Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods

Article

Accepted Version

Research Article
Proteins: Structure, Function and Bioinformatics
DOI 10.1002/prot.25360

Title: Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods

Short title: Template Based Modelling (McGuffin Group)

Liam J. McGuffin[1]*, Ahmad N. Shuid[1], Robert Kempster[1], Ali H.A. Maghrabi[1], John O. Nealon[1], Bajuna R. Salehe[1], Jennifer D. Atkins[1] and Daniel B. Roche[2,3]

1 - School of Biological Sciences, University of Reading, Reading, UK

2 - Institut de Biologie Computationnelle, LIRMM, CNRS-UMR 5506, Université de Montpellier, Montpellier, France

3 - Centre de Recherche en Biologie cellulaire de Montpellier, CNRS-UMR 5237, Montpellier, France

* To whom correspondence should be addressed

Liam J. McGuffin

School of Biological Sciences,

University of Reading,

Reading RG6 6AS,

UK

Email: l.j.mcguffin@reading.ac.uk

Tel: +44 (0) 118 378 6332

Fax: +44 (0)118 378 8106

Keywords: Protein Structure Prediction, 3D Modelling, Model Quality Assessment, Refinement, Estimates of Model Accuracy, Accuracy Self Estimate, QA, EMA, ASE

Received: May 31, 2017; Revised: Jul 12, 2017; Accepted: Jul 25, 2017

ABSTRACT

Our aim in CASP12 was to improve our Template Based Modelling (TBM) methods through better model selection, accuracy self-estimate (ASE) scores and refinement. To meet these aims we developed two new automated methods, which we used to score, rank and improve upon the provided server models. Firstly, the ModFOLD6_rank method, for improved global Quality Assessment (QA), model ranking and the detection of local errors. Secondly, the ReFOLD method for fixing errors through iterative QA guided refinement. For our automated predictions we developed the IntFOLD4-TS protocol, which integrates the ModFOLD6_rank method for scoring the multiple-template models that were generated using a number of alternative sequence-structure alignments. Overall, our selection of top models and ASE scores using ModFOLD6_rank was an improvement on our previous approaches. In addition, it was worthwhile attempting to repair the detected errors in the top selected models using ReFOLD, which gave us an overall gain in performance. According to the assessors' formula, the IntFOLD4 server ranked 3rd/5th (average Z-score > 0.0/-2.0) on the server only targets, and our manual predictions (McGuffin group) ranked 1st/2nd (average Z-score > -2.0/0.0) compared to all other groups.

INTRODUCTION

A principal research focus of our group has been the improvement of methods for Template Based Modelling (TBM) of proteins from their sequences, through accurate Model Quality Assessment (MQA). Improvements in the Estimates of Model Accuracy (EMA) lead to higher quality and more useful 3D models overall, and so the advancement of reliable MQA has always been at the core of our TBM strategy. The development of our IntFOLD tertiary structure (TS) prediction methods [1-3] stemmed from the development of the GenTHREADER[4] and nFOLD[5] methods, which integrated several Model Quality Assessment Programs (MQAPs) to improve the recognition of fold templates, initially for the CASP6 experiment. In parallel with nFOLD version 3 at CASP7, we developed the first version of the ModFOLD method, which focused purely on the quality assessment problem[6]. The ModFOLD server method was further improved through integration with clustering-based variant (ModFOLDclust)[7], which we subsequently used in CASP8, both for our predictions in the QA category[8] and to rank server models for our manual predictions in the TS category.

In the CASP9 experiment, assessors began to request that predictors include error estimates in Ångströms, in place of the temperature factor (B-factor) field, for each of their submitted 3D models. Since then, the CASP assessors have increasingly placed an emphasis on the importance of "B-factor" predictions, or Accuracy Self Estimate (ASE) scores as they are now termed. The analogy is that relying on a 3D protein model without quality assessment is like trusting the top hit of a BLAST search without knowing the E-value.

The first version of the IntFOLD server[9] integrated single template modelling with ModFOLDclust2 for ranking and ASE scores, which were included in the "B-factor" column of the model output files. In CASP9, the IntFOLD-TS method[1] gained attention for its high

performance in the assignment of model reliability/"B-factors", which was first evaluated in the TBM category[10].

The original IntFOLD server only used a single-template modelling approach and so with IntFOLD2-TS, for CASP10, our aim was to exploit our strengths in ASE scoring and use the scores to guide our multiple-template modelling protocols[2]. For CASP11, with our IntFOLD3-TS method[3] we added some extra sequence-structure alignment methods, but we used the same multiple-template modelling ranking and ASE scoring protocols that were used by IntFOLD2.

For our manual TS predictions in CASP8-CASP11 we used ModFOLD variants for ranking server models and adding ASE scores to our submitted models. We had some success with high cumulative GDT_TS rankings, but historically our main strength in the TS prediction category has been with our ASE scores. Until CASP12, we had always been able to accurately identify errors in the server models but we had not developed any reliable methods for fixing them.

In this paper, we describe our latest improvements to the IntFOLD-TS method (Version 4), which integrates the new ModFOLD6[11] variant, ModFOLD6_rank, for improved selection and ASE scoring. Additionally, in CASP12 for the first time we have made serious efforts at model refinement guided by ASE scores using our new ReFOLD[12] method.

METHODS

The IntFOLD4-TS prediction server

For CASP12, a bespoke version of the IntFOLD4 server was developed in order to return appropriately formatted results for the tertiary structure (TS) prediction category. Additionally, the local quality assessment predictions (ASE scores) were returned as

predicted distances in the B-factor column of each TS model file using our ModFOLD6[11] QA server (N.B. predictions in the EMA/QA category were also returned by our ModFOLD6 and ModFOLDclust2 servers, see our EMA methods paper in this issue for more details).

