UC Davis
UC Davis Previously Published Works

Title

Evaluation of free modeling targets in CASP11 and ROLL.

Permalink
https://escholarship.org/uc/item/9qp4n54c

Journal
Proteins: Structure, Function, and Bioinformatics, 84 Suppl 1(Suppl 1)

Authors
Kinch, Lisa
Li, Wenlin
Monastyrskyy, Bohdan
et al.

Publication Date
2016-09-01

DOI
10.1002/prot.24973

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

HHS Public Access
Author manuscript
Proteins. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:
Proteins. 2016 September ; 84(Suppl 1): 51–66. doi:10.1002/prot.24973.

Evaluation of free modeling targets in CASP11 and ROLL

Lisa N. Kinch[1,*], **Wenlin Li**[2], **Bohdan Monastyrskyy**[3], **Andriy Kryshtafovych**[3], and **Nick V. Grishin**[1,2]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center at Dallas, 6001 Forest Park Road, Dallas, Texas 75390-9050

[2]Department of Biophysics and Department of Biochemistry, University of Texas Southwestern Medical Center at Dallas, 6001 Forest Park Road, Dallas, Texas 75390-9050

[3]Genome Center, University of California, 451 Health Sciences Drive, Davis, California 95616

Abstract

We present an assessment of 'template-free modeling' (FM) in CASP11and ROLL. Community-wide server performance suggested the use of automated scores similar to previous CASPs would provide a good system of evaluating performance, even in the absence of comprehensive manual assessment. The CASP11 FM category included several outstanding examples, including successful prediction by the Baker group of a 256-residue target (T0806-D1) that lacked sequence similarity to any existing template. The top server model prediction by Zhang's Quark, which was apparently selected and refined by several manual groups, encompassed the entire fold of target T0837-D1. Methods from the same two groups tended to dominate overall CASP11 FM and ROLL rankings. Comparison of top FM predictions with those from the previous CASP experiment revealed progress in the category, particularly reflected in high prediction accuracy for larger protein domains. FM prediction models for two cases were sufficient to provide functional insights that were otherwise not obtainable by traditional sequence analysis methods. Importantly, CASP11 abstracts revealed that alignment-based contact prediction methods brought about much of the CASP11 progress, producing both of the functionally relevant models as well as several of the other outstanding structure predictions. These methodological advances enabled *de novo* modeling of much larger domain structures than was previously possible and allowed prediction of functional sites.

Keywords

protein fold prediction; structure comparison; alignment quality; ab initio; domain structure; protein structure; CASP11; CASP ROLL; free modeling

[*]Correspondence to: Lisa Kinch, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center at Dallas, 6001 Forest Park Road, Dallas, TX 75390-9050. lkinch@chop.swmed.edu.
Lisa N. Kinch and Wenlin Li contributed equally to this work.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

INTRODUCTION

CASP11 assessment of protein structure prediction intends to identify and evaluate the current state of the art methods in the field. The objective of the template- free modeling (FM) category is to assess methods that predict 3D structures from a given protein sequence without the explicit use of template structures available in the Protein Data Bank.[1] Historically, the most successful methodology for such *de novo* structure prediction from sequence assembles fragments into relatively small folds.[2–6] However, the past two FM assessments have noted the emergence of top-performing groups that incorporated the use of remote templates or selection and refinement of server models into their prediction methodology.[7,8] While the relative success of such methodology over the past decade suggests some advances in evaluating model quality and perhaps refining tertiary structure, it highlights the fact that the structure-folding problem remains unsolved.

In past CASP experiments, the FM category tended to have relatively few targets available for evaluation, with 30 FM domains in CASP9 and 20 domains in CASP10.[9,10] This dearth of targets tended to cause difficulty in assigning statistical significance to assessments of FM techniques and prompted the introduction of CASP ROLL into the CASP10 evaluation.[8] CASP ROLL collects potential FM targets year-round for the same double-blind prediction and assessment as the traditional biennial CASP tertiary structure evaluation. Although CASP ROLL significantly increased the number of targets for FM assessment in CASP10 (15 additional targets), fewer groups participated in the year-round prediction scheme (41 ROLL predictors vs. 147 traditional predictors). As such, for CASP11, the organizers chose to add several of the CASP ROLL targets to the traditional biennial experiment. With the help of these nine additional overlapping target structures (providing 13 domains for FM evaluation), the CASP11 FM category exhibited the largest number of targets for evaluation since its inception (45 target domains). Most of these target domains were designated for all groups (39), while six were for server only. The CASP11 ROLL category included 25 additional domains, which resulted in 38 evaluation units all together.

Given this large number of evaluation units, the CASP11 assessment was not predominantly based on a formal and comprehensive manual evaluation. Alternately, identification of interesting model predictions and group ranking relied on a battery of automated scores produced by the Prediction Center.[11] Only the top-scoring models were subjected to careful manual examination. This report outlines our evaluation procedure and its application to CASP11 FM and ROLL target predictions. We ranked server performance on all 45 FM target domains and all groups on 36 all-group FM target domains using a combination of scores that encompassed those used in previous CASP FM evaluations as well as two additional scores aimed to highlight local model quality. Such scores helped to identify several prediction models that could provide useful functional insights to experimentalists, an initiative highlighted by the CASP organizers as being an important part of the evaluation. We also provided insights into FM prediction methodology and outlined the progress of the fold prediction community as a whole.

The successful prediction of a 256-residue target domain (T0806-D1) exemplified one of the outstanding models provided for CASP 11 FM target domains, representing one of the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

largest correctly predicted FM folds in the history of CASP. Other successful predictions included three relative small single-domain targets T0824-D1 (110 residues), T0837-D1 (128 residues), and T0855-D1 (115 residues); a small N-terminal domain of a multidomain target T0793-D1 (109 residues); two domains of the same fold T0790-D2 (130) and T0791-D2 (139), and a C-terminal domain T0827-D2 (158 residues) fused to a TBM domain. The groups that provided these prediction models also tended to outperform in overall rankings (Baker lab manual group 64 and server group 184 for best models; Zhang lab Quark server group 499, Kihara lab group 333, and Lee lab LEER group 44). Notable CASP ROLL predictions include R0036 and R0021, and two of the same labs outperformed in this category (Baker Rosetta server and Zhang ab initio).

METHODS

Scores to identify top prediction models

To highlight top predictions we generated scores that (1) compare prediction models to random models (random ratio) and (2) compare prediction models to top templates (template ratio). For random model scores, we generated target specific ‘permutation shift’, ‘reverse chain’, and reverse chain permuted random models as previously described.[12] GDT_TS scores[13] for each random model with the initial target domain were averaged to get a random model score. The random ratio used for identifying notable model predictions corresponds to the ratio of the best server or manual group GDT_TS score to the random model score. For template ratios we divided the sequence independent LGA_S score[13] of the model by the LGA_S score of the template (both compared to the target).

Evaluating overall group prediction performance and significance

To provide a single measure to reflect overall prediction quality of each group, we combined six scores provided by the Prediction Center (GDT_TS,[13] TenS,[7,14] QCS,[15] ContS,[12] lDDT,[16] and MolProb[17]). Four of the scores were used in our evaluation of CASP9 FM predictions (GDT_TS, TenS, QCS, and ContS[7]) and two are introductions for CASP11 (lDDT and MolProb). We calculated *Z*-score sums (and averages) over all the targets for the six chosen scores. *Z*-scores were calculated as in previous CASPs.[7,14] The Prediction Center not only generated these comparison scores for every prediction, but also provided a web server that allowed us to combine various scores with different weights to produce ranks of first or best models from server only or all groups. The web server allowed rapid evaluation of group performance (ranking) according to previously published combined *Z*-scores analysis[16] and testing the robustness of various different scoring schemes for producing ranks.

Significance scores for CASP11 FM ranks included bootstraps and *t* tests similar to those used in previous CASPs.[7,18,19] We also carried out head-to-head trials of the group results, calculating the fraction of common targets for which one group outperformed the other. We repeated the overall win/loss counts for all-against-all pairwise comparisons produced by the CASP10 contact assisted assessment.[20] Briefly, we performed all-against-all pairwise comparisons for models from the same target and summed the numbers of win/loss cases for each group, as well as calculated the probability that a win/loss record was equal to or better

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

than that obtained by chance. All the evaluation tables are available via the link: http://prodata.swmed.edu/casp11/FM.

Evaluating groups that choose among server models

Many of the current top-performing methods use scoring schemes to choose among provided server models. In an attempt to identify and evaluate the performance of such strategies in absence of information about methodology, we used the GDT-TS clusters provided by the Prediction Center. The Prediction Center provided all against all GDT_TS comparisons of prediction models and clustered them at three GDT_TS cutoffs: 90, 80, and 70. To identify likely server models used by each prediction group, we searched for server models that cluster with each manual group at the lowest cutoff (70 GDT_TS), choosing the reference server model as the one with the closest GDT-TS. We calculated the difference between the GDT_TS of the manual group model to the target and the GDT_TS of their reference server model to the target, reporting negative values when the reference server model is closer to the target as compared to the manual model.

Evaluating progress: potential functional prediction and comparison to CASP10

To calculate family-based conservations for each target, we used two strategies to gather similar sequences at differing conservation stringencies. We ran three iterations of PSI_BLAST, filtering results for *E* value < 0.001 and identity less than two values: 40 or 60%. We generated multiple sequence alignments (MSAs) by aligning collected sequences to the target sequence using the BLAST alignments. MSAs were used to calculate conservations for each position in the target sequence using the program AL2CO[21] and map the conservations to the B-factor column of the PDB. Conservations were visualized using rainbow coloring of the B-factors in Pymol, from blue (variable) to red (conserved).

GDT_TS scores from the Karplus group server (SAM-T08[22]), which provides a stable method of predicting targets since CASP8, were used to compare CASP11 models to those from CASP10. Similar to our analysis of CASP9 progress,[7] we compared histograms of the SAM-T08 model GDT_TS scores from CASP11 and from CASP10 to evaluate the difficulty level of the targets. An estimate of progress was provided by normalizing the best model scores from CASP11 and CASP10 by the SAM-T08 server GDT_TS scores.

