D0. (**F**) 3bz5A, a regular LRR found by HHpred for T0713. (**G**) 3sb4A, a broken LRR obtained from structure search program Dali for T0653, and also the template used for model TS201_2. (**H**) Model TS201_2 for T0653-D1. (**I**) Model TS344_4 for T0690-D0. (**J**) Model TS317_4_1 for T0713-D0.

there were two exceptionally good predictions from one group, RaptorX-Roll (358), who was not a ROLL predictor (Table II). Other models were generally poor, especially so for models among the ROLL predictors. (See ROLLbestModels.pdf in Supporting Information.) This seems to be the case where numerical scores had difficulty in recognizing better ones among poor models. On the other hand, T0678 is a TBM target with known fold and at least one remote template.[4] Five models (Supporting Information Fig. SR3 and ROLLbestModels.pdf) looked more or less equally good visually but one turned out to have low scores when compared with the target quantitatively.

Naive models

Our idea of naive prediction first arose in a discussion with David Jones. Naive prediction is to serve as the baseline from which to estimate the added value of the *select-and-submit* strategy, where predictors assess the quality of prediction server models, choose one to five models as the best, and submit them perhaps after some refinement. Our naive predictor randomly selects five models from all available nonredundant server models for each target. Figure 6 and Table S3 in Supporting Information shows the odds of submitting a top model for the given number of targets, when tried 5000 times. For example, in the case of the CASP10 FM targets, the odds of submitting a top model for exactly one target is 35%, and for at least one target is over 46% (sum of the last four probabilities). There is a maximum in the distribution for the ROLL targets, meaning that the probability of getting zero or one hit is less than getting two. This is in contrast to the monotonic decrease observed for the CASP10 FM targets and is a consequence of the fact that there are larger number of server targets (26 for ROLL vs. 11 for CASP10 FM) and smaller number of servers (19 for ROLL vs. 68 for CASP10 FM). (See Supporting Information for an analytical expression of the probabilities in an ideal case.)

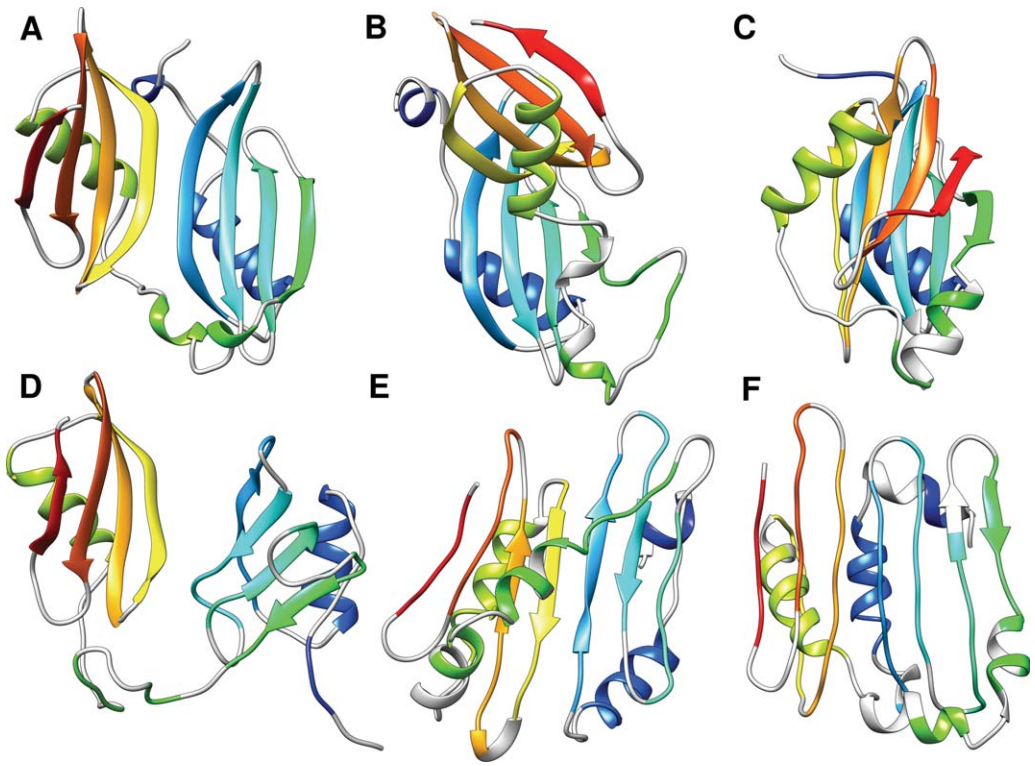

Figure 5

T0663-D0. Top row: target and two models with high scores that we did not select. Bottom row: three selected models. Models are rainbow-colored according to the secondary structure of the target and loops are in gray. (**A**) Target T0663-D0. (**B**) Model TS027_1 with the highest GDT_TS score (42.9). The two subunits are stacked on top of each other. (**C**) Model TS301_5 with the highest QCS and DFM scores. The two sheets face each other instead of forming one continuous sheet. (**D**) TS043_2, (**E**) TS152_1, and (**F**) TS152_4 have GDT_TS scores of 41.6, 23.2, and 22.0, respectively.

The number of hits that a predictor must have in order for the performance to be considered statistically significant at the 5% significance level is three for the CASP10 FM targets and seven (the probability of getting six or more by the naive method is 0.051, slightly over 0.05) for the ROLL targets. We hasten to add that this estimation of statistical significance applies only to the groups who use the *select-and-submit* strategy; it says nothing on the success of obtaining the top model for just one target, for example, by an *ab initio* procedure.

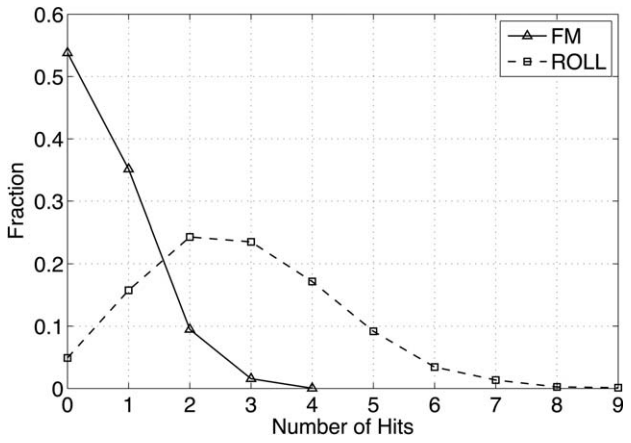

Figure 6

Distribution of the number of targets for which a best model was selected after 5000 naive predictions for the FM (solid line, triangles) and ROLL (broken line, squares) targets.

Best performing groups

Figure 7(A,B) lists the prediction groups who submitted at least one of the selected best models and the targets for which the group's models were among the best. We excluded T0756-D2 from ranking because it was the only FM target for which nonserver groups did not participate (server-only target). We also excluded from Figure 7(B) the three ROLL targets for which we could not select top tier models (see above).

We will refer to the targets for which least one server model was present in the selected cluster(s) (shaded in Fig. 7) as "server targets." There are 11 (58% of 19 total) FM and 26 (72% of 36 total) ROLL server targets. As shown in the bottom row of Figure 7(A,B), the number

