BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 14 2012, pages 1851–1857
doi:10.1093/bioinformatics/bts292

Structural bioinformatics

Advanced Access publication May 15, 2012

Improvement of 3D protein models using multiple templates guided by single-template model quality assessment

Maria T. Buenavista[1,2,3], Daniel B. Roche[1] and Liam J. McGuffin[1,*]

[1]School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AS, [2]Biocomputing Section, MRC Harwell, Harwell Oxford Campus, Didcot OX11 0RD and [3]Diamond Light Source, Beamline B23, Chilton, Didcot OX11 0DE, UK

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Modelling the 3D structures of proteins can often be enhanced if more than one fold template is used during the modelling process. However, in many cases, this may also result in poorer model quality for a given target or alignment method. There is a need for modelling protocols that can both consistently and significantly improve 3D models and provide an indication of when models might not benefit from the use of multiple target-template alignments. Here, we investigate the use of both global and local model quality prediction scores produced by ModFOLDclust2, to improve the selection of target-template alignments for the construction of multiple-template models. Additionally, we evaluate clustering the resulting population of multi- and single-template models for the improvement of our IntFOLD-TS tertiary structure prediction method.

Results: We find that using accurate local model quality scores to guide alignment selection is the most consistent way to significantly improve models for each of the sequence to structure alignment methods tested. In addition, using accurate global model quality for re-ranking alignments, prior to selection, further improves the majority of multi-template modelling methods tested. Furthermore, subsequent clustering of the resulting population of multiple-template models significantly improves the quality of selected models compared with the previous version of our tertiary structure prediction method, IntFOLD-TS.

Availability and implementation: Source code and binaries can be freely downloaded from http://www.reading.ac.uk/bioinf/downloads/.

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

Supplementary information: Supplementary data are available at *Bioinformatics* online. http://www.reading.ac.uk/bioinf/MTM_suppl_info.pdf

Received on March 1, 2012; revised on April 26, 2012; accepted on May 10, 2012

1 INTRODUCTION

Sequence-structure alignment methods are the springboard of template-based modelling (TBM). The 3D model output of single-template modeling methods is based on the top-ranked best available single template from the most similar protein of known structure (Tramontano and Cozzetto, 2011). Single-template models built for sequences with high-sequence identity (50–60%), between target and template, are known to provide highly accurate models; <30% sequence identity (the twilight zone), models must be built using more intensive fold recognition methods (Tramontano *et al.*, 2008).

Consensus-based fold recognition approaches such as LOMETS (Wu and Zhang, 2007), 3D-Jury (Ginalski *et al.*, 2003), 3D-SHOTGUN (Fischer, 2003), Pcons (Lundstrom *et al.*, 2001) and IntFOLD-TS (McGuffin and Roche, 2011; Roche *et al.*, 2011) improve the selection of single-templates by the use of alternative models from several independent servers or in-house version of those methods run as metaservers. The top-ranked consensus models, however, may not always be the best option due to sub-optimal alignments. It is important to consider optimal alignments both at the global (overall protein sequence) and local (regions such as domains) levels. In addition, the coverage of the target-template alignment must be considered when evaluating or selecting 3D models to guide experimental work. After all, the nearer-native a protein model is, the better information it can provide towards protein design for use in medicine, agriculture, biofuels and other fields of applications.

Multiple-template modelling methods that mix and match (Liu *et al.*, 2008), recombine templates (Contreras-Moreira *et al.*, 2003) or thread alternative target-template alignments (Peng and Xu, 2011a,b) have been produced in an effort to improve the quality of predicted 3D models. The advantage of using multiple-template methods is mainly attributed either to the increased alignment coverage or the incorporation of the best single template, which is dependent on alignment accuracy (Larsson *et al.*, 2008) and template complementarity (Chakravarty *et al.*, 2008; Peng and Xu, 2011a,b). Additionally, convergence (Chakravarty *et al.*, 2008; Martínez *et al.*, 2007; Wallner and Elofsson, 2003) is one problem identified in the use of multiple templates, which may arise from different alignments containing contradictory information. The quality of a model can be improved by combining complementary good templates, which is especially true for multi-domain targets; but model quality may deteriorate when there is one template that is closer to the target than all other available templates. In general, multiple-template modelling can improve on single-template models when it can provide an increase in coverage of the target, can give more useful information such as sequence and structure conservation and/or conserved distance restraints or can identify better alternative alignments for harder targets (Cheng, 2008).

Although the last Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition, CASP9, assessed the overall state-of-the-art TBM algorithms as high, the assessors still consider template selection, target-template alignment and

*To whom correspondence should be addressed.

quality prediction as the remaining limiting factors (Mariani *et al.*, 2011). The top-performing groups have attempted to address these limiting factors in numerous ways (Peng and Xu, 2011a). Various single-template and consensus methods incorporate Model Quality Assessment (MQA) methods to rank alternative models (Kryshtafovych *et al.*, 2011; McGuffin, 2010). PconsM (Larsson *et al.*, 2011), for example, uses ProQ (Wallner and Elofsson, 2003) for model quality assessment of the top six models drawn from the top six template sequences of Pcons.net (Wallner *et al.*, 2007). On the other hand, RaptorX (Peng and Xu, 2011b) uses TM-score (Xu and Zhang, 2010; Zhang and Skolnick, 2004) as a quality measure of the sequence-template alignment, from which the quality of its 3D model can be estimated. Additionally, MULTICOM (Wang *et al.*, 2010; 2011) combines multiple templates, alternative alignments and similar models guided by global model quality assessment. Furthermore, with I-TASSER (known as Zhang-Server in CASP) (Roy *et al.*, 2010; Zhang, 2008), the multi-template 3D models include global confidence scores, which are a measure of their overall prediction accuracy.