RESULTS

Community-wide server performance on FM targets

GTD_TS scores computed by the LGA program[13] have provided the basis for model predictions since early CASPs.[14] The server model GDT_TS score distributions plotted for each FM target domain (Fig. 1) were ordered and colored from more difficult (dark blue, left) to less difficult (light blue, right) according to the average server model GDT_TS scores. Numerous outliers were observed in the distributions, suggesting that some server model predictions outperformed the rest. For example, the average of all server models for target T0804-D2 was only 14.5 GDT_TS, yet three predictions extended toward much higher GDT_TS bins. Quark provided the top server model for this target domain (38.65 GDT_TS for TS499_5), followed by two models from the Zhang-server (TS277_5 and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TS277_3). Each of these models roughly captured the immunoglobulin-related topology of the target domain fold. Similar outlier server prediction examples were produced for targets T0790-D2 (TS184_3, GDT_TS 44.81), T0793-D1 (TS499_2, GDT_TS 48.27), T0855-D1 (TS041_1, GDT_TS 51.09), T0837-D1(TS499_1, GDT_TS 61.98), and T0775-D6 (TS184_3, GDT_TS 68.57). This score distribution provided the basis for using *Z* scores as a measure of performance in the CASP11 FM evaluation and suggested that a rigorous manual scoring component might not be necessary to evaluate the prediction models.

Top predictions: random and template score ratios

Given the numerous outliers observed in the server model GDT_TS score distributions (Fig. 1), we sought to reveal top performing predictions by comparing the best manual and server model scores to those of random models for each target. The distribution of random model GDT_TS scores for all FM target domains skewed to the right [Fig. 2(A)]. Targets with unusually high GDT_TS scores (T0775-D3 and T0775-D1, colored light gray in Fig. 2) corresponded to small domains of extended secondary structure that only form compact units in an oligomeric state (see Kinch *et al.* CASP11 Target Classification, this issue). Similar to findings from previous CASPs,[7,12] the random model GDT_TS scores for CASP 11 targets showed inverse correlation to domain length [Fig. 2(B)]. Thus, by considering random models ratios of top manual and server models [Fig. 3(A)], we tended to down weight the value of target domains with small and irregular folds.

With the exception of a single prediction (best server model for small and extended outlier target domain T0775-D1), the GDT_TS scores of top models outperformed the average GDT_TS of random models for all 45 CASP11 FM target domains. The top predictions beat the random models by an average of 2.2-fold for manual models and by twofold for server models, with one outstanding manual prediction (TS064_1) for target T0806-D1 improving random models by >6-fold. The best random ratio server prediction (TS499_1) for target T0837-D1 improved the random model score by >3-fold. The manual group TS317_1 slightly improved over the top server model for this target. Several additional manual prediction models tended to outperform when compared to random: TS333_1 on target T0804-D2, TS065_3 on target T0793-D1, TS064_2 on target T0824, and TS260_3 on target T0855-D1. With the exception of the outlier prediction for target T0824, the manual models slightly improved over top server models for these examples (produced by TS499_5, TS499_2, and TS041_1, respectively). One server provided a notable prediction (TS038_4) for target T0814-D1 that outperformed all the manual predictions. LGA template ratios highlighted predictions that outperform top templates [Fig. 3(B)]. The template ratio comparison emphasized similar outperforming predictions, with the same two standing out above the rest: TS064_1 prediction of target T0806-D1 and TS499_1 server with the corresponding TS317_1 manual predictions of target T0837-D1.

Top predictions: outstanding structure model examples

Figure 4 illustrates several of the outstanding structure models identified using random model and template ratios from Fig. 3, including the top manual prediction for target T0806 and the top server predictions for targets T0837 and T0855. While to top manual prediction for T0806 was unique among all provided models (GDT_TS score 60.55, with the next best

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

group score 34.38), the top server predictions on the other targets tended to closely resemble predictions from manual groups. Given the availability of all server predictions for use with manual prediction methods, the similarity suggests that some manual methods could successfully identify and perhaps marginally improve the top server models. For example, group 317 submitted a similar model as server 499 for T0837 with a slightly better GDT_TS (1.66 difference), and group 260 submitted a similar model as server 41 for T0855 with a better GDT_TS (3.26 difference).

The most outstanding FM target prediction based on both random model and template ratios was produced by the manual group 64 (BAKER) for target T0806. The model TS064_1 (Fig. 4A) maintained the correct topology over the entire 256-residue domain of target T0806-D1 (Fig. 4B), including an insertion of three α-helices into the Rossmann-like fold. Although numerous templates existed with similar Rossmann-like topologies, none could be identified by sequence. The closest structure template (25.0 LGA_S), uncharacterized protein AF0587 [PDB ID:2q07], adopted a uracil-DNA glycosylase-like fold [Fig. 4(C)]. Although this template retained the core three layer $\alpha/\beta/\alpha$ Rossmann-like topology of the target domain, it lacked the three-helix insertion as well as an additional C-terminal β-strand/α-helix that extended the β-sheet in the target structure. Given the difficulty of the target structure, we sought to discover the methodology used to provide such a prediction. Apparently, T0806 exemplified one of the few CASP11 targets with a sufficiently large enough family for predicting co-evolving residues from sequence alignments. Thus, incorporating contact-based prediction information from co-evolving residues into *ab initio* methods provided the unprecedented structure model for this 256-residue long CASP11 FM target domain.

The most outstanding server prediction based on both random model and template ratios was produced by group 499 (Quark) for target T0837-D1. The top server model TS499_1 and top manual model TS317_1 were almost identical [Fig. 4(D)], having a GDT_TS score of 92.38 (RMSD 0.98) between themselves. Both prediction models maintained the correct topology of the target fold over all seven α-helices [Fig. 4(E)]. The models also outperformed the structurally-related AhpD-like fold template (33.26 LGA_S) of gamma-carboxymuconolactone decarboxylase [PDB ID:2af7] [Fig. 4(F)]. This α-array template adopted roughly the same topology for all 7 α- helices, yet had alternate rotations for helix1–3 with helix 4–5 and helix 4–5 with helix 6–7.

On target T0855-D1, a top manual prediction TS260_3 was somewhat similar to an outstanding top server prediction TS041_1 [Fig. 4(G)], with a GDT_TS score of 60.0 (RMSD 4.1) between the two models. Both top predictions captured the entire new fold of target T0855-D1 [Fig. 4(H)] and significantly improved the top unrelated transmembrane fold of the voltage-dependent anion channel (VDAC) template [PDB ID:2k4t] [Fig. 4(I)], which retained only the β-meander of the target fold.

Combining automated scores for ranks

The server model distributions illustrated in Fig. 1 suggested that *Z* scores might provide a good system of evaluating prediction model performance. Indeed, notable server models that outperformed in our ratio tests were also outliers in the GST-TS score distributions (i.e.,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TS499_1 for T0837, TS499_5 for T0804-D2, and TS184_3 for T0790-D2). We therefore chose to expand on the GDT_TS scores by combining *Z* scores of different measures that intended to capture diverse aspects of the FM predictions. The Prediction Center reproduced four such scores that were combined in our previous FM evaluation of CASP9 FM targets: GDT_TS,[13] ContS,[12] TenS,[7,14] and QCS.[15] For the CASP11 FM scoring scheme, we chose to combine six scores with equal weights for final ranking: adding local distance difference test (lDDT), a superposition free score that well suited to assess local model quality;[16] and MolProb scores from a MolProbity server routinely used to evaluate the quality of crystal structures[17] to the four previously used scores.

While the suitability of measures used in the CASP9 FM assessment have been described previously,[7,12,14,15,20] the choice of incorporating lDDT and MolProb scores in the CASP11 FM evaluation aimed to promote methods that consider local model quality and local geometry, respectively. The lDDT score[16] was introduced as an evaluation component in CASP9 template-based modeling[23] to complement the rigid body superposition of GDT_TS and was subsequently included as a routine evaluation measure provided by the Prediction Center in CASP10.[11] Given the new CASP initiative to address the biological relevance of structure models in assessments, lDDT is well suited to compare functionally relevant regions of structure models. The other additional measure, MolProb, represents an aggregated penalty score produced by the MolProbity package[17] that considers the number of all-atom steric overlaps or Clash-score, the rotamer outlier score or percentage of side-chain conformations classified as rotamer outliers (Rot-out), and the percentage of backbone Ramachandran conformations in favored regions (Ram-fv). To determine the acceptable range of MolProb penalty scores, we compared their distribution in FM target structures (ranges from 0.5 to 3.25 bins, with a maximum frequency at 1 to 1.25) to their distribution in FM structure models (ranges from 0.5 to >5, with a relatively high maximum frequency at 3.5). The broad distribution of MolProb scores in FM models suggests that some methods account for local geometry of their models, while the majority of methods do not.

The combined FM scoring scheme should establish a more comprehensive and robust measurement of model quality, even in the absence of any formal manual analysis. Using summations of this combined scoring scheme, ranks on server-only targets (45 FM domains) using best models (Table I) highlight the top five server groups: BAKER-ROSETTASERVER (group 184), Zhang-Server (group 277), Quark (group 499), MULTICOM-NOVEL (group 41), and nns (group 38). With the exception of group277, these top servers produced all of the outstanding prediction models highlighted in Figure 3. Interestingly, the two Zhang group servers (group 499 and group 277) perform similarly. Their head-to-head pairwise comparison showed that group 277 beat group 499 for just over half of the FM targets (0.578), while group 499 tended to beat group 277 in providing outstanding predictions when compared to random models (i.e., T0804-D2, T0793-D1, and T0837-D1). Significance estimates of the combined scores using bootstraps and *t* tests suggested that the BAKER_ROSETTASERVER provided models that are significantly better than the rest of the servers (highlighted in gray, Table I), although the Zhang-server and Quark were not significantly different from the BAKER-ROSETTASERVER using only GDT_TS scores (http://prodata.swmed.edu/casp11/FM/). Ranks change when considering

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

first server models, with the top-performing Zhang-server, RBO-Aleph, and Quark being statistically indistinguishable by bootstrap and *t* tests.