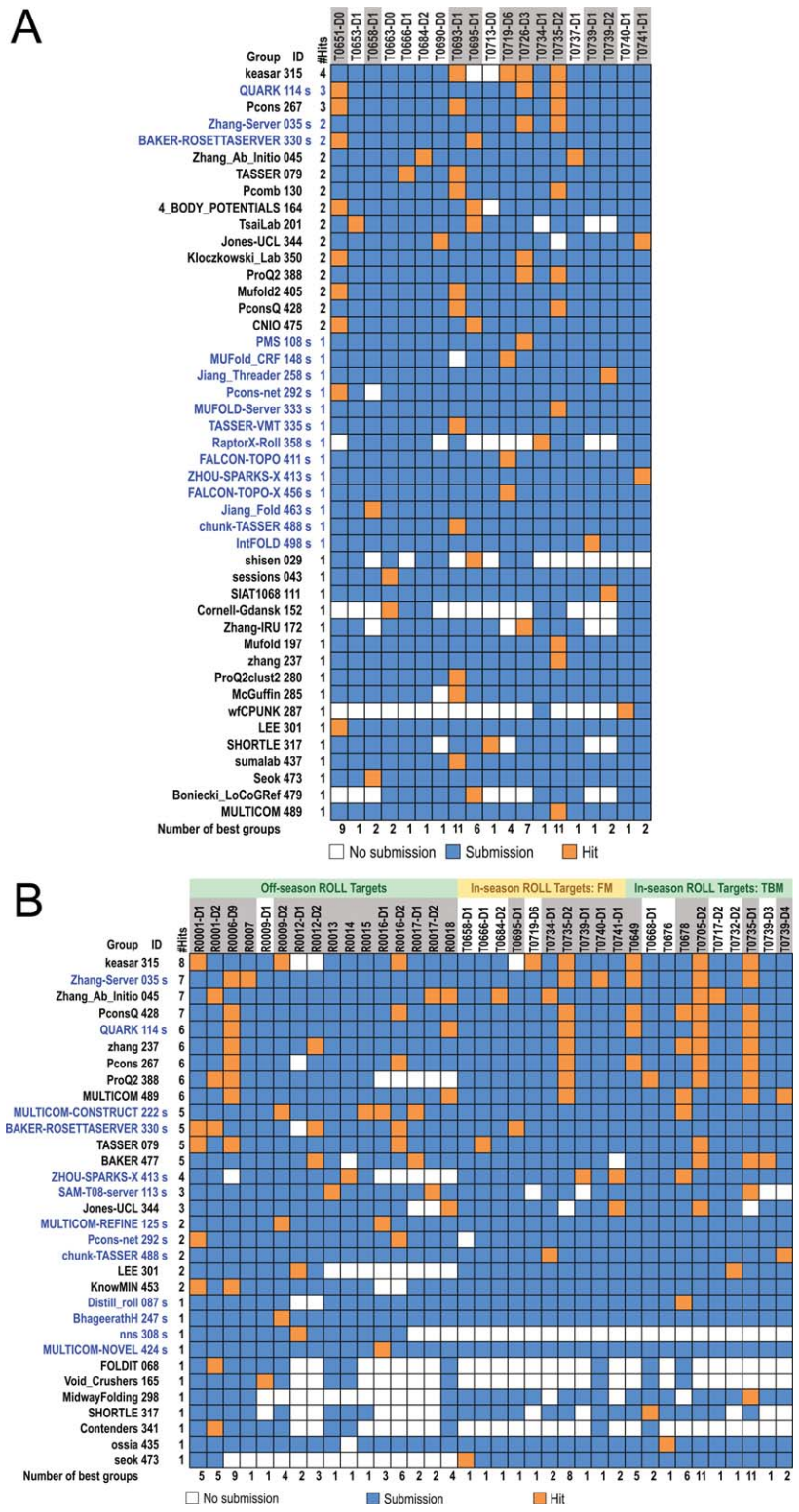

Figure 7

Group performance for the FM (**A**) and ROLL (**B**) categories. The #Hits column gives the total number of targets (hits) for which the group submitted a best model. The target names are shaded if a server produced a best model for the target ("server targets"). The prediction groups were sorted by #Hits. Server group names are in blue. Groups not listed here had no top submission. Orange and blue boxes mean that the group's model was or was not among the selected best for the target domain, respectively. White boxes mean no submission. The last row gives the number of groups (ties) who submitted a selected best model for each target.

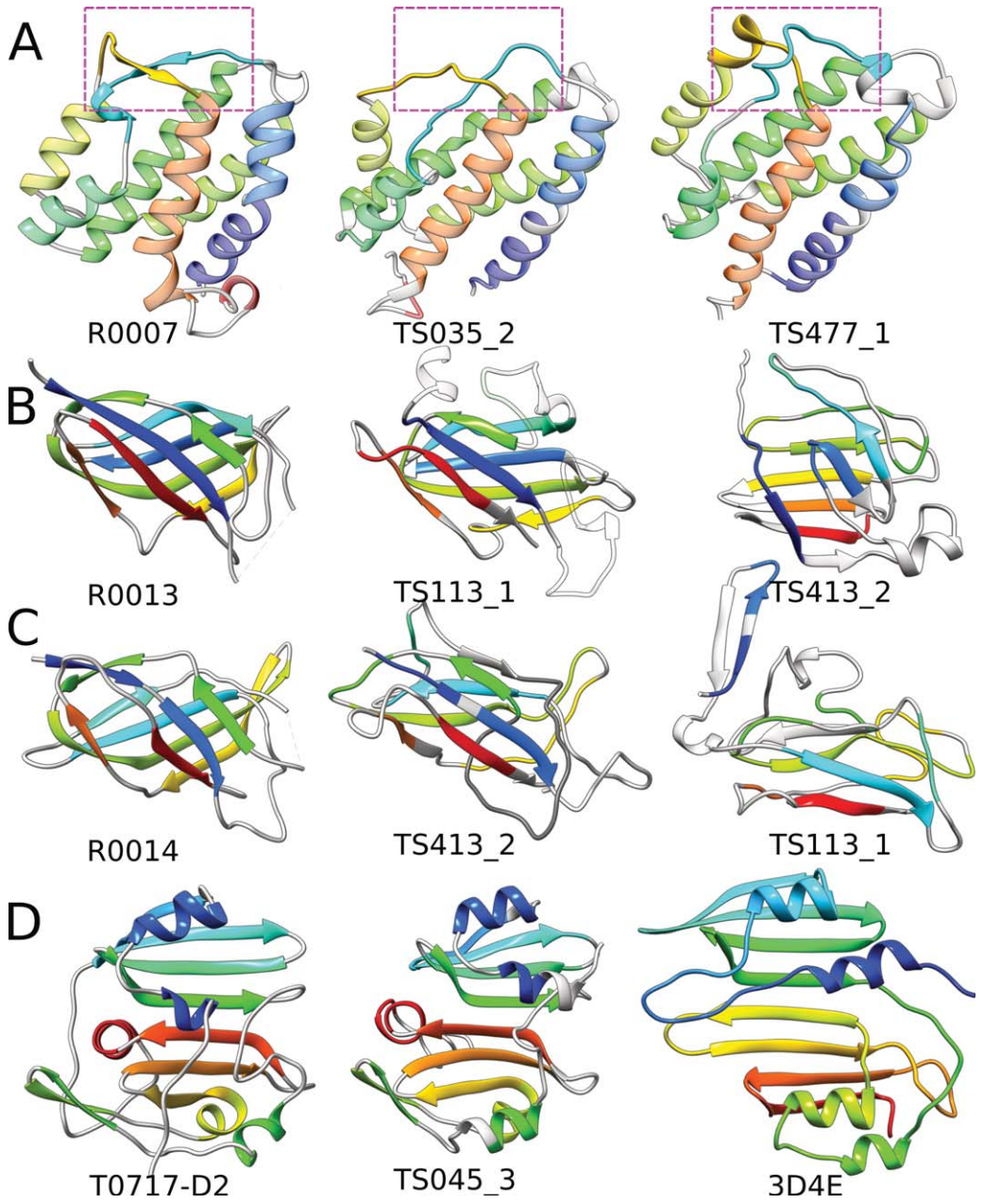

Figure 8

ROLL targets R0007, T0013, R0014, T0717-D2. (**A**) R0007 target structure (left), the visually selected best model TS035_2 (middle), and the model TS477_1 with the highest GDT_TS score (right). Magenta box highlights the crossover of the H1–H2 and H5–H6 linkers in the first two structures. The linkers did not cross in the GDT_TS top-scoring model. (**B** and **C**) Target structures R0013 and R0014 (left column) have the same fold. Each of two different groups (113 and 413) submitted a best model for one of the two targets (middle column), but a rather poor one for the other (right column, shown are the model with the highest GDT score from the given group). (**D**) Target structure of T0717-D2 (left), the selected best model (TS045_3, GDT_TS score 48.8, middle) and the template 3d4e found by HHpred (right). This template was used as parent for many models but the arrangement of the two substructures is tail-to-head, unlike the tail-to-tail arrangement of the target.

of groups submitting best models for a given target (number of ties) is high for server targets, 11 in four cases, but only one or at most two for the nonserver targets. This is mainly, although not entirely (see Tables II–IV), due to the large size of best clusters for server targets. This is another indication that good server models, and perhaps some not so good models as well, were selected by other predictors and submitted without detectable change at the RMSD 3 Å level.

In FM, Keasar (315) group submitted best models for four targets, which was the most among all predictors. All four models are in the same cluster with a server

model. Pcons (267) was the next best human group with three best predictions. Pcons best models are also all for "server targets" since this is a model quality assessment group who only selects from server models (see the CASP10 meeting abstract). QUARK (114s) was the best server with three best predictions, followed by Zhang-Server (035s) and BAKER-ROSETTASERVER (330s), each with two best predictions [Fig. 7(A)].

In ROLL, after excluding the three targets mentioned previously, the targets used for ranking were 20 in-season target domains and 16 off-season target domains. Keasar (315) ranked top, who submitted best models for eight targets. Three prediction groups, Zhang-Server (035s), Zhang_Ab_Initio (045), and PconsQ (428) were tied for second with seven targets each. QUARK (114s), Zhang (237), Pcons (267), ProQ2 (388), and Multicom (489) tied for third with six targets each. According to the CASP10 meeting abstract, Keasar, PconsQ, Pcons, ProQ2, and Multicom focus on quality assessment (QA) and selected their models from the server models. Keasar and Multicom also refined the selected models. Two other human predictors, Zhang_Ab_Initio (045) and Zhang (237), also used server models in the model generating process, but did not rely on them entirely. In particular, Zhang_Ab_Initio almost perfectly complemented Zhang-Server and had 13 different hits between them.

Results for individual targets

Here we briefly describe each target domain, the quality of the models we examined, and the considerations that went into selecting what we judged to be the best models. In the description of the models, we often cite the cluster ID, which is what we used to identify models during visual inspection. It is also the ranking by the composite score, sum of robust *Z*-scores (see Methods), starting from zero. Tables II–IV give the names of the models in all the selected best clusters. The FMTopClusterMembers.pdf and ROLLTopClusterMembers.pdf files in the Supporting Information give the names of the models in all clusters in the Top15Union set for all FM and ROLL targets. To save space, only a few selected targets are described subsequently. Other targets are described in the FMbestModels.pdf and ROLLbestModels.pdf files in the Supporting Information. (The short phrases in parentheses following the target names are only to remind one of the target structures.)