Our own IntFOLD-TS method (McGuffin and Roche, 2011; Roche *et al.*, 2011) integrates ModFOLDclust2 (McGuffin and Roche, 2009) into the core of its pipeline. This MQA tool re-ranks all alternative models, generated by four in-house versions of single-template methods, based on their global quality scores. The accuracy of this method addressed the strong emphasis of CASP9 on assessment of all-atom models producing predicted per-residue errors (Mariani *et al.*, 2011). However, while we were relatively successful at providing accurate local error estimates for our CASP9 models, we did not make any attempt at fixing the predicted errors prior to submission. In an attempt to exploit our accurate per-residue quality predictions, we have now explored the multiple-template modelling approach described here. Our approach is to use the ModFOLDclust2 local and global quality assessment scores, from the initial single template models, to inform target-template alignment selection, with the aim of consistently and significantly improving models through multiple-template modeling. The latest implementation of our template-based modelling server, IntFOLD2-TS, uses a further iteration of ModFOLDclust2 to rank the resulting pool of multiple-template models. Our new IntFOLD2-TS method shows a statistically significant improvement over the single-template methods and consensus single-template methods tested.

2 METHODS

2.1 Datasets and benchmarks

Single- and multiple-template modelling methods were evaluated on both the CASP8 and CASP9 datasets using the full chain sequences for each target. Official native 3D structures for full chains and domains were obtained from the CASP website: http://predictioncenter.org/download_area/. The observed model quality for the generated models was assessed using the TM-score program (Zhang and Skolnick, 2004) (http://zhanglab.ccmb.med.umich.edu/TM-score/) to generate structural alignment scores namely; TM-scores (Xu and Zhang, 2010); GDT_TS (Zemla *et al.*, 2001) scores and MaxSub (Siew *et al.*, 2000) scores. A statistical analysis of the relative performance of all the methods was carried out using the Wilcoxon signed-rank sum test with the R statistical package, version v2.12.1 (http://www.R-project.org). The statistical significance of the differences in observed model quality scores (TM-scores, GDT_TS-scores and MaxSub scores) was reflected by the calculated *P*-values produced by these tests. To generate the *P*-values, we used the R command: wilcox.test(*x*, *y*, paired=TRUE, alternative= "greater"), where *x* and *y* were vectors of observed model quality scores from different methods.

2.2 Single-template modelling methods and clustering or consensus methods

Nine different fold recognition methods were installed and run in-house to generate up to 10 sequence-to-structure alignments each, resulting in up to 90 alternative single-template-based models being generated for each CASP target. The fold recognition methods used were SP3 and SPARKS2 (referred to hereafter as SPK2) (Zhou and Zhou, 2005), HHsearch (Soding, 2004), COMA (Soding *et al.*, 2005) and the five alternative threading methods that are integrated into the current LOMETS (Wu and Zhang, 2007) package—QQQ, GGGd, GGGf, NNNd and SSSc (http://zhanglab.ccmb.med.umich.edu/I-TASSER/download/).

For each single-template method, the associated template libraries were screened so that they only included templates for structures that existed in the RCSB PDB (www.pdb.org) yearly snapshots (ftp://snapshots.rcsb.org/) taken prior to the start of each of the CASP experiments. For example, for the CASP8 target sequences, the template libraries for methods included only the structures that were available in the PDB on the January 7, 2008. Likewise, for the CASP9 targets, the template libraries included only the structures that were available in the PDB on the January 4, 2010. Furthermore, the UniProt (http://www.uniprot.org) sequence databases were also appropriately screened using a method similar to that carried out by Chubb *et al.* (2010). The sequence databases from the January 7, 2008, to the January 4, 2010, were recreated using the uniprot_trembl.fasta file and deposition date field found within the associated uniprot_trembl.dat file: ftp://ftp.uniprot.org/pub/databases/uniprot/.

In addition to the individual methods described earlier, clustering-based approaches were used to generate consensus single-template models. The LOMETS method was installed locally and used with default settings to select a consensus from a pool of 50 models from 5 methods: QQQ, GGGd, GGGf, NNNd and SSSc. In addition, the IntFOLD-TS40 method used the ModFOLDclust2 method with default parameters to select a consensus from a pool of 40 models from 4 methods: SP3 and SPARKS2, HHsearch and COMA. (N.B. This was similar to the approach we used for the IntFOLD-TS server method during CASP9; however, in this case, we only used single-template models and we used older template libraries and sequence databases—see previous paragraph.) Finally, the IntFOLD-TS90 method used the ModFOLDclust2 method with default parameters to select a consensus from a pool of 90 models produced by all 9 single-template methods (N.B. while methods such as LOMETS and IntFOLD make use of multiple templates, in this article, we define multi-template modelling methods as those that use several alignments to alternative templates to build an individual model for a given target.)

2.3 Alignment selection methods for multiple-template modelling

A number of simple target-template alignment selection protocols were followed and evaluated for the value added to each of the single-template modelling based methods described earlier. For each method, a multiple alignment file was generated, in PIR format, using the individual target-to-template alignments and Modeller v9.8 (Fiser and Sali, 2003) was subsequently used to generate multi-template models using the default parameters and scripts.

