CASP9 assessment of free modeling target predictions

Lisa Kinch,[1*] Shuo Yong Shi,[2] Qian Cong,[2] Hua Cheng,[2] Yuxing Liao,[2] and Nick V. Grishin[1,2]

[1] Howard Hughes Medical Institute, University of Texas, Southwestern Medical Center, Dallas, Texas 75390-9050

[2] Department of Biochemistry, University of Texas, Southwestern Medical Center, Dallas, Texas 75390-9050

ABSTRACT

We present an overview of the ninth round of Critical Assessment of Protein Structure Prediction (CASP9) "Template free modeling" category (FM). Prediction models were evaluated using a combination of established structural and sequence comparison measures and a novel automated method designed to mimic manual inspection by capturing both global and local structural features. These scores were compared to those assigned manually over a diverse subset of target domains. Scores were combined to compare overall performance of participating groups and to estimate rank significance. Moreover, we discuss a few examples of free modeling targets to highlight the progress and bottlenecks of current prediction methods. Notably, a server prediction model for a single target (T0581) improved significantly over the closest structure template (44% GDT increase). This accomplishment represents the "winner" of the CASP9 FM category. A number of human expert groups submitted slight variations of this model, highlighting a trend for human experts to act as "meta predictors" by correctly selecting among models produced by the top-performing automated servers. The details of evaluation are available at http://prodata.swmed.edu/CASP9/.

Proteins 2011; 79(Suppl 10):59–73.

Key words: protein-fold prediction; structure comparison; alignment quality; *ab-initio*; domain structure, CASP9.

INTRODUCTION

The goal of the biennial CASP assessment of protein structure prediction is to identify and evaluate the current state of the art methods in the field. The template free modeling (FM) category generally aims to assess *ab-initio* methods that predict 3D structures from a given protein sequence without the explicit use of template structures available in the Protein Data Bank. The most significant development of methodology for *de novo* structure prediction from sequence was introduced over a decade ago in CASP3, with the assembly of tertiary structures from selected fragments.[1] Fragment-based structure assembly methods have since been adopted and developed by a number of groups, whose *de novo* protein structure predictions tend to outperform in CASP evaluations of the FM category.[2–5] Despite this relative success, the "protein-folding problem" has remained unsolved, with results of the previous CASP8 FM category evaluation suggesting much room for improvement in *de novo* structure prediction methodologies.[6]

The CASP9 FM assessment included evaluating prediction models, quantifying these evaluations in a meaningful way, and using these measurements to produce group rankings that were subject to various tests of significance. Newly developed and previously applied automated methods played a crucial role in completing the assessment. Our report outlines the resulting evaluation procedure, the logic behind its development, and the results of its application to CASP9 FM target predictions. Based on these results, we highlight the progress and pitfalls of both the top performing prediction servers and the fold prediction community as a whole.

The assessment of the CASP9 FM category encompassed evaluations of 30 domains, which included 4 "server only" domains and 26 "human/server" domains. For the FM category alone, 16,971 predictions had to be evaluated, making manual judgment of all predictions impossible

The authors state no conflict of interest.
Grant sponsor: National Institutes of Health; Grant number: GM094575; Grant sponsor: Welch Foundation; Grant number: I-1505.
Lisa Kinch and Shuo Yong Shi authors contributed equally to this work.
*Correspondence to: Lisa Kinch, UT Southwestern, Biochemistry, 5323 Harry Hines Blvd, Dallas, TX 75390.
E-mail: lkinch@chop.swmed.edu
Received 11 April 2011; Revised 26 August 2011; Accepted 4 September 2011
Published online 14 September 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23181

within the required timescale of the assessment. We chose a subset of 16 FM domains that ranged in difficulty to score manually (scoring all prediction models with correct overall fold, see Methods) and compared these results to automatic scoring methods previously applied to CASP evaluations and a newly-developed automated scoring method aimed at mimicking the manual assessment. Given the drawbacks of relying upon a single measure to estimate the quality of all FM predictions, and the rough correlation of automated scores to manual scores, we decided to incorporate four different scores in our evaluation of overall prediction quality. The four scores include a combination of recognized structural comparison methods developed in our evaluation of CASP5[7] fold recognition (FR) targets (TenS), the newly-developed structural comparison method (QCS), a contact distance method similar to that used to evaluate CASP8[8] predictions (CS), and the GDT_TS method provided by the Prediction Center (GDT). By combining different measures that capture diverse aspects of predictions, we attempt to establish a comprehensive and robust measurement of model quality.

Comparing the overall prediction quality of different groups required combining scores generated with all measures and for all FM domains to produce a single value reflective of group performance, despite the fact that groups predicted varying numbers of targets. Given that each measure provided different types of scores and that each target varied in difficulty, combining scores required a meaningful rescaling. Similar to our previous assessment of CASP5 FR models,[7] we chose to rescale scores based on a comparison with the average prediction scores for individual target domains (*Z*-scores). The top performing groups were using various strategies to select among all submitted server models, prompting us to provide an additional evaluation that compared "human" predictions to the best server predictions. Thus we also applied individual target domain scaling based on a comparison to the top server score (server ratio). Scaled target domain scores (*Z*-scores or ratio scores) could then be combined across all FM targets using various strategies to produce values reflective of the overall performance of each group (ranks). Finally, predictions were compared to the best available templates to estimate the potential of FM methods to add value over template-based models. This comparison suggested the performance of a server (and a number of human experts who chose the correct server model) outshined the rest in CASP9.

METHODS AND RESULTS

FM target domains

The FM category traditionally includes difficult to predict domains that display no detectable (by sequence) similarity to available structures. The overlap between FM and Template-Based domains in CASP9 is extensive and difficult to define, perhaps due to increasing numbers and variations of available template domains. Assessors of the previous CASP8 chose to delineate the boundary between template-based and FM categories using structural similarity to the nearest known templates. While such similarity measures tend to indicate the general predictability of targets,[7,9] the performance of the predictor community as a whole provides the most direct estimate of difficulty. Our previous strategies for domain classification, which included combining a measure of target-template sequence similarity with an objective performance-based estimate of target difficulty in CASP5[7,9] and finding domain clusters that emerge naturally from average quality scores of the top 10 server models in CASP8,[2] did not establish clear boundaries between categories for CASP9 domains. We extensively modified these procedures to arrive at what we think was a reasonable compromised between the two trends in defining FM targets: the lack of detectable templates and prediction difficulty.[10] As a result, we defined the FM category to encompass 30 domains. All domains for which no template can be detected by sequence methods were taken (25 domains). In addition, domains for which homologous templates were detectable, but they greatly differed in structure, so that prediction quality was similarly poor to that of domains without templates, were included (5 domains). See target classification paper in this issue for details.[10]

Manual evaluation of predictions

Although a number of structure comparison methods, including GDT scores provided by the CASP Prediction Center,[11] can provide quantitative measures of prediction quality, the reliability of these scores tends to vary for poor models that fail to incorporate global structural features of targets. Because such models dominate the FM category, the power of existing automated methods to distinguish local structural features from globally good structure prediction becomes indiscriminate, and evaluating the overall quality of predictions requires manual inspection. For our evaluation, we chose to visually inspect all model predictions (1–5) for a diverse subset of FM domains and assign manual scores to each model based on a set of defined criteria for the target (see description below). We developed an automated method to mimic the manual evaluation, and we compared manual scores to those produced by the newly developed method. Finally, we applied a combination of structural comparison measures to our assessment of FM domains with a goal of capturing various aspects of prediction details with different comparison methods and providing robustness to method-specific pitfalls.

To allow visual inspection of FM predictions in a timely manner, we excluded a number of domains from the manual evaluation. Excluded targets consist of server-only domains (T0555, T0637, and T0639), structurally redundant domains (T0544_1, T0544_2, T0571_1, and T0571_2, they had close homologs among other targets), subunits with short helical segments that are easily distinguished by automated methods (T0547_3, T0547_4, and T0616), or

