A multi-template combination algorithm for protein comparative modeling

Jianlin Cheng

Address: Department of Computer Science, Informatics Institute, University of Missouri, Columbia, MO 65211-2060, USA
Email: Jianlin Cheng - chengji@missouri.edu

Abstract

Background: Multiple protein templates are commonly used in manual protein structure prediction. However, few automated algorithms of selecting and combining multiple templates are available.

Results: Here we develop an effective multi-template combination algorithm for protein comparative modeling. The algorithm selects templates according to the similarity significance of the alignments between template and target proteins. It combines the whole template-target alignments whose similarity significance score is close to that of the top template-target alignment within a threshold, whereas it only takes alignment fragments from a less similar template-target alignment that align with a sizable uncovered region of the target.

We compare the algorithm with the traditional method of using a single top template on the 45 comparative modeling targets (i.e. easy template-based modeling targets) used in the seventh edition of Critical Assessment of Techniques for Protein Structure Prediction (CASP7). The multi-template combination algorithm improves the GDT-TS scores of predicted models by 6.8% on average. The statistical analysis shows that the improvement is significant (p-value < 10^-4). Compared with the ideal approach that always uses the best template, the multi-template approach yields only slightly better performance. During the CASP7 experiment, the preliminary implementation of the multi-template combination algorithm (FOLDpro) was ranked second among 67 servers in the category of high-accuracy structure prediction in terms of GDT-TS measure.

Conclusion: We have developed a novel multi-template algorithm to improve protein comparative modeling.

Background

Protein structure prediction is one of the most important problems in structural bioinformatics [1-3]. Comparative (or homology) modeling is currently the most accurate and practical structure prediction method [4-19].

In general comparative modeling involves four steps [11,20,21]: (1) identify a homologous template protein for a target protein; (2) generate an alignment between the template and the target; (3) create a model based on the alignment and the template structure; (4) evaluate and refine the model. The two key factors determining the quality of comparative modeling are the template struc-
ture and the alignment accuracy [22]. Traditionally, automated comparative modeling methods use the top-ranked template and its alignment with the target protein to model its structure. This approach cannot always achieve the best results because it may not be able to select the best template and to generate the optimal alignment [23]. Previous research [6,24-28], particularly the human prediction [23,29-31] in the six edition of Critical Assessment of Techniques of Protein Structure Prediction (CASP6) [32-36], has shown that using multiple templates can often improve the quality of comparative modeling over a single template. Although human experts commonly select multiple templates and combine them manually to predict structure in their practice, multiple-template combination has not been widely used by automated servers until the latest CASP7 experiment, 2006. In CASP7, several servers including FOLDpro and 3Dpro [37,38], HHSearch [39], 3D-JIGSAW-POPULUS [13], MetaTasser [30,40,41], Zhang-Server [29], FAMS [42], used multiple templates to improve template-based structure prediction. Some of these methods such as Zhang-Server, FOLDpro, and 3Dpro show the good performance on the comparative modeling targets, particularly on the high-accuracy modeling targets. However, few of the automated multi-template methods are published so far.

A published multiple-template algorithm [29-31] tries to extract distance (or contact) restraints from multiple templates. The consistent distance restraints from multiple templates are used to guide structure modeling. This method is currently coupled with the in-house model assembly tool TASSER [29,31] and cannot be used with the popular, publicly available, comparative model generation tools such as Modeller [8,11,20,43,44], nest [17], SEGMOD-ENCAD [45], SWISS-Model [18,46,47], 3D-JIGSAW [13], and Builder [9]. Most of these tools take as inputs the alignments between templates and a target to generate structure models, instead of directly accepting distance restraints.

Thus, instead of extracting distance restraints from multiple templates as in [29-31], we develop a different, parametric algorithm to select templates and to combine their alignments directly. The multiple alignments between the target and the templates can be directly fed into the wildly used, standard comparative modeling tools such as Modeller [8] with the built-in multi-template modeling function, to generate models. The preliminary implementation of the method was ranked second in the automated high-accuracy structure prediction during the CASP7 community-wide experiment.

Furthermore, we systematically and rigorously compare the performance of the multiple- and single-template methods on the 45 comparative modeling targets of the CASP7 experiment. On average the multiple-template combination algorithm improves the GDT-TS score [48] of the predicted models by 6.8% over the single template approach. The pairwise statistical analysis shows that the improvement is significant. Thus, our experiment strictly demonstrates that the combination of multiple templates and their alignments can significantly improve comparative modeling over the single top template approach. Moreover, we compare the multi-template method against the ideal approach, which always uses the best, possible template in the Protein Data Bank [49]. The analysis shows that the multi-template combination algorithm can even achieve slightly better performance than the ideal approach on the 27 CASP7 comparative modeling targets. However, the improvement is not statistically significant.

**Results and Discussions**

We develop a pipeline for multi-template protein comparative modeling as shown in Figure 1. Given an input target (or query) protein, the pipeline uses PSI-BLAST [50] to search homologous structure templates. The target-template alignments ranked by PSI-BLAST e-values are combined with respect to the target protein. Briefly speaking, the algorithm always uses the most significant template-target alignment. The other significant alignments relative to the most significant one are also automatically included. The less significant template-target alignments are chosen only if they can align with a continuous region of the target that is not covered by the previously selected template-target alignments. And only the alignment fragments that align with the uncovered regions are used. The combined alignments and template structures are fed into Modeller [8] to generate structure models for the target protein. The details of the algorithm are described in the Methods section.