Ranking of manual groups on "all-group" targets (39 FM domains, Table II) highlighted the top performance of three (BAKER group 64, Kiharalab group 333, and LEER group 44) that were statistically indistinguishable according to bootstraps and *t* tests, together with a fourth manual group (Boniecki_pred group 32) that only provided predictions for 32 out of the 39 FM domains. When considering first models only, one of these groups (Kiharalab group 333) significantly outperformed the rest. The top-performing manual groups also tended to produce the outstanding predictions highlighted in Figure 3, with the top-ranked manual group 64 providing the most outstanding prediction model for CASP11 FM targets: a 256 residue-long target domain T0806-D1 with no sequence-related templates illustrated in Figure 4.

Insights into prediction methodology

Some of the top-performing methods in recent CASPs used partial templates to produce reasonable FM prediction models or used scoring schemes to choose and sometimes refine provided server models. We therefore sought to evaluate the performance of such methods on CASP11 FM targets. In absence of any provided information about methodology that was required for maintaining blind evaluations, such analysis required taking advantage of information provided by the prediction groups. For instance, groups provided template information in a "PARENT" line for each model. Such information could potentially be useful for evaluating the ability of servers to combine multiple partial templates into FM models. We limited such analysis to servers, as manual groups could have taken advantage of using multiple methods and might provide less rigorous template information. Manual groups had the added ability to choose among all server models. To evaluate the ability of such methods to score and potentially improve server models provided by the Prediction Center, we used a strategy to map manual models to their closest "reference" server models using clustering, evaluating the performance improvement over the reference model with GDT_TS differences.

To gain insight into the extent of methods that used templates, we counted the number of models provided by each server that declared a parent template or not for all target domains (FM and TBM) and for a limited set of templates categorized as FM-only (removing multidomain targets that span both categories). Considering all CASP11 targets, the servers range from always declaring parent templates (11 template-based servers) to declaring a variable number of parent templates (18 hybrid servers), to never declaring parent templates (15 N/A servers). We would like to assume that the N/A subset of servers (or manual predictors) utilize template free modeling predictions. However, most of them described their methodology as "template-based" in the CASP abstracts or primary citations: including SAM-T08, RaptorX, nns, FLOUDAS_SERVER, distill, Alpha-Gelly-Server, and 3D-Jigsaw. Others use multiple partial templates, yet do not declare any as parents. For example, the FALCON_TOPO server refers to its multiple partial templates as a "common framework" in the CASP11 abstracts. Finally, a smaller subset of N/A servers concentrated on predicting contacts (MULTICOM servers and myprotein-me). Most of the hybrid methods (13 out of

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

18 servers) declared a smaller proportion of parent templates for the subset of FM-only targets than they declared for TBM.

The outperforming Baker–Rosetta server declared parent templates for almost half (47%) of their models in the FM-only dataset. This relatively high number of template-based FM of models could suggest that the Baker Rosetta server successfully combined partial templates. However, inspection of their FM models revealed the high number of declared parent templates resulted from a strategy of using templates for model1 and model4 (and sometimes more) for the 18 FM-only targets. For most of these targets, the GDT_TS scores of the nontemplate based models far exceeded the template-based ones. For example, the nontemplate model prediction for T0790 (TS184_3) ranked number one (GDT_TS 24.34) while the template-based model (TS184_1) ranked 49 (GDT_TS 12.26). The template-based prediction for FM target T0761 provided one exception to this rule, with the model TS184_4 ranking number one (GDT_TS 38.27) for T0761-D2. The model for the entire target used two unrelated templates, assembling two α-helices with a β-meander from one of the templates into the C-terminal domain. This strategy also explained the poor performance of the Baker-Rosetta server on first models, where they fell behind the Quark and Zhang-servers. These additional outperforming servers declared 0 and 17% parent templates on the FM-only targets, respectively. Thus for CASP11, non-template based methods tended to outperform as servers.

Prompted by our observations of top-performing server model predictions having closely related manual model predictions, we attempted to map manual target submissions to their reference server models. All-against-all GDT_TS clusters provided by the Prediction Center were parsed for manual models that clustered with server models above GDT_TS 70. Unclustered manual models were considered unique. This somewhat conservative cutoff might have missed some manual models that aggressively refined server models. For example, the TS041 server model from Figure 4(G) did not cluster with the structurally similar TS260_3 manual model at this cutoff. Subsequent inspection of the abstracts after the CASP11 FM evaluation supported this cutoff, as group 260 described their model predictions as refinements of Quark models (group 499) and not MULTICOM_NOVEL models (group 041).

Using this clustering scheme as a guide, almost half of the manual groups participating in CASP11 tended to choose among server models [Fig. 5(A)], with 10 groups mapping 100% of their models and an additional 12 mapping >70%. One additional group with an intermediate percentage (41%) described their method as refinement of server models in the CASP11 abstracts (TS241). For the groups that tended to choose among server models, we calculated the tendency of their methods to improve server models through refinement [Fig. 5(B)]. Although the GDT_TS cutoff of 70 that we defined as an indicator of picking server models tends to limit the observed GDT_TS differences in Figure 5(B), one of the groups (TS153) consistently improved server models by almost 0.5 GDT_TS. Almost half (11 groups) improved server models (by an average of 0.1 GDT_TS), two groups picked server models as is, and the rest (10 groups) tended to worsen the models (by an average of 0.1 GDT_TS). The top-performing first model group (TS333) selected 100% of their models among servers and refined them on average to a marginally negative GDT_TS difference

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(−0.014 GDT_TS). Thus, the outperformance of group TS333 stemmed from their ability to pick but not necessarily refine the best server model for each FM target.

Target T0804-D2 exemplified one case of an outstanding selection by group 333 (Kiharalab). The GDT_TS score distribution of target T0804-D2 [Fig. 6(A)] highlighted a cluster of outlier models that outperformed the rest of the predictions. The T0804-D2 target structure [Fig. 6(B)] adopted the same β-sandwich domain in virus attachment proteins as the top canine adenovirus fiber head protein template [PDB ID:2j1k_f] [Fig. 6(C)], with the two having an LGA_S of 79. Group 333 selected the top server template TS499_5 [Fig. 6(D)], which roughly captured the domain in virus attachment proteins topology (GDT_TS 38.65), with a shift in alignment of the C-terminus and a failure to adopt the correct structure of 4 β-strands. The top manual template TS333_1 [Fig. 6(E)] slightly improved (GDT_TS 38.82) the top server model through refinement.

Utility of FM structure prediction models: suggested functions

We mapped family-based conservations of each residue position to the target structures and inspected the structures for clusters of conserved residues that could potentially serve as functional sites. Those structure prediction models that were of good enough quality to roughly position the conserved residue clusters were considered as useful for providing functional predictions. Contrary to our expectations, structure prediction models for two of the FM targets were adequate for providing useful functional information to experimentalists: T0836-D1 and T0824-D1. Each of these targets exemplified structures remotely related to their closest template fold in the evolutionary classification of protein domains (ECOD) database,[24] retaining secondary structure elements that define the core fold common to all existing members of the ECOD homology groups. The presumed homologous relationship of these two targets to existing ECOD groups, even in the absence of detected sequence similarity, was suggestive of function. However, given the remote relationships, clusters of conserved residues that map to similar positions as known active sites provided an additional level of support for both the homologous relationships as well as the suggested functions. In each case, migration of active site residues to alternate primary sequence positions that remain in close structural proximity likely contributed to the inability to detect the sequence similarity. Furthermore, inspection of CASP11 abstracts suggested both functionally relevant prediction models were produced by methodologies that incorporated alignment-based contact predictions.

Family-based conservations mapped to the T0836-D1 heme-binding protein of unknown function highlighted a potential active site in the target structure that included a typical Heme-coordinating His residue pointing into the center of a four-TMH helical bundle [Fig. 7(A), red spheres]. Furthermore, this core four-helical bundle of the target closely resembled the top template: the TMH domain of cytochrome *b* [PDB ID: 2fynA] [Fig. 7(B)], classified as a transmembrane heme-binding four-helical bundle in ECOD. The cytochrome b template bound two hemes in the center of the core four-helical bundle, with one of the heme-binding sites in a similar position as the mapped target active site. In comparison to the template heme-coordinating His residues, which were located on the second and fourth TMH of the core bundle, the presumed target heme-coordinating His residue migrated to the first TMH.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The top prediction TS065_4 by the Jones group retained the correct topology of the transmembrane heme-binding four-helical bundle and correctly placed the conserved His residue [Fig. 7(C), black spheres] that probably contributes to heme-binding function of the target.

The second example of prediction model utility for functional insights came from a potential active site marked by conserved residues mapped to the NucB DNase target T0824-D1 [Fig. 7(D), red spheres]. The fold of this target represented a significant deterioration of the top template endonuclease structure [PDB ID: 1g8t] [Fig. 7(E)], classified in ECOD as a His–Me finger endonuclease. The active site of the endonuclease marked by a bound Mg included a DxxH motif located near a conserved N (black spheres), in addition to a few positively charged residues that probably contribute to nucleotide binding (magenta spheres). The top structure prediction TS064_2 by the Baker group roughly placed the conserved active site DxD motif near the conserved N [Fig. 7(F), black spheres], residues that can coordinate Mg (or another metal) and mediate cleavage. The model also included several presumed nucleotide binding residues (magenta spheres) pointing toward the same side of the fold as the active site. Interestingly, existing members of the His-Me finger endonuclease superfamily have retained a conserver common core fold that includes an α-helix, β-strand, omega loop, β-strand, and α-helix surrounding the active site. This core fold was stabilized by extending the two β-strands into a larger β-sheet in the template structure, and was alternatively stabilized by a zinc finger for which the superfamily was named.

CASP ROLL PERFORMANCE

The performance of CASP11 ROLL participants using the FM style scoring scheme is summarized in Table III. The same top two groups outperformed using best models: BAKER-ROSETTASERVER (group 330) and Zhang *ab initio* (group 45), with models of the top-performing group 330 being significantly different by *t* test and bootstraps than group 45 using FM-style scoring and the two being statistically indistinguishable using GDT_TS. Each of these two groups also consistently provided top first models, as judged by their GDT_TS sums. However, many additional groups tend to have better average GDT_TS first models than the top groups, including Zhang, FOLDIT and Kaesar; and the first models of the top two ranked groups were not statistically different in significance tests.