T0653-D1 (LRR with opposite bend)

T0653-D1 is the only TBM/FM target in CASP10, which has a helical structure with 17 LRR-like motifs. Unlike most known LRR proteins, T0653-D1 is smoothly bent with the β-sheet side of the solenoid convex [Fig. 4(C)]. Majority of the models in Top15Union set have an LRR structure but are bent in the conventional manner, much like the sequence-findable template 3bz5A [Fig. 4(F)]. The purpose of the additional FM designation for this target was solely to see if any group predicted the opposite bend.

We found only two models, TS201_2 and TS335_1 (singleton Clusters 3 and 40, respectively), which were bent toward the β-sheet side, not smoothly as in the target, but by breaking at one or two places. We chose only the former because the latter had three pieces with two breaks, only one of which was bent in the fairly correct direction. These are two of some 54 models from 20 groups that were built using 3sb4 [Fig. 4(G)] as the template. This template was found by the crystallographer of this target by a structure similarity search using Dali,[22] but contains a break in the middle of the LRR structure as in TS201_2 [Fig. 4(H)].

T0658-D1 (nine-stranded barrel-like β-sandwich)

This 166-residue β-sandwich is made of six strands in one curved sheet and three in the other sheet to complete a barrel-like structure. There is a 35-residue loop between Strands 4 and 5 covering the outside of the larger sheet [Fig. 9(A)]. PSIBLAST finds several templates including the C-terminal domain residues 506–647 of 1wcqA [Fig. 9(B)], which has sequence identity of 29.2% and E-value of 2E−75. TMalign finds 1rxlA [Fig. SF2(C) in Supporting Information] with TMscore of 0.53 and RMSD of 3.2 Å, but this template is difficult to find from sequence search since the sequence identity is only 8.3% and the HHpred probability is only 22.8. Neither template has the long loop between Strands 4 and 5.

This is a difficult target and even the best models are poor in quality. Neither the representative of Cluster 0 [Supporting Information Fig. SF2(D)], which ranked best by GDT (24.7), QCS, CoDM, and DFM and second best by Handedness score, nor that of Cluster 18, which was the most popular with 13 members, had the β-sandwich architecture. Only 8 of the 30 clusters in the Top15Union set could be said to have a β-sandwich fold, but the β-sheets are not well developed in one of these (114_5 of Cluster 10), while the topology of the strands is poor in two others (TS223_1, TS237_3, and TS430_1 of Cluster 13 and TS490_1 of Cluster 3). We selected the six models in the remaining five clusters as the best models for this target. Five of these (from Clusters 1, 5, 6, and 7) are all from Seok group (473) and similar to each other [Supporting Information Fig. SF2(F)]. The other is TS463_1 (Cluster 4) from Jiang_Fold server [Fig. 9(C)] who used 1wcq as the "PARENT" for the model. All have large alignment errors, apparently stemming from mishandling the 35-residue loop that is missing in the template structures.

T0663-D0 (α/β tandem repeat, head-to-head)

T0663 is made of a two similar domains whose sequences share 65.4% similarity and 34.6% identity.

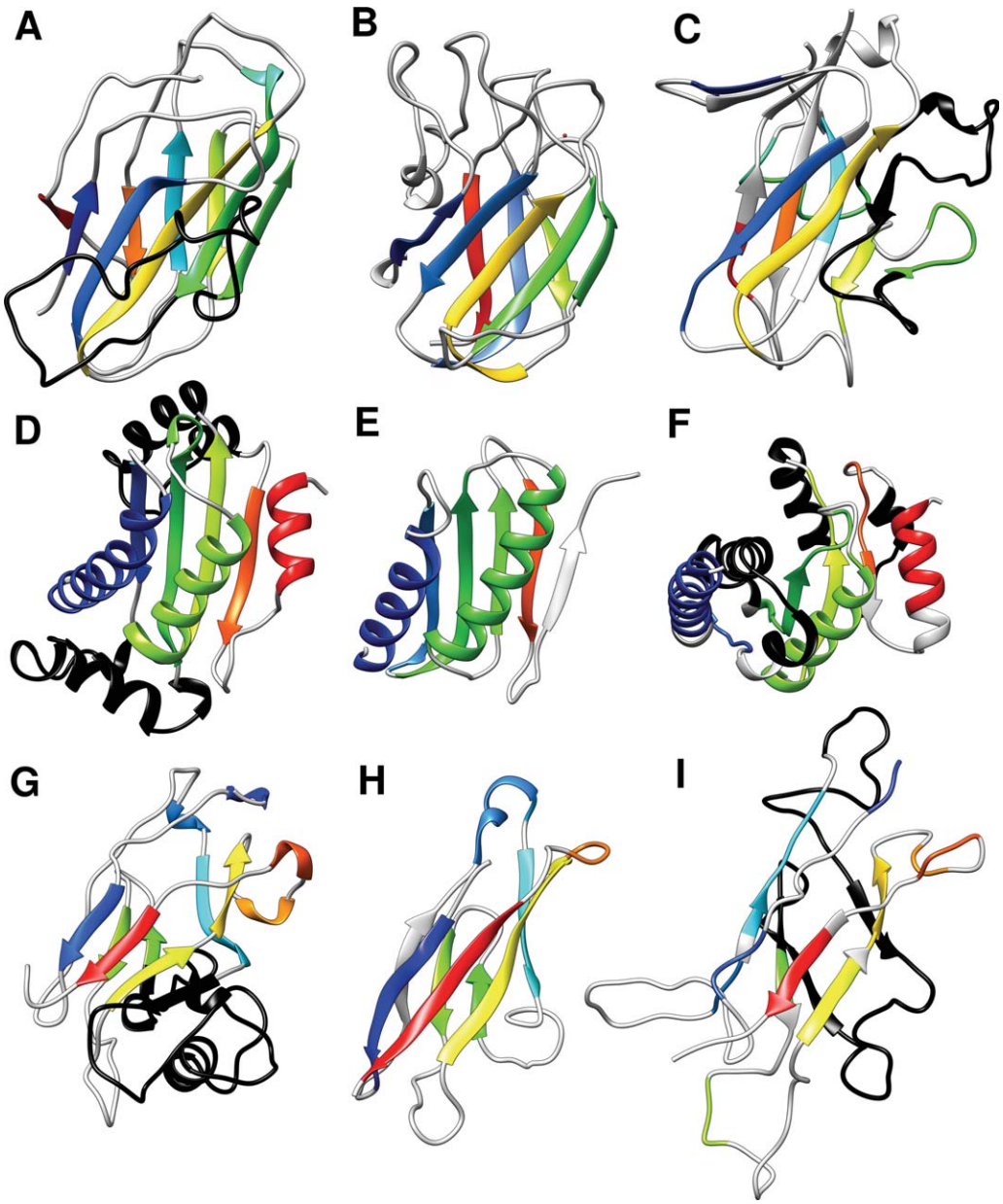

Figure 9

T0658-D1, T0684-D2, and T0719-D6 for which structural templates exist. All structures are rainbow colored according to the secondary structure of the target. The loops that are not in the template are in black. Left column shows the targets, middle the templates, and the right a selected model. (**A**) Target T0658-D1. (**B**) 1wcqA C-terminal domain residues 506–647, found by PSIBLAST. (**C**) Model TS463_1. (**D**) Target T0684-D2. (**E**) 3ib5A residues 275–381, found by TM-align. (**F**) Model TS045_3. (**G**) Target T0719-D6. (H) 3qrbA residues 1–100, found by TM-align. (**I**) Model TS315_2_6. (Detailed description for T0684-D2 and T0719-D6 are in FMbestModels.pdf in Supporting Information.)

Each domain consists of an N-terminal helix on a simple four-stranded antiparallel β-sheet and is a TBM target.[4] The two domains are arranged side-by-side forming one continuous eight-stranded β-sheet as is often the case for tandem repeats of such structures. However, how the two domains arranged in T0663 appears to be special. Usually, the two are arranged tail-to-head so that the last strand of the N-terminal domain is hydrogen bonded and antiparallel to the first strand of the second domain, as in T0644 and in the template 3u1wA found by Pfam. The two domains are related by a translation or a rotation around an axis that is outside of the molecule. In T0663 [Fig. 5(A)], the relation is head-to-head so that the *first* strand of the first domain, S1, is hydrogen bonded and antiparallel to the first strand of the second domain, S5. The two domains are related by a two-fold rotation axis that is perpendicular to the eight-stranded β-sheet. This arrangement requires a long linker between

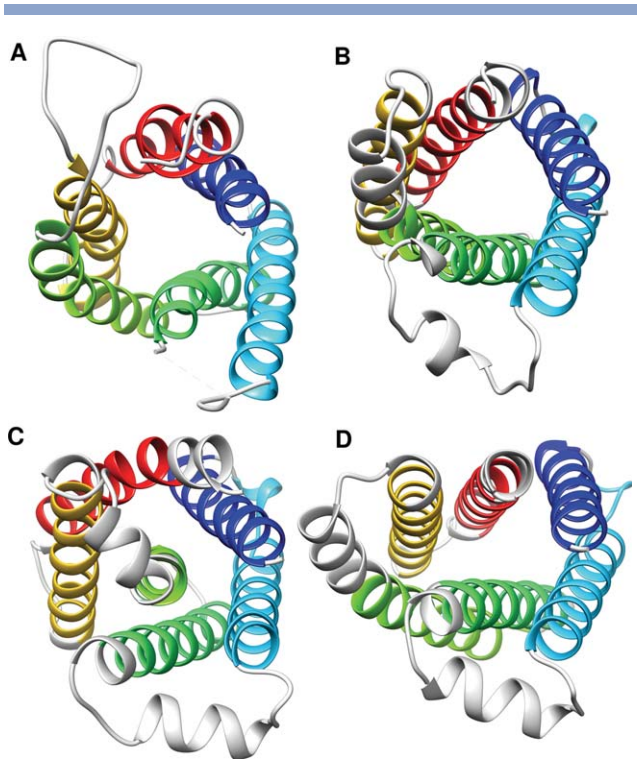

Figure 10

T0666-D1, a urea channel membrane protein. (**A**) Target T0666-D1, a six-helix barrel. (**B**) Model TS079_1 from TASSER is the manually selected best model. (**C**) Model TS475_2 with the highest GDT_TS score (33.8), representing 10 models in the cluster. The third helix is inside a five-helix barrel. (**D**) Model TS237_1 with the highest QCS score, representing 12 models in the cluster. The helices are arranged in two layers of three helices each.

the two domains, which spans the whole first domain. The D0 EU for this target was defined solely to assess the quality of the predictions of this unusual interdomain arrangement.