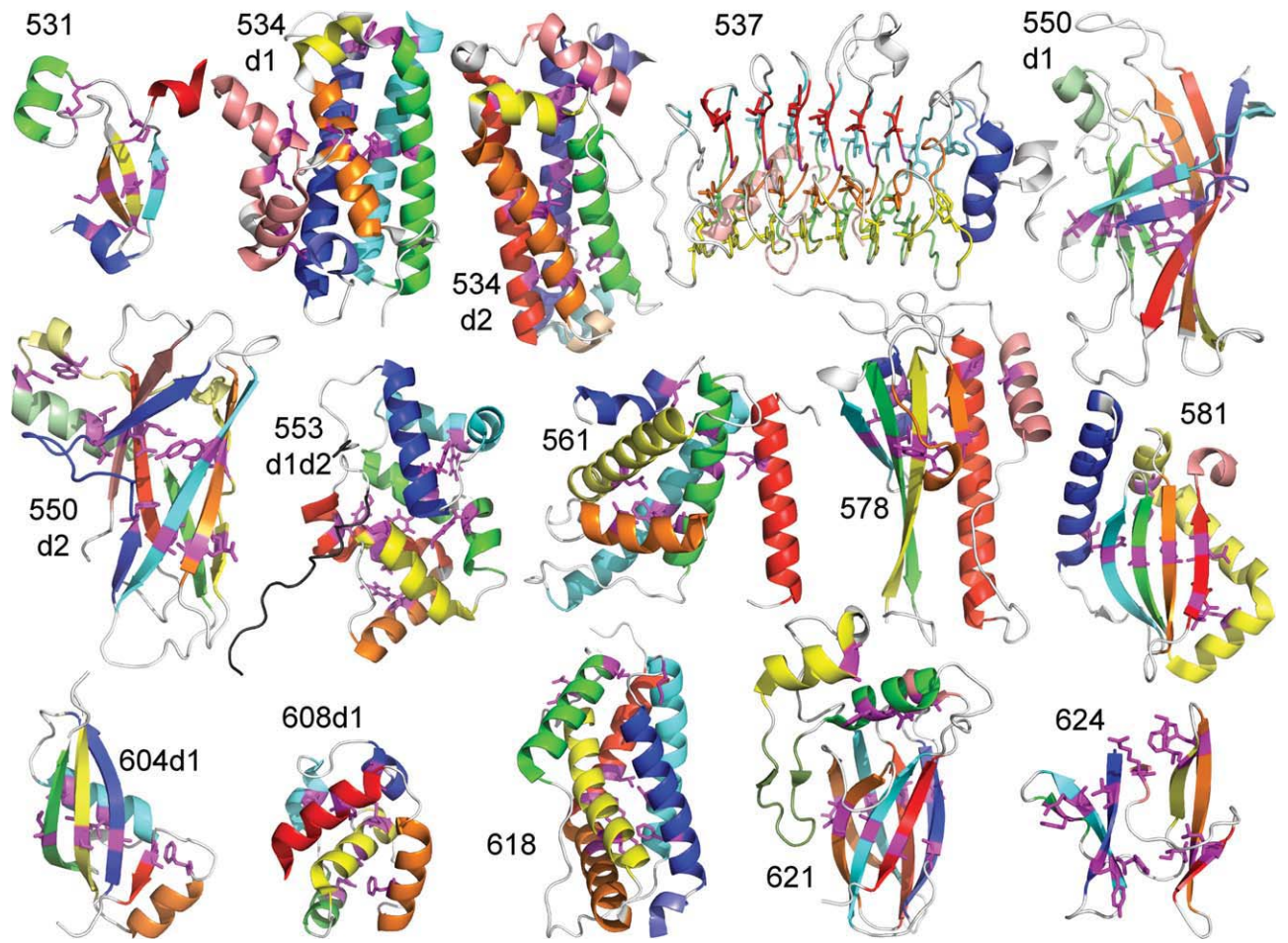

Figure 1

Manual FM target evaluation criteria. A diverse subset of FM targets (15 individual domains and 1 combined domain) provided the basis for manual evaluation. For each target, important secondary structures were colored, with connecting loops left white. Key interaction residues were highlighted with magenta sticks to evaluate the relative positions of secondary structures. These criteria provided the basis for side-by-side comparisons to all prediction models.

unusual domains that have abnormally poor prediction quality based on GDT (T0529_1 and T0629_2). The final manual evaluation subset included 16 domains (T0531, T0534_1, T0534_2, T0537, T0550_1, T0550_2, T0553_1, T0553_2, T0561, T0578, T0581, T0604_1, T0608_1, T0618, T0621, and T0624), with the domains from one target (T0553) combined into a single manual score (due to their tight association). For each chosen domain, a set of criteria aimed at evaluating global fold topology was developed to assess prediction quality. Criteria include size and orientation of secondary structure elements, key contacts between secondary structural elements, and any additional unusual structural features such as disulfide bonds in T0531.

To score predictions, all submitted models (1–5) were visualized using PyMOL scripts.[12] Each model was colored in rainbow according to secondary structure elements defined in the target structure. Key interactions and unusual structure features defined in the target structures were highlighted in models by displaying residues as magenta sticks. The final defined criteria for each manually evaluated target domain are illustrated in Figure 1. For some larger target domains with more than six secondary structure elements (like T0534d1), peripheral decorations to the core fold were colored with pastels, or some sequence continuous elements were colored with a single color. Similarly, the repeats of target 537 were colored the same along the same surface of the fold to ease counting and identification of their relative position in models. Each prepared model was then visually compared to the target structure side-by-side, without superposition. Not to be biased by computational evaluation, models were assessed in order of the assigned group numbers. As the scoring strategy and definition of key elements were developed and sometimes modified during the evaluation procedure, the scoring and/or definitions of the first evaluated target domain may differ somewhat from the last evaluated target domain. Specifically, the first evaluated target (T0531) scored secondary structure elements in a more detailed

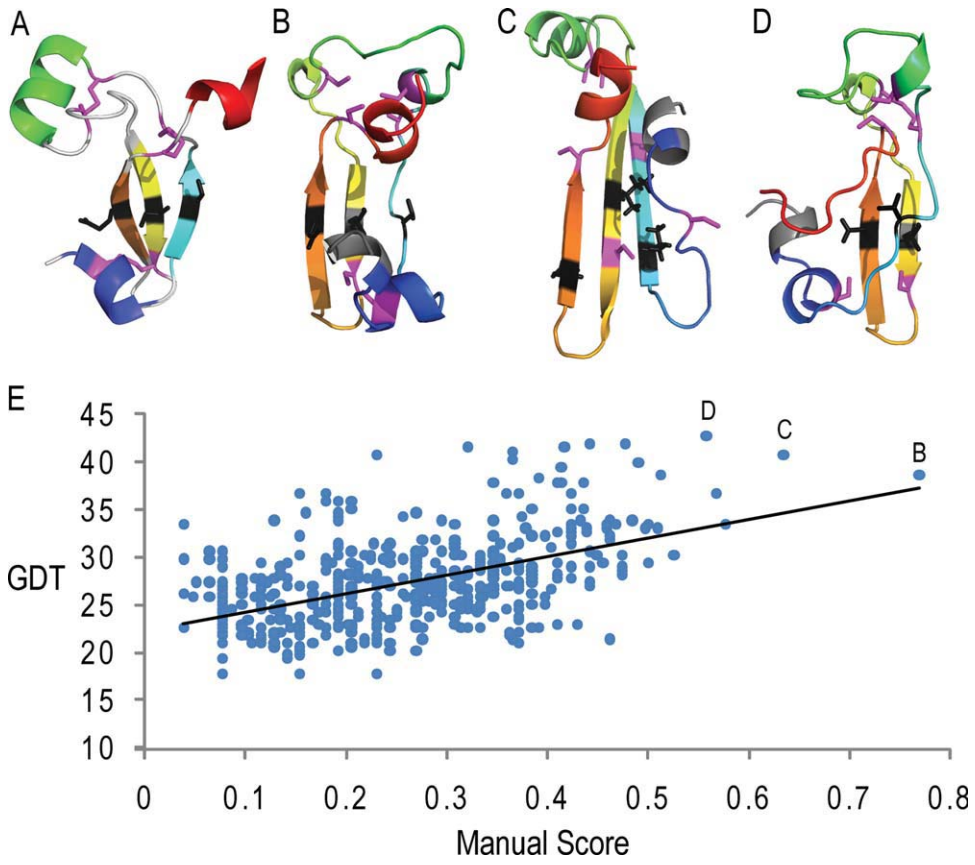

Figure 2

Target 531 manual top-scoring predictions. **A**: Target 531 is colored with secondary structure criteria as in Figure 1. Key contact residues highlight the orientation and shift of the β-strand meander (black) and special structural features specific to this target (magenta disulfide bonds). **B**: Rank 1 prediction model (399_4) is colored in rainbow and illustrates the correct features of the β-strand meander (black residues) and the relative placement of the flanking helices. **C**: Rank 2 prediction model (55_1) is colored as in B and includes a better prediction for the second helix (green), but a shifted second β-strand. **D**: Top ranked GDT model (399_4) is colored as in (B) and is missing a key feature of the core β-meander (cyan strand). **E**: Correlation plot of all target 531 non-zero manual scores with GDT scores shows a positive correlation ($R = 0.53$), albeit with different ranking of top prediction models (labeled according to illustrated examples).

way then the remaining targets and included significantly more partial scores for poor models. In the interest of time, for subsequent targets, poor models missing a significant portion of the fold topology were simply scored zero. Ultimately, scores were assigned essentially as follows: (1) each correct secondary structure element scores a point with an additional point for correct boundaries, (2) relative placement of each secondary structure with respect to neighboring secondary structures scores a point for correct shift and a point for correct angle, (3) correct key residue contacts score a point, and (4) any additional considerations get a point. Scores were recorded as a percentage of the maximum points assigned to each target.

