Template-based protein structure modeling is commonly used for protein structure prediction. Based on the observation that multiple template-based methods often perform better than single template-based methods, we further explore the use of a variable number of multiple templates for a given target in the latest variant of TASSER, TASSERVMT. We first develop an algorithm that improves the target-template alignment for a given template. The improved alignment, called the SP₃ alternative alignment, is generated by a parametric alignment method coupled with short TASSER refinement on models selected using knowledge-based scores. The refined top model is then structurally aligned to the template to produce the SP₃ alternative alignment. Templates identified using SP₃ threading are combined with the SP₃ alternative and HSEARCH alignments to provide target alignments to each template. These template models are then grouped into sets containing a variable number of template/alignment combinations. For each set, we run short TASSER simulations to build full-length models. Then, the models from all sets of templates are pooled, and the top 20–50 models selected using FTCOM ranking method. These models are then subjected to a single longer TASSER refinement run for final prediction. We benchmarked our method by comparison with our previously developed approach, pro-sp₃-TASSER, on a set with 874 easy and 318 hard targets. The average GDT-TS score improvements for the first model are 3.5 and 4.3% for easy and hard targets, respectively. When tested on the 112 CASP9 targets, our method improves the average GDT-TS scores as compared to pro-sp₃-TASSER by 8.2 and 9.3% for the 80 easy and 32 hard targets, respectively. It also shows slightly better results than the top ranked CASP9 Zhang-Server, QUARK and HHpredA methods. The program is available for download at http://cssb.biology.gatech.edu/.

Key words: template-based modeling; threading; alignment; SP₃; TASSER.

ABSTRACT

Template-based protein structure modeling is commonly used for protein structure prediction. Based on the observation that multiple template-based methods often perform better than single template-based methods, we further explore the use of a variable number of multiple templates for a given target in the latest variant of TASSER, TASSERVMT. We first develop an algorithm that improves the target-template alignment for a given template. The improved alignment, called the SP₃ alternative alignment, is generated by a parametric alignment method coupled with short TASSER refinement on models selected using knowledge-based scores. The refined top model is then structurally aligned to the template to produce the SP₃ alternative alignment. Templates identified using SP₃ threading are combined with the SP₃ alternative and HSEARCH alignments to provide target alignments to each template. These template models are then grouped into sets containing a variable number of template/alignment combinations. For each set, we run short TASSER simulations to build full-length models. Then, the models from all sets of templates are pooled, and the top 20–50 models selected using FTCOM ranking method. These models are then subjected to a single longer TASSER refinement run for final prediction. We benchmarked our method by comparison with our previously developed approach, pro-sp₃-TASSER, on a set with 874 easy and 318 hard targets. The average GDT-TS score improvements for the first model are 3.5 and 4.3% for easy and hard targets, respectively. When tested on the 112 CASP9 targets, our method improves the average GDT-TS scores as compared to pro-sp₃-TASSER by 8.2 and 9.3% for the 80 easy and 32 hard targets, respectively. It also shows slightly better results than the top ranked CASP9 Zhang-Server, QUARK and HHpredA methods. The program is available for download at http://cssb.biology.gatech.edu/.

INTRODUCTION

Despite considerable efforts to develop accurate template free approaches, template-based protein structure modeling is still the only reliable protein structure modeling method.¹ Template-based modeling involves (1) template identification, usually by threading; (2) construction of the target-template alignment; (3) model generation based on this alignment; and (4) refinement of the template models with the goal of generating closer structures to the target native state than that provided by the template. The accuracy of the final model depends on the alignment accuracy between the template and target as well as the quality of the subsequent refinement. An ideal threading method should identify the best templates, namely, those with the best structural similarity to the native structure. In practice, this is not always the case.² Furthermore, the target-template alignment may not be optimal, especially for hard targets.³ Hard targets are those with poor quality threading alignments and good/poor quality structure alignments, respectively. Advances have been made in increasing the accuracy of template identification and threading alignment by going from single sequence alignments⁴ to profile-profile alignments, ⁵–¹² to machine learning, ¹³,¹⁴ and meta-threading,⁶,¹⁵,¹⁶ Recent studies show that using multiple templates yields better models than those obtained from a single best template approach,¹⁷–²²