2.3.1 Method 1 (multi1) The top two alignments generated by the single-template modelling method were used to generate a multi-template model for each sequence to structure alignment method. In the analysis carried out by Larsson *et al.* (2008), performance was observed to peak when just the top two target-to-template alignments were used to generate multi-template models, so this was the rationale behind this method.

2.3.2 Method 2 (multi2) For this method, the top alignment was used along with any subsequent alignments, if greater than or equal to 40 new residues were covered, and if the overlapping regions of covered residues was less than or equal to 20 residues. If no subsequent alignments were identified, then a model was built using the top alignment only. This is similar to the alignment selection successfully used by Hildebrand *et al*. (2009).

2.3.3 Method 3 (multi3) This method made use of the ModFOLDclust2-predicted per-residue errors to guide the alignment selection process if target-template alignments were found to overlap. The top alignment and any subsequent alignments were included if any overlapping residues in the model (built using the subsequent alignment) were predicted to result in fewer local error scores than those occurring in the models built using the preceding alignments. The mean local error scores were taken for the overlapping regions covered by the new residues and then compared with the best mean local scores from the equivalent local regions previously covered. If the mean local errors for the overlapping regions were found to be reduced, then the alignment was included. If no subsequent alignments were identified, a model was built using the top alignment only.

2.3.4 Method 4 (multi4) This method was based on the observations made by Larsson *et al*. (2008) that when extra templates do not add any extra coverage to the aligned target sequence, there is little advantage to be gained from multiple-template modelling. Thus, for this method the top alignment was used and any subsequent alignments were also included, but only if they increased coverage of the target by at least 1 residue.

2.3.5 Methods 5, 6, 7 and 8 (multi5 to multi8) Another observation made by Larsson *et al*. (2008) was that the global predicted model quality (particularly for the first two alignments) was one of the most important factors for successful multi-template modelling. Therefore for methods 5, 6, 7 and 8, we repeated methods 1, 2, 3 and 4, respectively. However, prior to alignment selection, the alignments for each of the nine single-template methods were firstly re-ranked based on the ModFOLDclust2 predicted global model quality scores.

2.4 Iteration of clustering for IntFOLD2-TS: running ModFOLDclust2 on the resulting set of multi-template models

The multi-template modelling alignment selection methods described earlier resulted in the generation of a new population of 84 models for each target (eight new models each for SP3, SPARKS2, HHsearch, COMA, QQQ, GGGd, GGGf, NNNd and SSSc, plus four new models each for LOMETS, IntFOLD-TS40 and IntFOLD-TS90). These models were then assessed using ModFOLDclust2 and the top-ranked models were designated as the IntFOLD2-TS predictions. A flowchart of the complete modelling pipeline is shown in Supplementary Figure 1.

3 RESULTS

Three structural alignment scoring methods—the GDT_TS (Global Distance Test Total Score), the TM-score (Template Modelling score) and the MaxSub score—provide the performance metrics for the benchmarking. These scores, generated by the TM-score method, show how similar two protein structures are to each other, in this case, the model and the experimentally determined native structure. For clarity and brevity, the results shown in this article are the benchmarking results against the CASP9 full chain dataset. However, the full analyses for both the CASP8 and CASP9 full chain and domain datasets (with and without FM domains) are shown in the supplementary data. Furthermore, in the supplementary data, an analysis of the upper limits for multi-template modelling based on perfect model selection is shown. In this case, the TM-score between the model and native structure is used to determine the global quality for each model. The perfect local quality scores are calculated as the distances between equivalent CA atoms based on the TM-score superposition between the model and native structures. Although all structural alignment scores were used for benchmarking, the main tables and figures shown here were based on the GDT_TS score, which has been used as an official CASP measure since CASP4.

The 9-fold-recognition methods that represent single-template modelling methods are as follows: SP3, SPARKS2 (SPK2), HHsearch, COMA, QQQ, GGGd, GGGf, NNNd and SSSc. The consensus methods that cluster-specific single-template methods and generate single-template models are represented here by LOMETS, IntFOLD-TS40 and IntFOLD-TS90. The multi-template methods applied to sets of alignments from individual methods are denoted by the subscript '_multi*' (e.g. sp3_multi3 is the multi-template modelling Method 3 applied to the SP3 alignments).

3.1 Single-template and consensus modelling methods

An all-against-all pairwise comparison matrix of 12 tertiary structure modelling methods is presented in Table 1. Calculated Wilcoxon signed-rank sum test *P*-values that populate the matrix were computed based on the GDT_TS scores of CASP9 full chain targets. A significant difference between any two methods is established by *P*-value of ≤ 0.05 which suggests an improved performance by the method.

Given the ≤ 0.05 *P*-value threshold, it can be said that the top three performers, as shown in Table 1 (values in boldface), are SP3, IntFOLD-TS40 and IntFOLD-TS90. The SP3 method outperformed 8 out of 11 of the other single-template models. However, SP3 did not perform significantly better than HHsearch (and *vice versa*) and only IntFOLD-TS40 and IntFOLD-TS90 significantly outperformed the method. IntFOLD-TS40 *P*-values were better in 10 out of 11 methods, and IntFOLD-TS90 significantly outperformed all methods. Slight variations of results can be seen in the *P*-values of the rest of the full chain and domain datasets of CASP8 and CASP9 and using different scoring metrics (Supplementary Tables). However, the consensus method, IntFOLD-TS90, was the consistent top performer, showing statistically significant improvements over methods, across all datasets.