Most of the predictions had the two domains modeled well individually, but none had them arranged correctly with respect to each other. In many models, including those with the best GDT_TS and QCS scores [Fig. 5(B,C)], the two domains interact so that a part or whole of the C-terminal β-sheet interacts with the N-terminal β-sheet like in a β-sandwich. A few other models are also sandwiches, but with the helices between the β-sheets. Still others have the two domains separated with the two β-sheets nearly facing each other as in the template 3nqzA found by HHpred. Among the models that do have the β-sheets side-by-side, many have the two domains arranged tail-to-head or head-to-tail. Only six singleton clusters had the correct head-to-head orientation. In two of these (TS237_4 and TS172_4), one domain is flipped with respect to the other so that the interacting strands S1 and S5 are parallel to each other and the helices are on opposite sides of the β-sheet. In another (TS081_3), the two domains are separated and the strand directions are nearly orthogonal to one another. We chose the remaining three models [TS043_2, TS152_1, and TS152_4 from Clusters 36, 45, and 50, Fig. 5(D–F), respectively] as the best models for this target because they had the topology of the interaction essentially correct. Each of these has a major defect: the domains are separated and the two β-sheets do not form one single β-sheet in TS043_2; S4 is not in the sheet in TS152_1; and S5 is not in the sheet in TS152_4, making S1 to interact with the parallel strand S6.

T0666-D1 (six-helix bundle urea channel)

T0666-D1 is one of the only two membrane proteins we have in CASP10. It is an up-and-down six-helix bundle with a small channel in the center [Fig. 10(A)]. All 36 clusters in the Top15Union set predicted 6 helices, but most had them arranged without the central channel, which is essential for the acid-activated urea channel activity of this protein.[23] There were two main types among the models with the up-and-down six-helix bundle: one helix surrounded by five others [Fig. 10(C)] or a two-layer structure with three helices in each layer [Fig. 10(D)], such as the top QCS cluster with only a narrow space in the center. Only one, TS079_1, had a barrel-like structure with a central channel [Fig. 10(B)]. One of the "PARENT" templates it used is 1e12, a chloride pump membrane protein (see Discussion).

T0690-D0 and T0713-D0 (kinked LRR)

Both targets are LRR structure with a kink in the middle where the front and the back solenoids form a nearly 90° angle, with the β-sheet sides of the solenoids on the convex side of the kink [Fig. 4(D,E)]. The front and back solenoids are each defined as a TBM target while the D0 EU is solely to assess the prediction of the kink and the relative orientation of the two domains.

The quality of predictions of the whole molecule is generally poor for both targets—the maximum GDT_TS scores are around 30 and the median GDT_TS below 25. Most models had a regular LRR in smooth horseshoe shape without a kink and the β strands were on the concave side, as in many templates found by sequence search. TS344_4 was the only model that recognized the unusual kink and put the β-sheets of the LRRs on the convex side of the kink for T0690-D0 [Fig. 4(I)]. The model with the top GDT and QCS scores and a few others for T0690-D0 recognized the kink, but folded the two segments perpendicularly, similar to the template 3sb4A that they found [Fig. 4(G)]. For target T0713-D0, we selected three models from Clusters 0, 2, and 18 as the best because they were the only ones that had kinked LRR with at least one segment with the β-sheet on the convex side of the kink [Fig. 4(J)]. They are all from

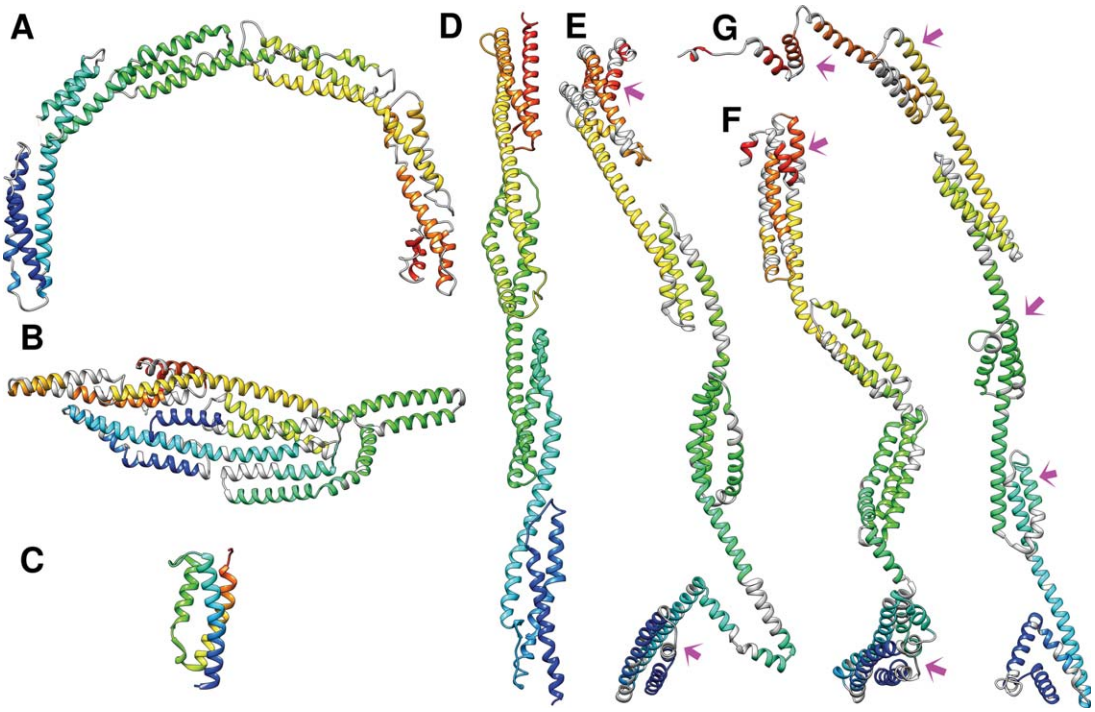

Figure 11

The largest FM target T0695-D1. Target and the models are colored according to the secondary structure of the target. Templates are rainbow colored. (**A**) Target T0695-D1 contains six repeating units of up-and-down three-helix bundle. All repeating units have the same, left-handed chirality. (**B**) Model TS172_4 with the highest GDT_TS score. (**C**) The three-helix bundle of a template 1gvnA, which has the same chirality as that of the repeating units of the target. (**D**) Template 1hciA of the spectrin fold used by 31 models. All repeating units in this structure have the opposite chirality to that of the target. The three representative models of the best clusters picked manually each have at least four up-and-down three-helix bundles in an open structure and at least two units of correct chirality indicated by the magenta arrows. They are (**E**) TS330_4, (**F**) TS479_4, and (**G**) TS201_1.

Shortle group (317) and share some similarities. One of these (TS317_1_1) ranked below 15 by all scores except the Handedness score (Table II).

T0695-D1 (rainbow-shaped three-helix bundle repeats)

This 535-residue single chain target is the largest ever in the FM or New Fold category. Understandably, the average and the best GDT_TS scores are also the lowest among all targets. It consists of six repeating units bent like a rainbow [Fig. 11(A)]. Each repeating unit is an up-and-down left-handed three-helix bundle as in the structural template, 1gvnA [Fig. 11(C)]. The repeating units are connected by extending the third helix of a unit to the first helix of the next unit, as in spectrin domain repeats [e.g., 1hciA, Fig. 11(D)], which however has the three-helix bundles with the opposite chirality.

Although most models had the helical secondary structures predicted correctly, not all recognized the three-helix bundle repeating units, few had the correct chirality for the helix bundles, and none had them arranged in a semicircular fashion as in the target. Therefore, we judged models on the basis of the existence and the chirality of the helix bundles and decided to ignore the overall shape of the molecule, which could change in any case because of the flexibility inherent in such a large open structure. The six highest ranking models by GDT_TS score all had several repeating units packed together [Fig. 11(B)]. The three clusters of six models that we selected as the best had five or six clearly separated three-helix bundle units, at least two of which with the correct chirality [Fig. 11(E–G)]. They all ranked within top five by CoDM and top 15 by Deformation score. Model TS201_1 [Cluster 9, Fig. 11(G)] ranked top by CoDM. It has five of the six three-helix bundles, four of which with correct chirality. Most of the 30 models that were built using 1hciA as the template had the wrong chirality for the helix bundle repeats. The four models in Cluster 6 [Fig. 11(E)] were exceptions; they had five units, two of which had the correct chirality. Six three-helix bundles are recognizable in TS479_4 [Fig. 11(F)], of which two have the correct chirality.