Comparison of manual scores to automated GDT score: target 531

Examples of top-scoring predictions according to this manual scheme are illustrated for target T0531 in Figure 2. When compared using manually assigned criteria for T0531 [Fig. 2(A)], a human expert group prediction [Fig. 2(B), group 399 model 4] ranked highest, followed by a server prediction [Fig. 2(C), group 55 model 1]. The top manually ranked predictions include the core β-sheet meander in the correct topology with the flanking helices placed correctly relative to the sheet, albeit in altered orientations. While the rank 1 model is missing a portion of the second flanking helix, it correctly positions all of the core strands. In contrast, the rank 2 model includes all of the second flanking helix. However, the second strand is shifted with respect to the others, causing the sheet to appear elongated with respect to the target. The top scoring model according to GDT (Group 399, model 5, GDT 42.74) misplaces its first strand, appears compacted with several steric clashes in the core, and is missing the C-terminal flanking helix [Fig. 2(D)]. When compared to GDT scores for this target, the manual scores show positive correlation

[Fig. 2(E), $R = 0.53$] Similar trends occur for the remaining manually assessed targets.

New automated score (QCS) to mimic manual inspection

Given the number and composition of FM predictions, the manual evaluation was subjective (reflected by scoring inconsistencies) and failed to distinguish the local structural features that are captured by automated scores such as GDT [reflected by numerous ties, vertical lines in Fig. 2(E)]. To overcome these problems, we developed an automated Quality Control Score (QCS) to mimic the manual evaluation. The new measure captures the global structure features of FM predictions in an objective way, while discriminating between close models by using local structure features. Briefly, QCS defines the boundaries of secondary structure elements (SSEs) and essential contacts between SSEs in the target and propagates the definitions to the model according to residue number. With these definitions, QCS simplifies the target and the model, respectively, into a set of SSE vectors and several key contacts, which allows direct comparison between the target and the model for the following four global features: (1) the correct prediction of SSE boundaries measured by the length of SSE vectors, (2) the global position of SSEs represented by the distances between the centers of SSEs and the center of the whole protein, (3) the angles between SSE pairs, and (4) the distances between the Cα atoms in the key contacts that reflects the relative packing and interaction between SSEs. A score negatively correlating with the difference between the model and the target is assigned for each feature. Moreover, to characterize the quality of models in details, the contact score and percentage DSSP agreement between the model and the target is calculated as well. Finally, these two local feature scores, together with the four scores measuring global properties, are averaged to represent the overall measurement of the model quality.

Good global structure features of models revealed by QCS: target T0561

Overall, the QCS score shows a slightly better correlation with manual scores ($R = 0.8$) than it does with GDT ($R = 0.77$) and for some cases, reveals models with good structural features that were missed by other scores. One such example is represented by a QCS—favored model 382_5 [Fig. 3(A)] predicted for Target T0561 [Fig. 3(B)], whose features can be compared to those of the GDT favored model 324_5 [Fig. 3(C)]. Overall, the global topology of the QCS model 382_5 [Fig. 3(D)] agrees exactly to that of the target [Fig. 3(E)], while the GDT model 324_5 [Fig. 3(F)] includes two helices at the C-terminus that are packed on the opposite face in a different orientation than in the target structure. Evolutionary evidence suggests that the two C-terminal helices play a role in the function, which further diminishes the quality of the GDT favored model. Moreover, by inspecting the three key residue pairs mediating SSE interactions [Fig. 3(G–I)], 382_5 gets all three correct, while 324_5 gets only one. Apparently, by paying attention to the global features, QCS has revealed models with superior global topology and interactions.

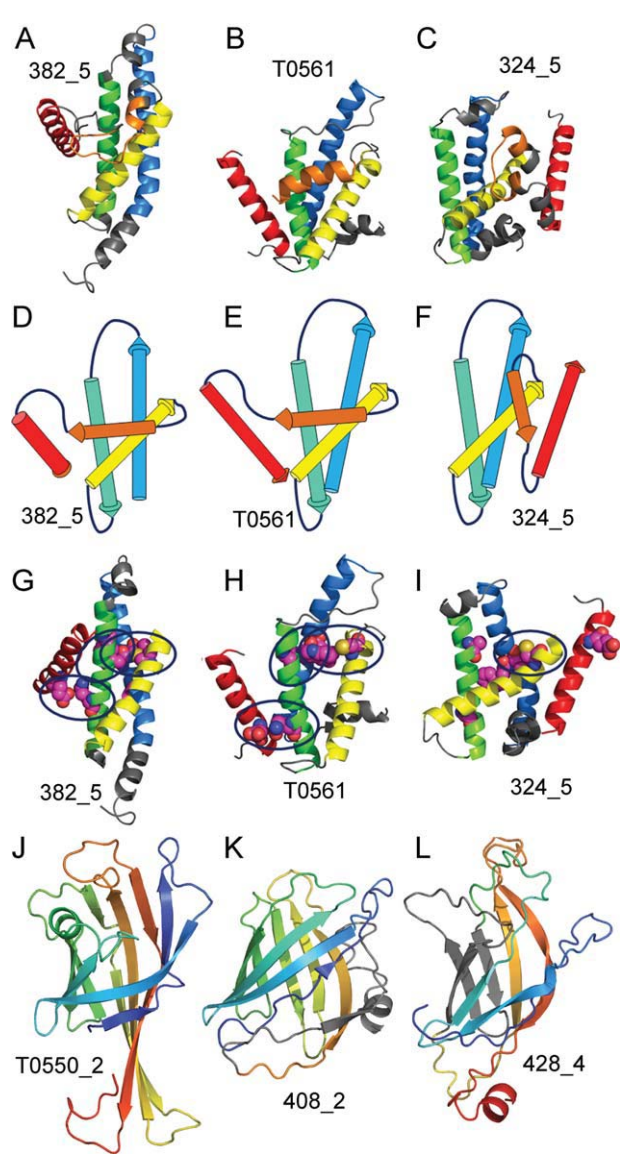

Figure 3

QCS reveals good structural features. A: QCS selects the best prediction model TS382_5 for (**B**) the target T0561, (**C**) while GDT selects a different prediction, TS324_5, as the best model. Topology diagrams are displayed for (**D**) TS382_5, (**E**) T0561, and (**F**) TS324_5. Identified SSE interactions are colored (magenta) (**G**) for QCS selected model TS382_5, (**H**) for target T0561, and for (**I**) GDT selected model TS324_5. **J**: Target 550d2 forms an 8-stranded β-meander barrel with longer C-terminal 3 helices. **K**: QCS favored model TS408_2 forms a β-meander barrel with an extra strand (gray), while (**L**) TenS favored model TS428_4 forms a similarly shaped barrel with C-terminal strands of similar length but an altered topology (gray).

Combining ten scores (TenS) to reduce the noise in automated assessment

As described in the CASP5 evaluation of difficult FR targets,[7] using the input from multiple evaluation methods tends to increase the significance of scores, as the shortcomings of individual methods are essentially averaged away. We chose to apply this concept to the FM targets of CASP9 using individual scores developed to evaluate sequence and structural characteristics of FR predictions in CASP5 (Dali,[13] CE,[14] secondary structure overlap (SOV),[15] LGA,[11] and Mammoth[16]). The results of the methods tend to diverge for dissimilar structures (SCOP superfamily/family pairs under about 10% sequence identity),[17] and may thus detect different aspects of structural predictions for the various FM targets with marginal predictions. In CASP5, ten scores were deliberately selected to improve the quality of numerical evaluation, with six structure similarity measures and four alignment quality *Q* ratios (described below). Thus, the score system is balanced in terms of the number of sequence and structural measures. However, due to the relatively poor performance of CE on FM targets, we substituted this measure with TMalign[18] to complete the TenS score used in our current evaluation of FM target domains.