The most outstanding prediction for ROLL was for the up and down α-helical bundle of target R0034-D1 [Fig. 8(A)]. The top-performing ROLL prediction model TS045_1 [Fig. 8(B)] from the Zhang *ab initio* server included all five of the α-helices of the target domain in the correct topology with correct alignment over most of the structure (residues 40–110) but the last α-helix being broken. The closest spider silk N-terminal domain template [PDB ID: 2lpj] (classified as a PWI domain in ECOD) also included all five α-helices in the same topology [Fig. 8(C)], but the prediction model outperformed the top template by 1.4-fold. One top prediction by the BAKER_ROSETTASERVER was for the CASP ROLL target R0021 [Fig. 8(D)], which adopted an eight-stranded β-meander barrel similar to lipocalin/streptavidin folds. The top-performing prediction model TS330_4 from the BAKER-ROSETTASERVER [Fig. 8(E)] correctly predicted the β-barrel, but placed a peripheral α-helix on the wrong side. The closest lipocalin-like nitrophorin 4 template [PDB ID: like]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[Fig. 8(F)] adopted a somewhat elongated β-barrel compared to the template with a shorter α-helix placed on the same side as the target α-helix.

Progress and pitfalls

The exceptional performance of several FM prediction models in CASP11 suggested an overall improvement of prediction methodology in comparison to CASP10. To fairly compare the performances of the two CASPs, we first needed to ensure the overall difficulty level of CASP10 FM targets (domains) was similar to the overall difficulty level of CASP11 FM targets (domains). To evaluate the difficulty level of targets over time, we took advantage of prediction models produced by a server whose methodology has not changed since CASP8 (SAM-T08). The GDT-TS distribution of SAM-T08 models for CASP10 target domains overlaps with its distribution for CASP11 target domains [Fig. 9(A)]. The overall difficulty of FM target domains for each CASP, measured as the average of SAM-T08 GDT_TS over all FM targets, was comparable (average 19.98 GDT_TS for CASP10 vs. average 20.52 GDT_TS for CASP11). To compare the performance of FM prediction models, we normalized the top model performance by the SAM-T08 performance (GDT_TS ratio) for each FM target from both CASPs. The distribution of these normalized performance ratios skewed toward higher levels for CASP11 targets with respect to that of CASP10 targets [Fig. 9(B)]. For CASP11 FM targets, an average 2-fold enhanced performance of top models over SAM-T08 models was observed, with many skewing toward 2.5-fold. This average ratio was the same as the top-performing ratio from CASP10 (2-fold), which had an average performance enhancement of top models over SAM-T08 models of 1.6-fold. Taken together, these distributions highlight the relative outperformance of top models in CASP11 with respect to those of CASP10, suggesting a significant improvement in FM prediction methodologies over the past 2 years.

Despite the progress observed in CASP11 on many FM target domains, several pitfalls remain. Multidomain targets still provide a significant challenge to FM prediction methodologies, especially when the domains repeat, such as the concanavalin A-like duplication of target T0808-D2 or the immunoglobulin-related triplication of target T0814, which also contains swapped secondary structure elements and an alternate domain topology. The poor performance of servers on the duplicated T0808-D2 (average GDT_TS 9.8), despite the TBM classification of the T0808-D1, landed this target as the second most difficult among FM domains. Compared to the conserved concanavalin A-like core topology, the T0808-D2 domain contained numerous elaborations to the core fold (over half of the sequence) that include the unusual insertion of a β-hairpin into the center of each jelly-roll sandwich β-sheet as well as a C-terminal extension that extends one sheet by three β-strands and the other by 2. The duplicated nature of such folds might have allowed rapid evolution of one or more of the domains. Potentially, such rapid evolution could evade the knowledge-based potentials of existing structures on which so many of the FM methodologies rely. Knowledge-based potentials also fail to predict structures with atypical characteristics. The CASP11 phage tail target domains, which have extended secondary structure elements that only form as a trimer (i.e., six domains in target T0775 and three domains in target T0779), all exhibited overall poor performance. CASP11 FM targets included additional difficult target domains classified as obligate multimers. Finally, complex topologies of large folds

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with numerous long-range contacts such as that found in the most difficult 345-residue single-domain target T0777-D1 (average server GDT_TS 9.6) remain challenging.

Acknowledgments

Support: National Institutes of Health GM094575

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
2. Lee J, Lee J, Sasaki TN, Sasai M, Seok C, Lee J. *De novo* protein structure prediction by dynamic fragment assembly and conformational space annealing. Proteins. 2011; 79:2403–2417. [PubMed: 21604307]
3. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004; 383:66–93. [PubMed: 15063647]
4. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997; 268:209–225. [PubMed: 9149153]
5. Xu D, Zhang Y. *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins. 2012; 80:1715–1735. [PubMed: 22411565]
6. Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L, Skolnick J. Analysis of TASSER-based CASP7 protein structure prediction results. Proteins. 2007; 69(Suppl 8):90–97. [PubMed: 17705276]
7. 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. [PubMed: 21997521]
8. Tai CH, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. Proteins. 2014; 82(Suppl 2):57–83.
9. 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. [PubMed: 21997778]
10. Taylor TJ, 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. [PubMed: 24123179]
11. 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. [PubMed: 24038551]
12. 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. [PubMed: 20157476]
13. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003; 31:3370–3374. [PubMed: 12824330]
14. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. Proteins. 2003; 53(Suppl 6):395–409. [PubMed: 14579328]
15. 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; 27:3371–3378. [PubMed: 21994223]
16. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013; 29:2722–2728. [PubMed: 23986568]
17. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

crystallography. Acta Crystallogr Sect D Biol Crystallogr. 2010; 66(Pt 1):12–21. [PubMed: 20057044]

18. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins. 2009; 77(Suppl 9):18–28. [PubMed: 19731382]
19. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins. 2003; 53(Suppl 6):352–368. [PubMed: 14579324]
20. Taylor TJ, Bai H, Tai CH, Lee B. Assessment of CASP10 contact-assisted predictions. Proteins. 2014; 82(Suppl 2):84–97. [PubMed: 23873510]
21. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics. 2001; 17:700–712. [PubMed: 11524371]
22. Karplus K. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res. 2009; 37:W492–W497. Web Server issue. [PubMed: 19483096]
23. 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. [PubMed: 22002823]
24. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014; 10:e1003926. [PubMed: 25474468]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

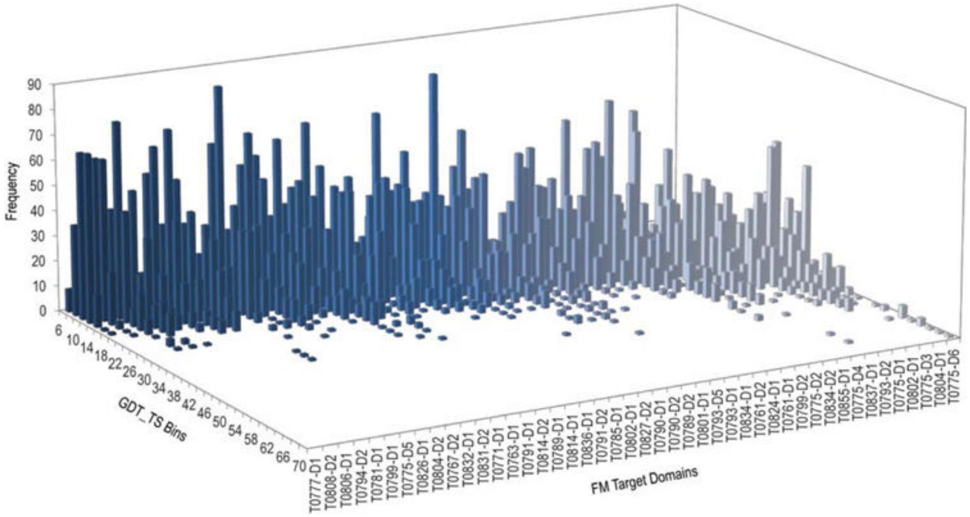

Figure 1.

Overall performance on FM targets. A three-dimensional graph depicting server model GDT_TS score distributions (first two coordinates) for each FM target domain plotted in the third coordinate. Targets are labeled, ordered by the average server GDT_TS, and colored in bluescale from more difficult (dark blue) to less difficult (light blue).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

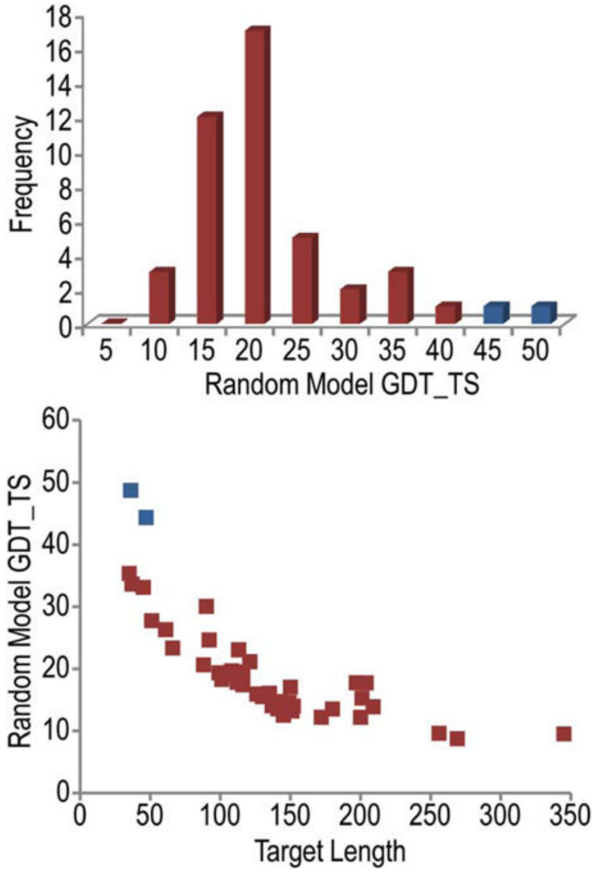

Figure 2.