T0717-D2 (α/β tandem repeat, tail-to-tail)

T0717-D2 is made of two substructures [Fig. 8(D)] with rotational symmetry. The N-terminal subdomain has a helix followed by a four-stranded antiparallel sheet.

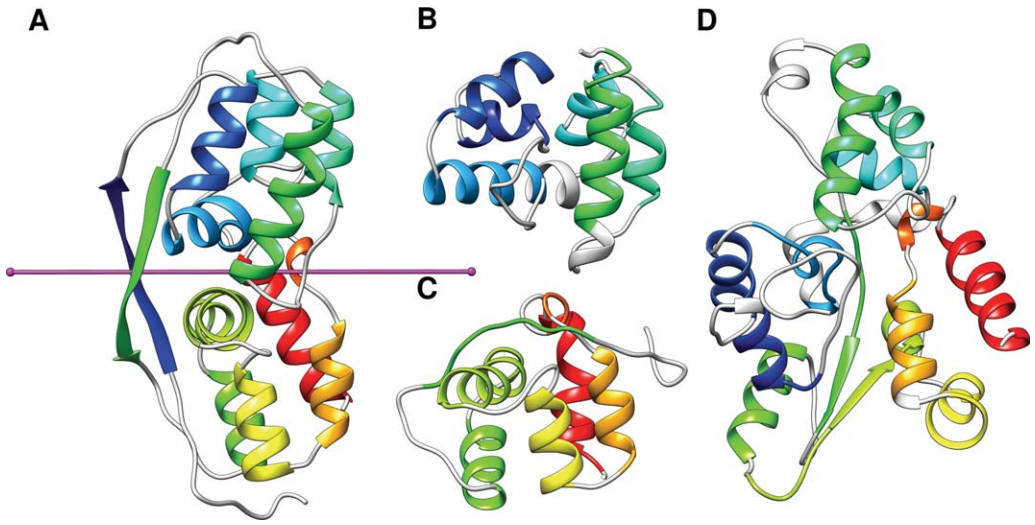

Figure 12

T0734-D1. The target and models are rainbow colored according to the target secondary structure. Target T0734-D1 (**A**) has a two-fold symmetry axis, shown in magenta, as calculated by using the program SymD.[25] Models TS358_2_1 (**B**) and TS358_1_2 (**C**) are the selected best models. Each matches only one of the N- and C-terminal helix bundles of the target. The model with the highest GDT_TS, QCS, Handedness and Deformation scores is TS172_5 (**D**), which we did not select.

The C-terminal subdomain contains a short helix followed by a β-hairpin and a short helix, then another three-stranded antiparallel sheet. The last strand of N-terminal substructure forms hydrogen bonds with the last strand of C-terminal substructure to have a seven-stranded sheet. Many predictions have individual subdomains correctly modeled, but in a tail-to-head arrangement like the template found by HHpred, 3d4e [Fig. 8(D)]. Only TS045_3 (Cluster 0) has seven-stranded sheet and the two subdomains arranged tail-to-tail [Fig. 8(D)]. Its GDT_TS score is 48.8, about 15 higher than the best GDT_TS score of models from all other groups.

T0734-D1 (tandem repeat of a strand plus a short five-helix bundle)

This is a tandem repeat—the sequences of residues 1–115 and 116–214 share 56% similarity and 30% identity. The two repeats have similar structures, arranged in a two-fold symmetric manner [Fig. 12(A)]. A ribbon of two twisted β-strands, one from each repeat, holds two "globes." Each "globe" is made of an up-and-down four-helix bundle and a transverse helix. The transverse helices from the two repeats are parallel to each other and form the central part of the interface between the two repeats.

There was no model that had the β-ribbon with the proper residues, no model that had two approximately correct "globe" structures, and no model that had them arranged as in the target structure. TS172_5 [Fig. 12(D)] ranked top by GDT_TS, QCS, Handedness, and Deformation scores but it had only two-helix hairpins and one of the strands correctly oriented and wrong topology for the rest of the structures. We selected TS358_2_1 [Fig. 12(B)] and TS358_1_2 [Fig. 12(C)], both from the same server group RaptorX-Roll, as the best models. Each is a model of only the N- or the C-terminal "globes," respectively. Although missing the other "globe" and the β-ribbon is a major defect of these models, each had the correct topology of the "globe" that it modeled, including the transverse helix.

T0741-D1 (long, nearly closed β-ribbon with N- and C-terminal extensions in the middle)

This is one of the most difficult targets with the maximum and median GDT_TS scores of only 17.20 and 11.80 respectively. The residues 73–149 form a long twisted two-stranded β-ribbon, which is bent into a U-shape with two hairpin turns [Fig. 13(A)]. The N- and C-termini of this β-ribbon are near each other at the base of one of the arms of U. The N- and the C-terminal extensions participate in forming two three-stranded sheets, one overlapping with a base part of the U and the other outside of the ribbon, holding the N- and C-termini together.

Although this is an entirely β protein, surprisingly, two thirds of models in Top15Union set are entirely or mostly α-helical. No model had the entire β-ribbon, which is the most characteristic feature of this target structure. The models we selected as best are TS413_5 and TS344_2. Zhou-Sparks-X server found 3bghA as the template [Fig. 13(B)] and dramatically improved it to produce model TS413_5 [Fig. 13(C)]: one leg of the β-ribbon is modeled rather well, although the rest of the structure is

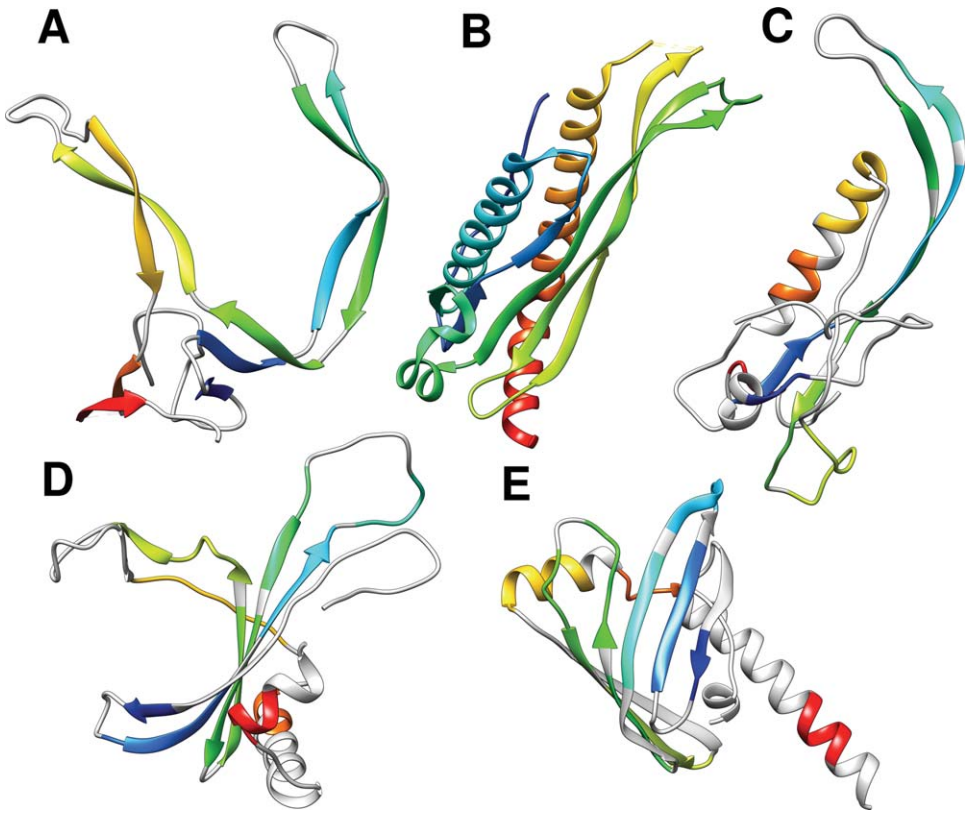

Figure 13

T0741-D1. The target and models are rainbow colored according to the target secondary structure. (**A**) Target T0741-D1. (**B**) Template 3bghA which model TS413_5 used. (**C**) The selected model TS413_5 from Zhou-Sparks-X server has the N terminal half of the U-shape twisted beta-hairpin correct. The structure of the rest of the domain is wrong. (**D**) Another selected model TS344_2 has two β-ribbons, which are partially separated. (**E**) The GDT_TS top rank model TS381_1, which we did not select.

entirely wrong. It was ranked best by QCS and second best by GDT_TS. Both legs of the β-ribbon are discernable in TS344_2 [Fig.13(D)], albeit with large alignment errors, but they are folded together to form a large four-stranded sheet instead of forming an open U-shaped structure. Only two other models (TS344_5 and TS152_1) have detectible β-ribbon parts, but they are much inferior to those mentioned previously. Model TS381_1 has the highest GDT_TS score, but the strands are arranged to form an up-and-down half β-barrel [Fig. 13(E)].