The first structural component of the TenS score is GDT, which has served as the preferred CASP evaluation method since its introduction by the prediction center in CASP3.[19] GDT measures the global structure quality of a model by counting the percentage of superimposed residues falling within four superposition cutoff distances (1, 2, 4, and 8 Å) in four different superpositions. The second component, SOV, is another evaluation method offered by the prediction center. SOV measures the overlap between observed (target) and predicted (model) secondary structure element assignments. In our calculation, DSSP[20] was used to delineate the types of secondary structures both in targets and models, and the eight types of secondary structures returned from DSSP were converted to the applicable input for SOV (helix, strand, and coil). In addition to these two methods suggested by the prediction center, three conventional structure comparison methods (Dali,[13] TM-align,[18] and Mammoth[16]) were introduced to our scoring system to capture the diverse aspects of structural features in models. In contrast with the sequence-dependent GDT analysis, these three methods were developed to search for the optimal rigid structural alignment. As one of the most widely used structure superposition programs, Dali evaluates the similarity of intra-molecular contact patterns of two structures. Dali reports two scores: a raw score and a *Z*-score. Though the latter is commonly used in structural comparison, FM target predictions tend to result in low Dali *Z*-scores that are hard to differentiate. To overcome this problem, we incorporate the Dali raw score into TenS. The TM-align method displays comparable accuracy to Dali, although it minimizes the intermolecular C_α atom distance between two structures. The significance of structure similarity is reported by a TM-score, which we used as one component of our scoring system. The third structural evaluation method (Mammoth) was originally developed to compare model conformations to an experimental structure and works well in the detection of remote homology. The Mammoth alignment score ($-\ln(E)$) was taken as our component score. We also developed a sequence-dependant intramolecular contact distance score in CASP5 (see evaluation paper[7] for details), which represents the final structural component score. To include alignment quality measures as components of TenS, *Q* scores were calculated as the fraction of correct aligned residues for sequence alignments produced by the sequence-independent structural superposition methods (Dali, TM-align, Mammoth, and LGA-4 structural analysis mode with default distance cutoff 5.0 Å). Since different methods yield different results and have their own pros and cons, we combined all the ten component scores into a single score (TenS). Scores were combined with equal weights after conversion to *Z*-scores with the mean (μ) and standard deviation (σ) computed on predictions disregarding scores below $\mu - 2\sigma$ from the entire sample (*Z*-score). The final TenS score can be viewed as a combined index score that weighs the respective merits of the ten different scores and attenuates the noise caused by each individual score.

TenS scores favor local model quality: T0550d2

As expected, our scores produced different rankings for various targets and captured models that displayed different qualities of FM target structures. For example, ranks for Target T0550d2 [Fig. 3(J)], an 8-stranded barrel formed by a β-meander similar to a streptavadin-like fold, differed according to the score used for evaluation. QCS favored the inclusion of a correct global fold in top-ranking models. One such model (408_2) is illustrated in Figure 3(K). While this prediction suffered from alignment problems in the C-terminus, it included a barrel formed by a 9-stranded β-meander [one too many strands, colored gray in Fig. 3(K)]. Alternatively, TenS favored more local model qualities in top ranks. The top ranking TenS server prediction (428_4) displayed good alignment and superposition of 5 out of 8 β-strands of a more correctly shaped barrel, with the remaining strands being in an incorrect topology [Fig. 3(L), incorrect topology colored gray].

Combining scores to produce ranks: sum/average of *Z*-scores

Individual scores emphasize specific aspect of structural predictions, and different scores lead to different ranking preferences. Based on our experience with evaluating structural predictions,[7,8] scores need to be

combined in a meaningful way to compare the overall prediction quality of different groups. Because the FM Targets exhibit different levels of difficulty (average server GDT ranges from 8.4 to 31.7), we chose to rescale prediction scores according to the average prediction quality of each target by using *Z* scores calculated as described in previous section. Rescaled *Z*-scores for the automated TenS and QCS methods were summed together with those of the traditionally recognized GDT score (also a component of TenS) and a contact distance score CS (also a component of QCS) that was similar in concept to the method reported to perform well on FM targets in CASP8[6] to produce a single value reflective of overall performance (ComS). For each participating group, these *Z* scores could be either summed over all targets or averaged so that groups are not penalized for omitting targets. Predictors were allowed to submit up to five models for each target. For our assessment, we chose to evaluate both "first models" and "best models" to ascertain both the best quality models produced by given methods (best models) and to evaluate how well methods assessed the quality of their models (first models). The various scores are reported as sortable spreadsheets that include all individual FM targets as well as the average/sum over all FM targets for each group's best models and first models (http://prodata.swmed.edu/CASP9/evaluation/Evaluation.htm).

A summary of the top-performing human expert and server groups according to the average ComS score is illustrated in Figure 4(A). Groups are ranked according to best model scores (black), and groups appear to perform significantly worse according to first model scores (gray). The top two human expert groups (96 and 408) tend to consistently outperform the rest no matter which score is used for ranking (see discussion and Table II below). The average server-only ComS scores for best models (on server-only targets) are also better than scores for first models [Fig. 4(B)]. The top three server groups (380, 428, and 321) significantly outperform the remaining servers on best models, while the top two servers (380 and 428) also do a relatively good job at choosing first models compared to the rest of the servers.

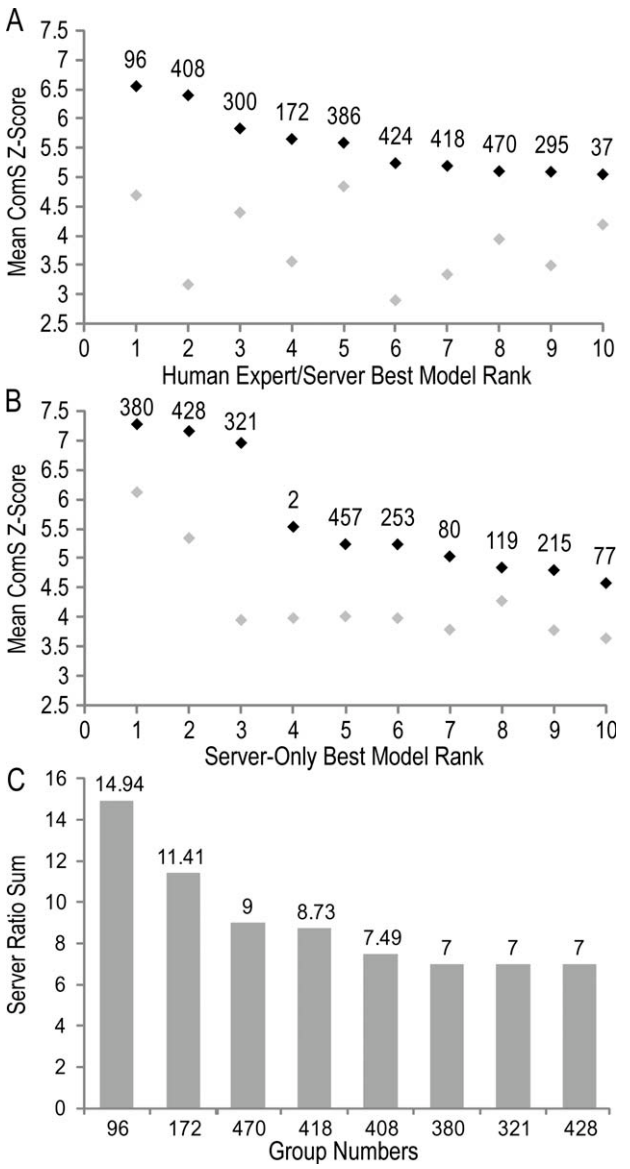

Figure 4

Group performance. **A**: Average best model ComS *Z*-scores (black diamonds) for the top ten ranked of all groups on FM domains designated as human/server are compared to Average ComS *Z*-scores on models designated as first (gray diamonds). **B**: Average best ComS *Z*-scores (black diamonds) for the top ten ranked of Server-only groups on FM domains designated as Server-only are compared to Average ComS *Z*-scores on models designated as first (gray diamonds). **C**: Column graph illustrates best Server ratio scores (indicated above column) summed over all FM domains designated as human/server. Groups (96, 172, 470, and 408) that outperformed top servers (380, 321, and 428) overall are included in the graph.

Combining scores to produce ranks: comparison to top server model

Over the course of evaluating CASP9 FM predictions, we observed that top-performing human expert groups were acting as "meta-predictors" by choosing and refining the best among all models provided by the automated servers. Groups applied similar strategies to evaluate server models using various energy functions to rank and sometimes refine top models. This observation motivated us to evaluate groups by a comparison to the top scoring server models. We ranked predictions using each of the main scores (TenS, QCS, GDT, and CS), chose the top-scoring server model, and scored all models as a ratio to the chosen server model. As a note, each of the scores may rank a different server model at the top to provide the basis for ratios. To focus on the best models, ratio scores below 1 were ignored, and the remaining scores were averaged for each target. A sum of the target