Random model scores. (**a**) A histogram of random model GDT_TS scores (red bars) skews to the left where outlier targets (blue bars) with noncompact folds have unusually high GDT_TS scores. (**b**) A scatter plot of random model GDT_TS scores for each FM target domain (*y* axis) and their corresponding target lengths (*x* axis) illustrates the dependence of random model scores on target length. Outlier sequences from panel A are in blue.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

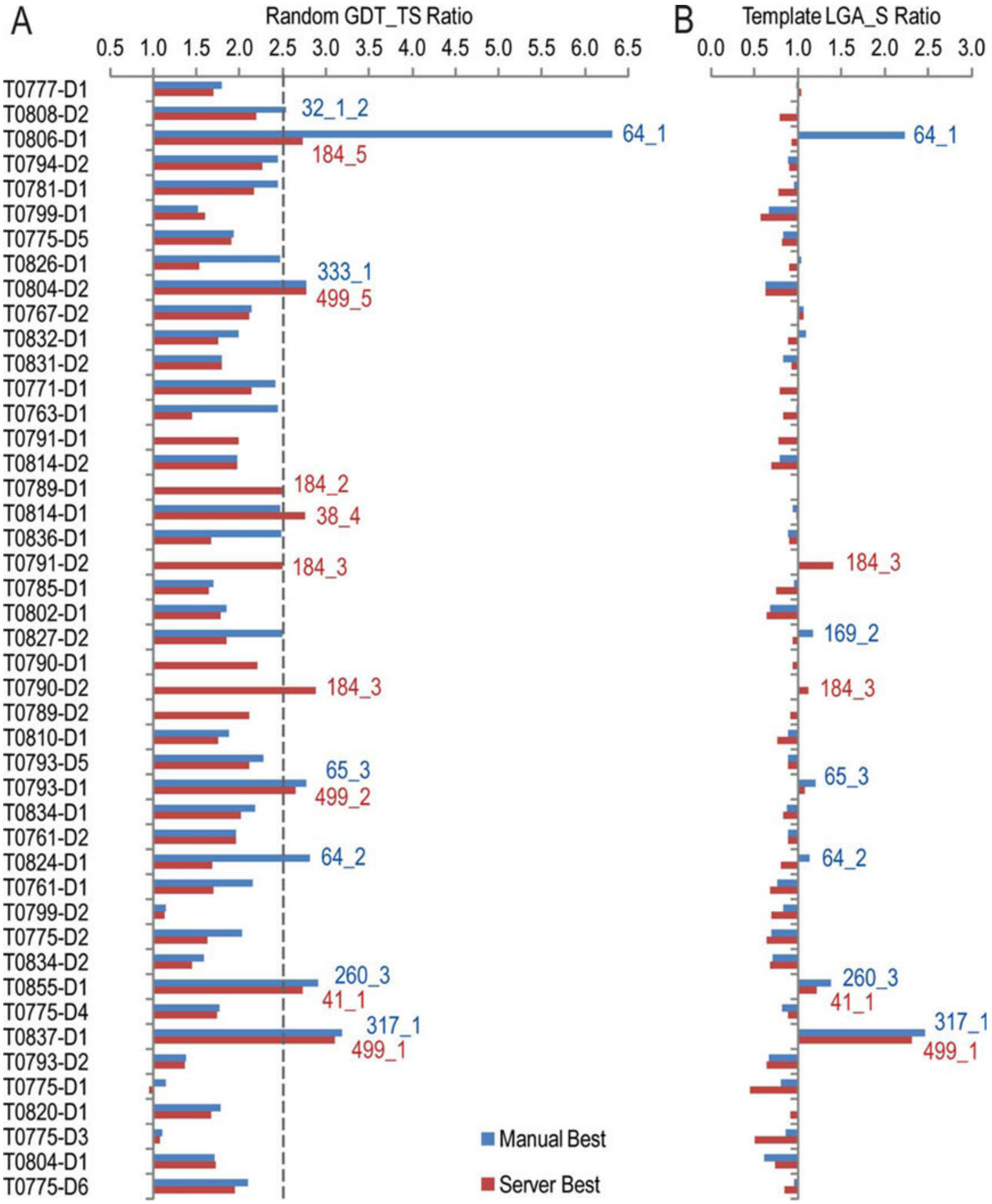

Figure 3.

Top prediction model highlights. Bar graphs illustrate top manual models (blue bars) and server models (red bars) for all FM templates ordered according to difficulty from top (low average GDT_TS for best server models) to bottom (high average GDT_TS of best server models). (**a**) A random model ratio compares the best prediction model GDT_TS to the random model average GDT_TS, with the *Y* axis marking the equivalence ratio and an arbitrary dashed line marking 2.5-fold improvement. Group models outperforming the 2.5-fold ratio are labeled (group number_model number_*doman*). *Domains* are only indicated

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where groups split them. (**b**) A template ratio compares the top prediction model LGA_S to the top template LGA_S for all FM targets (labeled below). Group models with LGA_S scores that beat the top template LGA_S score by at least 1.1- fold are labeled.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

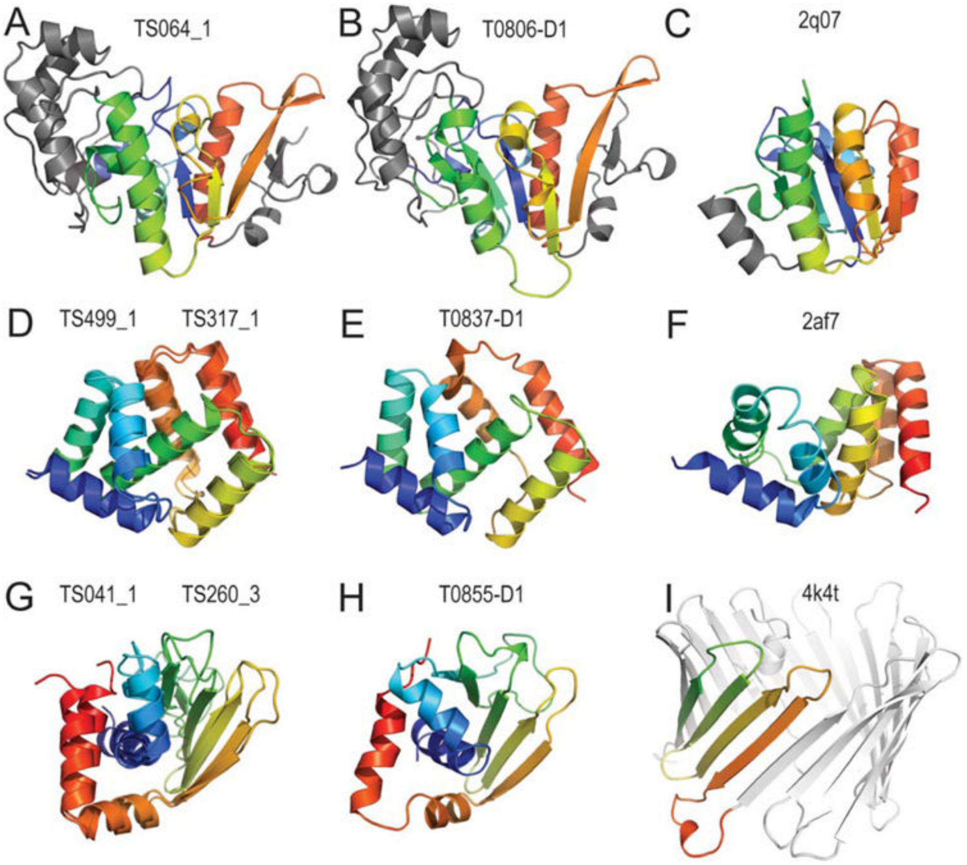

Figure 4.

Top prediction model examples. (**a**) Top random ratio manual model TS064_1 compared to (**b)** the target T0806-D1 structure shows the correct prediction of the entire fold. The model also outperforms (**c)** the top template of uncharacterized protein AF0587 [PDB ID:2q07], which retains the core three-layer Rossmann-like topology but lacks the 3-helix insertion as well as an additional C-terminal β-strand/α-helix. (**d**) Top server prediction TS499_1 superimposed with the top manual prediction TS317_1 compared to (**e)** the target T0837-D1 structure shows the correct prediction of the entire fold. The model also outperforms (**f)** the top template [PDB ID:2af7], which has roughly the same topology but with differences in interactions of the α-helices. (**g)** The top manual model TS317_1 for superimposed with the top server model TS041_1 capture the entire fold of **h)** the target T0855-D1 structure and improve over (**i)** the top unrelated template [2k4t], which retains only the β-meander of the target fold.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

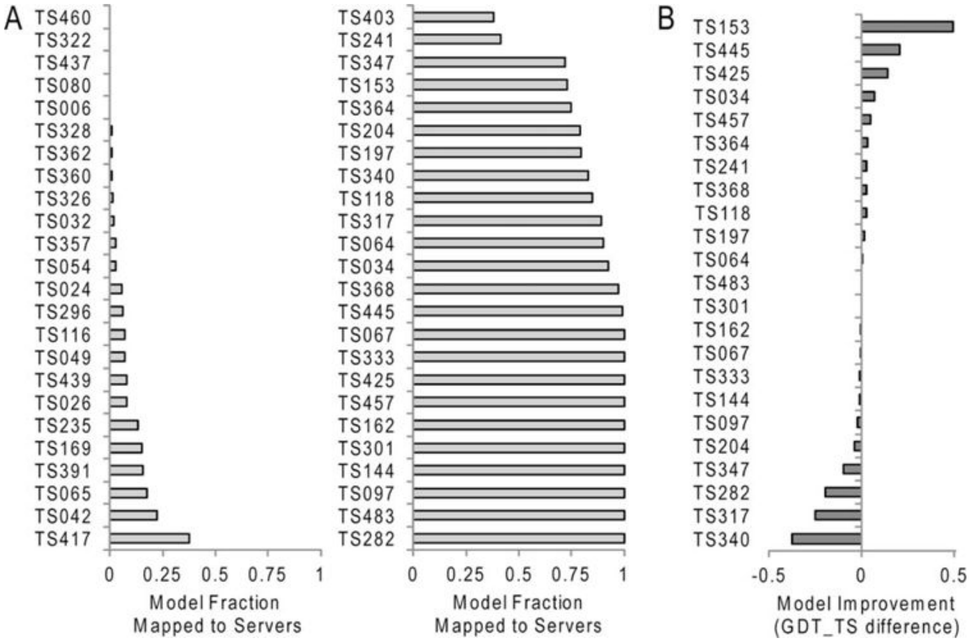

Figure 5.