Our multi-template combination method was first blindly tested in the CASP7 experiment, 2006. Since then we systematically evaluate the algorithm on the CASP7 comparative modeling targets. A target is classified as a comparative modeling target if a structure template covering all the domains of the target can be found by PSI-BLAST search as in [51]. Here, we firstly compare the multi-template combination algorithm with a single-top-template algorithm on all the comparative modeling targets. Secondly, we compare two approaches on the high-accuracy comparative modeling targets. Thirdly, we compare the multi-template combination algorithm against the ideal approach that always uses the best template. Fourthly, we compare the multi-template combination method against the other automated methods and the human predictors in the CASP7 experiment. Finally, we discuss why the multi-template approach improves model quality.
The models of the multi-template combination algorithm were generated during the CASP7 period when the structures of the targets were not known. For the comparison we use Modeller [8] to generate 3D structure models for the single top template and the best template respectively, based on the alignments generated by PSI-BLAST during the CASP7 experiment. We use LGA [48], a sequence-dependent structure alignment tool, to evaluate the models against the experimental structures to get GDT-TS scores.

Comparison with the Single Top Template Approach on the CASP7 Comparative Modeling Targets

We compare the multi-template combination algorithm with the single-template algorithm on the 45 CASP7 comparative modeling targets, for which PSI-BLAST can identify at least two significant templates. The other two comparative modeling targets (T0326 and T0328) that have only one template found by PSI-BLAST are excluded. The single-template algorithm always uses the most significant template with the lowest e-value of PSI-BLAST.

Table 1 shows the GDT-TS scores of 45 targets using the single- and multi-template methods, respectively. The number of templates used by the multi-template combination algorithm ranges from 2 to 39. The average number of templates used is 12.4. According to the results, the multi-template combination improves GDT-TS score for the majority of cases (38 out of 45 targets) as shown in Figure 2, consistent with the previous human prediction experiment [23,31]. The average score of using multiple templates is 71.15 versus 66.59 of using the single most significant template. The average improvement of GDT-TS score is 6.8% (raw score increase = 4.56).

We conduct a paired t-test (t-value = 4.39, the degree of freedom = 44) on the GDT-TS scores of 45 targets. The p-value of getting an average score difference >= 4.56 is 3.5 \times 10^{-5} under the hypothesis that the difference is 0. So combining multiple templates and their alignments significantly improves the quality of comparative modeling over the single-template approach.

Comparison with the Single Top Template Approach on the CASP7 High-Accuracy Modeling Targets

A special category of template-based modeling (i.e. high-accuracy modeling), where models have GDT-TS scores > 80 compared to experimental structures, is particularly
useful for understanding protein function at the atomic level. To emphasize its importance, CASP7 dedicated a category of high-accuracy template-based modeling to evaluate methods on the targets for which there is at least one template with LGA-S score > 80 and at least one method produced a model with GDT-TS > 80. Using this criteria, CASP7 classifies 28 domains from 24 protein targets into the high-accuracy modeling category. Among them T0326 and T0328 have only one template. The templates for T0311 and T0367 cannot be found by PSI-BLAST. The structure of T0334 is not released at the time of writing the paper. So we exclude these five targets and select the remaining 23 high-accuracy domains to compare the multiple-template combination algorithm with the single-template approach.

Table 2 reports the results of the multi- and single-template methods on the 45 comparative modeling targets of CASP7. The average GDT-TS score for the multi- and single-template approaches is 86.7 and 81.0 respectively. The average difference is 5.7. We conduct a paired t-test on the scores (t-value = 2.51, the degree of freedom = 22). The p-value of getting an average difference >= 5.7 is 0.01 under the GDT-TS scores of 45 comparative modeling targets (multi-template versus single-template).

For 38 out of 45 targets, the multi-template approach yields higher GDT-TS scores than the single-template approach. The dots above the line represent the targets where the multi-template method yields higher scores, on the line where two methods yields the same scores, and below the line where the single-template method yields higher scores.

Table 1: The results of the multiple- and single-template methods on the 45 comparative modeling targets of CASP7.

| Target Id | Temp Num | Multi  | Single | Multi – Single |
|-----------|----------|--------|--------|----------------|
| T0288     | 27       | 83.8   | 75.0   | 8.8            |
| T0290     | 14       | 97.3   | 90.8   | 6.5            |
| T0291     | 6        | 78.6   | 91.4   | -12.8          |
| T0292     | 37       | 69.8   | 67.0   | 2.8            |
| T0293     | 15       | 32.6   | 33.6   | -1.0           |
| T0294     | 18       | 81.8   | 68.1   | 13.7           |
| T0295     | 2        | 83.0   | 76.3   | 6.7            |
| T0297     | 8        | 62.8   | 62.9   | -0.1           |
| T0298     | 39       | 70.8   | 50.2   | 20.6           |
| T0302     | 26       | 80.1   | 69.9   | 10.2           |
| T0303     | 20       | 68.7   | 59.0   | 9.7            |
| T0305     | 5        | 93.0   | 91.5   | 1.5            |
| T0308     | 19       | 90.5   | 74.8   | 15.7           |
| T0310     | 6        | 55.7   | 66.5   | -10.8          |
| T0313     | 12       | 80.4   | 74.7   | 5.7            |
| T0315     | 4        | 94.7   | 83.8   | 10.9           |
| T0316     | 9        | 17.9   | 17.1   | 0.8            |
| T0317     | 8        | 81.7   | 79.6   | 2.1            |
| T0318     | 4        | 58.9   | 57.4   | 1.5            |
| T0322     | 22       | 68.5   | 55.5   | 13.0           |
| T0323     | 20       | 57.6   | 53.1   | 4.5            |
| T0324     | 18       | 79.1   | 57.1   | 22.0           |
| T0329     | 22       | 63.0   | 48.8   | 14.2           |
| T0330     | 23       | 62.7   | 44.9   | 17.8           |
| T0332     | 2        | 82.9   | 80.7   | 2.2            |
| T0337     | 4        | 52.9   | 49.5   | 3.4            |
| T0338     | 17       | 49.8   | 51.7   | -1.9           |
| T0339     | 7        | 76.7   | 77.9   | -1.2           |
| T0340     | 4        | 90.5   | 90.4   | 0.1            |
| T0341     | 6        | 67.2   | 66.9   | 0.3            |
| T0345     | 3        | 95.1   | 95.0   | 0.1            |
| T0346     | 8        | 98.0   | 89.7   | 8.3            |
| T0359     | 15       | 82.5   | 80.9   | 1.6            |
| T0362     | 12       | 73.6   | 73.4   | 0.2            |
| T0364     | 23       | 71.7   | 68.6   | 3.1            |
| T0366     | 3        | 92.6   | 90.2   | 2.4            |
| T0371     | 6        | 61.7   | 59.4   | 2.3            |
| T0373     | 12       | 62.6   | 61.5   | 1.1            |
| T0374     | 5        | 62.5   | 57.7   | 4.8            |
| T0375     | 6        | 57.4   | 54.1   | 3.3            |
| T0376     | 14       | 64.3   | 64.7   | -0.4           |
| T0379     | 6        | 63.5   | 60.9   | 2.6            |
| T0380     | 9        | 63.7   | 56.7   | 7.0            |
| T0381     | 2        | 57.6   | 56.8   | 0.8            |
| T0384     | 11       | 61.8   | 60.7   | 1.1            |