Our IntFOLD4-TS method was developed with the aim of identifying, and then attempting to fix, the local errors in an initial pool of single template models via iterative multi-template modeling. The method attempts to exploit our previous CASP successes in accurately predicting local errors in our models[1] by taking the global and local per-residue errors into consideration during the multiple template selection stage[2].The pipeline can be broken down into two major stages: i) single template modelling with ASE scoring and ii) QA guided multiple template modelling with ASE scoring (Figure 1).

For the single template modelling first major stage, 14 different fold recognition methods were installed and run in-house, generating up to 10 sequence-to-structure alignments each; this resulted in up to 140 alternative single-template based models being generated for each CASP target. The following fold recognition methods were used: SP3[13], SPARKS2[13], HHsearch[14], COMA[15], SPARKSX[16], CNFsearch[17] and the 8 alternative threading methods that are integrated into the current LOMETS package[18] (PPA, dPPA, dPPA2, sPPA, MUSTER, wPPA, wdPPA and wMUSTER). At the end of the first stage, all single-template models were assessed using ModFOLDclust2[19] in order to assign global and local model quality scores.

In the second major stage, the single-template model quality scores, and other criteria involving template coverage, were used to select sequence-structure alignments for building multiple-template models[2]. The overall aim was to select appropriate target-template alignments that would minimise local errors in the final models. The Multiple Template Modelling (MTM) stage included 4 main alternative alignment selection methods (multi1-multi4) for 3D model building. The first method, multi1, simply used the top 2 alignments

according to the template ranking. The multi2 method used the top ranked alignment and any subsequent alignments if there were >= 40 new residues covered and <= 20 residues were overlapping. The multi3 method used the top ranked alignment and any subsequent alignments, but only if the overlapping region was predicted to increase local model quality. Finally, the multi4 method used the top ranked alignment and any subsequent alignments, but only if the coverage was increased by at least 1 residue. Four additional variants on these methods (multi5-multi8) repeated multi1-multi4, respectively, however, the alignments for each of the single-template methods were firstly re-ranked based on the ModFOLDclust2 predicted global model quality scores. These MTM approaches were first introduced in our IntFOLD2-TS method and they are fully described and benchmarked in our paper published in 2012[2].

The alternative MTM alignment selection methods resulted in the generation of a new population of up to 124 multi-template models for each target. Additionally, I-TASSER_LIGHT [20] (I-TASSER4.4 run in "light mode" with wall-time restricted to 5h; for sequences <600 residues) and HHpred[21] were used to generate 3 models each, which were then added into the final pool of alternative multi-template models for ranking. In the final stage of the method, the ~130 models in the final reference set were then evaluated using our ModFOLD6_rank[11] QA method and the top 5 ranked models were submitted as the final IntFOLD4-TS predictions (see also our EMA paper in this issue for more details about our ModFOLD6_rank method).

The McGuffin group TS predictions

Our manual TS prediction protocol is shown in Figure 2. Initially, for each target, the server models were ranked according to their ModFOLD6_rank global quality scores. The top ranked initial model was then selected and submitted to the ReFOLD pipeline, which has

now been implemented as a server[12]. The ReFOLD method consisted of three protocols. The first protocol used a rapid iterative strategy (i3Drefine[23]) and the second protocol employed a more CPU/GPU intensive molecular dynamic simulation strategy (using NAMD[24]) to refine each starting model. Refined models generated from each protocol were assessed and ranked using ModFOLD6_rank. The third protocol was a combination of the first 2 approaches, where the top ranked model from the 2nd protocol was further refined using i3Drefine. Finally, all of the refined models generated by each of these protocols and the starting model were pooled and re-ranked again using ModFOLD6_rank and the final top 5 models were selected and submitted. For each model, the ModFOLD6 predicted per-residue error scores were added into the B-factor column for each set of atom records

In addition to our independent submissions, we were also the submitters for the collaborative WeFOLD group, wfRosetta-ProQ-ModF6, which also made use of ModFOLD6_rank for final model ranking and ASE scoring (paper in preparation).

Benchmarking methods prior to CASP12

Prior to CASP12, we benchmarked our ModFOLD6 methods against our existing methods to gauge their ability to rank CASP11 server models and their ASE scoring capability. We then built the ModFOLD6_rank method into the IntFOLD4 server and continuously benchmarked both the IntFOLD4-TS and ModFOLD6 servers using the independent CAMEO resource[25].

RESULTS

Benchmarking prior to CASP12

Prior to CASP, our main priority was to benchmark our new methods to confirm that they were working as intended for model selection, and importantly, outperforming our older

server TS methods (IntFOLD2-TS[2] and IntFOLD3-TS[3]) and QA methods (ModFOLD4[26] and ModFOLD5). In CASP11, we used the IntFOLD3-TS method for our server TS predictions and ModFOLD5 (which was similar in performance to ModFOLD4), in order to select the top server models from our manual submissions.

Figure 3 shows our in-house benchmarking of the ModFOLD6_rank method compared with other QA methods for model selection using CASP11 data. The results show that the ModFOLD6_rank method obtains higher cumulative GDT_TS scores for model selection (∑GDT=44.42) than its component methods[11,27] as well as outperforming the older ModFOLD5_single and ModFOLDclust2 methods (∑GDT_TS=40.06 and ∑GDT_TS=42.68 , respectively), which were used in CASP11 for QA and model selection.