Prediction methodology insights: selection and refinement of server models. (**a**) Bar graphs in left and center panels map fraction of prediction models for each manual group that cluster with any server model above GDT_TS 70. (**b**) Bar graph in right panel illustrates average GDT_TS improvement of manual models with respect to the closest mapped server models.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

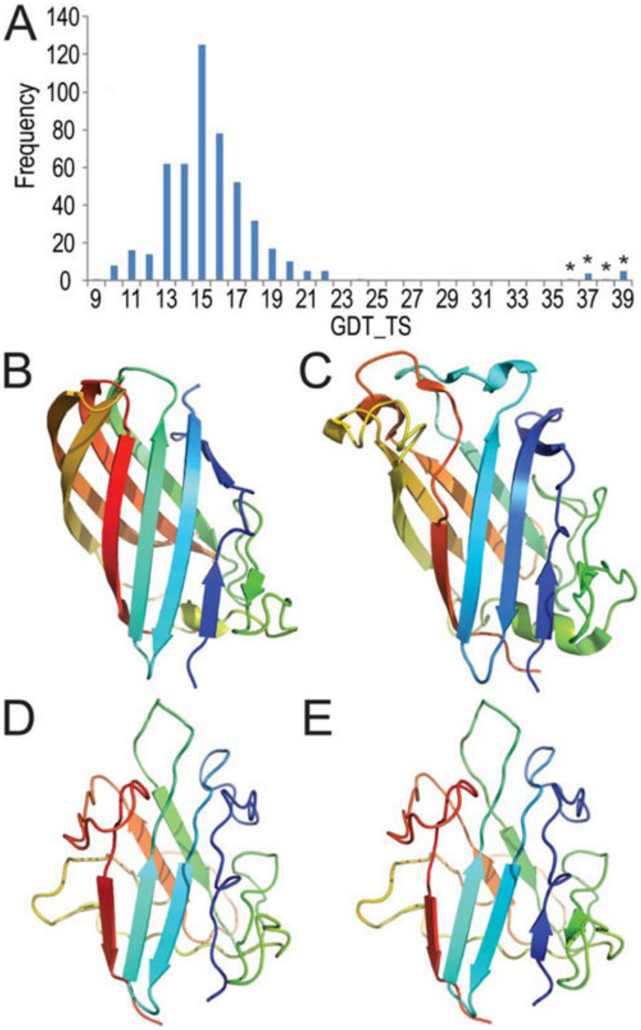

Figure 6.

First model performance. (**a**) The score distribution of target T0804-D2 highlights a cluster of outlier models (marked by *) that outperform the rest according to GDT_TS. **(b)** The target structure T0804-D2 adopts the same fold as (**c**) the top template [PDB ID:2j1k_f] with an LGA_S of 79. (**d)** The top server template TS499_5 (GDT_TS 38.65) roughly captures the topology, with a shift in alignment of the C-terminus and a failure to adopt the correct structure of 4 β-strands. (**e)** The top manual template TS333_1 (GDT_TS 38.82) slightly improves the top server model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

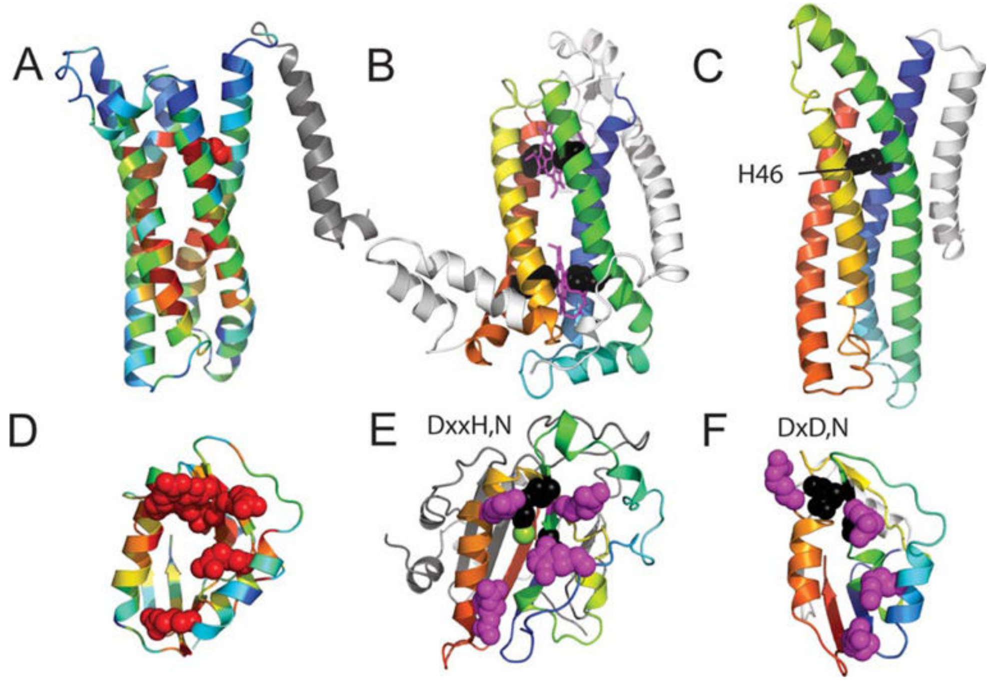

Figure 7.

Models useful for function prediction. (**a**) Residue conservations depicted in rainbow from blue (variable) to red (conserved) are mapped to the four-helical TMH bundle of the T0836-D1 heme-binding protein of unknown function. Conserved residues highlight the potential active site (red spheres) of the target structure, which adopts the same core fold as (**b)** the top template [PDB ID: 2fyn] classified as a transmembrane heme-binding four helical bundle. The template bound heme (magenta stick) is coordinated by four His residues (black sphere). **(c)** The top prediction TS065_4 correctly places a conserved His residue (black spheres, numbered according to the CASP target) that probably contributes to heme binding of the target. (**d)** A potential active site (colored as above) is marked by conserved residues mapped to the T0824-D1 NucB DNase, which represents a deterioration of (**e)** the top template [PDB ID: 1g8t] classified as a His-Me finger endonuclease. Active site (black spheres, motif labeled) and nucleotide binding (magenta spheres) residues are highlighted. (**f)** The top prediction TS064_2 roughly places conserved active site (motif labeled) and nucleotide binding residues in the correct sites.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

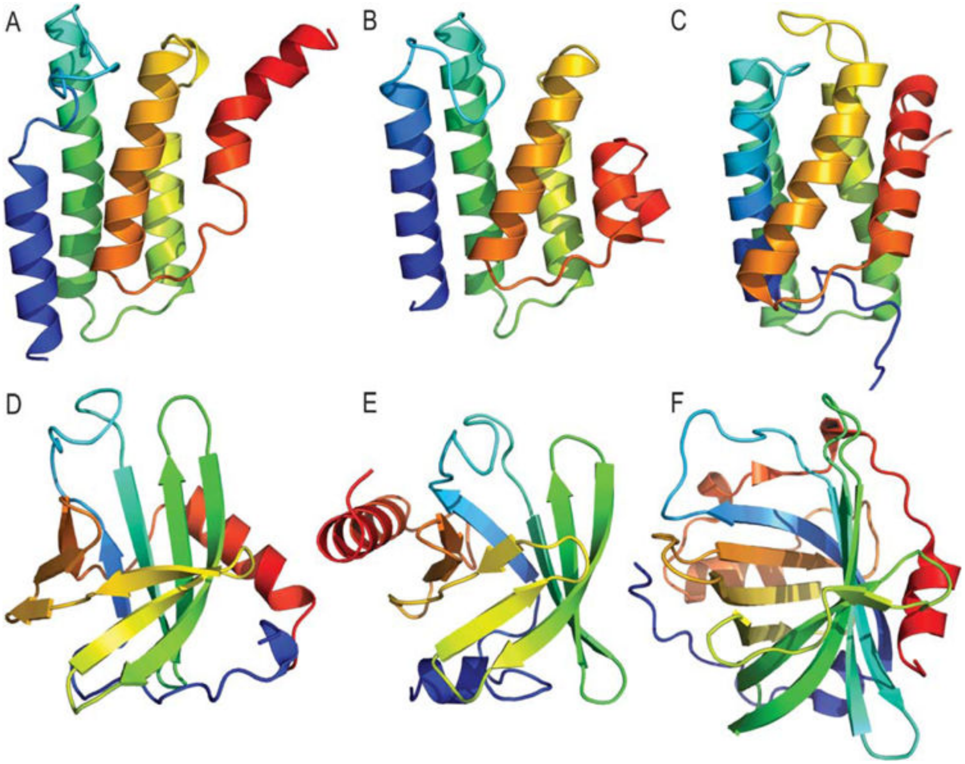

Figure 8.