R0001 (homotrimer)

R0001 is a unit of tightly interacting homotrimer (Fig. 3). In each monomer, the N-terminal domain (D1) is composed of nine strands (S1–S9) and a short helix. S1, S2–S6, and S7–S9 are part of three different twisted sheets in the trimer. This makes it hard to correctly predict D1 as a stand-alone structure. We focused on whether the overall structure was correctly preserved and how the secondary structure elements were predicted. The best noncollapsed structure is TS292_4 (Cluster 0), although S5 and S6 are not aligned with S2–S3 and the short helix is disordered. The five similar models in Cluster 1 have all the secondary structural elements correct, but the helix folded back and is in contact with the sheet. These six models were selected as the best for R0001-D1.

The C-terminal domain (D2) of R0001 is a β-sandwich composed of two four-stranded sheets, with a special strand topology. We selected two clusters as the best models for R0001-D2. One is TS045_5 (Cluster 5), a β-barrel structure that has the correct topology for six of the eight strands. The other is Cluster 0 of four similar β-sandwich structures, which are composed of two four-stranded sheets, with four of the eight strands in correct relative positions.

R0007 (crossover of linkers)

R0007 is essentially an up-and-down bundle of six helixes (H1–H6). H1 is bent in the middle. The linkers connecting H1–H2 and H5–H6 cross each other and form an antiparallel two-stranded sheet, highlighted with magenta box in Figure 8(A). The crossover makes it topologically more difficult to unravel the protein from the N-terminal relative to the C-terminal side of the

protein. Many predictors predicted the six-helix bundle correctly. But only a few predicted the crossover of H1–H2 linker and H5–H6 linker correctly. We selected TS035_2 [Cluster 5, Fig. 8(A)] as the best model because it is not only a six-helix bundle, but also has the correct crossover of H1–H2 and H5–H6 linkers. In TS477_1, the top GDT_TS scored model, the linkers do not cross over so that the N-terminal is topologically easier to unravel than the C-terminal [Fig. 8(A)].

R0009 (homopentamer)

R0009 is a unit of a homopentamer [Fig. 3 (A,C)]. The head domain (D1) has 1 strand and two helices connected with long loop. It extends, wraps around, and interacts with four other head domains to form the head part of the pentamer. Similar to R0001-D1, it is hard to predict the correct structure of R0009-D1 without oligomeric context. We selected TS165_2 [Cluster 7, Fig. 3 (C)], which resembles R0009-D1 best. Interestingly, the structure was obtained in a different context—intimately interacting with a malformed D2 domain.

R0009-D2 is composed of eight strands (S1–S8), which form a β-sandwich of two four-stranded sheets, and a short helix at the N-terminus [Fig. 3 (C)]. We selected Clusters 0 and 1 as the best models, because Cluster 0 has the correct β-sandwich structure, although with a large alignment error and S1 out of position; while Cluster 1 is similar to Cluster 0, except that it correctly predicted the secondary structure of an additional short helix.

R0013 and R0014 (two structurally similar targets)

R0013 and R0014 share a similar structure. They are composed of a β-sandwich with seven strands, three in one layer, and four in the other [Fig. 8(B,C)]. The sequence identity between R0014 to R0013 is 20.5%. TS113_1 (Cluster 0) and TS113_2 (Cluster 1) are clear winners for R0013, because they get the relative positions of all the strands correctly, which no other predictions did. We selected TS113_1 as the top model of R0013 [Fig. 8(B)], as it is visually slightly better than TS113_2. For R0014, we selected TS413_2 (Cluster 1) as the top model because it has the seven strands in correct topology, except S3 is shifted and the missing residues between S3 and S4 in the target became a loop wrapping around S1 and S7 to form a knot [Fig. 8(C) middle panel residues in dim gray]. It is interesting that groups 113 and 413 submitted the best model for only one and different one of the two similar targets.

DISCUSSION

Roll target selection

In the regular CASP experiment, target domains are classified as FM or TBM manually during the evaluation process when both the target and the model structures are known. In contrast, the off-season ROLL targets must be selected before any of the structures are known. Nonetheless, the ROLL chain filter that the Prediction Center designed was highly effective in spotting FM and hard TBM target domains when applied to the CASP10 targets. The number of CASP10 targets that passed the ROLL chain filter was 22, of which 19 included at least one FM or TBM-hard domain. The filter missed only one FM (T0756-D2) and one TBM-hard (T0690-D1) domains, if FM D0 and decoration/linker domains are excluded.

The off-season ROLL targets include domains that are TBM, as well as FM, targets when classified by the same procedure used to classify the CASP10 targets. This is because an off-season ROLL chain can contain more than one domain, some of which are TBM, while other(s) FM, targets. Even a single-domain ROLL chain can be a TBM target because ROLL chain selection criteria were not identical to the FM/TBM classification procedure used for CASP10 targets. Since ROLL experiment was to focus on FM targets, we initially classified the off-season target domains manually in the same manner as we classified the CASP10 target domains and intended to use only the FM target domains. However, a concern was expressed about excluding the TBM-hard target domains. These are the difficult TBM domains for which predictions' maximum GDT score is less than 50.[4] Usually the TBM-hard target domains do have at least one template, but it is either poorly similar to the target structure or difficult to find because of low sequence similarity. Unlike in-season TBM-hard target domains, which will be evaluated by TBM assessors, the off-season TBM-hard target domains would not be evaluated at all if we evaluated only the FM target domains. Therefore, we decided to expand the target list by including TBM-hard target domains in the evaluation, but with a slight modification of the definition of TBM-hard—we used 90 percentile GDT cutoff rather than the maximum GDT. This was not to reject targets for which a few (not more than 10% of the predictors) did exceptionally well.

Automatic classification (90%GDT < 50) has the advantage of being simple and easy. However, it lets in targets such as T0739-D3 and T0739-D4, which have a well-known fold (β-helix) but low GDT_TS score (Fig. 1). Many of its models appear to be template-based and the low GDT_TS score arises from alignment shifts. We realize that many CASP10 FM models are also template-based (see subsequently), but the automatic classification and the expansion of the target base would increase such incidences and would tend to diffuse the focus on true template-free modeling. A related but different problem is the occurrence of targets such as T0671-D2 and T0678. After excluding obviously bad models, we were left with a large majority of models in the Top15Union set that were of similar quality by visual inspection (see

Results), presumably because they were all based on one or a few templates. In order to distinguish these models, a quantitative measure needs to be used and selecting just the top and ignoring next best models seem inappropriate.

Visual inspection is necessary for FM model evaluation

Notwithstanding exceptional cases like those mentioned previously for ROLL targets, the quality of models for FM targets is generally poor and the models often do not resemble the target structure. In such cases, a similarity score based on structural superposition like GDT_TS will be unreliable. Generally, when two structures have different merits and different deficiencies, it can be unclear what standards should be used to judge one model as better or worse than the other. The principle we used during visual assessment was to establish a few important features/aspects of the target, then judge the models in terms of how well the model reproduced the selected features. The features we focused on varied from target to target (see individual target descriptions in the Results section and in FMbestModesl.pdf and ROLLbestModels.pdf files in the Supporting Information). For example, for the D0 targets, the emphasis is on the interdomain relation and not the quality of individual domains, which would normally be the concern for other targets (see Fig. 5 and Results). In the case of T0653, which is an LRR smoothly bent in a nonconventional fashion and the only target with both TBM and FM designation,[4] it is the direction of the bend rather than the local structural fidelity of the LRRs that is the object of the FM evaluation.

It is difficult for any one automated score function to accurately reproduce human judgments on FM model quality. The model with the top GDT_TS score was selected as best visually for only four of the 20 FM target domains, the model with the best QCS score was picked for only eight domains, and the bests by each of the three other scores for five or less domains (Table II). There are eight FM targets for which the visually chosen best models did not score best by any of the five score functions. The cluster number is the rank by the sum of the robust *Z*-scores (see Methods), which is a composite score of the five scores. This score was only slightly better than the QCS score: the models with the best composite score, which are those in Cluster 0, were picked as best for 10 targets. The numbers are better for the ROLL targets; for example, top QCS models were selected as best for 21 of the 36 ROLL targets. This is presumably because ROLL targets include TBM targets, the models for which tend to be more similar to the target. Thus, good score functions help, but we believe that visual inspection should still be the final arbiter when the models are of poor quality.

Novel features of the evaluation procedure

As in previous CASP FM experiments, the challenge assessors face is how to filter nearly 10,000 models in such a way that it is unlikely that good predictions are excluded, yet the remaining set is small enough for a small group of experts to visually inspect them in a reasonable time. The approach we adopted is one way to do this, namely first cluster models on pairwise RMSD and then filter the cluster representatives with five score functions independently to collect a short list of cluster representatives for visual inspection.

The CASP9 assessors[3,6] also examined many models, but used a composite score of at least 10 different score functions as the sole measure of model quality. Here, we used only five different score functions, but our purpose was different. Our score functions were not intended to produce a single quality measure for each model, but to serve as parallel, not serial, filters. The five scores are quite different from each other and will each emphasize different aspects of the structure. The union of best strategy enables each score by itself to rescue models with some good features that other score functions miss.

Table II provides some experimental justification for this union of best strategy. Of the 46 clusters picked as best in Table II, seven entered the Top15Union by only one of our five scores, and another seven by only two scores. Moreover, no single score function found all clusters in Table II. GDT found 83%, QCS 61%, handedness 67%, CoDM 52%, and DFM found 59% of the clusters. There also are cases where each score greatly under-ranks a good prediction. Supporting Information Figures S3 and S4 provide further support for the union of best strategy. They show the correlations between pairs of scores (our five scores plus sequence dependent TM[24]) for TBM, TBM-hard, and FM targets. The correlation between scores is high for TBM models but much lower for FM predictions. In the case of T0734, for example, the correlation coefficient is at most 0.6 (between GDT and TM) and much lower than that for other pairs of scores (Supporting Information Fig. S3). The low correlation for the FM target models will enhance the variety of models with different good features in the union of top models. It also implies that an objectively good FM prediction will not necessarily score high under all or even most of the score functions used.