3.2 Multiple-template modelling based on original alignment rankings from the single-template modelling methods

The multiple-template models referred to as multi1, multi2, multi3 and multi4 used information drawn from the top-ranked and subsequent ranked alignments based on the original single-template method rankings. The improvement in model quality from each multi-template modelling method is referred to here as the added-value.

Figure 1 shows a plot of the relative added-value (or cost) of multi-template modelling arrived at by totalling all the GDT_TS scores of all targets for each method. The sum of each single-template method is then subtracted from the sum of each corresponding multiple-template method as well as from the sum of each consensus method. This difference, whether positive or negative, represents the added-value or cost, respectively, of using each multiple-template method over single-template modelling.

Table 1. Calculated pairwise *P*-values for single template and consensus methods based on the CASP9 full chain GDT_TS scores

Method	SP3	SPK2	HHsearch	COMA	QQQ	GGGd	GGGf	NNNd	SSSc	LOMETS	IntFOLD-TS40	IntFOLD-TS90
SP3	1.0000	**1.81E-003**	3.01E-001	**3.00E-003**	**2.91E-002**	**2.97E-002**	**7.62E-003**	**4.55E-003**	**2.77E-004**	**1.75E-002**	9.57E-001	0.9998
SPK2	9.98E-001	1.0000	9.16E-001	1.15E-001	8.15E-001	6.78E-001	5.98E-001	5.36E-001	1.62E-001	8.27E-001	0.9996	1.0000
HHsearch	7.00E-001	8.48E-002	1.0000	**1.07E-003**	4.45E-001	2.39E-001	1.44E-001	1.19E-001	**3.35E-002**	3.38E-001	9.76E-001	0.9999
COMA	9.97E-001	8.86E-001	9.99E-001	1.0000	9.92E-001	8.90E-001	9.26E-001	8.87E-001	7.86E-001	9.83E-001	1.0000	1.0000
QQQ	9.71E-001	1.86E-001	5.56E-001	**8.12E-003**	1.0000	3.84E-001	9.18E-002	**1.66E-002**	5.95E-002	5.19E-001	9.95E-001	1.0000
GGGd	9.71E-001	3.23E-001	7.62E-001	1.10E-001	6.17E-001	1.0000	2.02E-001	1.08E-001	1.00E-001	6.76E-001	9.93E-001	1.0000
GGGf	9.92E-001	4.03E-001	8.57E-001	**7.48E-002**	9.09E-001	8.00E-001	1.0000	3.88E-001	2.03E-001	9.16E-001	9.99E-001	1.0000
NNNd	9.95E-001	4.65E-001	8.82E-001	1.14E-001	9.84E-001	8.92E-001	6.14E-001	1.0000	3.22E-001	9.60E-001	0.9996	1.0000
SSSc	0.9997	8.38E-001	9.67E-001	2.15E-001	9.41E-001	9.00E-001	7.99E-001	6.80E-001	1.0000	9.63E-001	9.99E-001	1.0000
LOMETS	9.83E-001	1.74E-001	6.63E-001	**1.68E-002**	5.00E-001	3.26E-001	**8.45E-002**	**4.07E-002**	**3.73E-002**	1.0000	9.97E-001	1.0000
IntFOLD–TS40	**4.38E-002**	**3.97E-004**	**2.46E-002**	**4.08E-005**	**4.57E-003**	**7.17E-003**	**9.50E-004**	**3.91E-004**	**5.46E-004**	**2.97E-003**	1.0000	9.97E-001
IntFOLD–TS90	**1.69E-004**	**9.08E-007**	**7.27E-005**	**3.43E-008**	**4.64E-008**	**1.90E-009**	**2.62E-009**	**7.99E-010**	**3.69E-009**	**6.80E-008**	**3.30E-003**	1.0000

H_o: The method in the row produces models that are equal or lower in quality than those produced by the method in the column. H_1: the method in the row produces higher quality models than the method in the column. *P*-values ≤0.05 indicate significant differences (in boldface). The Wilcoxon signed-rank sum test *P*-values were calculated using GDT_TS scores of 117 CASP9 full chain targets.

Table 2. Assessment of the added value provided by each multiple-template modelling method (CASP9 full chain dataset)

Method	Multi1	Multi2	Multi3	Multi4
SP3	9.47E-001	7.89E-001	**9.81E-004**	9.67E-001
SPK2	8.21E-001	9.09E-001	**8.20E-005**	9.97E-001
HHsearch	**1.87E-005**	1.99E-001	**2.65E-005**	**4.50E-005**
COMA	**3.01E-004**	2.64E-001	**2.52E-007**	**5.79E-004**
QQQ	**5.02E-003**	1.00E+000	**1.11E-004**	4.07E-001
GGGd	**1.58E-003**	5.00E-001	**5.81E-005**	1.22E-001
GGGf	**3.47E-005**	1.40E-001	**2.32E-004**	**1.12E-004**
NNNd	**1.22E-003**	5.00E-001	**4.86E-005**	5.48E-002
SSSc	**8.75E-003**	8.14E-001	**6.42E-006**	7.25E-002
LOMETS	**3.20E-007**	3.95E-001	**3.22E-007**	7.64E-002
IntFOLD-TS40	5.46E-002	**1.47E-002**	**4.39E-003**	5.62E-001
IntFOLD-TS90	**1.80E-002**	9.69E-001	5.47E-001	4.14E-001

H_o: The method in the column produces models that are equal or lower in quality than those produced by the method in the row. H_1: the method in the column produces higher quality models than the method in the row. *P*-values ≤0.05 indicate significant differences (in boldface). The Wilcoxon signed-rank sum test*P*−values were calculated using GDT_TS scores of 117 CASP9 full chain targets.