Table I
Ranks

Group	Name	ComS best	TenS best	QCS best	GDT best	CS best	ComS best Avg	ComS first Avg	Server ratio best	Manual best Avg	PCA
TS096	ZHANG	**170.18**	46.51	38.46	**46.02**	**43.32**	6.55	4.69	**14.29**	0.537	2.888
TS408	KEASAR	166.16	**47.69**	**39.10**	45.09	38.69	6.39	3.17	7.17	**0.538**	2.398
TS172	BAKER	146.89	41.00	34.37	40.98	34.39	**5.65**	3.56	10.49	0.493	2.254
TS470	ELOFSSON	148.01	40.64	33.99	41.40	37.28	5.10	3.94	9.00	0.461	2.192
TS418	ZHANG_AB_INITIO	150.60	44.74	36.78	43.33	33.79	5.19	3.34	8.40	0.465	2.177
TS037	FAMS-ACE3	146.45	41.94	32.83	41.48	37.04	5.05	4.19	5.01	0.491	2.055
TS386	MUFOLD	139.64	38.73	32.97	37.12	34.22	5.59	**4.84**	6.07	0.418	2.024
TS380	QUARK	139.04	39.62	31.19	40.59	32.55	4.63	3.65	7.00	0.418	1.903
TS428	Zhang-Server	135.50	38.12	30.85	39.86	34.97	4.52	3.20	7.00	0.485	1.901
TS490	MULTICOM	141.76	38.89	34.36	38.63	35.34	4.89	4.13	4.01	0.443	1.895
TS113	FAMSSEC	141.24	39.67	33.08	39.27	34.85	4.87	4.05	3.01	0.459	1.850
TS424	BATES_BMM	136.25	40.16	31.02	37.28	34.68	5.24	2.90	4.03	0.492	1.797
TS088	SPLICER	141.86	41.87	34.49	39.30	33.80	4.89	3.15	3.02	0.452	1.795
TS295	KNOWMIN	132.41	38.05	32.74	36.79	32.82	5.09	3.49	5.06	0.394	1.773
TS300	4BODY_POTENTIALS	110.73	31.61	26.71	31.97	28.21	5.83	4.40	6.00	0.469	1.731
TS399	CHICKEN_GEORGE	122.67	34.33	33.40	32.78	29.65	4.72	3.45	4.20	0.459	1.638
TS321	ROSETTASERVER	126.04	34.02	30.68	34.39	31.91	4.20	2.17	7.00	0.377	1.570
TS016	SEOK-META	116.01	29.62	28.97	30.02	29.23	4.46	4.31	6.07	0.357	1.565

averages over all FM models (which were rarely much higher than 1) captures the number of times each group outperforms the servers [Fig. 4(C)]. For this analysis only five groups (96, 172, 470, 418, and 408) performed better than the top servers (380, 321, and 428), who did best on 7 out of 26 evaluated domains each. Impressively, the top human expert group (96) outperformed top server models on more than half (14) of the evaluated domains.

Ranks and significance

Table I summarizes the top combined expert human and server (bold) prediction groups ranked by each of a chosen subset of our evaluation scores (ComS, TenS, QCS, GDT, CS, ServerRatio, and Manual) and methods for combination (sum or average, first or best). The groups are ranked in the table according to the first principal component computed on these nine scores (last column), and the top ten from each method are highlighted gray. Groups are included in the table only if they performed in the top ten judged by any of the scores (i.e., atleast one gray highlight), The table ends with the principal component score rank of the first group with no top ten highlights, and due to this cutoff, some of the more differing scores (first model, server ratio, and manual) have a portion of their top ten ranked groups omitted, as those groups performed worse in the overall rankings. The top three servers highlighted in Figure 4 perform among the top groups: Zhang-server (428), Quark (380), and ROSETTASERVER (321). The human expert groups who developed these servers also tended to outperform: the Zhang group (96 and 418), who developed Zhang-server and Quark, and the Baker group (172), who developed ROSETTASERVER. Also noted consistently among the top was the Keasar group (408).

Since CASP4[21] it has been generally appreciated that the quantitative assessment of the reliability of the top ranking is necessary. To test whether the prediction quality of the highest scoring groups can be reliably distinguished from the remaining scores, we sought to evaluate the statistical significance of the results using paired Student's *t*-test and bootstrap selection of *Z*-Scores. *T*-test was performed on ComS scores between paired samples of FM targets common to both groups with the probability value estimated by an incomplete β function based on *t* value. Our *P* value was derived from a one tailed *t*-test, which is justified because we were testing whether one group is significantly better than the lower-rank group in the direction of observed effect. The two instinct assumptions for the *t*-test are the sample data should be normally distributed and the variance should be equal, and these may not be necessary satisfied in CASP9. Therefore, we also introduced the nonparametric bootstrap method to evaluate the rank significance (Table II). In the bootstrap procedure, the sum score for each group was calculated from a random selection of *N* CombS over predicted targets common to both groups, where *N* equals to the number of common targets between the two groups. This routine was repeated 1000 times with returns for each pair of groups, and the number of times one group outperformed the other indicated the significance of the difference in rank, the larger the better.

Significant outperformance of the top three servers (Zhang-server, Quark, and ROSETTASERVER), when compared to each of the remaining servers is supported by both tests (>90% confidence, see bootstrap and *T*-test tables here: http://prodata.swmed.edu/CASP9/evaluation/

Table II
Significance

	96	408	172	470	418	37	386	380	428	490	113	424	88	295	300	399	321	16	94	60	407	382
96	—	0.62	0.90	0.99	0.94	1	1	1	1	0.99	0.99	0.95	0.98	1	0.97	1	0.97	1	1	1	1	1
408	26	—	0.87	0.88	0.79	0.93	0.80	0.98	0.99	0.93	0.95	1	0.91	1	0.42	0.99	1	1	1	0.99	1	1
172	26	26	—	0.47	0.43	0.50	0.40	0.66	0.75	0.64	0.58	0.72	0.57	0.79	0.34	0.89	0.87	0.98	0.92	0.70	0.93	0.90
470	26	26	26	—	0.41	0.53	0.86	0.91	0.90	0.68	0.77	0.73	0.67	0.88	0.54	0.94	0.86	0.99	0.99	0.97	1	0.99
418	26	26	26	26	—	0.63	0.69	0.84	0.88	0.74	0.72	0.78	0.73	0.92	0.39	0.95	0.88	0.99	0.97	0.93	0.98	0.98
37	26	26	26	26	26	—	0.59	0.81	0.90	0.65	0.80	0.73	0.61	0.85	0.34	0.94	0.85	1	0.99	0.96	0.99	0.98
386	25	25	25	25	25	25	—	0.64	0.78	0.64	0.57	0.95	0.50	0.89	0.51	0.88	0.98	0.98	0.96	0.95	1	0.99
380	26	26	26	26	26	26	25	—	0.77	0.43	0.41	0.58	0.41	0.70	0.22	0.87	0.73	0.97	0.96	0.82	0.99	0.91
428	26	26	26	26	26	26	25	26	—	0.33	0.29	0.51	0.34	0.56	0.25	0.80	0.69	0.96	0.91	0.64	0.95	0.85
490	26	26	26	26	26	26	25	26	26	—	0.53	0.65	0.49	0.69	0.09	0.95	0.78	0.97	0.99	0.79	0.94	0.89
113	26	26	26	26	26	26	25	26	26	26	—	0.60	0.48	0.68	0.38	0.88	0.77	0.99	0.98	0.92	0.99	0.96
424	26	26	26	26	26	26	25	26	26	26	26	—	0.37	0.57	0.00	0.77	0.82	0.85	0.77	0.70	0.81	0.79
88	26	26	26	26	26	26	25	26	26	26	26	26	—	0.71	0.22	0.88	0.79	0.92	0.88	0.77	0.95	0.91
295	26	26	26	26	26	26	25	26	26	26	26	26	26	—	0.06	0.71	0.66	0.89	0.75	0.56	0.81	0.84
300	19	19	19	19	19	19	19	19	19	19	19	19	19	19	—	1	1	0.99	0.95	0.96	0.96	1
399	26	26	26	26	26	26	25	26	26	26	26	26	26	26	19	—	0.45	0.66	0.58	0.26	0.65	0.46
321	26	26	26	26	26	26	25	26	26	26	26	26	26	26	19	26	—	0.69	0.57	0.56	0.66	0.55
16	26	26	26	26	26	26	25	26	26	26	26	26	26	26	19	26	26	—	0.40	0.17	0.52	0.33
94	26	26	26	26	26	26	25	26	26	26	26	26	26	26	19	26	26	26	—	0.22	0.63	0.46
60	23	23	23	23	23	23	22	23	23	23	23	23	23	23	16	23	23	23	23	—	0.85	0.48
407	26	26	26	26	26	26	25	26	26	26	26	26	26	26	19	26	26	26	26	23	—	0.33
382	26	26	26	26	26	26	25	26	26	26	26	26	26	26	19	26	26	26	26	23	26	—

domainscore_sum/server-best-Z.html). Below these top server groups, the statistical significance of the differences in ranking was marginal. When comparing all groups, the significance of the rankings was less clear (Table II). The top two human expert groups (Zhang and Keasar) tended to outperform the rest in bootstraps and paired *T*-tests (>90% confidence). A second larger pack of groups that includes the top three servers (418, 470, 172, 37, 88, 490, 113, 386, 380, 424, 428, 295, 321, 300, and 399) tended to outperform the rest and also made the rankings cut in Table I. Group 300 (4BODY_POTENTIALS) was a case of particular interest, submitting models for only 19 FM targets. Our assessment demonstrated that this group ranked 28 (best model) taking into account all FM target sums, whereas it ranked 4 (best model) in terms of average score, indicating a relatively good performance on the subset of submitted models.