CASP ROLL outperformance. (**a**) The CASP ROLL Target R0034-D1 adopts an up and down α-helical bundle containing five α-helices. (**b**) The top-performing prediction model (TS045_1) includes all five α-helices in the correct topology, with correct alignment over most to the structure (residues 40–110) and the last α-helix being broken. (**c**) The closest template [PDB ID: 2lpj] includes all 5 α-helices in the same topology. (**d**) The CASP ROLL Target R0021 adopts an eight-stranded β-meander barrel. (**e**) The top-performing model (TS330_4) correctly predicts the β-barrel, but places a peripheral α-helix on the wrong side of the barrel. (**f**) The closest template [PDB ID: 1ike] classified as lipocalin adopts a somewhat elongated β-barrel compared to the template.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

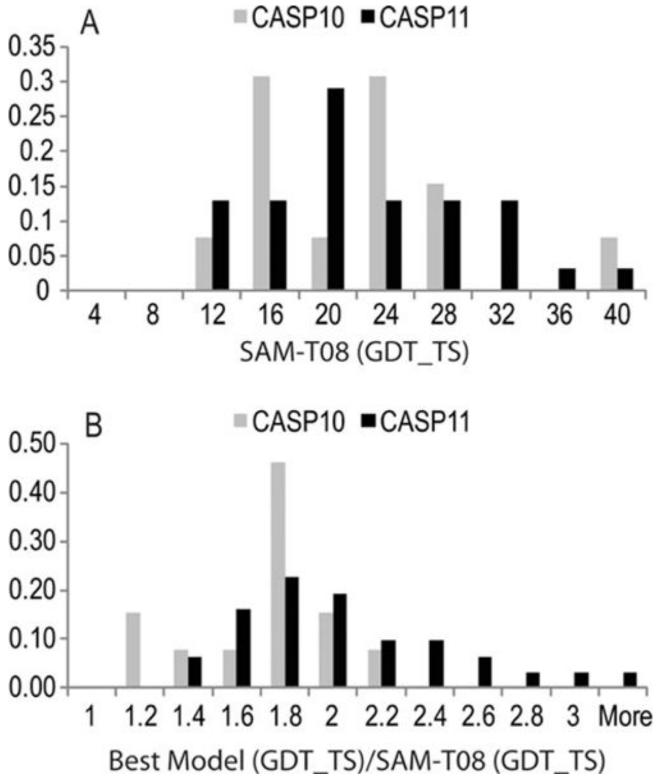

Figure 9.

Progress. (**a**) Distributions of SAM-T08 GDT_TS scores on FM targets from CASP10 (gray bars) and CASP11 (black bars) suggest similar target difficulties. (**b**) Distributions of normalized performance ratios (best model GDT_TS/SAM-T08 GDT_TS) for CASP11 (blackbars) skew toward higher performance than those for CASP10 (gray bars).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table I

FM Server Group Performance

			Best FM-style scoring				First FM-style scoring				Best win/loss	
Group code	Group name	Dom	SumZ	SumR	AvgZ	AvgR	SumZ	SumR	AvgZ	AvgR	WinF	P values
184	**BAKER-ROSETTASERVER**	45	**74.87**	**1**	**1.66**	**1**	27.01	4	0.60	4	0.960	0.0E+00
277	**Zhang-Server**	45	59.11	2	1.31	2	**43.80**	**1**	**0.97**	**1**	0.921	0.0E+00
499	**QUARK**	45	57.86	3	1.29	3	41.44	2	0.92	3	0.907	5.36e−311
41	MULTICOM-NOVEL	45	37.43	4	0.83	4	7.74	13	0.17	13	0.826	5.5E−188
38	nns	45	31.41	5	0.70	6	22.62	5	0.50	5	0.809	4.0E−167
216	myprotein-me	45	29.97	6	0.67	7	17.28	8	0.38	8	0.774	1.9E−129
479	**RBO_Aleph**	40	18.88	7	0.72	5	28.41	3	0.96	2	0.748	3.2E−94
8	MULTICOM-CONSTRUCT	45	15.23	8	0.34	8	15.39	9	0.34	9	0.690	6.7E−62
345	FUSION	45	13.68	9	0.30	9	12.16	12	0.27	12	0.695	4.4E−65
420	MULTICOM-CLUSTER	45	11.00	10	0.24	10	18.51	7	0.41	7	0.638	4.8E−33
268	MULTICOM-REFINE	45	7.67	11	0.17	11	14.66	10	0.33	10	0.643	2.6E−35
410	Pcons-net	45	7.26	12	0.16	12	0.31	17	0.01	19	0.634	4.3E−31
50	RaptorX	45	2.42	13	0.05	13	12.44	11	0.28	11	0.590	5.8E−15
228	BhageerathH	45	1.81	14	0.04	14	6.86	14	0.15	14	0.577	1.5E−11
11	Seok-server	45	1.05	15	0.02	15	20.73	6	0.46	6	0.567	5.7E−09
263	STRINGS	45	1.01	16	0.02	16	−4.51	23	−0.10	26	0.571	6.0E−10
160	ZHOU-SPARKS-X	45	0.94	17	0.02	17	−3.26	21	−0.07	24	0.569	1.9E−09
349	Distill	45	−0.90	18	−0.02	19	−4.23	22	−0.09	25	0.556	9.1E−07
436	Alpha-Gelly-Server	45	−1.57	19	−0.03	20	−15.45	27	−0.34	29	0.524	1.9E−02
251	TASSER-VMT	45	−5.41	20	−0.12	22	−9.53	25	−0.21	27	0.500	5.0E−01
210	BioSerf	42	−8.20	21	−0.05	21	−5.00	24	0.02	17	0.542	2.4E−04
454	eThread	45	−8.43	22	−0.19	24	−16.26	28	−0.36	30	0.472	9.9E−01
212	FFAS-3D	45	−9.03	23	−0.20	25	−2.75	20	−0.06	23	0.446	1.0E+00
452	FALCON_EnvFold	45	−10.03	24	−0.22	26	−0.68	18	−0.02	21	0.459	1.0E+00
381	FALCON_MANUAL	45	−10.45	25	−0.23	27	1.30	15	0.03	16	0.448	1.0E+00
335	FALCON_TOPO	45	−10.71	26	−0.24	28	0.98	16	0.02	18	0.445	1.0E+00
414	FALCON_MANUAL_X	45	−10.80	27	−0.24	29	−0.83	19	−0.02	22	0.442	1.0E+00

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

			Best FM-style scoring				First FM-style scoring				Best win/loss	
Group code	Group name	Dom	SumZ	SumR	AvgZ	AvgR	SumZ	SumR	AvgZ	AvgR	WinF	P values
466	RaptorX-FM	38	−13.68	28	0.01	18	−11.29	26	0.07	15	0.583	3.3E−11
300	PhyreX	43	−17.43	29	−0.39	31	−17.47	29	−0.31	28	0.387	1.0E+00
145	BioShell-server	44	−17.63	30	−0.36	30	−20.64	32	−0.42	31	0.399	1.0E+00
117	raghavagps-tsppred	45	−21.94	31	−0.49	33	−36.52	38	−0.81	39	0.331	1.0E+00
492	slbio	45	−25.44	32	−0.57	34	−25.43	33	−0.57	34	0.337	1.0E+00
110	MATRIX	42	−25.72	33	− 0.47	32	−64.28	43	−0.83	40	0.348	1.0E+00
448	FLOUDAS_SERVER	45	−30.57	34	−0.68	35	−19.08	30	−0.42	32	0.249	1.0E+00
73	SAM-T08-server	31	−30.83	35	−0.15	23	−28.17	36	−0.01	20	0.516	1.2E−01
133	IntFOLD3	45	−34.07	36	−0.76	37	−20.29	31	−0.45	33	0.232	1.0E+00
156	Atome2_CBS	41	−36.95	37	−0.71	36	−35.28	37	−0.67	37	0.261	1.0E+00
237	chuo-fams-server	42	−43.10	38	−0.88	38	−43.27	40	−0.89	42	0.187	1.0E+00
171	MUFOLD-Server	44	−43.23	39	−0.94	39	−36.67	39	−0.79	38	0.176	1.0E+00
346	HHPredA	45	−44.17	40	−0.98	40	−27.30	35	−0.61	36	0.151	1.0E+00
279	HHPredX	45	−45.13	41	−1.00	42	−27.23	34	−0.61	35	0.160	1.0E+00
22	3D-Jigsaw-V5_1	36	−53.76	42	−0.99	41	−49.74	41	−0.88	41	0.157	1.0E+00
206	PSF	24	−71.02	43	−1.21	43	−63.49	42	−0.90	43	0.095	1.0E+00
193	FFAS03	27	−74.65	44	−1.45	44	−69.75	44	− 1.25	44	0.044	1.0E+00

Highlighted values are indistinguishable by significance tests, bold values denote top performers.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table II

FM Manual Group Performance

			Best FM-style scoring				First FM-style scoring				Best win/loss	
Code	**Group**	**Dom**	**SumZ**	**SumR**	**AvgZ**	**AvgR**	**SumZ**	**SumR**	**AvgZ**	**AvgR**	**WinF**	***P* values**
64	**BAKER**	39	**49.99**	**1**	**1.28**	**1**	27.15	9	0.70	13	0.868	0.0E+00
333	**Kiharalab**	39	44.54	2	1.14	2	**44.20**	**1**	**1.22**	**1**	0.900	0.0E+00
44	**LEER**	39	40.18	3	1.03	4	36.70	2	0.94	2	0.872	0.0E+00
132	ProQ2-refine	39	40.00	4	1.03	5	31.58	5	0.81	8	0.874	0.0E+00
338	ProQ2	39	36.67	5	0.94	6	33.88	3	0.87	3	0.850	0.0E+00
290	MULTICOM	39	35.42	6	0.91	7	21.56	13	0.55	22	0.839	0.0E+00
358	Skwark	39	34.79	7	0.89	8	24.75	10	0.63	14	0.799	0.0E+00
425	Seok-refine	39	34.33	8	0.88	10	33.28	4	0.85	4	0.840	0.0E+00
67	CNIO	39	33.40	9	0.86	12	15.66	18	0.40	26	0.802	0.0E+00
347	Wallner	39	30.01	10	0.77	13	31.57	6	0.81	9	0.807	0.0E+00
204	Zhang	39	29.52	11	0.76	14	29.28	8	0.75	12	0.799	0.0E+00
282	PML	39	29.43	12	0.75	16	19.15	15	0.56	21	0.766	1.5E−254
169	LEE	39	26.82	13	0.69	20	23.43	12	0.60	17	0.762	6.5E−248
328	RosEda	38	26.63	14	0.75	17	29.32	7	0.82	6	0.750	1.3E−218
310	MUFOLD-R	37	23.58	15	0.75	18	24.62	11	0.77	11	0.772	3.0E−253
368	Seder1	37	22.06	16	0.70	19	4.00	31	0.22	30	0.749	1.4E−211
301	Boniecki_pred	32	21.70	17	1.12	3	11.36	24	0.79	10	0.869	0.0E+00
445	Seder2	37	20.54	18	0.66	21	−18.73	48	−0.02	42	0.760	1.3E−231
483	LmtdSeder	37	20.01	19	0.65	22	−16.46	45	−0.01	41	0.750	5.0E−213
65	Jones-UCL	38	19.85	20	0.58	26	14.21	20	0.43	25	0.689	3.3E−123
317	Keasar	39	18.52	21	0.47	29	6.81	29	0.17	35	0.689	1.2E−126
197	wfMix-KFb	36	17.34	22	0.65	23	−15.69	44	0.08	37	0.750	2.1E−207
118	wfMix-KFa	36	15.82	23	0.61	25	2.78	32	0.24	29	0.735	1.2E−181
439	Pareto	39	15.06	24	0.39	31	5.92	30	0.15	36	0.670	4.0E−102
360	Gong3701	31	10.89	25	0.87	11	10.40	25	0.85	5	0.764	2.9E−200
54	NEFILIM	32	10.20	26	0.76	15	12.35	22	0.82	7	0.776	4.5E−226
97	Rluethy	30	8.69	27	0.89	9	−18.02	47	0.00	40	0.810	5.0E−274