The Top15Union sets probably contain all best models for all targets

It is not possible to prove that we did not miss any worthwhile models without visually examining all submitted models. However, having identified best models by examining all models in the Top15Uinon set, we can examine how many we would have missed had we used a stricter rank cutoff criterion. The minimum ranks described in the Results indicate that we would have

retained all selected best models using the cutoff value of 13, except one (T0678), which is a TBM target and for which we needed 15 to retain the last of the five clusters (see above). These observations do not prove but strongly indicate that all top models are probably included in the Top15Union set for all targets.

Ranking by the first tier membership count

An obvious concern with visual inspection as the final arbiter is that the results are difficult to quantify and tend to be subjective. However, it is our experience, both in CASP6 and in this CASP, that the first tier of best FM models can be identified in a reproducible manner. First, we select as best more than one model with different list of merits and deficiencies if we judge that a strong case cannot be made for choosing one and not the other. All models in a cluster are included if its representative is selected. Secondly, we only identify the first tier of models and do not attempt to differentiate the remainder, which is more difficult. The ranking is, therefore, elitist; it is based on the best models submitted for each target and no credit is given to the second best. We feel that this is not a serious drawback because even the best models were poor for most FM targets. Nevertheless, we tended to be generous in selecting the "best" models to include all potentially competitive models. Finally, we relied on four examiners, who judged independently using only the cluster numbers, which varied from target to target so that they were blinded not only on the identity of predictors, but even on the frequency with which a predictor ID showed up in the first tier list. The final decision was by consensus among the four examiners, with one person (BL) who had the final say when the consensus could not be reached. The procedure is not simple or straightforward. However, we all understand the reasons for our choices and, when we revisited our choices, only one or two cases required a modification.

Target size dependence

The length of the CASP10 FM domain ranges from 58 to 535 residues. Six are longer than 200 residues including three D0 targets. Also, unseen in previous CASPs, three chains, T0713, T0719 and T0739, are longer than 700 residues, which contributed 4 FM domains. (For T0713, the eventual structure contained only 374 residues although the predictors had to wrestle with the full 739 residues.) Generally, one would expect that larger domains are more difficult to predict. Also, chain length may matter since a large chain may contain more than one domain and domain parsing presents an additional opportunity for error. However, we found little correlation between the score and the length of the protein (Supporting Information Fig. S5). The maximum GDT_TS score correlates somewhat better with the domain size than the whole chain length, with correlation coefficients of 0.271 and 0.017, respectively.

Prediction of oligomeric structures

Two of the most challenging targets in this ROLL exercise are R0001-D1 and R0009-D1, whose structures are most probably not maintained outside of their tightly interacting oligomeric state. Successful prediction of these structures requires simultaneous prediction of the oligomeric structure of the protein, which remains challenging.

Most successful was selection/refiner of server models, not an *ab initio* folder

Since good templates do not exist or are difficult to find for the FM targets, one would not normally expect that prediction techniques that rely on selecting among existing structures to do well with these targets. However, as mentioned in the Results section, the predictor who submitted the winning models for most targets was Keasar, who selected best models from among the server models and refined them (CASP10 meeting abstract and private communication). Keasar's five winning models for four targets were in the same clusters as the seven models from six different servers. The fact that this group was more successful than the servers indicates the group's prowess at spotting good models. We also note that the estimated odds of picking best models for four targets by random trial is merely 0.02% (Fig. 6; Supporting Information Table S3) and that Keasar is the top performer for the ROLL experiment as well (see Results and subsequently). On the other hand, the fact that this group's models were in the same cluster as the server models indicates that Keasar's refinement did not produce noticeable changes at the 3.0 Å RMSD level.

Use of templates for template-free modeling targets

Keasar's success makes it clear that selecting from server models and possibly refining the selected model(s) is a highly viable strategy. CASP9 FM assessors[3] noted the emergence of this type of predictors, whom they called "metapredictors." We also saw indications of template usage for these supposedly template-free modeling targets from visual inspection of individual models (e.g., T0658-D1) and from the pattern of errors in the model detectable in the position-specific alignment plots that the Prediction Center provides. Templates are difficult to find for these targets. But many predictors found templates and improved on them to produce the best models for the target. Some examples are TsaiLab produced TS201_2 for T0653-D1 apparently using 3sb4 as the template (Fig. 4); for model TS330_4 from BAKER-ROSETTASERVER for T0695-D1, three parents are listed,

Table V
Top FM Target Predictors Categorized[a]

Group name	ID	#Hits	Type[a]	Group name	ID	#Hits	Type[a]
TsaiLab	201	2	IC	keasar	315	4	R
Jones-UCL	344	2	IA	Zhang_Ab_Initio	045	2	O
Cornell-Gdansk	152	1	IA	Mufold2	405	2	R
LEE	301	1	IT	TASSER	079	2	R
Seok	473	1	IC	Boniecki_LoCoGRef	479	1	R
sessions	043	1	IT	Mufold	197	1	R
wfCPUNK	287	1	IA	MULTICOM	489	1	R
Zhang-IRU	172	1	IA	SHORTLE	317	1	R
Combined[b]		8	I	zhang	237	1	O
				Combined[b]		10	R, O
QUARK	114	3	SC				
BAKER-ROSETTASERVER	330	2	ST	Pcons	267	3	Q
Zhang-Server	035	2	ST	4_BODY_POTENTIALS	164	2	Q
chunk-TASSER	488	1	SC	CNIO	475	2	Q
FALCON-TOPO	411	1	SC	Kloczkowski_Lab	350	2	Q
FALCON-TOPO-X	456	1	S	Pcomb	130	2	Q
IntFOLD	498	1	ST	PconsQ	428	2	Q
Jiang_Fold	463	1	ST	ProQ2	388	2	Q
Jiang_Threader	258	1	S	McGuffin	285	1	Q
MUFold_CRF	148	1	S	ProQ2clust2	280	1	Q
MUFOLD-Server	333	1	ST	Combined[b]		5	Q
Pcons-net	292	1	ST				
PMS	108	1	ST	shisen	029	1	N
RaptorX-Roll	358	1	ST	SIAT1068	111	1	N
TASSER-VMT	335	1	ST	sumalab	437	1	N
ZHOU-SPARKS-X	413	1	ST				
Combined[b]		11	S				

The prediction groups who submitted at least one top model are grouped into different types (type) and sorted by the number of targets for which the group submitted a best model (#Hits).
[a]Group type codes are I: independent of server models; S: server; R: select from server models and refine; O: omnivorous—select from templates and/or server models and refine; Q: select from server models and submit; and N: classification unknown. The I and S groups are further classified, when known, as A: *ab initio*; T: template-based; and C: combined approach—use both *ab initio* and template-based. The group type assignments reflect our best judgment based on the methods described in the CASP10 meeting abstract. The assignments, particularly among the IA, IT, IC, R, and O types, can be erroneous or debatable as they depend on the extent and manner with which the templates and server models are used.
[b]The number of targets for which any of the groups of the indicated type submitted a top model.

each of which is not as good as their model (Fig. 11); Zhou-SPARKS-X server produced TS413_5 for T0741-D1 by improving on the template 3bghA (Fig. 13); and TASSER, who refines from server models according to the CASP10 meeting abstract, nonetheless reports five templates including the chloride pump 1e12A as the parents of their model TS079_1 for T0666 (Fig. 10, see subsequently).

In order to compare prediction strategies and to see how successful they were, we labeled the prediction methods by five broad types according to their relation to the servers (Table V). We then grouped the predictors who submitted at least one winning model according to these prediction types as best as we could. These groupings were made according to the predictor's description of their methods in the CASP10 meeting abstract, but the description in the abstract is sometimes insufficiently detailed and some techniques are inherently difficult to classify. What follows is, therefore, meant to describe trends and not to be treated as a report of numerically accurate statistics.

We see in Table V that there were nine quality assessment type (Q Type in Table V) predictors who successfully selected at least one best server model and submitted it/them without modification. As a group,

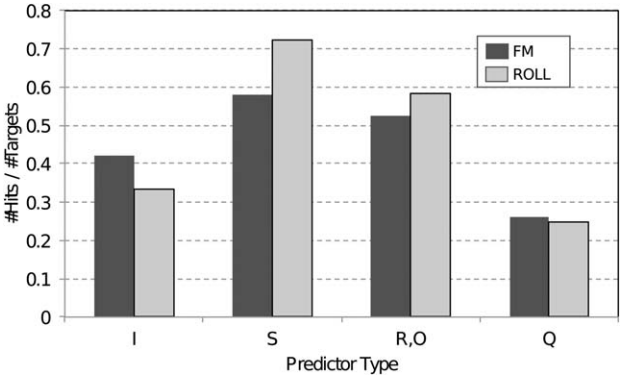

Figure 14
Fraction of the number of hits (#Hits) by different type of predictors for FM and ROLL targets. Predictor type codes are I: independent of the servers; S: server; R: select from server models and refine; O: omnivorous—select from templates and/or server models and refine; Q: select from server models and submit.

however, they were not particularly successful, selecting best models for only 5 of the 11 "server targets."