Highlights and pitfalls of FM predictions

We highlight the predictions of several interesting FM targets to illustrate the progress and failures of CASP9 predictions. Figure 5 shows the performance of the top ten prediction models (black circles) as measured by GDT, with respect to the top scoring available template (white triangles). Target domains are ordered according to their similarity with the closest template. The FM category included a variety of domains encompassing a wide range of difficulty levels, ranging from the most difficult targets (GDT 18.5 to the closest template for T0529_1, with top models approaching GDT 12) to the easiest targets (GDT 82.7 to the closest template for T0547_3, with the top models approaching GDT 72). Notably, several predictions more closely resembled the target structure than any available template (581, 604d1, and both domains of two related templates 553 and 544) and are discussed below. The two easiest target domains (both from T0547) can be identified as a continuous sequence insertion and extension of an alignment to a closely related template homolog with a PLP-binding TIM barrel inserted into a eukaryotic ODC-like Greek key β-barrel. One of the domains (T0547_3) forms a three-helix bundle inserted into the TIM barrel, while the other forms a helical pair extension at the C-terminus. Impressively, the top performing prediction models accurately reflect the relative position and orientation of the helices. On the opposite end of the spectrum, the two most difficult targets belong to multi-domain structures, with one (T0529_1) being relatively large with high contact order and the other (T0629_2) possessing an unusual and highly elongated fold (see Discussion below).

Highlight: prediction model beats closest template (T0531 and T0604d1)

When compared to the closest templates, the quality of top predictions stands out above the rest for one domain target (Fig. 5, T0581). Although a number of predictions significantly improve over the closest template (44% GDT increase), these models can be traced to a single server prediction (model 4) by ROSETTA [Fig. 6(A)]. The ROSETTA model includes all of the core components of the target α+β (HEEH*EE) sandwich fold [Fig. 6(B)]; including an unusual helix H* that is kinked in two places. The kinks result in a central helix with two perpendicular helical extensions on either end facing almost opposite directions, like an S, that dictates the curvature of the sheet. While the closest template struc-

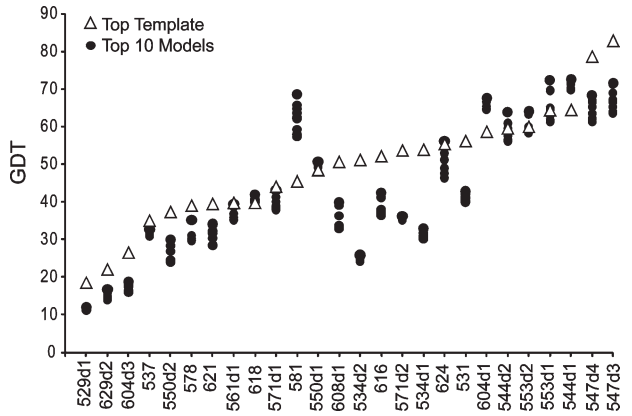

Figure 5

Top models compared to top template performance. The top ten best prediction model GDT scores (black circles) for each FM domain (labeled below) are plotted together with the GDT score for the best available template (white triangles). Target domains are ordered according to their similarity with the closest template.

ture includes a similarly curved sheet, the respective template helix H* is not kinked, and the two helices form an extended interaction on the back side of the sheet that is not present in the target [Fig. 6(C)]. Another remarkable aspect of this ROSETTA model is the presence of the four-stranded sheet, since PSI-PRED secondary structure predictions[22] dictate a mainly helical domain, with a single predicted β-strand (strand 3). Apparently, ROSETTA can overcome this incorrect prediction by extending the predicted strand into a sheet (using neighboring less confidently predicted helical segments). Among all of the five ROSETTA models, the winning model 4 is the only one with a four stranded sheet. Most of the predictions for this target are extended or entirely helical (558 out of 625 prediction models or 89% score zero in the manual assessment).

One multidomain target (T0604) includes a domain (T0604_1) where top prediction models improve over the best template (15% GDT improvement). This N-terminal domain is easily split from the second domain, whose template can be identified with sequence-based methods. Although it represents one of the most common folds (ferredoxin-like), no single template correctly reflects all of the secondary structure orientations and interactions (58.5 GDT to the closest template). Two different servers (Zhang Server and PRO-SP3-TASSER) produce relatively close prediction models [Fig. 6(D)] that include most of the core components of the fold, including the first three strands of the four-stranded sheet and two correctly positioned helices oriented perpendicularly to each other on one side of the sheet [Fig. 6(E)]. This target domain has the most human expert groups (19) outperforming the two best server models. Despite the distance of the closest template, which correctly orients the two helices but forms a shorter sheet lacking the correct curvature [Fig. 6(F)], the relatively good performance of the prediction community (23 groups improve over the closest template GDT) may reflect the ease of predicting large fold families.

Pitfall: multidomain problem (T0604 and T0534)

Although predictions were good for the first domain of Target 604, the third domain of this target represents one of the worst predicted domains among the FM targets. The poor performance of groups on this target domain (highest GDT is 18.7) reflects its fusion to the second domain (604d2, TBM category), whose closest Flavin adenine dinucleotide(FAD)/nicotinamide adenine dinucleotide (NAD)(P)-binding domain template is easily detected by sequence. The sequence-based detection methods extend the alignment of the top scoring template to include an unrelated domain from the "HI0933 insert domain-like" SCOP fold at C-terminus. Instead, a template with a more distantly related FAD/NAD(P)-binding domain possesses an FAD-linked reductase C-terminal domain with all of the core components of the target 604 C-terminal domain (see classification). A few groups correctly identified the template. For example, the top-scoring GDT group (166_5) used the FAD-linked reductase C-terminal domain template (2gb0). However, the numerous insertions present in the target structure (47% of the sequence) preclude good-scoring models. Although an apparently identifiable template is present for T0604_3, the poor quality of predictions and the absence of almost half the structure in the template make it better suited for evaluation in the FM category. We imposed a similar FM categorization of target 621 (top models around 30 GDT), which possesses a difficult to identify galactose-binding domain-like fold with a relatively large, somewhat extended helix-β-hairpin-helix insertion (closest template is below GDT 40, HHpred probability score below 20).

In CASP9, the correct prediction of multiple domain proteins in general remained a challenging task. One FM targets contained a very difficult to predict domain organization, having a four helix up-and-down bundle inserted into another all-helical bromodomain-like fold [Fig. 6(G), red and blue, respectively]. In addition to this discontinuous domain boundary, the sequence reported for this target included an N-terminal signal peptide that further confounded predictions (Target T0608_1 also included a signal peptide having similar effects on performance). The top ranking groups for this Target performed significantly worse than the closest templates (38% and 49% lower GDT for T0534_1 and T0534_2, respectively). With one exception, the top ranking groups correctly place only two out of four helices for each domain. Interestingly, our manual scoring scheme captures