Results from the latest critical assessment of protein structure prediction (CASP9),²³ show that the top ranked automatic servers all used multiple template information in template-based modeling.²³ The Zhang-Server,²⁴ which performed well across all levels of target difficulty, employed a locally installed meta-server, LOMETS,²⁵ that uses more than eight individual threading methods for template identification/target alignment and TASSER¹⁷ for refinement. HHpredA,²⁶,²⁷ which performed well for easy targets, but moderately well for hard targets, used an updated method for hidden Markov profile generation and a neural network trained to filter distance restraints from multiple templates,²³ The final models are built by MODELLER²⁸ using filtered distance restraints. RaptorX,²⁹,³⁰ however, performed
well for hard targets and slightly worse for easy targets. RaptorX and its variants RaptorX-MSA and RaptorX-Boost employed an outstanding alignment method that uses a boost tree to fit the alignment matching scores and a new method of generating multiple template-sequence alignments. Most methods used a fixed number of multiple templates for a given target to produce the intermediate or final models. For example, standard TASSER uses 20, 30, or 50 templates as input for easy (those targets with good quality threading alignments), or hard targets (those having templates with poor quality threading alignments but often having good structural alignments). Pro-sp3-TASSER, that performed fairly well in CASP8 and CASP9, uses a few different sets with a fixed number of templates for generating TASSER inputs.

Obvious lessons learned from the top performing approaches in recent CASPs are that the use of multiple templates is an effective way of improving template-based modeling methods and that further improvement can be achieved by improving target-template alignments and filtering or selecting those templates that are optimally compatible with the target. For example, the outstanding performance of Zhang-Server and the template-based part of QUARK could be attributed partly to the consensus information from many threading methods that serve as a filter to remove poor quality templates and bad alignments, the inclusion of template fragments by segment threading, and finally TASSER refinement. The use of better alignments and a neural-network trained filter for distance restraints in HHpredA makes it stand out among other multiple template-based methods, especially for easy targets.

In this work, we follow a similar scheme as in our previous pro-sp3-TASSER approach and use multiple sets of templates for a given target. In contrast to pro-sp3-TASSER that uses five different component threading scores, our new approach TASSER (TASSER with variable number of multiple templates) uses SP3 threading for template identification. To build better models, we developed an algorithm to improve the target-template alignment for a given template. Alignments generated by HHSEARCH are also included, since our tests show that they are slightly better than the default SP3 threading alignments for easy targets and could provide complementary sampling of alignment space. Then, short TASSER simulations build a pool of structures derived from different combinations of templates and their associated target alignments. The rationale of using subsets of a given multiple template set is to explore different combinations of templates to sample distance restraints or contacts to find the subset that are optimally compatible with the target. FTCOM serves as an effective filter equivalent to the neural-network trained distance restraint filter in HHpredA to select the top 20–50 models that are then subjected to longer TASSER refinement. Table I summarizes the major methodological differences between current TASSER and previous pro-sp3-TASSER methods. We benchmarked TASSER against pro-sp3-TASSER on a set of 874 easy and 318 hard targets. We also tested TASSER on the 112 CASP9 targets and compare the results to top performing CASP9 methods, Zhang-Server, QUARK, HHpredA, RaptorX, Seok-server, MULTICOM_CLUSTER, and BAKER-ROSETTASERVER. We show that TASSER represents a promising approach to template-based modeling.

### METHOD

The overall flowchart of TASSERVMT is shown in Figure 1(a). Figure 1(b) shows the SP3 alternative alignment generation procedure that is part of Figure 1(a). In what follows, we shall describe the approach in more detail.