Average 12.42  71.15  66.59  4.56

Column 1 lists the CASP7 protein target IDs. Column 2 lists the number of templates used by the multi-template combination algorithm. Column 3 lists the GDT-TS scores of the multi-template algorithm for the targets. Column 4 lists the GDT-TS scores of using the most significant single template with the lowest e-value. Column 5 lists the GDT-TS score difference between the multiple- and single-template approaches.
hypothesis that there is no difference between the multi- and single-template methods. According to the standard 0.05 threshold, the difference is significant. Among 20 out of 23 high-accuracy targets, the multi-template combination method generates better models than the single-template method as shown in Figure 3.

Comparison with the Best Template Approach on the CASP7 Comparative Modeling Targets

A very challenging problem of comparative modeling is to improve the model accuracy over the best templates [14,52,53]. The series of the community-wide experiments from CASP1 to CASP6 show that few methods can consistently improve the model accuracy over the best templates or even the top ranked templates. A couple of recent methods [29,54,55] and the latest CASP7 experiment show that in some cases comparative modeling and refinement can improve model quality over the best templates. Thus, here we test if our multiple template combination algorithm can improve structure modeling over the best templates on the comparative modeling targets.

We use the best templates for the comparative modeling targets provided at the CASP7 web site. We select 27 targets whose best templates can be identified by PSI-BLAST to compare the multi- and best template methods. We also use the PSI-BLAST alignment between the best templates and the targets to generate structure models.

As shown in Table 3 and Figure 4, among 27 targets, the multi-template combination method produces better models for 16 targets, worse models for 10 targets, and the same quality model for 1 target. Thus, the multi-template algorithm produces better models for 62% of the targets. However, on average, the GDT-TS score increase of using multiple templates over the best possible template is not significant.

Comparison with the other Servers and Human Predictors on the CASP7 High-Accuracy Modeling Targets

We compare the accuracy of our multi-template combination algorithm against the other servers that participated in the CASP7 community-wide experiment. Two of our
servers, FOLDpro and 3Dpro, along with 65 other servers were evaluated in the category of high-accuracy structure prediction in CASP7. FOLDpro and 3Dpro used the same multi-template combination algorithm, but run on different versions of non-redundant sequence databases. Table 4 reports the official total GDT-TS scores of the first models of top 20 (out of 67) servers on the 28 high-accuracy domains in CASP7. The GDT-TS scores are directly taken from the official CASP7 evaluation. The data is kindly provided by Dr. Yang Zhang at his web site [56]. The results show that our methods FOLDpro and 3Dpro using multiple-template combination algorithm were ranked second and third respectively. The performance of our methods that use the simple combination of PSI-BLAST alignments is very close to the best method (Zhang-Server) that extracts distance restraints from multiple templates used in conjunction with a more sophisticated and complicated model generation tool TASSER.

Furthermore, using the more strict measure GDT-HA specially designed for high-accuracy models, FOLDpro and 3Dpro are ranked third and fifth according to the official CASP assessment [57,58]. We also compare the performance of automated servers with the human predictors. The comparison is not fair because human predictions started from the server predictions and took much longer time (about three weeks of human versus two days of server). However, it is still interesting to see what values human predictions can add on the high accuracy targets.

Table 5 reports the top 10 predictors among the 116 human and 67 server predictors in the high-accuracy structure modeling in CASP7. The data is kindly provided by Dr. Yang Zhang at his website [59]. The results show that three automated servers Zhang-Server, FOLDpro, and 3Dpro yielded the performance comparable to the best human predictors that used much longer time and took a pool of server predictions as inputs. Our servers FOLDpro and 3Dpro were ranked fifth and sixth, respectively. Zhang-Server is better than all human predictors except for Zhang human predictor from the same group. FOLDpro is better than 113 out of 116 human predictors.

Since the main goal of this paper is to demonstrate the effectiveness of using multiple templates instead of evalu-
The correct usage of multiple homologous templates in general but, not always can improve comparative modeling [60,61]. As the reviewer point out, the effectiveness of multi-template modeling may correlate with the number of templates, the structural consistency amongst templates, and query-template similarity. Clarifying their relationship can help decide when to use multiple templates.

Figure 5 shows a good example (T0315, length = 257). The best template for the target is protein 1J6O in the PDB, whose Root Mean Square Distance (RMSD) with the experimental structure (2GCX) is 1.33 Å for 264-residue long aligned region. The other good template is 1YIX having RMSD 1.44 Å for 248-residue long aligned region. The sequence identity and PSI-BLAST e-value is 81% and 10^{-153}. The RMSD between other most significant templates (2SRC, 1Y57, 1KSW) and the experimental structure is 3.22 Å for 250-residue aligned regions. These three templates can improve model quality.