The histogram for GDT_TS scores in Figure 1 shows that the multi3 method adds value to each of the single-template modelling methods. The highest relative added-value is for COMA; the lowest, for QQQ. While the single highest added-value calculated is from the multi4 method for HHsearch, this method also recorded the highest cost for SPK2. Fluctuations between overall added value and cost were noted for both multi1 and multi4. The added values or costs associated with using multi2 were all marginal. Across CASP8 and CASP9 datasets, similar trends in the added-value or cost associated with each multi-template modelling method were noted (see Supplementary Figures).

The added-value is the essential criterion providing answer to the question 'Which is the best multi-template modelling method?'. Table 2 informs not only on the significant differences between paired multiple-template and single-template or clustering/consensus methods but also identifies the best multiple-template selection approach. Again, a *P*-value ≤0.05 suggests a significant difference between two methods and the method with ≤0.05 *P*-value possesses a better model quality. Similar tables for other datasets and scoring metrics are shown in the Supplementary Tables.

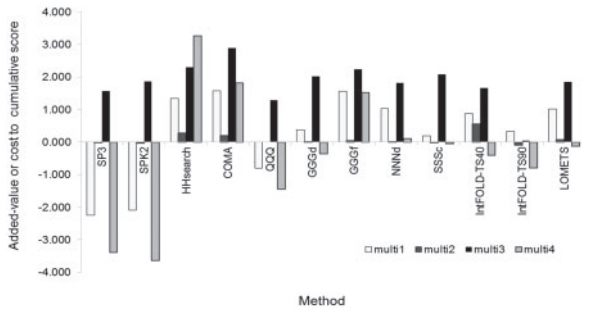

Fig. 1. Added-value or cost to cumulative GDT_TS scores by using each multiple-template modelling method (CASP9 full chain dataset)

The 'multi3' method consistently outperformed almost all the single-template and consensus/clustering methods across all datasets. The only exception was IntFOLD-TS90 for CASP9 full chain targets (Table 2). Likewise, multi1 and multi4 methods performed well over some single-template and consensus/clustering methods although their performance was inconsistent across CASP8 and CASP9 datasets. Multi2 trailed the other three methods in terms of performance; and in some cases did not improve on the performance of any methods (Supplementary Tables).

3.3 Multiple-template modelling methods using ModFOLDclust2 re-ranked alignments

The histogram in Figure 2 is interpreted in the same way as Figure 1, but in this case, the target-template alignments from each single-template method are first re-ranked according to the ModFOLDclust2 global scores for the resulting models (IntFOLD-TS40, IntFOLD-TS90 and LOMETS already use clustering to rank models by predicted global quality, so there is no need to re-rank). In Figure 2, note that while multi7 generally leads all other multiple-template modelling methods, the performance of multi6 is a close second. Although fluctuations in performance are shown by multi5 and multi8, there is an improvement in added-values for both methods compared with multi1 and multi4. Furthermore, multi7, which uses per-residue errors to guide target—template alignments consistently improves over all single-template methods.

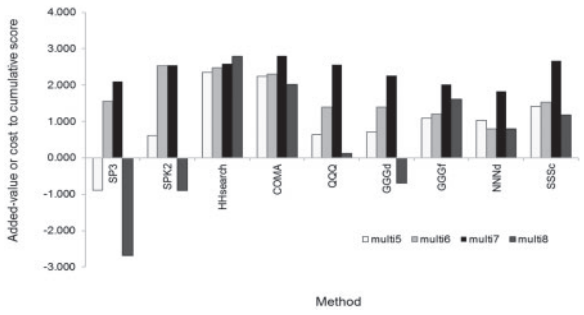

Fig. 2. Added-value or cost to GDT_TS scores by using multiple-template modelling methods with alignment re-ranking (CASP9 full chain dataset)

Table 3. Re-ranking of single-template alignments using global quality assessment scores from ModFOLDclust2 improves multiple-template modelling methods (CASP9 full chain dataset)

Method	Multi5	Multi6	Multi7	Multi8
SP3	9.27E-001	1.51E-001	**1.87E-002**	9.84E-001
SPK2	8.28E-002	**5.28E-004**	**5.58E-004**	3.77E-001
HHsearch	**1.02E-003**	**2.03E-003**	**1.97E-003**	**5.45E-004**
COMA	**2.05E-003**	**5.55E-004**	**3.45E-005**	**4.50E-004**
QQQ	**1.81E-003**	**2.83E-003**	**9.90E-005**	**1.53E-002**
GGGd	**6.20E-003**	**8.68E-003**	**1.38E-003**	6.70E-002
GGGf	**3.41E-004**	**1.33E-002**	**1.33E-003**	**1.22E-004**
NNNd	**8.66E-004**	6.33E-002	**5.88E-004**	**9.90E-003**
SSSc	**3.85E-004**	**2.79E-003**	**2.02E-006**	**1.45E-003**

H_o: The method in the column produces models that are equal or lower in quality than those produced by the method in the row. H_1: the method in the column produces higher quality models than the method in the row. *P*-values ≤ 0.05 indicate significant differences (in boldface). The Wilcoxon signed-rank sum test *P*-values were calculated using GDT_TS scores of 117 CASP9 full chain targets.