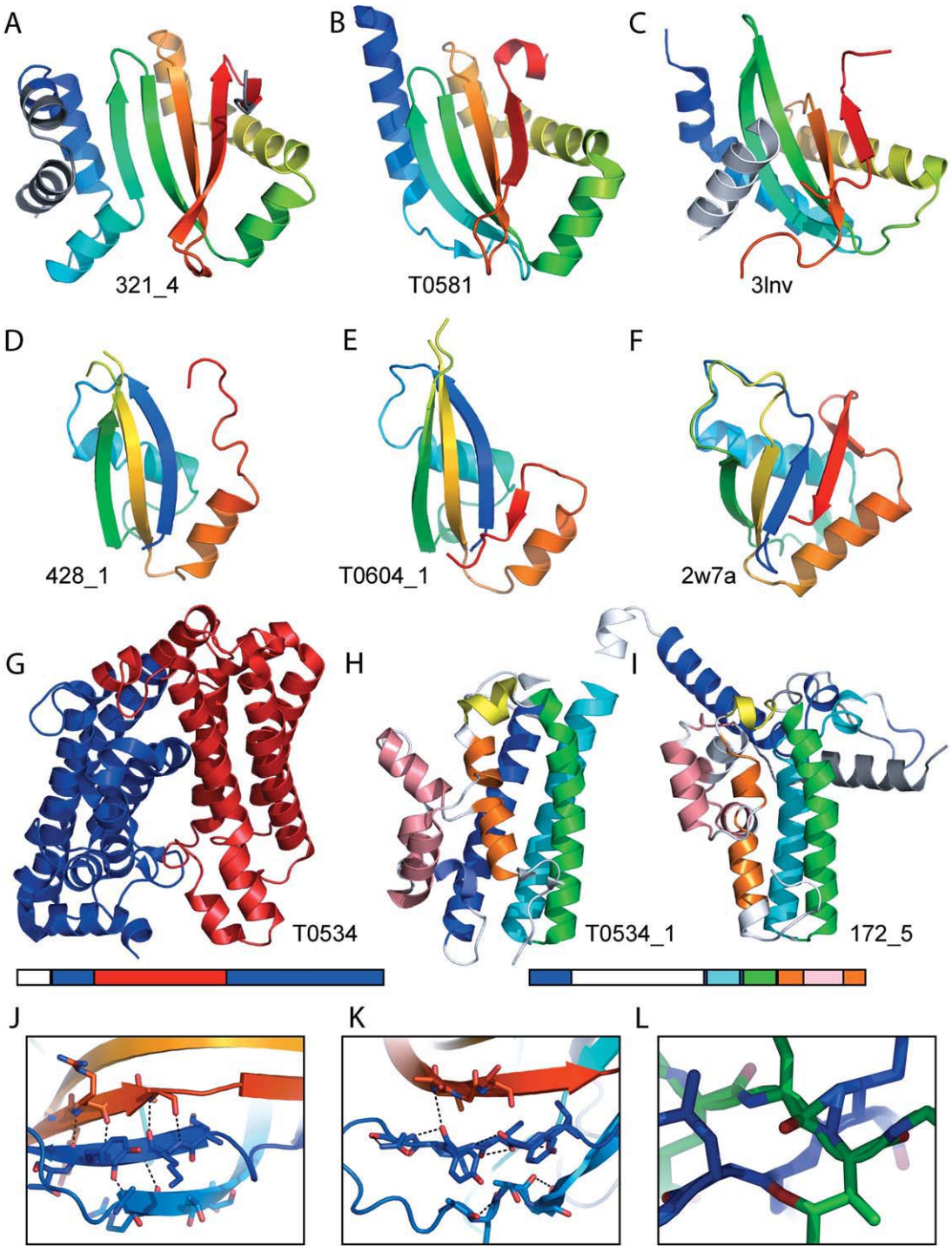

Figure 6

Highlights and pitfalls of FM predictions. **A**: The best server prediction (321_4) for (**B**) the target T0581 is compared to (**C**) closest template structure represented by the C-terminal domain of fatty acyl-AMP ligase (3lnv). **D**: The best server prediction (428_1) for (**E**) the target T0604_1 is compared to (**F**) the closest template structure represented by the structure of human LINE-1 ORF1P central domain (2w7a). **G**: Target T0534 has discontinuous domain bounds resulting from a four helix up-and-down bundle (red) inserted into another all-helical bromodomain-like fold (blue). Domain organization is mapped to sequence below. **H**: T0534_1 includes a three helical insertion (salmon) in the center of one helix (orange) that is captured by (**I**) a prediction model (172_5). Domain organization is mapped to sequence below. **J**: Target T0550_2 β-sheet displays typical backbone hydrogen bonds between adjacent β-strands. **K**: A top server prediction model (428_4) for T0550_2 displays incorrect β-sheet hydrogen-bonding patterns. **L**: Another server model (2_5) has steric clashes between loop backbones and sidechains.

a prediction by the Baker group (172_5) that includes some interesting features of the discontinuous target 534d1 domain [Fig. 6(H)]. Despite the presence of some overlapping secondary structures and the misplacement of the first helix [Fig. 6(I), blue], the Baker model places the second two helices correctly (cyan and green) and includes a correctly positioned helical loop (yellow), and a correctly broken fourth helix (orange) interacting with the third helix (green).

Pitfall: physically unrealistic models

Many top FM predictions displayed poor local structural quality, and some of the top-performing servers produced a number of physically unrealistic models. For example, numerous secondary structure regions in server models have incorrect backbone orientations. Instead of forming hydrogen bonds between neighboring β-strand backbones to form a typical sheet like that seen in Target T0550d2 [Fig. 6(J)], corresponding strand-like elements from a top-performing server model (428_4) form hydrogen bonds between the backbones of consecutive residues [Fig. 6(K)]. In addition to poorly defined backbones, steric clashes were also frequently observed in models. A server model produced for target T0621 (2_5) includes two loops whose backbones cross so closely that PyMOL draws bonds between atoms from the backbone of one loop to the backbone and side chain of another [Fig. 6(L)]. These frequent examples of poor model quality suggest that merely applying some form of model refinement to structures produced by a number of FM methods should improve models and must be included in the last stages of the prediction pipelines.

DISCUSSION

Emergence of the "meta-predictor" and potential directions for consensus methods

Unfortunately, our evaluation process falls short in identifying promising new methods that may ultimately drive future significant progress in the field and merely ranks the performance of groups on a defined set of target domains. The setup of the CASP experiment has driven participants to fully automate their prediction process. Groups that provide the most successful prediction servers seem to be successful engineers of computational pipelines that combine the best available techniques for each prediction category. Due to the ever increasing number of targets and the pressure to predict all targets to get the best scores, human expert predictors that showed promise in past CASPs have ceased to participate. A new type of predictor has emerged in CASP9: the human expert "meta-predictor," analogous to the "meta-server" of CASP5.[7] The strategy of applying energy functions to pick among all the models provided by servers appeared to outperform other methods. However, the inability of the same groups to assign "first" models suggests room for improvement for these types of methods. Interestingly, nine FM domains had nine or more human expert groups outperform the top servers. The top server predictions for all of these domains were provided by the top performing servers in the past CASPs[2,3,6]: either ROSETTASERVER (3 domains) or Zhang-Server (6 domains). These data suggest that human expert selection of models was perhaps influenced by the reputation of the servers in addition to the energy that favored the models.

By including a manual evaluation of a subset of FM targets, we noticed that despite the presence of a number of different secondary structure arrangements and interactions, a predominance of the predictions displayed the same correct local cores. For example, a majority of predictions for the target T0531 illustrated in Figure 2 included the β-hairpin formed by the second and third strands of the meander. These secondary structures form the most local interactions of the structure core, being separated by a short loop. Manual scoring suggests that a significant portion of the predictions (60 out of 550 predictions with non 0 scores, 11%) placed these two strands adjacent to each other in the correct register. Inspection of the position dependent alignment for this target provided by the prediction center supports this observation, with a number of the top-performing groups correctly aligning this section of the structure. Similar correctly-identified local cores were present in predictions of most of the FM target domains. For example, a β-hairpin and short helical segment in T0550_2 (residue ranges 233–261), a helix/β-hairpin in T0578 (residue ranges 1–47), and a three-strand β-meander in T0624 (residue ranges 34–60). Although the reasons behind the presence of these local cores in CASP9 predictions remain unclear, perhaps they could provide a basis for developing future structure prediction methodologies. If short locally interacting secondary structures could be identified through consensus methods, then the degrees of freedom become lower for the remainder of the structure. Freezing local cores may allow fragment-based assemblies to more fully sample the remainder of the structure or may provide enough constraints for physics-based methods to tackle larger structures.

Current CASP assessment procedure hinders evaluation of methodology

Similar to previous CASPs, a significant overlap between prediction categories remains. One of the main aims of CASP is to evaluate current state of the art methods according to categories: template-based modeling or template-FM. However, since the top servers provide pipelines of different techniques (both template based and template free), such a methods-based evaluation is impossible. To compound the problem in CASP9, many

human experts picked among these models without noting the server sources. The CASP assessment is necessarily blind to methods to avoid any kind of bias in establishing performance. However, this blindness resulted in a great deal of time spent trying to establish which techniques were actually "template-free" and worthy of mention. Perhaps the best indicator of methodology performance arose from comparing top predictions to closest available templates. We assume making significant improvements over the closest templates to be the "template-free" portion of the methodology. Some of these improvements came in the form of energy refinement of top server models (sometimes template-based server models) and some came from the same fragment-based assembly methods that have outperformed in the FM/new fold categories since the initial development of Rosetta in CASP3.[1,23] By comparing these top predictions to the best server models (ServerRatio score), we attempted to identify the source of the top-performing models used for refinement. Once the source was identified, we could note the "template" comment in the prediction file to exclude template-based predictions. Due to the ambiguity of this procedure in guessing which server models served as sources for refinement and whether or not a template was utilized, our assessment of FM methodology itself as used by predictors remained largely unsuccessful and lagged behind establishing the overall group performance and rankings.

CASP9 "winners"

Two target domains represent the highlights of the CASP9 FM category. The known *ab initio* method Rosetta developed by the Baker group, which has not changed in a significant way since CASP8, provided the most outstanding prediction (prediction 321_4 for target T0581) in CASP9. Although a number of groups selected this model for submission, the ROSETTASERVER could not distinguish it as the best. On the contrary, the Baker group did select this model as best among their Rosetta models (172_1). However, their refinement of the initial server model (GDT 64.7 for ROSETTASERVER 321_4) moved the prediction further from the target (GDT score 48.2 for Baker group 172_1). A second notable server prediction also improved over the closest template and likely resulted from a template-free method, as none of the noted templates corresponded to a ferredoxin-like fold (the templates applied to the other domains for target T0604). Although the prediction for this target (T0604_1) displayed less of an improvement over the closest template, the Zhang-server correctly designated the best model as first (428_1), and the human expert Zhang group refined the model to a higher GDT score (96_1, improved by 5%). Thus, these two groups (Baker and Zhang) have developed servers (Zhang-server, Quark, and ROSETTASERVER) that both significantly outperformed the remaining servers in CASP9 and improved over the closest available template for at least one of the FM targets. As servers provided the basis for top human expert predictions, we consider predictions by the Baker and Zhang servers as the "winners" of CASP9.