However, both templates have some bad regions that do not align well with the experimental structure. Figure 5(1) and Figure 5(2) show two different bad loop regions for the two templates, respectively. Interestingly, the two bad regions are corrected in the model generated by using multiple templates as in Figure 5(3). In addition to the obvious improvement in the two regions, the backbone of most other regions of the model are also closer to the experimental structure than the two templates. The RMSD between the model and the experimental structure is 0.88 Å for 248-residue long aligned regions. This example clearly shows that combining the complementary good templates can improve model quality.

The other good example is a two-domain target T0324, where the improvement of using multiple templates on the second domain (4-helix bundle) is drastic. The GDT-TS score of the second domain is increased by 48 (Table 4). A close examination reveals that the top template does not provide the right orientations for the helices, which are corrected by the other templates.

However, multi-template combination may occasionally decrease the model quality when there is a very good template that is much closer to the target than all other templates. One such an example is T0291 (length = 310). The most significant and best template is 1JPA, whose RMSD with the experimental structure (2GSF) is 0.72 Å for 264-residue long aligned region. The sequence identity and PSI-BLAST e-value is 81% and 10^{-153}. The RMSD between other three significant templates (2SRC, 1Y57, 1KSW) and the experimental structure is 2.33 Å for 262-residue aligned region, 2.33 Å for 248-residue aligned region, and 3.20 Å for 250-residue aligned region respectively. These three templates are much more different from the target structure than the best template. However, because the alignments between these three templates and the experimental structure are very significant (e-value < 10^{-142} and sequence identity is 81%), the RMSD is increased.

Table 4: The total GDT-TS scores of the top 20 out of 67 servers on the 28 high-accuracy comparative modeling domains in CASP7.

| Predictors       | Rank | GDT-TS | ZScore |
|------------------|------|--------|--------|
| Zhang-Server     | 1    | 2415   | 17.7   |
| FOLDpro          | 2    | 2389   | 16.6   |
| 3Dpro            | 3    | 2379   | 15.9   |
| UNI-EID expm     | 4    | 2350   | 13.9   |
| CIRCLE           | 5    | 2341   | 12.7   |
| RAPTOR           | 6    | 2328   | 12.6   |
| ROBETTA (61,95,96)| 7    | 2328   | 12.1   |
| beautshotbase    | 8    | 2328   | 11.9   |
| FAMS             | 9    | 2327   | 12.0   |
| FUNCTION          | 10   | 2321   | 11.9   |
| HHpred1          | 11   | 2314   | 11.2   |
| Pcons6           | 12   | 2309   | 11.0   |
| Huber-Torda-Srv  | 13   | 2306   | 10.8   |
| RAPTOR-Ace       | 14   | 2300   | 10.7   |
| SP3 [63]         | 15   | 2295   | 10.4   |
| HHpred2          | 16   | 2294   | 10.6   |
| SPARKS2          | 17   | 2293   | 10.2   |
| HHpred3          | 18   | 2291   | 10.3   |
| beautshot        | 19   | 2288   | 10.9   |
| SP4 [63]         | 20   | 2287   | 9.8    |

Column 1 lists the predictor names, column 2 the ranks, column 3 the total GDT-TS scores and column 4 Z-Scores. For a model of each target, Z-score is the normalized GDT-TS score: (x - u)/σ calculated as in [52,90]. Here, x is the GDT-TS score of the model; u is the average GDT-TS score of all predicted models for the target; σ is the standard deviation. For each predictors, the Z-scores for all targets are summed into a total Z-Score to compare them as shown in the table.

Table 5: The total GDT-TS scores of the top 10 out of 183 predictors (67 servers + 116 human predictors) in the category of the high-accuracy structure modeling in CASP7.

| Predictors       | Rank | GDT-TS | Z-Score |
|------------------|------|--------|---------|
| Zhang            | 1    | 2425   | 17.7    |
| Zhang-Server*    | 2    | 2415   | 17.0    |
| fms-ace          | 3    | 2396   | 15.9    |
| TASSER           | 4    | 2390   | 15.9    |
| FOLDpro*         | 5    | 2389   | 15.9    |
| 3Dpro*           | 6    | 2379   | 15.4    |
| fms-multi        | 7    | 2378   | 15.0    |
| CIRCLE-FAMS      | 8    | 2368   | 14.1    |
| hPredGrp         | 9    | 2368   | 14.1    |
| CHIMERA          | 10   | 2361   | 13.5    |

* denotes server predictors. Z-score is defined in the caption of Table 4.

However, both templates have some bad regions that do not align well with the experimental structure. Figure 5(1) and Figure 5(2) show two different bad loop regions for the two templates, respectively. Interestingly, the two bad regions are corrected in the model generated by using multiple templates as in Figure 5(3). In addition to the obvious improvement in the two regions, the backbone of most other regions of the model are also closer to the experimental structure than the two templates. The RMSD between the model and the experimental structure is 0.88 Å for 248-residue long aligned regions. This example clearly shows that combining the complementary good templates can improve model quality.

However, multi-template combination may occasionally decrease the model quality when there is a very good template that is much closer to the target than all other templates. One such an example is T0291 (length = 310). The most significant and best template is 1JPA, whose RMSD with the experimental structure (2GSF) is 0.72 Å for 264-residue long aligned region. The sequence identity and PSI-BLAST e-value is 81% and 10^{-153}. The RMSD between other three significant templates (2SRC, 1Y57, 1KSW) and the experimental structure is 2.33 Å for 262-residue aligned region, 2.33 Å for 248-residue aligned region, and 3.20 Å for 250-residue aligned region respectively. These three templates are much more different from the target structure than the best template. However, because the alignments between these three templates and the experimental structure are very significant (e-value < 10^{-142} and sequence identity is 81%), the RMSD is increased.
tity > 40%), these three templates together with the best one are combined to generate models for the target.