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

			Best FM-style scoring				First FM-style scoring				Best win/loss	
Code	**Group**	**Dom**	**SumZ**	**SumR**	**AvgZ**	**AvgR**	**SumZ**	**SumR**	**AvgZ**	**AvgR**	**WinF**	***P* values**
296	Seok	39	6.36	28	0.16	34	6.92	28	0.18	34	0.585	9.8E−27
340	Handl	33	5.60	29	0.53	27	−2.38	35	0.29	28	0.682	3.4E−100
162	McGuffin	39	4.40	30	0.11	37	8.23	27	0.21	31	0.562	3.8E−15
241	SHORTLE	37	2.70	31	0.18	33	17.59	16	0.58	19	0.552	1.9E−10
438	QA-RecombineIt H	38	1.55	32	0.09	40	20.44	14	0.59	18	0.546	5.7E−09
116	2PG	33	0.83	33	0.39	30	−14.84	42	−0.09	43	0.683	1.3E−100
144	Mufold	39	0.19	34	0.00	42	11.55	23	0.30	27	0.510	1.2E−01
80	MeilerLab	36	−0.42	35	0.16	35	−5.34	37	0.08	38	0.590	1.2E−27
457	wfKeasar-PTIGRESS	29	−1.51	36	0.64	24	−17.01	46	0.18	33	0.758	1.1E−178
434	QA-RecombineIt_H2	36	−2.01	37	0.11	38	15.87	17	0.61	15	0.556	1.1E−11
364	QA-RecombineIt_WFH	36	−2.17	38	0.11	39	15.63	19	0.60	16	0.552	3.0E−10
153	wfAll-MD-RFLB	34	−2.63	39	0.22	32	−15.54	43	−0.16	45	0.616	9.3E−43
42	TASSER	39	−2.88	40	−0.07	44	0.08	33	0.00	39	0.475	1.0E+00
362	BioShell	30	−3.04	41	0.50	28	−1.02	34	0.57	20	0.698	4.5E−108
482	wfMix-KPa	36	−3.68	42	0.06	41	13.89	21	0.55	23	0.541	3.6E−07
276	FLOUDAS_A4	39	−4.03	43	−0.10	45	−4.78	36	−0.12	44	0.491	8.8E−01
56	wfMix-KPb	36	−7.51	44	−0.04	43	9.95	26	0.44	24	0.512	7.2E−02
235	FLOUDAS_A3	39	−9.46	45	−0.24	47	−12.92	40	−0.19	46	0.424	1.0E+00
391	Chicken_George	37	−10.37	46	−0.17	46	−14.28	41	−0.23	47	0.465	1.0E+00
326	FLOUDAS_A2	39	−10.72	47	−0.27	48	−10.91	38	−0.28	48	0.416	1.0E+00
157	FLOUDAS_A1	39	−12.97	48	−0.33	49	−24.05	53	−0.37	50	0.382	1.0E+00
32	Legato	39	−19.44	49	−0.50	51	−21.49	51	−0.55	55	0.322	1.0E+00
442	WfCPUNK	39	−20.18	50	−0.52	52	−21.30	50	−0.55	54	0.319	1.0E+00
155	Cornell-Gdansk	39	−21.03	51	−0.54	53	−21.05	49	−0.54	53	0.307	1.0E+00
322	Bilab	33	−26.28	52	−0.43	50	−25.70	55	−0.42	51	0.349	1.0E+00
26	Bates_BMM	39	−26.96	53	−0.69	54	−11.80	39	−0.30	49	0.273	1.0E+00
63	KIAS-GDANSK	39	−28.90	54	−0.74	55	−26.98	56	−0.69	57	0.242	1.0E+00
403	wfAll-Cheng	20	−35.40	55	0.13	36	−33.95	59	0.20	32	0.583	3.6E−14
49	DELCLAB	39	−36.14	56	−0.93	58	−32.37	58	−0.83	61	0.186	1.0E+00
133	IntFOLD3	39	−36.72	57	−0.94	59	−25.16	54	−0.65	56	0.183	1.0E+00

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Code	Group	Dom	Best FM-style scoring				First FM-style scoring				Best win/loss	
			SumZ	**SumR**	**AvgZ**	**AvgR**	**SumZ**	**SumR**	**AvgZ**	**AvgR**	**WinF**	***P* values**
34	WfHHPred-PTIGRESS	37	−37.35	58	−0.90	57	−22.61	52	−0.50	52	0.198	1.0E+00
73	SAM-08-server	26	−38.36	*	−0.48	*	−34.63	*	−0.27	*	0.356	1.0E+00
156	Atome2_CBS	35	−38.46	59	−0.87	56	−36.93	60	−0.83	60	0.211	1.0E+00
357	STAP	39	−42.75	60	−1.10	61	−32.10	57	−0.82	59	0.126	1.0E+00
237	chuo-fams-server	36	−43.12	61	−1.03	60	−40.99	61	−0.97	64	0.156	1.0E+00
417	Chuo-fams	37	−51.69	62	−1.29	65	−45.17	63	−1.09	65	0.090	1.0E+00
437	ALAdeGAP	33	−52.22	63	−1.22	64	−41.23	62	−0.89	63	0.091	1.0E+00
460	Victoria	37	−55.39	64	−1.39	66	−45.73	65	−1.13	66	0.059	1.0E+00
430	WY-C	28	−55.42	65	−1.19	63	−45.46	64	−0.84	62	0.123	1.0E+00
24	Dppred	17	−63.46	66	−1.14	62	−57.91	66	−0.82	58	0.145	1.0E+00
6	Sun_Tsinghua	24	−65.95	67	−1.50	67	−64.35	67	−1.43	67	0.050	1.0E+00

Highlighted values are indistinguishable by significance tests, bold values denote top performers.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table III

CASP ROLL Group Performance

			Best FM-style scoring				First FM-style scoring				Best win/loss	
Group code	Group name	Dom	SumZ	SumR	AvgZ	AvgR	SumZ	SumR	AvgZ	AvgR	WinF	*P* values
330	**BAKER-ROSETTASERVER**	33	26.78	1	1.11	1	7.60	1	0.53	3	0.893	2.6E−75
45	**Zhang_Ab_Initio**	33	17.73	2	0.84	2	0.35	2	0.31	8	0.813	1.8E−46
81	MULTICOM-CLUSTER	32	−10.18	3	0.06	14	−9.79	4	0.07	16	0.548	1.9E−02
222	MULTICOM-CONSTRUCT	32	−14.60	4	−0.08	16	−20.95	9	−0.28	22	0.476	8.4E−01
125	MULTICOM-REFINE	32	−15.78	5	−0.12	18	−7.76	3	0.13	13	0.458	9.7E−01
424	MULTICOM-NOVEL	32	−16.21	6	−0.13	19	−16.28	6	−0.13	20	0.452	9.8E−01
488	chunk-TASSER	38	−16.22	7	−0.43	22	−14.89	5	−0.39	24	0.342	1.0E+00
315	Keasar	23	−17.12	8	0.56	3	−19.13	7	0.47	4	0.800	3.2E−28
113	SAM-T08-server	32	−28.44	9	−0.51	23	−19.36	8	−0.23	21	0.308	1.0E+00
87	DistilLroN	24	−30.07	10	−0.09	17	−23.56	10	0.18	9	0.480	7.7E−01
381	SAM-T06-server	30	−34.72	11	−0.62	24	−26.74	11	−0.36	23	0.261	1.0E+00
114	QUARK	13	−43.07	12	0.53	5	−45.45	13	0.35	7	0.793	9.5E−22
435	Ossia	35	−43.80	13	−1.08	25	−36.26	12	−0.83	25	0.155	1.0E+00
35	Zhang-Server	13	−45.85	14	0.32	10	−47.72	17	0.18	12	0.705	3.3E−11
124	PconsD	14	−46.60	15	0.10	13	−46.15	14	0.13	14	0.614	1.9E−04
292	Pcons-net	14	−47.92	16	0.01	15	−49.25	21	−0.09	19	0.582	5.7E−03
68	FOLDIT	11	−47.95	17	0.55	4	−46.28	15	0.70	2	0.730	8.5E−12
85	Anthropic_Dreams	11	−48.15	18	0.53	6	−48.84	19	0.47	5	0.791	2.3E−18
165	Void_Crushers	11	−48.51	19	0.50	8	−49.11	20	0.44	6	0.749	1.3E−13
237	Zhang	10	−50.74	20	0.53	7	−48.53	18	0.75	1	0.789	1.1E−16
344	Jones-UCL	10	−51.97	21	0.40	9	−54.16	22	0.18	10	0.671	1.3E−05
438	FALCON-Server	14	−52.91	22	−0.35	20	−47.11	16	0.06	17	0.403	1.0E+00
413	ZHOU-SPARKS-X	10	−54.63	23	0.14	12	−56.63	24	−0.06	18	0.634	7.6E−04
341	Contenders	8	−57.62	24	0.30	11	−58.58	25	0.18	11	0.707	5.6E−08
22	FALCONX	10	−60.23	25	−0.42	21	−55.27	23	0.07	15	0.339	1.0E+00
88	panther	17	−65.50	26	−1.38	26	−63.00	26	−1.19	26	0.073	1.0E+00

Highlighted values are indistinguishable by significance tests, bold values denote top performers.