#### SP3 alternative alignment generation

For a given template, our goal is to improve the target-template alignment over that provided by SP3 threading. Targets are classified as easy if the SP3 threading scores
Z-score of the top template is \( \geq 6.0 \), as medium if \( 4.5 < Z\text{-score} < 6.0 \), and as hard if \( Z\text{-score} \leq 4.5 \).\(^{34}\) We refer them as SP\(^3\) easy, medium, and hard. In practice, here we lump the medium and hard targets in the same category as hard for the prediction protocol. This classification is for prediction purposes and depending on predicted target difficulty allows for the use of different parameters in TASSER simulations and different template choices. (In contrast, in the results section, easy and hard targets are defined for assessment purposes using the actual model quality of the predicted models.) In prosp3-TASSER,\(^{22}\) we generate alternative alignments only for SP\(^3\) hard targets using a parametric alignment generation method\(^{35}\) and select the top alignment simply by TASSER-QA\(^{36}\) without any refinement. Here, we also generate alternative alignments for easy targets and employ short TASSER refinement.

Given a target sequence and a template structure, we use a parametric alignment method\(^{35,37}\) to generate an ensemble of alignments. This is realized by a grid search on the five dimensional parameter space \((w_0, w_1, w_{\text{struc}}, w_{\text{shift}}, S_{\text{shift}})\) of the SP\(^3\) threading score,\(^{34}\) where \(w_0, w_1\) are gap penalties, \(w_{\text{struc}}\) is the secondary structure score weight, \(w_{\text{struc}}\) is weight of the structurally derived profile of the template, and \(S_{\text{shift}}\) is a shift parameter. Each parameter is sampled with 0.0, 0.5, 1.0, 1.5, and 2.0 times the original values except for the gap opening penalty, \(w_0\), which is sampled with 0.5, 1.0, 1.5, 2.0, and 2.5 times the original value. Because a gap-opening penalty of zero could lead to unrealistic, highly gapped alignments, this value of \(w_0\) value is not sampled. Alignments are ranked according to their coverage (fraction of target residues aligned), with the top 500 distinct alignments kept. We then employ MODELLER\(^{28}\) to build the full-length models and add the side-chain, and apply FTCOM ranking\(^{38}\) and the DFIRE statistical potential\(^{39}\) to select the top 21 models from these 500 models. The reason to use FTCOM and DFIRE is that they are not consensus-based scores so that they could potentially select a few good models. We apply the two independent components of FTCOM,\(^{38}\) fragment comparison and template comparison, separately; each component selects seven models; we then use DFIRE to select another seven models. A short TASSER refinement run (limited to 5 h) is conducted using the selected 21 models as input and the top model from the top cluster centroid is selected by SPICKER\(^{40}\)

![Flowchart of TASSER\(^{\text{VMT}}\)](image)

(a) Flowchart of TASSER\(^{\text{VMT}}\). (b) Flowchart of SP\(^3\) alternative alignment generation.
clustering on the TASSER trajectories. We then use TM-align\textsuperscript{31} to generate the structural alignment to the template. The resulting template alignment comprises the new target-template alignment that we term the “SP\textsuperscript{3} alternative alignment” of the corresponding target-template pair. We also employed the HHSEARCH alignment, as an independent alignment to the SP\textsuperscript{3} alternative alignment for each template because it is slightly better than the SP\textsuperscript{3} default alignment for easy targets, and it provides alignments that could potentially be complementary to the SP\textsuperscript{3} alternative alignments for both easy and hard targets.

**Template identification and construction of a variable number of multiple template sets**

In this work, we use SP\textsuperscript{3} threading to identify templates.\textsuperscript{34} We use four different output scores to rank templates to explore different combinatorial space of templates: (1) the default SP\textsuperscript{3} rank score, that is, the raw threading score minus the reverse threading score; (2) the raw threading score; (3) the raw threading score/alignment length; (4) the raw threading score/target length. For each target, four sets of ranked templates are collected from the top (10 for easy, 20 for hard targets) ranked templates corresponding to the above four scores. We denote them as (T\textsubscript{1}, \ldots, T\textsubscript{10}), and (T\textsubscript{1}, \ldots, T\textsubscript{20}), with i = 1, 2, 3, and 4 denoting the templates from the four scores for easy and hard targets, respectively. T\textsubscript{i}\textsuperscript{HHM} and T\textsubscript{k}\textsuperscript{sp3a} denote the template alignment of the k\textsuperscript{th} ranked template T\textsubscript{k} in each i = 1, 2, 3, and 4, using the HHSEARCH and the above generated SP\textsuperscript{3} alternative alignments, respectively.