There were also nine select-and-refine types (R and O types) who selected at least one best server model and refined them. This group includes Keasar and produced best models for 10 targets, one less than the number for all servers as a group. However, they include three "nonserver targets," two contributed by Zhang_Ab_Initio and one by TASSER. The former predictor uses both server models and templates and the best models for the two "nonserver targets" could have come from using templates. TASSER is an R type according to the CASP10 meeting abstract, but also lists five templates as parents of their model for T0666-D1 (see above). These templates are what TASSER-VMT server used for their five server models. Whether they started from templates or server models, they improved them mainly by creating a channel in the middle of this transmembrane structure (see Results). It should also be noted that the select-and-refine type, successful as they were as a group, failed to recognize four of the 11 best server models.

The independent human groups, who are *ab initio* and template-based modelers, were not impressive individually: the best ones produced best models for only one target except for TsaiLab and Jones-UCL who produced two. However, as a group, their performance is respectable—eight of them produced best models for eight targets. Four of these were "nonserver targets." Best servers, who are *ab initio* and template-based predictors as well, performed similarly: most produced best models for only one target and even the very best, QUARK, for only three. But servers as a group did better than independent predictors since there are more servers who had at least one hit and they produced best models for 11 targets compared to 8 for the independents.

The performance of ROLL predictors is summarized in Supporting Information Table S4 in the Supporting Information. The numbers of winning models submitted for the ROLL targets by the ROLL predictors of different types are compared with those for the FM targets by the FM predictors in Figure 14. It shows that different predictor types performed rather similarly between the ROLL and FM experiments. One noticeable difference is that the ROLL servers, fewer in number (13 vs 16 for FM, Supporting Information Table S4), were more successful as a group than they were for the FM targets, whereas the independent human groups did less well for ROLL than for FM. The relative prowess of the servers may be related to the all-year-round format of the ROLL experiment—servers are ready all year round, but not all human groups may have been able to exert concentrated effort to the prediction during the short CASP off-season, which was between December 2011 to April 2012 for this first ROLL experiment.

Comparison of best predictors for ROLL and FM targets

The ROLL and CASP10 FM target sets are largely distinct—only 10 of the 36 targets are common. Also, the number of predicting groups is very different, 50 and 147 for the ROLL and FM exercises, respectively. Yet the rankings of top prediction groups are largely similar. Keasar has most hits (the targets for which the group's model was one of the selected best) in both exercises. Seven of the nine ROLL predictors with six or more hits (Keasar, Zhang-Server, Zhang_Ab_Initio, PconsQ, QUARK, Pcons, and ProQ2) are among the 16 FM predictors with two or more hits. Under the naive prediction null model, six and two hits are both just below the 5% significance level for the respective category (see Results). Although not identical, the similarity increases confidence in the ability of the top predictors to produce consistently better models and in the evaluation procedure adopted.

Quality of the best FM predictions

Three targets, T0737-D1, T0739-D1, and T0739-D2, have potentially new folds. The best models for the first two do resemble the target structure, but no model for T0739-D2 has the correct fold of the target (see individual target descriptions). The maximum GDT_TS scores, 41, 36, and 38, respectively, are relatively high presumably because of the small size of these domains and, in the case of T0737-D1, because of the helical fold. In comparison, the maximum GDT_TS scores for the six NFh (New Fold, hard) targets in CASP6, for example, were all below 31.[9] There are four other targets (T0659-D1, T0684-D2, T0719-D6, and T0735-D2) for which potentially useful template structures do exist, but the maximum GDT scores (25, 25, 27, and 42, respectively) for three of them are lower than those for the three mentioned previously. They are comparable to those for the NFh targets in CASP6.[9]

Four targets (T0651-D0, T0693-D1, T0726-D3, and T0756-D2) are decorations or linkers to other domains. Not many such targets existed in previous CASP experiments. Even the best models are poor for these targets (see individual target descriptions), except for T0756-D2, for which a surprisingly good model was submitted.

There are nine targets that are made of substructures or subdomains of well-known folds. The focus of evaluation for these as FM targets is not the structure of the individual substructures, but the relation between them. In two of these, the substructures are helices and the evaluation is on helix packing, which is a familiar topic. One is the urea channel protein T0666-D1, for which an impressive model was submitted (see above). The other is an $\alpha-\alpha$ superhelix-like protein T0740-D1, for which the best models clearly resemble the target structure. For other targets, however, the relations between the

substructures are more novel and the quality of predictions is rather disappointing. For example, the three targets (T0695-D1, T0741-D1, and T0734-D1) with the lowest maximum GDT_TS scores (17, 17, and 24, respectively) all have substructures of known folds. In the case of T0695-D1, the poor GDT_TS score is excusable since this is an open structure with six three-helix bundle subdomains and the overall shape of the molecule may be variable in solution. T0741-D1 and T0734-D1 are made of two or three substructures, but can also be considered as unique new folds. T0690-D0, T0713-D0, and T0653-D1 are LRR structures, first two kinked in the middle and the last smoothly bent with the β-sheet side convex. None of the predicted models for these structures satisfactorily reproduced these unusual features. T0663-D0 is another structure with two subdomains arranged in an unusual manner, which was not accurately reproduced in any model.

CONCLUSIONS

In CASP10 FM category and ROLL experiment, no single prediction group dominated. Even the most successful one submitted best models for only four of the 19 FM targets and eight of the 36 ROLL targets. Many, if not most, good models appear to have been produced by template-based modeling or the related technique of server model selection and refinement. Prediction of structures without a template, including prediction of novel interdomain relations, remains difficult. However, the presence of six or more *ab initio* folders among those who presented at least one best model gives hope that more progress will be made in the future on true template-free modeling. On the technical side, it appears possible to identify all top tier models by visual inspection through clustering, parallel prefiltering with well-chosen score functions, and a rapid visual aid tool.

ACKNOWLEDGMENTS

We thank the CASP10 organizers, John Moult, Anna Tramontano, Torsten Schwede Krzysztof Fidelis, and Andriy Kryshtafovych for providing us the opportunity to participate in the CASP10 experiment as an assessor; David Jones for the idea of doing the naive model experiment; and Andriy Kryshtafovych for the invaluable support and comprehensive data. We also thank the CASP prediction participants and the target structure providers, particularly Hartmut Luecke for valuable information on the urea channel protein, T0666. QCS program was provided by Nick Grishin's team and we thank Lisa Kinch and Qian Cong for their help in implementing the program and for sharing their experience. We thank Conrad Huang, Eric Pettersen, Elaine Meng, Tom Goddard, and Thomas Ferrin for their dedicated, professional, and timely support on implementing our idea and developing the EvalScore plug-in in UCSF Chimera to streamline the process of manual structural comparison. All molecules were visualized and some analyses performed using the Chimera package.

REFERENCES

1. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69 Suppl 8:57–67.
2. Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y. Assessment of CASP8 structure predictions for template free targets. Proteins 2009;77 Suppl 9:50–65.
3. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins 2011;79 Suppl 10:59–73.
4. Taylor T, Tai CH, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee B. Definition and classification of evaluation units for CASP10. Proteins 2014;82(Suppl 2):14–25.
5. Tress ML, Ezkurdia I, Richardson JS. Target domain definition and classification in CASP8. Proteins 2009;77 Suppl 9:10–17.
6. Kinch LN, Shi S, Cheng H, Cong Q, Pei J, Mariani V, Schwede T, Grishin NV. CASP9 target classification. Proteins 2011;79 Suppl 10:21–36.
7. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
8. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370–3374.
9. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins: Struc Funct Genet 2005;61 Suppl 7:67–83.
10. Cong Q, Kinch LN, Pei J, Shi S, Grishin VN, Li W, Grishin NV. An automatic method for CASP9 free modeling structure prediction assessment. Bioinformatics 2011;27(24):3371–3378.
11. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF chimera—a visualization system for exploratory research and analysis. J Comput Chem 2004;25(13): 1605–1612.
12. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82(Suppl 2):7–13.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–3402.
14. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951–960.
15. Chung N, Zhang XD, Kreamer A, Locco L, Kuan PF, Bartz S, Linsley PS, Ferrer M, Strulovici B. Median absolute deviation to improve hit selection for genome-scale RNAi screens. J Biomol Screening 2008;13(2):149–158.
16. 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 2003;53 Suppl 6:395–409.
17. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61 Suppl 7:27–45.
18. 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.
19. Okabe A. Spatial tessellations: concepts and applications of Voronoi diagrams, Vol. xii. Chichester; New York: Wiley; 2000. 671 pp.
20. Barber CB, Dobkin DP, Huhdanpaa H. The Quickhull algorithm for convex hulls. ACM Trans Math Softw 1996;22(4):469–483.
21. Taylor TJ, Vaisman, II. Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. Phys Rev E: Stat Nonlin Soft Matter Phys 2006;73(4 Pt 1):041925.

22. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16(6):566–567.
23. Strugatsky D, McNulty R, Munson K, Chen CK, Soltis SM, Sachs G, Luecke H. Structure of the proton-gated urea channel from the gastric pathogen Helicobacter pylori. Nature 2013;493(7431):255–258.
24. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–2309.
25. Kim C, Basner J, Lee B. Detecting internally symmetric protein structures. BMC Bioinformatics 2010;11:303.