The RMSD between the model and the experimental structure is 2.63 Å for 271-residue long alignment, which is better than two templates (2SRC and 1KSW), but worse than the best template (1JPA). This example shows that combining multiple templates may not help if one template is much closer to the true structure than all other templates.

**Why does the Multi-Template Approach Work in General?**

We consider the following factors contributing to the improvement. First, statistically, the average model generated from multiple templates is better than the single top template on average. This is due to the ability of Modeller extracting spatial restraints from multiple templates and weighting them effectively. The weighting scheme can weight the most likely distance restraints more, resulting in picking correct aligned regions from different templates in most cases [8]. The effective combination of good aligned regions of different templates can improve comparative modeling, particularly in the cases where multiple templates provide the complementary good coverage of the target.

Second, multiple templates contain sequence and structure conservation and variation information (e.g. conserved distance restraints), which is not available in a single template. The evolutionary information is often useful to improve both secondary and tertiary structure prediction [58,61-63]

Third, PSI-BLAST can generate good local alignments for homologous proteins (comparative modeling), particularly for close homologs (easy comparative modeling). In fact, PSI-BLAST, a profile-sequence local alignment method, generates better alignments for the comparative modeling (or easy) targets than the profile-profile alignment methods (e.g., ClustalW [64], T-Coffee [65], COACH [66], and Palign [67]) we tested, which is consistent with the previous observations (Dr. Kimmen Sjölander, personal communication). However, profile-profile alignment methods are shown to produce better alignments on hard targets (<= 20% identity) [39] and to have stronger fold recognition power than profile-sequence alignment methods [37,39,66-74].

Fourth, the ranking of homologous templates by PSI-BLAST e-values for a target protein is also reasonably good, although not perfect. The greedy combination of PSI-BLAST templates and alignments into a multiple alignment centered on the target protein is effective for comparative modeling as shown in our experiments.

**Figure 5**

An good example (CASP7 target T0315) of using multiple templates to improve model quality. (1) The superimposition of the experimental structure (PDB code: 2GZX) and the best template (PDB code: 1J6O). Blue and red lines represent the backbone of the experimental and template structures, respectively. One bad region is identified. (2) The superimposition of the experimental structure and a good template (PDB code: 1YIX). One bad region is identified. (3) The superimposition of the experimental structure and the model generated by 3Dpro during CASP7, based on multiple templates. Two bad regions in (1) and (2) are corrected in the model (3). Most other regions of the model are also closer to the experimental structure than the two templates.
Conclusion
In this study, we have developed a novel and effective algorithm of selecting and combining multiple templates and their alignments generated by PSI-BLAST, which significantly improves the quality of comparative modeling over the traditional single-template approach on the CASP7 benchmark. The alignment files of both the single-template approach and the multi-template approach are available [see Additional files 1 and 2]. The comparative modeling web server of using multiple templates is accessible at the MULTICOM website [75].

Methods
Multi-Template Combination Algorithm
We develop a novel and effective multi-template combination algorithm to select and combine template-target alignments for comparative modeling (Figure 1). The algorithm uses as inputs the template proteins identified by the PDB-BLAST approach [50,68], similar as ISS [76] and DOUBLE-BLAST approaches [77].

The PDB-BLAST approach searches for structure templates for a target protein in two steps. First, it uses PSI-BLAST [50] to search the target protein against the NCBI non-redundant sequence database [78] to build a profile. The e-value threshold (-h option) for building iterative profiles is set to 10^-10; the number of iterations (-j option) is set to 3; and the e-value threshold (-e option) for inclusion in the final profile is set to 0.001. Second, it uses PSI-BLAST to search the profile against a template library compiled from the proteins in the Protein Data Bank [49] to identify structure templates homologous to the target protein. The number of iterations in this step is set to 5 and all other parameters to the default values. We select the template-target alignments with e-value < 0.001 for the combination.

Each returned template protein may have one or more local alignments with the target protein. Each alignment is associated with an e-value that measures its similarity significance. We use the logarithm of the e-value as the significance score. The smaller the score, the more significant is the alignment.

The only inputs to the algorithm are the template-target alignments and the associated significance scores. Thus, with some modification, the algorithm can be applied to other template-based structure prediction methods with different scoring schemes (e.g. z-score in threading). Figure 6 briefly describes the algorithm. The algorithm always uses the most significant template-target alignment. The other significant alignments whose significance score is less than $\sigma$ and close to the score of the most significant template-target alignment within the threshold $\delta$ are also automatically included. The less significant template-target alignments are used only when they can align with a continuous region of the target whose size is bigger than $\rho$ and which is not covered by the previously selected template-target alignments. And only the alignment fragments that align with the uncovered regions are excised and included. $\rho$ controls the size of the selected fragments, which can be used to avoid selecting very small fragments.

One template-target alignment may provide several fragments that align with disjoint, uncovered regions of the target. The alignment fragments can be extended at both ends up to $r$ residues if possible, which may create a linker to overlap with the other fragments from the same template or the alignments from other templates. After the extension, the overlapped or non-gapped alignment fragments excised from the same template-target alignment are combined into one alignment. The join of fragments can introduce long-range constraints, which is useful for structure modeling.

The alignments between the target and the selected templates are combined into a multiple alignment using the target sequence as an anchor, the same as the central star multiple alignment approach [79,80] employed by the construction of position specific scoring matrix (PSSM) [81,82] in PSI-BLAST [50]. The multiple alignment together with the template structures are fed into comparative modeling tools such as Modeller to generate structure models for the target protein. Modeller weights the spatial restraints extracted from multiple templates according to sequence identity appropriately [8], resulting in generating a better average model than a single template in most situations.