HHsearch and COMA derived most of the improvement from multi-template modelling based on the CASP9 dataset (Figure 2). COMA was the most consistently improved using all multi-template modelling methods for both the CASP8 and CASP9 targets. QQQ, GGGd, GGGf, NNNd and SSSc were the least improved overall, and in some cases, did worse when benchmarked against the CASP8 dataset. The cost to SP3 and SPK2 from multiple-template modelling was reduced by re-ranking alignments, and SPK2_multi5 showed a positive added-value compared with SPK2_multi1. Improvements are evident across all methods shown by higher cumulative scores; however, it is the multi7 method (the multi3 method using re-ranked alignments) that shows consistent improvements according all plots. The results in Table 3 suggest that re-ranking alignments by using ModFOLDclust2 global quality scores is providing additional value to each of the multiple template modelling methods tested.

3.4 IntFOLD2-TS—a consensus approach using multi-template models

The IntFOLD2-TS method is a consensus approach that uses ModFOLDclust2 to rank the multiple-template models produced using all methods (multi1–multi7) and all target-template alignment methods (a total of 84 models). Here, the results are shown variously as cumulative model quality scores in Figure 3 (Supplementary Figures) and as calculated *P*-values for the pairwise Wilcoxon signed-rank sum tests shown in Table 4 (Supplementary Tables).

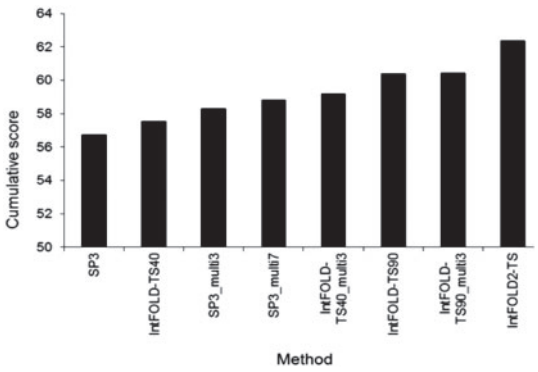

Fig. 3. Cumulative GDT_TS scores for single-template, multiple-template and consensus modelling methods (CASP9 full chain dataset)

Figure 3 shows the incremental improvements to cumulative GDT_TS scores from the multi-template modelling methods multi3 and multi7, put in context within the consensus single-template approaches and the consensus multi-template approach used by the IntFOLD2-TS method. The cumulative score for SP3 single-template modelling provides the baseline score. The models built from SP3 alignments are improved by 2.76% using the multi3 multi-template modelling method, beyond the level achieved by the consensus single-template approach used by IntFOLD-TS40. A further improvement in the cumulative score (3.70%) can be seen using the multi7 method where alignments are re-ranked using ModFOLDclust2, prior to selection. In addition, IntFOLD-TS40 and IntFOLD-TS90 methods have improved scores using the multi3 method. However, the highest cumulative score is for IntFOLD2-TS (9.97%), which is gained by using ModFOLDclust2 to cluster the multi-template models from all methods. A similar trend is shown in Supplementary Tables using alternative datasets and scoring metrics with the IntFOLD2-TS method providing the highest cumulative score in all tests.

The pairwise matrix of *P*-values in Table 4 shows the significance of differences between methods. Significant improvements ($P < 0.05$) over the baseline method SP3 are shown by using the consensus single-template methods (IntFOLD-TS40 and IntFOLD-TS90) and the multiple-template modelling approaches (SP3_multi3 and SP3_multi7). Furthermore, the new IntFOLD2-TS method provides significantly higher quality models than all other methods (Supplementary Tables).

The incremental refinements brought about by consensus/clustering method and the multiple-template methods shown in Figure 3, are illustrated structurally for an example target (T0623) in Figure 4. The PyMOL (http://www.pymol.org) cartoon views of superpositions depict where the improvements of the models (grey) have been made relative to the native structure (white). The target's native structure was solved by X-ray diffraction (PDB ID 3NKH).

4 DISCUSSION

In this article, we evaluate alternative methods for the selection of target-template alignments for multiple-template modelling, using the alignments generated by a number of alternative single-template modelling methods. We explore the use of ModFOLDclust2 local and global model quality assessment scores as a discriminatory measure to refine the selection of better target-template alignments

Table 4. Calculated pairwise *P*-values for single-template, multiple-template and iteratively clustered multiple-template modelling methods with SP3 as a baseline (CASP9 full chain GDT_TS scores)

Method	SP3	SP3_multi3	IntFOLD-TS40	SP3_multi7	IntFOLD-TS40_multi3	IntFOLD-TS90	IntFOLD-TS90_multi3	IntFOLD2-TS
SP3	1.0000	9.99E-001	9.57E-001	9.82E-001	9.99E-001	0.9998	9.99E-001	1.0000
SP3_multi3	**9.81E-004**	1.0000	3.95E-001	4.05E-001	8.90E-001	9.92E-001	9.81E-001	1.0000
IntFOLD-TS40	**4.38E-002**	6.07E-001	1.0000	6.93E-001	9.96E-001	9.97E-001	9.96E-001	1.0000
SP3_multi7	**1.87E-002**	5.97E-001	3.08E-001	1.0000	9.26E-001	9.81E-001	9.60E-001	1.0000
IntFOLD-TS40_multi3	**1.19E-003**	1.10E-001	**4.39E-003**	7.44E-002	1.0000	7.75E-001	8.65E-001	0.9998
IntFOLD-TS90	**1.69E-004**	**8.31E-003**	**3.30E-003**	**1.89E-002**	2.26E-001	1.0000	4.54E-001	9.99E-001
IntFOLD-TS90_multi3	**1.19E-003**	**1.88E-002**	**4.22E-003**	**3.99E-002**	1.36E-001	5.47E-001	1.0000	9.98E-001
IntFOLD2-TS	**3.83E-008**	**1.34E-005**	**1.82E-007**	**1.40E-005**	**1.83E-004**	**1.30E-003**	**1.84E-003**	1.0000