The ModFOLD6_rank method was built into the IntFOLD4-TS pipeline and then benchmarked against our previous servers using the CAMEO resource[25]. The direct comparison of the performance of the IntFOLD4-TS server versus the other servers is shown in Table 1. According to the CAMEO-3D results, using 12 months of data and a common subset of 500 targets, IntFOLD4-TS is shown to outperform our own older servers (IntFOLD2-TS & IntFOLD3-TS), and all but one other server – Robetta (a direct 12 month comparison of IntFOLD4-TS and Robetta on more targets is shown in Table S1). The CAMEO analysis using 6 months of data shows the same ranking of servers (Table S2). In addition to benchmarking IntFOLD4-TS with CAMEO-3D, the ModFOLD6 server is continually benchmarked with CAMEO-QE in terms of ASE/local score predictions, and is verified to outperform our older method, ModFOLD4, as well as most others.

CASP12 results - TBM and TBM/FM performance comparison with other groups

Tables 2 and 3 show the relative performance of our group methods versus other top 20 groups on the manual and server-only targets respectively. Table 2 shows that our manual prediction group (McGuffin) is ranked as the top group by SUM Zscore (>0.0) and AVG

Zscore (>0.0) and 2nd top by other rankings. Our WeFOLD group (wfRosetta-ProQ-ModF6; paper in preparation), which used ModFOLD6_rank for final model selection and ASE scoring, is also shown to be best out of the WeFOLD groups by all rankings, the 5th best group overall by AVG Zscore (>0.0), and within the top 10 overall for all score rankings.

On the 'all groups' targets, our IntFOLD4 server is competitive with the manual groups, ranking as the 11th best group overall by SUM Zscore (>-2.0) and the 3rd best server group by all rankings (Table 2). The relative performance of IntFOLD4 is also reflected on the 'server only' targets where it ranks 3rd best by SUM Z-score (>-2.0) and AVG Z-score (>-2.0) and 5th by other rankings (Table 3).

CASP12 results - comparison of top models produced at different stages of the IntFOLD4-TS pipeline

One of the questions asked at the CASP12 meeting was: "Which stages of the IntFOLD4-TS pipeline are worthwhile and actually show improvements?" The results in Table 4 and supplementary Tables S3-S9 show our in-house comparison of the IntFOLD4-TS performance versus the single-template modelling stages and reference multiple-template modelling methods. The cumulative scores of the top models from the IntFOLD140 single-template stage, which uses models from all 14 sequence-structure alignment methods, is shown to outperform the other single-template model ranking stages, which use fewer input alignments (IntFOLD60 and LOMETS4.4). This result indicates that adding more alignment methods leads to higher performance on average, according to the GDT-TS scores (Table 4, Table S6), GDT_HA scores (Tables S3 & S7), TMscores (Tables S5 & S8) and MaxSub scores (Tables S5 & S9).

Furthermore, the cumulative scores from the single-template methods are shown to be outperformed by those of the multiple-template modelling methods, indicating that using multi-templates for building models is a worthwhile stage of the pipeline. In addition, the

models from IntFOLD4-TS are of higher quality overall than those from the other multiple-template methods (IntFOLD3-TS, HHpred and I-TASSER_LIGHT) according to all scores. The pairwise t-tests results show that the IntFOLD4 top models are of significantly higher quality than those from each of the single-template model ranking stages, according to all scores. In comparsion with the multi-template methods, the pairwise t-tests on the data in Table 4 and supplementary Tables S1-S7 do not provide evidence that the improvement in scores over those from HHpred is statistically significant, however the scores are shown to be significantly better than those from the I-TASSER_LIGHT and IntFOLD3-TS methods.

DISCUSSION

For our CASP12 TS predictions, our main aims were to: i) increase the diversity of models, using more sequence-structure alignment methods, ii) improve the ranking of models and increase the quality of the top selected models, iii) improve detection of local errors in models (ASE scores) and finally, iv) fix the detected errors in the models.

What we did differently in CASP12 compared with CASP11

We developed 3 new methods to help us meet our aims: i) the IntFOLD4 server (group number 405) for TS predictions, ii) the ModFOLD6 variants (groups 201, 072, 360) for model ranking/selection and ASE, and iii) the ReFOLD method for repairing models using QA guided iterative refinement (N.B. ReFOLD was used as a standalone in-house tool during CASP12, but is now a publicly available server).

For our IntFOLD4 server TS predictions, the major improvement over the IntFOLD3-TS method (used in CASP11) was the integration of our new ModFOLD6_rank model quality assessment method for final model selection and ASE, instead of using ModFOLDclust2. Additionally, the sequence-structure alignment method CNFsearch was

added to the initial single-template modelling stage and we added HHpred models to the final ranking stage. The ReFOLD method was not incorporated into the IntFOLD4-TS server pipeline, as it was not fully automated during the prediction season.

For our manual predictions (McGuffin, group 017), we also used the ModFOLD6_rank method, for model selection and ASE scoring, but in addition, we used our ReFOLD iterative refinement method, which was guided by the model quality scores. Conversely, in CASP11 we only used the ModFOLDclust2 method for server model selection and ASE scores and we did not carry out any refinement.

What went right?

Our IntFOLD4-TS server is a significant improvement over our previous automated approach, IntFOLD3-TS (Table4 & Tables S1-S7), and this result is reflected in the independent CAMEO benchmark results. At the time of writing, IntFOLD4 ranks as the 2nd best 3D server according to the CAMEO lDDT scores based on pairwise comparisons (it is outperformed by only 1 public server in the benchmark - Robetta). Our ModFOLD6_rank method significantly improves upon our previous approaches for quality assessment, particularly for top model selection (Figure 3) and for providing local error estimates[11]. On average, the ReFOLD method provided a further significant improvement in model quality of our submitted modes for the regular (T0) targets[12]. Figure 4A & B show two examples of TBM/FM targets where we successfully selected from among the best server models and then further improved them with refinement. On average, mostly things went right – overall our selection of top models and ASE scoring was successful and it was worthwhile attempting to repair the errors with ReFOLD.

What went wrong?