The algorithm is fully parametrized (Figure 6). All parameters can be tuned. In our experiments, we set $\sigma$ to -20, $\delta$ to 12, $\rho$ to 5, and $r$ to 5. The parameters were tuned on the CASP6 targets and blindly tested on the CASP7 targets. The most important parameters are $\sigma$ and $\delta$ because the combination of the whole significant templates contributes most to the improvement of structure modeling. $\rho$ and $r$ usually play a less important role because short uncovered regions of a target protein are usually well handled by the loop building module of Modeller.

Limitation and Future Work
The multi-template combination algorithm developed here is very effective for comparative modeling where target and template proteins have strong homologous relationship and their alignments are rather accurate. But the method does not always produce good results for hard template-based structure prediction, i.e. protein fold recognition.
Multi-Template Selection and Combination Algorithm

**Parameters**

- $\sigma$: minimum significance score.
- $\delta$: maximum score difference.
- $\rho$: minimum cover length.
- $\tau$: maximum extension length.

**Inputs**

A list of template-target alignments with the associated significance scores.

**Initialization**

Rank alignments according to the significance scores in the increasing order.
Set the set of selected template-target alignments ($\phi$) to empty.
Remove the top template-target alignment and add it into $\phi$.
Set the smallest significance score ($\omega$) to the score of the top ranked template-target alignment.

**Repeat**

Remove the template-target alignment with the significance score $x$ from the top of the list.

If $x < \sigma$ and $x - \omega < \delta$
Add the template-target alignment into $\phi$.

else
Select template fragments that align with a continuous region (size > $\rho$)
of the target that is not covered by the selected templates in $\phi$ if available.
Extend the ends of selected alignment fragments up to $\tau$ residues if possible
Join the overlapped alignment fragments into one fragment alignment.
Add the selected alignment fragments into $\phi$.

**Until**

No template-target alignment is left in the list.

**Output**

Transform the pairwise alignments in $\phi$ between the target and the selected templates into
a multiple alignment using the target sequence as an anchor (the central star method).

*Figure 6*
Multi-Template Selection and Combination Algorithm.
We had applied the similar algorithm to the hard fold recognition targets during CASP7 using the global alignments generated by COACH [66] (results not shown here). The algorithm works well when the structure templates and their alignments with the target are consistent, but performs poorly when structure templates or sequence-structure alignments have a lot of spatial inconsistency (particularly in unalignable regions). The models generated from multiple inconsistent target-template alignments usually contain a lot of atom-atom clashes – a quantitative indicator of spatial inconsistency. One possible reason is that the current version of Modeller cannot well resolve a large amount of conflicting distance restraints within multiple templates as also observed in [83].

Thus, although multiple templates are useful in general, a more sophisticated way of combining them and removing inconsistency is required to achieve better performance for fold recognition and threading, where the template-target relatedness and alignment are uncertain and less reliable than comparative modeling. The spatial inconsistency may be removed either in model reconstruction phase as in TASSER or in alignment optimization phase as in [83].

One possible direction is to use structure alignment tools such as DALI [84], SSAP [85], VAST [86], CE [87], and TM-align [88] to check the structure consistency between templates and to remove inconsistent templates and alignments (work in progress).

Authors’ contributions
JC designed the algorithm, wrote the program, carried out the experiments, and authored the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1
The description of supplemental materials. This file describes how to use supplemental materials. It is a text file that can be viewed by any text viewer. Click here for file
[http://www.biomedcentral.com/content-supplementary/1472-6807-8-18-S1.txt]

Additional file 2
The compressed file of supplemental materials. The file can be decompressed by winzip on Windows or by tar xzf on Linux. The file includes query-template alignments of both the single-template approach and the multi-template approach that were used to generate the structure models for the CASP7 comparative modeling targets. Click here for file
[http://www.biomedcentral.com/content-supplementary/1472-6807-8-18-S2.gz]

Acknowledgements
The work is partially supported by a MU new faculty start-up grant, a MU research board grant, and the MU Bioinformatics Consortium. JC is grateful to Dr. Pierre Baldi for the support during his PhD research at University of California Irvine. JC thanks Dr. Arne Elofsson, Dr. Björn Wallner, Dr. Yang Zhang, and Dr. Dylan Chivian for releasing their evaluation results at the ForCASP website [89]. JC thanks Dr. Jeffrey Skolnick, Dr. Johannes Söding, Dr. Jinbo Xu, Xin Gao, and Dr. Jooyoung Lee for the helpful discussions. JC also thanks CASP7 organizers and all CASP7 participants for the valuable experimental data.

References
1. Vindup D, Melamud E, Moult J, Sander C: Completeness in structural genomics. Nature Struct Biol 2001, 8:559-566.
2. Lambernt S: A tour of structural genomics. Nature Rev Genet 2001, 2:801-809.
3. Westbrook J, Feng Z, Chen L, Yang H, Berman H: The protein data bank and structural genomics. Nucleic Acids Res 2003, 31:489-491.
4. Browne W, North A, Phillips D, Brew K, Vanaman T, Hill R: A possible three-dimensional structure of the hen’s egg-white lysozyme. J Mol Biol 1969, 42:65-86.
5. Blundell T, Sibanda B, Sternberg M, Thornton J: Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987, 326:347-352.
6. Greer J: Comparative modeling methods: Application to the family of the mammalian serine proteases. Proteins 1990, 7:317-334.
7. Levitt M: Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 1992, 226:507-533.
8. Sali A, Blundell T: Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993, 234:779-815.
9. Koehl P, Delarue M: Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 1994, 239:249-275.
10. Sali A: 100,000 protein structures for the biologist. Nat Struct Biol 1998, 5:1029-1032.
11. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Mello F, Sali A: Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000, 29:291-325.
12. Sali A: Target practice. Nat Struct Biol 2001, 8:482-484.
13. Bates P, Kelley L, MacCallum R, Sternberg M: Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-FSSM. Proteins 2001, 45(Suppl 5):39-46.
14. Tramontano A, Leplae R, Morea V: Analysis and assessment of comparative modeling in CASP4. Proteins 2001, 22-38.
15. Kolinski A, Betancourt M, Kihara D, Roskiewicz P, Skolnick J: Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001, 44:133-149.
16. Laberty C, Leonard N, Bolle X, Depiereux E: ESpred1D: Prediction of proteins 3D structures. Bioinformatics 2002, 19(9):1250-1256.
17. Petrey D, Xiang Z, Tang C, Xie L, Gimpelev M, Mitros T, Soto C, Goldsmith-Fischman S, Kernytsky A, Schlessinger A: Using multiple structure alignments, structure of based on that of hen’s egg-white lysozyme. J Mol Biol 1969, 8:347-352.
18. Schwede T, Kopp J, Guex N, Peitsch M: SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003, 31(13):3381-3385.
19. Tramontano A: Protein Structure Prediction: Concepts and Applications Wiley-VCH; 2006.
20. Sanchez R, Sali A: Advances in comparative protein-structure modeling. Curr Opin Struct Biol 1997, 7:206-214.
21. Petrey D, Honig B: Protein structure prediction: inroads to biology. Mol Cell 2005, 20:811-819.
22. Venclovas C: Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. Proteins 2003, 53:380-388.
23. Venclovas C, Margelevisius M: Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. Proteins 2005, 61(suppl 7):99-105.