H_o: The method in the row produces models that are equal or lower in quality than those produced by the method in the column. H_1: the method in the row produces higher quality models than the method in the column. *P*-values ≤0.05 indicate significant differences (in boldface). The Wilcoxon signed-rank sum test *P*-values were calculated using GDT_TS scores of 117 CASP9 full chain targets.

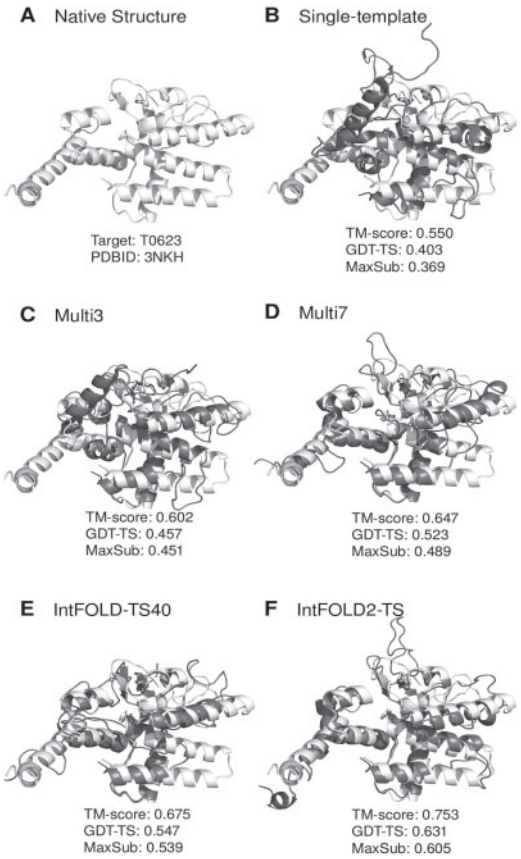

Fig. 4. SP3 model quality is incrementally improved by multi-template methods with alignment selection refined by the use of ModFOLDclust2. All images were rendered in PyMOL

prior to multi-template modelling. Although there is a basic nuance that using multiple templates will translate into a better model, this is shown not to be true in all cases and success is dependent on the target-template selection method used.

Comprehensive analyses of our multiple-template models showed that there are appropriate combinations of templates that will lead to improvement of a model quality over that of the single-template model. When that combination is not found, then multiple-template modelling will not provide a better option over single-template modelling and may sometimes lead to a deterioration of model quality. The aim of multiple-template modelling should not be to simply 'mix and match' segments of templates to achieve more coverage; it should, most importantly, be to improve model quality.

From benchmarking the alternative alignment selection approaches on the CASP8 and CASP9 full chain and domain datasets, we have found that the most consistent way to significantly improve models is to make use of accurate predicted local model quality scores to guide alignment selection. Furthermore, using accurate global quality prediction scores, for re-ranking the order of alignments prior to selection, significantly improves most of the multi-template modelling methods tested. This corroborates the findings of Larsson *et al*. (2008). The upper theoretical limits of these strategies are investigated through use of perfect model quality assessment scores (Supplementary Tables and Figures). The results indicate that there is room for improvement and it is therefore worthwhile putting further effort into developing more accurate model assessment to guide multi-template modeling.

Overall, SP3 proved to be a competitive single-template method when tested on the CASP9 datasets and it provides a useful baseline single template modelling method. The increased performance of both IntFOLD-TS40 and IntFOLD-TS90 relative to SP3 points again to the usefulness of generating consensus single-template models using ModFOLDclust2. However, these simple consensus methods only make use of the ModFOLDclust2 global scores and do not exploit the accurate per-residue error predictions. According to the CASP9 assessors accurate per-residue error predictions are crucial indicators of a model's usefulness. Here, we are now capitalizing on the accuracy of ModFOLDclust2 for predicting model quality at both the global and per-residue level, using identified errors in single template models to guide alignment selection for multi-template modelling.

The multi3 and multi7 methods were observed to be the best performing methods and they consistently and significantly improved model quality for all of the original sequence-structure alignment methods, on all datasets and according to all model quality scoring metrics. The multi3 works by simply rejecting any subsequent target-template alignment after the top hit, if the overlapping local regions from the alignment are not predicted to improve the local model quality according to ModFOLDclust2. The multi7 method uses essentially the same approach but the target-template alignments for each method are firstly re-ordered using the ModFOLDclust2 global scores.

In comparison, the methods multi2, multi4, which were based on improving alignment coverage, and the multi1 method were not as consistent. The multi1 and multi4 methods worked well for some sequence-structure alignment methods such as HHsearch. However, these methods often lead to lower quality models, on average, particularly when applied to the SPK2 and SP3 alignments. The HHsearch method, which generates optimal local alignment, to alternative templates, may benefit more from any additional coverage by alternative templates, whereas the SP3 and SPK2 methods will attempt to thread the full sequence even if some local regions are not adequately covered by a single template.