If CASP12 groups are ranked on the TBM & TBM/FM targets using GDT_TS based metrics alone, then our methods are no longer at the very top; the McGuffin group rankings drop from 1st/2nd place to 5th place overall, and the IntFOLD4 server drops from 3rd/5th to 9th/10th place on the server only targets (http://predictioncenter.org/casp12/zscores_final.cgi). We clearly got an extra ranking boost from our ASE scores and we are grateful that the assessors placed an emphasis on it in their scoring formula.

Despite our significant progress with ModFOLD6_rank[11], the method is still sub optimal in terms of model ranking, which led to some poorly chosen IntFOLD4 final models and initial server models used by the McGuffin group. Additionally, the IntFOLD4 server is outperformed on the FM targets by many methods, and presently, the server does not integrate our refinement protocol (http://predictioncenter.org/casp12/zscores_final.cgi).

Our refinement of 3D models is still a bit hit and miss; although we have improved upon our identification of errors (we know which parts of a model are likely to be incorrect), we still can't consistently repair them. The current ReFOLD approach is very inconsistent on easier targets with higher quality starting models, as there is less room for improvement and detecting smaller changes in quality is more difficult[12]. Figure 4C shows an example where our initial model selection was sub optimal and then our refinement process made the model even worse. Figure 4D shows an example where we managed to successfully refine the starting model, however we should have selected our own server model instead, which had a higher GDT_TS score.

CONCLUSIONS

What we learned and our future plans

Focusing on model selection, refinement and maximising ASE scoring is clearly a worthwhile strategy. The emphasis by the assessors on the ASE performance is clearly very important. A model (or parts of it) could be very unreliable, even if it is from a good method, in which case, it could be worse than useless if biologists rely upon it to inform their experiments. Biologists are therefore becoming increasingly sophisticated in their use of 3D protein models and we have observed a steady growth in citations and users of our model quality assessment servers[11]. The major contribution to our group performance came from the ModFOLD6_rank method, i.e., improved model selection and ASE scoring. Additionally, we gained a small overall performance boost in GDT_TS by attempting refinement for the T0 targets using ReFOLD [12].

We will continue to work on improving our model ranking, selection and ASE scoring with future versions of ModFOLD. With ReFOLD, we will work on detecting smaller changes in improvements to models (for good quality starting models). In addition, we will work on a more focused ReFOLD protocol, which makes better use of our strengths in ASE scoring, by concentrating on improving only the low quality residues in the starting models. The future versions of the ModFOLD and ReFOLD methods will be integrated with later versions of the IntFOLD server. For our manual predictions, we plan to continue to make use of the server models, and likewise, we will use future versions ModFOLD for selection and ReFOLD for refinement.

FUNDING

This work was supported by studentships from: the Malaysian Government (to A.N.S.), the Saudi Arabian Government (to A.H.A.M.), the Tanzanian Government through the Institute of Finance Management (IFM) (to B.R.S.), and the University of Reading and the Diamond

Light Source Ltd. (to J.D.A.). This work was also supported by IBC ANR Investissements D'Avenir (to D.B.R.).

REFERENCES

1. McGuffin LJ, Roche DB. Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. Proteins 2011;79 Suppl 10:137-146.
2. Buenavista MT, Roche DB, McGuffin LJ. Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. Bioinformatics 2012;28(14):1851-1857.
3. McGuffin LJ, Atkins JD, Salehe BR, Shuid AN, Roche DB. IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. Nucleic Acids Res 2015;43(W1):W169-173.
4. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19(7):874-881.
5. Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins-Structure Function and Bioinformatics 2005;61:143-151.
6. McGuffin LJ. Benchmarking consensus model quality assessment for protein fold recognition. Bmc Bioinformatics 2007;8:345.
7. McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 2008;24(4):586-587.
8. McGuffin LJ. Prediction of global and local model quality in CASP8 using the ModFOLD server. Proteins 2009;77 Suppl 9:185-190.
9. Roche DB, Buenavista MT, Tetchner SJ, McGuffin LJ. The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. Nucleic Acids Res 2011;39(Web Server issue):W171-176.
10. 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.
11. Maghrabi AHA, McGuffin LJ. ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. Nucleic Acids Res 2017.
12. Shuid AN, Kempster R, McGuffin LJ. ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. Nucleic Acids Res 2017.
13. Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP6. Proteins 2005;61 Suppl 7:152-156.
14. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.

15. Margelevicius M, Venclovas C. Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. Bmc Bioinformatics 2010;11.
16. Yang YD, Faraggi E, Zhao HY, Zhou YQ. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011;27(15):2076-2082.
17. Ma JZ, Wang S, Zhao F, Xu JB. Protein threading using context-specific alignment potential. Bioinformatics 2013;29(13):257-265.
18. Wu ST, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35(10):3375-3382.
19. McGuffin LJ, Roche DB. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics 2010;26(2):182-188.
20. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5(4):725-738.
21. Meier A, Soding J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. Plos Comput Biol 2015;11(10).
22. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Bioph Biom 2000;29:291-325.
23. Bhattacharya D, Nowotny J, Cao R, Cheng J. 3Drefine: an interactive web server for efficient protein structure refinement. Nucleic Acids Res 2016;44(W1):W406-409.
24. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26(16):1781-1802.
25. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T. The Protein Model Portal--a comprehensive resource for protein structure and model information. Database (Oxford) 2013;2013:bat031.
26. McGuffin LJ, Buenavista MT, Roche DB. The ModFOLD4 server for the quality assessment of 3D protein models. Nucleic Acids Res 2013;41(Web Server issue):W368-372.
27. Uziela K, Wallner B. ProQ2: estimation of model accuracy implemented in Rosetta. Bioinformatics 2016;32(9):1411-1413.