24. Blundell T, Sternberg M: Computer-aided design in protein engineering. Trends Biotechnol 1985, 3:228-235.

25. Sutcliffe M, Haneef I, Carney D, Blundell T: Knowledge based modeling of homologous proteins. Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1987, 1:377-384.

26. Blundell T, Barlow D, Sibanda B, Thornton J, Taylor W, Tickle I, Sternberg M, Pitts J, Haneef I, Hemmings A: Three-Dimensional structural aspects of the design of new protein molecules. Phil Trans Roy Soc Lond Ser A 1986, 317:333-344.

27. Overington J, Johnson M, Sali A, Blundell T: Tertiary structural constraints on protein evolutionary diversity; templates, key residues and structure prediction. Proc Roy Soc Lond Ser B 1990, 245:105-124.

28. Al-Lazikani B, Shiehmaner F, Honig B: Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to SH2 domains of Janus kinase. PNAS 1998, 95(26):14796-14801.

29. Zhang J, Joachimiak J: Automated Structure Prediction of Weakly Homologous Proteins on a Genomic Scale. PNAS 2004, 101:7579-7589.

30. Skolnik J, Kihara D, Zhang Y: Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004, 56:502-518.

31. Zhang Y, Arakaki A, Skolnik J: TASSER: an automated method for the prediction of protein tertiary structure in CASP6. Proteins 2005, 61(suppl 7):91-98.

32. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT: Critical assessment of methods of protein structure prediction (CASP): round II. Proteins Suppl 1997, 1:2-6.

33. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT: Critical assessment of methods of protein structure prediction (CASP): round III. Proteins Suppl 1999:29.

34. Moult J, Fidelis K, Zemla A, Hubbard T: Critical Assessment of Methods of Protein Structure Prediction (CASP): Round IV. Proteins 2001, 45(suppl 5):2-7.

35. Moult J, Fidelis K, Zemla A, Hubbard T: Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 2003, 53(suppl 6):334-339.

36. Moult J, Fidelis K, Tramontano A, Rost B, Hubbard T: Critical assessment of methods of protein structure prediction (CASP) – round VI. Proteins 2005 in press.

37. Cheng J, Baldwin AL: Machine Learning Information Retrieval Approach to Protein Fold Recognition. Bioinformatics 2006, 22:1456-1463.

38. Cheng J: Machine Learning Algorithms for Protein Structure Prediction University of California Irvine; 2006. [Ph.D Thesis]

39. Soding J: Protein homology detection by HMM-HMM comparison. Bioinformatics 2002, 19:951-960.

40. Pandit S, Zhang Y, Skolnik J: TASSER-Lite: an automated tool for protein comparative modeling. Biophys J 2006, 91:4180-4190.

41. Zhou H, Pantal S, Lee S, Borrogsero J, Chen H, Wroblewska L, Skolnik J: Analysis of TASSER-based CASP7 protein structure prediction results. Proteins 2007, 69:90-97.

42. Ogata K, Umeysama H: An automatic homology modeling method consisting of database searches and simulated annealing. J Mol Graphics Mod 2000, 18:258-272.

43. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M: Evaluation of comparative protein modeling by MODELLER. Proteins 1995, 23:318-326.

44. Fiser A, Do R, Sali A: Modeling of loops in protein structures. Protein Science 2000, 9:1753-1773.

45. Levitt M: Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 1992, 226(2):595-617.

46. Guex N, Peitsch M: SWISS-MODEL and Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997, 18:2714-2723.

47. Schwede T, Diemand A, Guex N, Peitsch M: Protein structure computing in the geomic era. Res Microbiol 2000, 151:107-112.

48. Zemla A: LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research 2003, 31:3370-3374.

49. Berman H, Westbrook J, Feng Z, Gililand G, Bhat T, Weissig H, Shindyalov I, Bourne P: The Protein Data Bank. Nucleic Acids Res 2000, 28:235-242.

50. Akschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997, 25(17):3389-3402.

51. Tress M and OR Grana IE, Lopez G, Valencia A: Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005, 61:27-45.

52. Tramontano A, Morea V: Assessment of homology based prediction in CASP5. Proteins 2003, 52(suppl 6):352-368.

53. Wallner B, Elofsson A: All are not equal: A benchmark of different homology modeling programs. Protein Science 2005, 14:1315-1327.

54. Zhang Y, Skolnick J: The protein structure prediction problem could be solved using the current PDB library. Proc Nat Acad Sci USA 2005, 102:9219-9224.