The multi2 method is a bit more conservative than the multi1 and multi4 methods; in most instances, using the method leads to a marginal improvement in model quality. In some cases, there is a cost associated with using the multi2 method; however, it is never as detrimental as using multi1 and multi4 methods. In many cases, the multi1, multi2 and multi4 alignment selection methods can be made more consistent by firstly re-ranking target-template alignments using ModFOLDclust2 (multi5, multi6 and multi8, respectively), but again this does not work in all cases.

Multi-template modelling using the target-template alignment methods described here leads to the generation of a new population of models that are of higher model quality on average than the population of single template models. This is demonstrated by the improved performance achieved by the IntFOLD2-TS method which uses ModFOLDclust2 to rank the population of 84 multi-template models, versus the IntFOLD-TS90 method which only ranks the initial population of single-template models. The IntFOLD2-TS method has been incorporated into the new IntFOLD server (Roche *et al.*, 2011), which will be tested during the CASP10 prediction season, where it is registered as IntFOLD2, as well as being continuously assessed by the CAMEO3D server (http://www.cameo3d.org/), where it is currently listed as development server12. The server will shortly be made freely publicly available via http://www.reading.ac.uk/bioinf/IntFOLD/. In addition, we freely provide the source code and executables for the multi-template modelling methods described here for use with any set of sequence-to-structure alignments and associated quality assessment file provided in CASP QMODE2 format.

Funding: University of Reading (to M.T.B and D.B.R); Medical Research Council, Harwell (to M.T.B.); Diamond Light Source Ltd. (to M.T.B)

Conflict of Interest: none declared.

REFERENCES

Chakravarty,S. *et al.* (2008) Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC. Struct. Biol.*, **8**, 31.

Cheng,J. (2008) A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.*, **8**, 18.

Chubb,D. *et al.* (2010) Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. *Bioinformatics*, **26**, 2664–2671.

Contreras-Moreira,B. *et al*. (2003) In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. *J. Mol. Biol.*, **328**, 593–608.

Fischer,D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.

Fiser,A. and Sali,A. (2003) Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.,* **374**, 361–491.

Ginalski,K. *et al.* (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Hildebrand,A. *et al.* (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77**, 128–132.

Kryshtafovych,A. *et al.* (2011) Evaluation of model quality predictions in CASP9. *Proteins*, **79**, 91–106.

Larsson,P. *et al.* (2008) Using multiple templates to improve quality of homology models in automated homology modeling. *Prot. Sci.*, **17**, 990–1002.

Larsson,P. *et al.* (2011) Improved predictions by Pcons.net using multiple templates. *Bioinfromatics.*, **27**, 426–427.

Liu,T. *et al.* (2008) Improving the accuracy of template-based predictions by mixing and matching between initial models. *BMC Struct. Biol.*, **8**, 24.

Lundstrom,J. *et al.* (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Prot. Sci.*, **10**, 2354–2362.

Mariani,V. *et al.* (2011) Assessment of template based protein structure predictions in CASP9. *Proteins*, **79**, 37–58.

Martínez,L. *et al*. (2007) Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, **8**, 306.

McGuffin,L.J. (2010) Model quality prediction. In: Rangwala,H. and Karypis,G. (eds.) *Introduction to Protein Structure Prediction: Methods and Algorithms*. Wiley, New Jersey, pp. 323–342.

McGuffin,L.J. and Roche,D.B. (2009) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, **26**, 182–188.

McGuffin,L.J. and Roche,D.B. (2011) Automated tertiary structure prediction with accurate local model quality assessment using the intfold-ts method. *Prot.: Struct. Funct. Bioinformatics*, **79**, 137–146.

Peng,J. and Xu,J. (2011a) A multiple-template approach to protein threading. *Prot.: Struct. Funct. Bioinformatics*, **79**, 1930–1939.

Peng,J. and Xu,J. (2011b) Raptorx: exploiting structure information for protein alignment by statistical inference. *Prot.: Struct. Funct. Bioinformatics*, **79**, 161–171.

Roche,D.B. *et al.* (2011) 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.*, **39**, W171–W176.

Roy,A. *et al.* (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.

Siew,N. *et al.* (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Soding,J. (2004) Protein homology detection by HMM-HMM comparison, *Bioinformatics*, **21**, 951–960.

Soding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.

Tramontano,A. and Cozzetto,D. (2011) Evaluation of protein structure prediction methods: issues and strategies. In: Kolinski,A. (ed.). *Multi-scale Approaches to Protein Modelling*. Springer, New York, pp. 315–339.

Tramontano,A. *et al.* (2008) The assessment of methods for protein structure prediction. In: Zaki,M. and Bystroff,C. (eds.). *Protein Structure Prediction*. Humana Press, New Jersey, pp. 43–57.

Wallner,B. and Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci.*, **12**, 1073–1086.

Wallner,B. *et al*. (2007) Pcons.net: protein structure prediction meta server. *Nucleic Acids Res.*, **35**, W369–W374.

Wang,Z. *et al*. (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics*, **26**, 882–888.

Wang,Z. *et al*. (2011) APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics*, **27**, 1715–1716.

Wu,S. and Zhang,Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.*, **35**, 3375–3382.

Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.

Zemla,A. *et al.* (2001) Processing and evaluation of predictions in CASP4. *Proteins*, **45**, 13–21.

Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins,* **57**, 702–710.

Zhou,H. and Zhou,Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins*, **61**, 152–156.