FIGURE LEGENDS

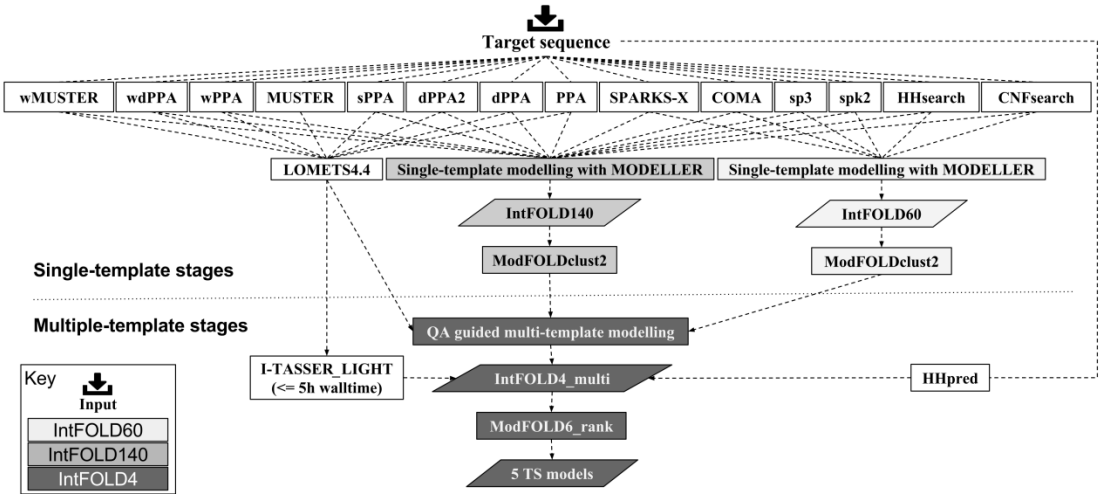
Figure 1. Flowchart outlining the principal stages of stages of the IntFOLD4-TS prediction pipeline. Rectangles show processes, parallelograms show datasets. The only input is the target sequence. The initial single-template modelling stages start with 14 sequence-structure alignment methods (8 from the LOMETS[18] package & 6 others as described in the main text[13-17]). Single-template models are built from the various alignment methods using MODELLER[22] (creating the IntFOLD60, IntFOLD140 model datasets) and then ranked with ModFOLDclust2[19]. LOMETS4.4 is also used to rank the backbone models produced by its own component threading methods. The multiple-template modelling stages include QA guided multi-template modelling (using the scores from ModFOLDclust2) in order to generate a set of multi-template models. Additionally, models from HHpred[21] and I-TASSER_LIGHT[20] are added to the final IntFOLD4_multi set for evaluation. The ModFOLD6_rank method[11] is used for ASE and final model selection.

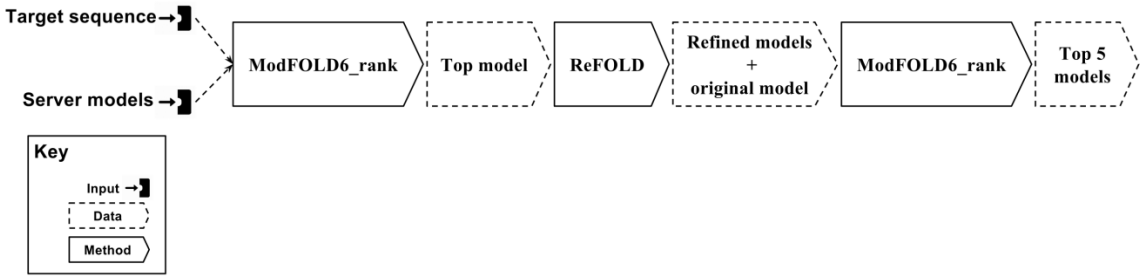
Figure 2. Flow of data and methods used by the McGuffin group for making TS predictions. The initial server models for each target were ranked using the ModFOLD6_rank method. The top ranked server model was then refined using ReFOLD to produce a set of alternatives, which were then further scored and ranked against the original top model using ModFOLD6_rank.

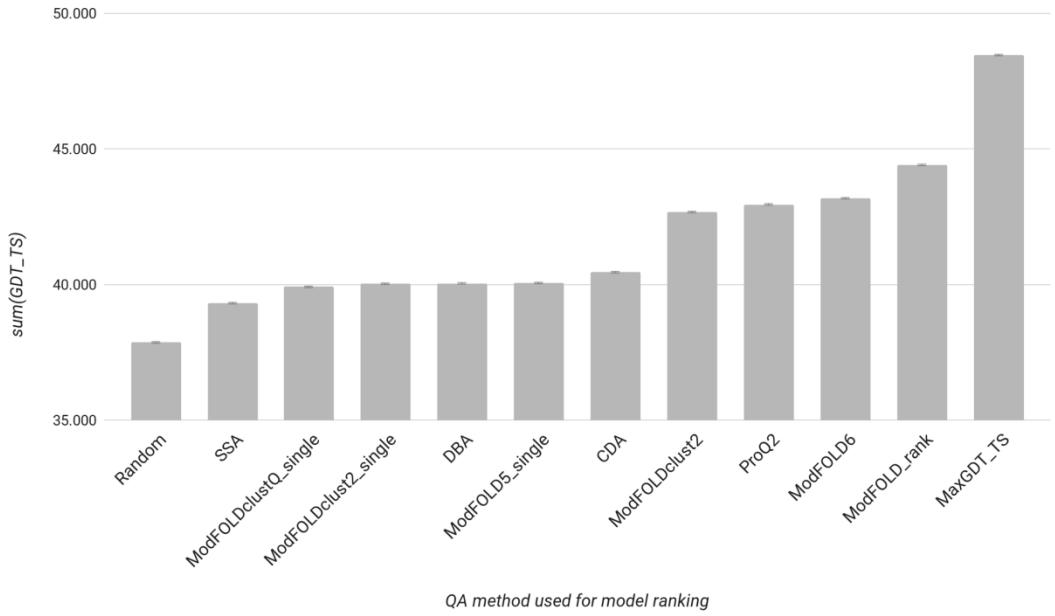
Figure 3. Benchmarking the performance of QA methods for model selection using CASP11 data, prior to CASP12. ModFOLD6_rank versus other global scoring methods: SSA, Secondary Structure Agreement; DBA, Disorder B-factor Agreement; CDA, Contact Distance Agreement[11]. Cumulative GDT scores for the top selected models from the QA targets (models from QA round1 and round2 combined, 84 targets with structure). The