Knowledge-based potentials fail predict atypical structures

Despite these highlights, a predominance of pitfalls challenged CASP9 predictions. In addition to those pitfalls previously discussed, two difficult targets classified as new folds (see classification paper) provided examples of structures with atypical characteristics that knowledge-based prediction methods would fail to identify. The first target (T0529_1) is quite large (339 residues in the domain, 561 including the second domain) and displays a significant number of non-local contacts with high contact order. For example, the N-terminal helix forms a three helix bundle with the two C-terminal helices. Two N-terminal sections of the target sequence (residues 8–57 and residues 58–128) wrap around the circumference of the structure in opposite directions, making few local contacts outside the helical backbone. Only a small portion of the structure (residues 180–266) forms a five-helix array (see classification paper) with local contacts that could be classified as a typical structural domain. However, the helices in this array are relatively short with poorly predicted secondary structure (only 3 are predicted correctly). Presumably, the numerous additional decorations and the lack of correctly predicted secondary structures mask the presence of this small core for most structure prediction methods. Additionally, since this domain is long, most predictors tried to partition it in a large number of small domains, which obviously resulted in poor models for a single domain sequence segment. The second atypical new fold forms an elongated tail fiber from non-local β-strand interactions. In addition to the non-globular nature of the domain (roughly 175 Å long, with a diameter of 15 Å), the elongated strands organize around seven iron atoms coordinated by clusters of histidine residues. The histidines are contributed from three HXH motifs, with one motif from each of three chains in a trimer. Although one might predict the target forms, a trimer based on the N-terminal TBM domain (the template forms a trimer), the presence of a C-terminal extended region in the N-terminal domain template that also has histidine motifs caused a similar alignment extension problem as described for T0604_3.

Purification tags and signal peptides hinder *ab-initio* methods

Finally, a number of submitted sequences included purification tags and signal peptides. These extensions of

the structure domains (especially the hydrophobic nature of the signal peptide) provided additional complications for prediction methods. For the N-terminal domain of target 608 (T0608_1), most prediction methods attempted to place the signal peptide as the central helix of a helical array. Although the target domain forms a helical array with a topology analogous to the helices of lysozyme, replacing the central core helix with the hydrophobic signal peptide sequence destroys most of the contacts of the real structure and results in low scores. In some of the top-performing predictions for this domain (for example 172_5), the signal peptide is either decorated by the two N-terminal helices (172_5) or is split into two helices to form self-interactions (147_1), allowing the core helix some of its native contacts. Signal peptides display a strong sequence motif, having a defined stretch of hydrophobic residues near the N-terminus. Several good programs exist to predict the presence of such sequences, for example SignalP.[24] Routinely using such programs to predict and remove signal peptides would probably improve the performance of most FM prediction methods over such targets.

CASP9 Progress

The availability of a single server by the Karplus group (SAM-T08[25]) that has not changed since CASP8 provided a unique opportunity for comparison of current predictions with those of the previous CASP. Frozen SAM-T08 server GDT scores should provide a consistent difficulty estimate of the target domains from the present and the past CASPs. Inspection of SAM-T08 server GDT scores for FM-defined domains in CASP9 (26 domains) suggested that those targets in CASP8 with scores below GDT 46 were somewhat matched in difficulty (28 domains). A histogram of target domain GDT scores produced by the SAM-T08 server illustrates the distribution of target difficulties in each CASP [Fig. 7(A)], with three of the CASP9 target domains being somewhat more difficult than those of CASP8 (T0529_1, T0604_3, and T0629_2). The distributions are quite similar, with a median that is shifted lower for CASP9 targets, indicating that current targets are relatively more difficult than those in the previous CASP. Despite the increased difficulty, groups tended to perform better on these matched targets in CASP9, as estimated by the SAM-T08 server relative performance [Fig. 7(B), best model/SAM-T08 GDT ratios]. The distribution of performance on difficulty matched targets suggests a slight overall increase in performance of the prediction community as a whole. Despite the improved relative performance of CASP9, the question remains as to whether this better performance actually reflects significant advances in prediction technology or if it merely reflects the public release of server predictions to the community. CASP9 FM targets include two good-performing outliers with respect to the overall

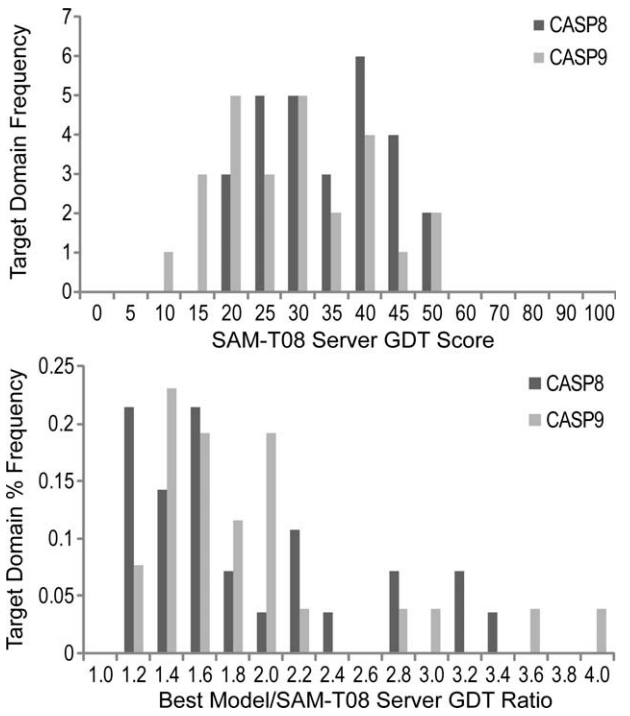

Figure 7

CASP9 performance compared to CASP8 by Sam-T08 server. **A**: Distribution of SAM-T08 GDT scores for the defined CASP9 FM targets (gray bars) suggests CASP8 targets with SAM-T08 GDT scores below 46 are matched in difficulty (black bars). Bars representing the most difficult CASP9 targets are labeled by domain name. **B**: Distribution of best model GDT/SAM-T08 GDT ratio scores for all FM targets in CASP9 (gray bars) and for matched targets in CASP8 (black bars). Bars corresponding to Top-performing CASP9 target domains are labeled.

performance distribution (T0547_3 and T0581). As discussed above, the T0547_3 domain folds into a small three-helix bundle. The relatively good performance may reflect the limited number of ways these helices can associate into a bundle and the relative ease of splitting out the insertion. The top-performing group (75_2, RAPTORX-FM) limited their prediction to only this segment and declared no template, suggesting that the group correctly identified the insertion and applied a template-free method to produce the best prediction. The second outlier formed the basis of the CASP9 winning prediction (discussed above). Although the next best CASP9 target domain (T0604) tended to outperform the best template, the relative performance estimated by the SAMT-08 ratio did not outperform the best targets of CASP8. Those servers that drove the success of the prediction community should be recognized: Rosetta and Zhang/Quark servers.

ACKNOWLEDGMENTS

Authors thank the CASP9 organizers, John Moult, Anna Tramontano, Krzysztof Fidelis and Andriy Kryshtafovych for asking them to be a part of the CASP experience (again). Eight years must have been long enough

for them to forget the headaches that the authors caused. Authors greatly appreciate input and discussions from many CASP participants and the other assessors, in particular Torsten Schwede, numerous discussions with whom were crucial in formulating and refining our ideas. Andriy Kryshtafovych provided constant computational support, expertly analyzed and modified target structures, computed scores of models and promptly addressed all the questions arising in the process of assessment.

REFERENCES

1. Simons KT, Bonneau R, Ruczinski I, Baker D. *Ab initio* protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37 (Suppl 3):171–176.
2. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 2009;77 (Suppl 9):89–99.
3. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins 2009;77 (Suppl 9):100–113.
4. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 2007;69 (Suppl 8):118–128.
5. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69 (Suppl 8):108–117.
6. 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.
7. 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.
8. Shi S, Pei J, Sadreyev RI, Kinch LN, Majumdar I, Tong J, Cheng H, Kim BH, Grishin NV. Analysis of CASP8 targets, predictions and assessment methods. Database (Oxford) 2009;2009:bap003.
9. Kinch LN, Qi Y, Hubbard TJ, Grishin NV. CASP5 target classification. Proteins 2003;53 (Suppl 6):340–351.
10. 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.
11. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
12. DeLano WL. The PyMOL Molecular Graphics System. 2002. DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org
13. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566–567.
14. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
15. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
16. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
17. Sauder JM, Arthur JW, Dunbrack RL, Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40:6–22.
18. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
19. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999; (Suppl 3):22–29.
20. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
21. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure 2002;10:435–440.
22. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
23. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999; (Suppl 3):149–170.
24. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004;340:783–795.
25. Karplus K. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res 2009;37(Web Server issue):W492–W497.