55. Misura K, Chivian D, Rohl C, Kim D, Baker D: Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006, 103:5361-5366.

56. CASP7 high accuracy GDT-TS results (server) [http://zhang.bioinformatics.ku.edu/casp72.html]

57. Battey J, Kopp J, Bordoli L, Read R, Clarke N, Schwede T: Automated server predictions in CASP7. Proteins 2007, 69:68-82.

58. Read R, Chavali G: Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 2007, 69:27-37.

59. CASP7 high accuracy GDT-TS results (human and server) [http://zhang.bioinformatics.ku.edu/casp732.html]

60. Wu S, Zhang Y: LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Research 2007, 35:3375-3382.

61. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka M, Bhat D, Chivian D, Kim D, Sheffer W, Malmstrom L, Wollacott A, Wang C, Andre I, Baker D: Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 2007, 69:118-128.

62. Rost B, Sander C: Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994, 19:55-72.

63. Zhou H, Zhou Y: Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005, 58:321-328.

64. Thompson J, Higgins D, Gibson T: CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673-4680.

65. Notredame C, Higgins D, Heringa J: T-Coffee: A novel method for multiple sequence alignment. J Mol Biol 2000, 302:215-217.

66. Edgar R, Sjolander K: COACH: profile-profile alignment of protein families using hidden Markov models. Bioinformatics 2004, 20:1309-1318.

67. Ohlson T, Wallner B, Elofsson A: Profile-profile methods provide improved fold-recognition. A study of different profile-profile alignment methods. Proteins 2004, 57:188-197.

68. Rychlewski L, Jaroszewski L, Li W, Godzik A: Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000, 9:232-241.

69. Yona G, Levitt M: Within teh twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002, 315:1257-1275.

70. Mitelman D, Sadreyev R, Grishin N: Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. Bioinformatics 2003, 19:151-153.

71. Ginalsiki K, Paj J, Wyrwicz L, von Grothuss M, Bujnicki J, Rychlewski L: ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 2003, 31:3804-3807.

72. Sadreyev R, Grishin N: COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003, 326:317-336.

73. Wallner B, Fang H, Ohlson T, Frey-Skott J, Elofsson A: Using evolutionary information for the query and target improves fold recognition. Proteins 2004, 54:342-350.
74. Marti-Renom M, Madhusudhan M, Sali A: Alignment of protein sequences by their profiles. Protein Sci 2004, 13:1071-1087.
75. MULTICOM [http://casp.nec.missouri.edu/multicom/multicom.html]
76. Park J, Teichmann S, Hubbard T, Chothia C: Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997, 273:349-354.
77. Capuglossus K, Barrett C, Hughley R: Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998, 14(10):846-56.
78. Pruitt K, Tatusova T, Maglott D: NCBI Reference sequence project: update and current status. Nucleic Acids Res 2003, 31:24-37.
79. Gusfield D: Efficient methods for multiple sequence alignment with guaranteed error bounds. Bulletin of Mathematical Biology 1993, 55:141-154.
80. Gusfield D: Algorithms on Strings, Trees and Sequences: Computer science and computational biology New York: Cambridge University Press; 1997.
81. Gribskov M, McLachlan M, Eisenberg D: Profile analysis: detection of distantly related proteins. PNAS 1987, 84:4335-4338.
82. Henikoff S, Wallace J, Brown J: Finding protein similarity with nucleotide sequence database. Methods in Enzymology 1990, 183:111-132.
83. Joo K, Lee J, Lee S, Seo J, Lee S, Lee J: High accuracy template based modeling by global optimization. Proteins 2007, 69:83-89.
84. Holm L, Sander C: Protein Structure Comparison by Alignment of Distance Matrices. J Mol Biol 1993, 233:123-138.
85. Taylor W, Flores T, Orengo C: Multiple protein structure alignment. Protein Sci 1994, 3:1858-1870.
86. Gibrat JF, Madej T, Bryant SH: Surprising similarities in structure comparison. Curr Opin Struct Biol 1996, 6:377-385.
87. Shindyalov IN, Bourne PE: Protein Structure Alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 1998, 11:739-747.
88. Zhang Y, Skolnick J: TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 2005, 33:2302-2309.
89. ForCASP [http://www forcasp org]
90. Tress M, Eskurdia I, Grana O, Lopez G, Valencia A: Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005, 61:27-45.
91. Wu S, Skolnick J, Zhang Y: Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biology 2007, 5:7.
92. Zhang Y: Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007, 69:108-117.
93. Debe D, Danzer J, Goddard W, Poleksic A: STRUCTFasta: Protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. Proteins 2006, 64:960-967.
94. Xu J, Li M, Kim D, Xu Y: RAPTOR: Optimal Protein Threading by Linear Programming. J Bioinformatics and Computational Biology 2003, 1(1):95-117.
95. Kim D, Chivian D, Baker D: Protein structure prediction and analysis using the Robetta server. Nucleic Acids Research 2004, 32(Web Server Issue):W526-W531.
96. Chivian D, Kim D, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss C, Bonneau R, Rohl C, Baker D: Automated prediction of CASP-3 structures using the Robetta server. Proteins 2003, 53(56):524-533.
97. Fischer F: 3DS and 3DS 3D-SHOTGUN Meta-Predictors in CAFASP3. Proteins 2003, 53:517-523.
98. Wallner B, Elofsson A: Pcons: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 2005, 21:424-425.
99. Torda A, Procter J, Huber T: Wurst: a protein threading server with a structural scoring function, sequence profiles and optimised substitution matrices. Nucleic Acids Res 2004, 32:W532-W535.
100. Xu J, Yu L, Li M: Consensus fold recognition by predicted model quality. Asian-Pacific Bioinformatics Conference (APBC) 2005:105-116.
101. Zhou H, Zhou Y: Quantifying the effect of burial of amino acid residues on protein stability. Proteins 2004, 54:315-322.