maximum possible GDT_TS (MaxGDT_TS) is the cumulative score obtained by selecting the best model available for every target. The error bars show the Standard Error in GDT_TS ($\sigma/\sqrt{n}$, where σ is the standard deviation and n is the number of targets (84)).

Figure 4. Examples of what went right (**A**&**B**) and what went wrong (**C**&**D**) with our manual TBM & TBM/FM predictions. Left panels, refined model with the ModFOLD6 accuracy self-estimate (ASE) displayed using the temperature colour scheme (blue = residues close to the native structure, red= residues far from the native structure). Middle panels, superposition of the top selected server model (cyan), refined model (magenta) and native structure (green). Right panels, GDT plots comparing top selected server models (cyan) with the ReFOLD refined models (magenta). (**A**) T0912TS017_1-D2 (McGuffin TS1) - Perfect initial model selection (best model = GOAL_TS1), ASE score = 82.28, and successful refinement with an improvement on the initial model (GDT_TS from 62.95 to 65.36) (**B**) T0892TS017_1-D1 (McGuffin TS1) - Excellent initial model selection (2nd best server model = Zhang-Server_TS3), ASE score = 80.95, and successful refinement (GDT_TS improved from 79.71 to 82.25). (**C**) T0942TS017_1-D1 (McGuffin TS1) suboptimal initial model selection (QUARK_TS1), ASE score = 86.09, and unsuccessful refinement (GDT_TS declined from 79.91 to 79.33). (**D**) T0896TS017_1-D2 (McGuffin TS1) ASE score = 81.45, refinement successful (GDT_TS improved from 42.12 top 45.25), but suboptimal initial model selection (GOAL_TS2), our own IntFOLD4 TS1 model was better overall (GDT_TS=46.50).







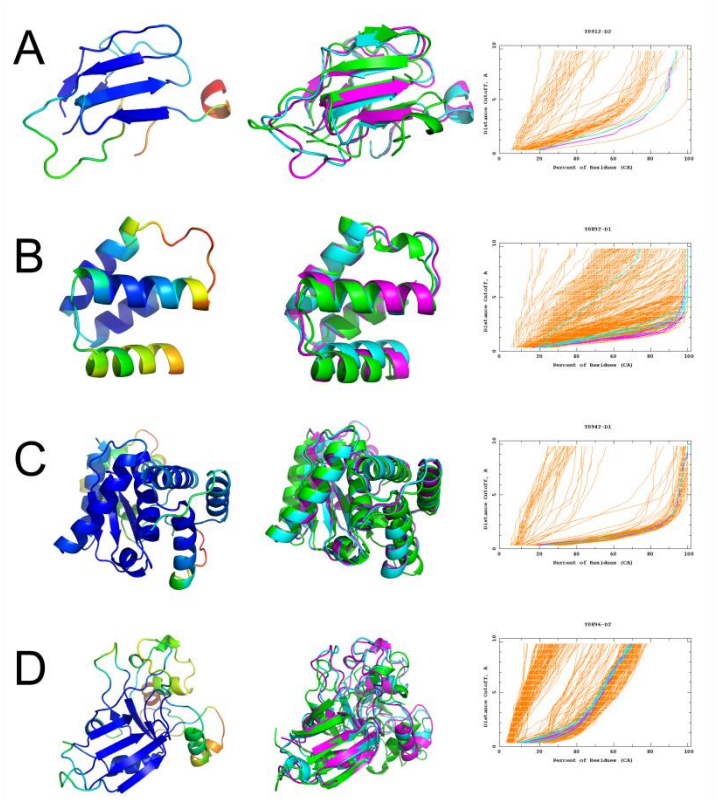


Table 1. Performance of IntFOLD4-TS versus other servers. CAMEO-3D: Common Subset Comparison, 1-year Performance (2016-05-13 - 2017-05-06) (500 targets - 10 methods). IntFOLD4_TS is the reference server (listed as server58, or IntFOLD4-TSb on CAMEO). Data are from http://www.cameo3d.org/. The table is sorted by difference in Average lDDT score.

	Average lDDT		Average CAD score		Average lDDT-BS	
Server Name	**Dif.**	**Ref.**	**Dif.**	**Ref.**	**Dif.**	**Ref.**
Robetta	-1.63	70.9	-0.02	0.7	2.73	68.86
IntFOLD4-TS	0	69.27	0	0.68	0	71.6
RaptorX	0.82	68.45	0	0.67	4.36	67.24
IntFOLD3-TS	1.74	67.53	0.02	0.66	3.02	68.57
IntFOLD2-TS	1.98	67.28	0.02	0.66	2.64	68.96
HHpredB	2.09	67.17	0	0.67	2.59	69.01
SWISS-MODEL	3.82	65.44	0.04	0.64	1.1	70.5
SPARKS-X	5.26	64.01	0.03	0.64	5.54	66.06
Princeton_TEMPLATE	9.36	59.91	0.09	0.59	15.14	56.46
NaiveBLAST	11.57	57.7	0.12	0.56	11.15	60.45

Table 2. Official CASP12 TBM performance comparison with Top 20 groups - All groups on 'all groups' targets. Analysis on the models designated as "1". Assessors' Formula for TBM + TBM/FM domains (GDT_HA + (SG + lDDT + CAD)/3 + ASE), sorted by SUM Z score (>0.0). * Denotes server groups. Data are from http://predictioncenter.org/casp12/zscores_final.cgi.

GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
17	**McGuffin**	**29**	**27.1618**	**2**	**0.9366**	**2**	**27.882**	**1**	**0.9614**	**1**
417	VoroMQA-select	29	27.1702	1	0.9369	1	27.7907	2	0.9583	2
4	Zhang	29	26.6779	3	0.9199	3	26.8982	3	0.9275	3
203	ProQ2	29	23.9338	4	0.8253	4	25.6378	4	0.8841	4
73	Wallner	29	23.5157	5	0.8109	5	24.1663	5	0.8333	6
439	MULTICOM	29	23.363	6	0.8056	6	23.839	6	0.822	7
252	**wfRosetta-ProQ-ModF6**	**28**	**19.1779**	**10**	**0.7564**	**8**	**23.6082**	**7**	**0.8432**	**5**
243	Seok-refine	29	20.553	9	0.7087	10	23.4151	8	0.8074	8
479	Zhang-Server*	29	22.9713	7	0.7921	7	23.3177	9	0.8041	9
450	LEEab	29	12.1215	15	0.418	17	22.9844	10	0.7926	10
11	LEE	29	11.2964	18	0.3895	20	22.2071	11	0.7658	11
183	QUARK*	29	20.8679	8	0.7196	9	21.484	12	0.7408	13
239	wfAll-Cheng	27	14.0968	14	0.6703	11	19.5346	13	0.7235	14
320	raghavagps	28	14.1009	13	0.575	13	19.204	14	0.6859	15
64	Jones-UCL	29	14.4308	12	0.4976	15	17.9234	16	0.618	18
384	wfMESHI-Seok	29	10.6692	19	0.3679	21	17.674	19	0.6094	20
405	**IntFOLD4***	**29**	**14.6733**	**11**	**0.506**	**14**	**16.3385**	**28**	**0.5634**	**32**
23	Seok	28	11.6229	16	0.4865	16	16.0142	30	0.5719	30
498	AP_1	29	11.4258	17	0.394	19	15.9422	31	0.5497	34
250	Seok-server*	29	10.48	20	0.3614	23	15.5036	34	0.5346	35

Table 3. Official CASP12 TBM performance comparison with Top 20 groups - Server groups on 'all groups' + 'server only' targets. Analysis on the models designated as "1". Assessors' Formula for TBM + TBM/FM domains (GDT_HA + (SG + lDDT + CAD)/3 + ASE), sorted by SUM Z score (>-2.0). Data are from http://predictioncenter.org/casp12/zscores_final.cgi.

GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
479	Zhang-Server	57	53.2687	1	0.9345	1	53.558	1	0.9396	1
183	QUARK	57	51.2984	2	0.9	2	51.8455	2	0.9096	2
405	**IntFOLD4**	**57**	**38.1512**	**3**	**0.6693**	**3**	**40.7005**	**5**	**0.714**	**5**
250	Seok-server	57	36.5368	4	0.641	4	41.7567	4	0.7326	4
236	MULTICOM-CONSTRUCT	57	34.5974	5	0.607	5	37.0262	8	0.6496	8
287	MULTICOM-CLUSTER	57	34.3201	6	0.6021	6	37.661	7	0.6607	7
345	MULTICOM-NOVEL	57	29.3235	7	0.5144	8	33.4039	9	0.586	10
220	GOAL	57	24.4642	8	0.4292	9	40.0205	6	0.7021	6
5	BAKER-ROSETTASERVER	57	22.6251	9	0.3969	10	43.5567	3	0.7642	3
119	HHPred0	57	10.6627	10	0.1871	11	21.3322	14	0.3742	17
349	HHPred1	57	10.2547	11	0.1799	12	21.102	15	0.3702	18
251	myprotein-me	57	7.6733	12	0.1346	13	25.0498	12	0.4395	13
48	ToyPred_email	57	6.0518	13	0.1062	14	27.6669	10	0.4854	11
92	RaptorX	57	5.921	14	0.1039	15	27.6238	11	0.4846	12
313	HHGG	57	4.1865	15	0.0734	16	24.6569	13	0.4326	14
425	FALCON_TOPOX	57	-5.3579	16	-0.094	18	16.0002	21	0.2807	23
77	FALCON_T	57	-5.3673	17	-0.0942	19	16.483	19	0.2892	21

	OPO						2			
16	FFAS-3D	57	-7.8423	18	-0.1376	21	14.3766	23	0.2522	26
446	YASARA	56	-8.1083	19	-0.1091	20	20.5859	17	0.3676	19
464	tsspred2	57	-10.2517	20	-0.1799	22	12.695	27	0.2227	30

Table 4. Comparative performance of the single-template and multiple-template modelling stages of the IntFOLD4-TS server pipeline. GDT_TS scores of the top models selected for the TBM + TBM/FM domains. IntFOLD60 & IntFOLD140 refer to the top models selected by ModFOLDclust2 from the datasets shown in Figure 1. LOMETS, LOMETS version 4.4. IT4, top model from I-TASSER version 4.4 with default parameters in light mode (<= 5h walltime). HHpred, top HHpred model from HHsuite version 3.0.0. IntFOLD3, top model selected from the final pool of models using the ModFOLDclust approach used by IntFOLD3-TS. The row labelled "t-test v IntFOLD4" shows the p-value for the paired samples t-tests for the scores from the method in the column versus those from the IntFOLD4 column. *The IT4 method was not run for targets >600 residues, so for T0912 the top LOMETS model score is used.

		Single-template methods			Multiple-template methods			
Domain class	Target	IntFOLD60	IntFOLD140	LOMETS	IT4	HHpred	IntFOLD3	IntFOLD4
FM/TBM	T0868-D1	0.2220	0.1659	0.1681	0.3427	0.2457	0.2522	0.1918
FM/TBM	T0884-D1	0.2148	0.2324	0.2500	0.2641	0.1901	0.2535	0.3134
FM/TBM	T0890-D1	0.3963	0.3415	0.3079	0.3384	0.1890	0.3384	0.3415
FM/TBM	T0892-D1	0.0833	0.4384	0.2862	0.4928	0.3587	0.4348	0.4928
FM/TBM	T0894-D2	0.1296	0.1343	0.1481	0.1759	0.2315	0.1296	0.3611
FM/TBM	T0896-D1	0.0843	0.0988	0.0000	0.0930	0.2384	0.0988	0.0988
FM/TBM	T0896-D2	0.3488	0.3200	0.2888	0.0588	0.1300	0.3250	0.3238
FM/TBM	T0898-D2	0.2000	0.1045	0.2273	0.3045	0.4500	0.3955	0.3045
FM/TBM	T0901-D1	0.2937	0.2511	0.2500	0.2433	0.2478	0.2444	0.2489
FM/TBM	T0909-D1	0.2793	0.3026	0.0323	0.2605	0.2905	0.2905	0.2905
FM/TBM	T0912-D2	0.1205	0.3313	0.3223	0.3223*	0.3313	0.3373	0.3223
FM/TBM	T0943-D1	0.1290	0.0000	0.0000	0.1935	0.3347	0.1250	0.3105
FM/TBM	T0945-D1	0.0660	0.2627	0.2400	0.2487	0.3507	0.2387	0.2487
TBM	T0860-D1	0.0735	0.5239	0.5276	0.5184	0.4724	0.5092	0.5184
TBM	T0861-D1	0.7965	0.7965	0.8277	0.7837	0.8365	0.8638	0.8269
TBM	T0865-D1	0.4073	0.7137	0.7137	0.3589	0.5645	0.5403	0.5645
TBM	T0867-D1	0.8918	0.8654	0.8149	0.6611	0.8774	0.8726	0.8798

TBM	T0872-D1	0.4205	0.4063	0.4261	0.4716	0.4063	0.4261	0.4886
TBM	T0877-D1	0.4384	0.4384	0.3908	0.4225	0.4366	0.4437	0.4296
TBM	T0879-D1	0.5489	0.5489	0.4693	0.4727	0.5398	0.5784	0.5784
TBM	T0881-D1	0.4047	0.4047	0.3861	0.3354	0.4703	0.4245	0.4059
TBM	T0882-D1	0.3766	0.3766	0.5380	0.4272	0.5759	0.3703	0.5063
TBM	T0885-D1	0.3728	0.4276	0.2632	0.4912	0.3553	0.4737	0.5307
TBM	T0889-D1	0.6715	0.6715	0.6130	0.5962	0.6893	0.6611	0.6674
TBM	T0891-D1	0.6875	0.6875	0.7031	0.6853	0.7232	0.7165	0.6920
TBM	T0893-D1	0.4418	0.4418	0.3048	0.3048	0.3322	0.2979	0.4110
TBM	T0893-D2	0.5621	0.5621	0.5695	0.5636	0.6568	0.5843	0.5932
TBM	T0895-D1	0.1146	0.4458	0.4979	0.4500	0.5063	0.4771	0.4750
TBM	T0902-D1	0.3604	0.3139	0.3139	0.3268	0.3766	0.3160	0.2955
TBM	T0903-D1	0.6759	0.7330	0.7130	0.7130	0.6512	0.7461	0.7461
TBM	T0912-D1	0.2530	0.3804	0.2512	0.2512*	0.2729	0.2699	0.2633
TBM	T0913-D1	0.3506	0.3698	0.3698	0.3491	0.3913	0.3772	0.3772
TBM	T0917-D1	0.5185	0.5185	0.4955	0.5192	0.5492	0.5301	0.5301
TBM	T0920-D1	0.4346	0.4525	0.4206	0.4136	0.4829	0.4470	0.4813
TBM	T0920-D2	0.0365	0.0411	0.0103	0.0491	0.3824	0.0411	0.0434
TBM	T0921-D1	0.4112	0.3895	0.4112	0.4330	0.4076	0.4565	0.4112
TBM	T0922-D1	0.5135	0.5135	0.5034	0.5034	0.5439	0.5980	0.5980
TBM	T0928-D1	0.2610	0.2559	0.2617	0.2683	0.3028	0.2808	0.3174
TBM	T0942-D1	0.1040	0.1171	0.5650	0.1315	0.6012	0.6012	0.5997
TBM	T0942-D2	0.2710	0.2757	0.0572	0.2757	0.3435	0.3435	0.3353
TBM	T0943-D2	0.2886	0.2897	0.2897	0.2757	0.3809	0.2349	0.2942
TBM	T0944-D1	0.4792	0.4792	0.4101	0.3676	0.4881	0.5089	0.5119
TBM	T0946-D2	0.3573	0.3679	0.2217	0.2476	0.1769	0.3125	0.3172
TBM	T0947-D1	0.3600	0.3600	0.3929	0.3800	0.3786	0.4100	0.3800
TBM	T0948-D1	0.1493	0.5084	0.4765	0.4715	0.5017	0.4832	0.5101
	Total	15.4516	17.1520	16.2540	16.3858	18.7612	18.1769	**18.9180**
	t-test v IntFOLD4	**7.93E-05**	**6.59E-03**	**3.14E-05**	**1.15E-04**	3.97E-01	**2.66E-02**	