To explore the combinatorial space of multiple templates, for easy targets, the following sets of template alignments are used in subsequent model building: (1) sets of (T\textsubscript{1}\textsuperscript{HHM}, (T\textsubscript{1}\textsuperscript{HHM}, T\textsubscript{2}\textsuperscript{HHM}), \ldots, (T\textsubscript{10}\textsuperscript{HHM}, T\textsubscript{10}\textsuperscript{HHM})); (2) sets of (T\textsubscript{i}\textsuperscript{sp3a}, (T\textsubscript{i}\textsuperscript{sp3a}, T\textsubscript{j}\textsuperscript{sp3a}), \ldots, (T\textsubscript{i}\textsuperscript{sp3a}, T\textsubscript{10}\textsuperscript{sp3a})); (3) sets of (T\textsubscript{i}\textsuperscript{HHM}, T\textsubscript{j}\textsuperscript{sp3a}, T\textsubscript{k}\textsuperscript{sp3a}, \ldots, (T\textsubscript{i}\textsuperscript{HHM}, T\textsubscript{20}\textsuperscript{sp3a})); with i = 1, 2, 3, and 4. For hard targets, the template alignment sets are (T\textsubscript{i}\textsuperscript{HHM}, T\textsubscript{j}\textsuperscript{sp3a}), (T\textsubscript{i}\textsuperscript{HHM}, T\textsubscript{j}\textsuperscript{sp3a}, T\textsubscript{k}\textsuperscript{sp3a}, \ldots, (T\textsubscript{i}\textsuperscript{HHM}, T\textsubscript{20}\textsuperscript{sp3a}), with i = 1, 2, 3, and 4. In these combinations, only templates within the same i (=1, 2, 3, and 4) are combined. The total template alignment sets for an easy target is 120 and for a hard target is 80.

It should be noted that the above sets of template alignment represent only a small portion of all possible combinations. For 10 ranked templates and a given alignment method, the 10 sets we used consist of <1% of all possible 1023 combinations. However, we have found that among all sets with k (\leq 10) templates, the one we used is the most significant one because it contains the best-ranked k templates of the 10 templates.

**Building models by short TASSER simulations**

For each set of multiple templates (120 for easy and 80 for hard targets), we run short TASSER\textsuperscript{17} simulations (limited to 10 h to control the total modeling time) to build full length models. Each TASSER simulation contributes up to five models from the top five cluster centroids by SPICKER\textsuperscript{40} clustering on the TASSER trajectories. The total number of models built for each easy target is 600 and 400 for each hard target. We found that for hard targets, building more models from the templates sets (T\textsubscript{i}\textsuperscript{HHM}, T\textsubscript{j}\textsuperscript{HHM}, T\textsubscript{2}\textsuperscript{HHM}, \ldots, T\textsubscript{10}\textsuperscript{HHM}), (T\textsubscript{i}\textsuperscript{sp3a}, T\textsubscript{j}\textsuperscript{sp3a}, T\textsubscript{2}\textsuperscript{sp3a}, \ldots, T\textsubscript{10}\textsuperscript{sp3a}) with i = 1, 2, 3, and 4 does not give obvious improvement. Therefore, these sets of template alignments are not used for hard targets but are used for easy targets where an increase in the number of template sets did yield some improvement. The reason could be that for easy targets, good models dominate in the additional ensemble of models; whereas for hard targets, bad models dominate and lead to more false positives in the next step of model selection.

**TASSER model refinement**

After the pool of models is generated by short TASSER simulations, the top 20/50 models for each SP\textsuperscript{3} easy/hard target are selected by FTCOM,\textsuperscript{38} a model assessment method that combines fragment comparison and template comparison scores. The selected models are subsequently input into TASSER for refinement. After around 20 h of simulation, TASSER trajectories were clustered with the SPICKER method\textsuperscript{40} and top five cluster centroids were used for the final prediction. Possible clashes were removed and the main-chain, side-chain atoms built with Pulchra.\textsuperscript{42} The longer TASSER refinement simulation time is limited to 20 h based on the requirement of 72 h of total modeling time for servers as required in CASP.\textsuperscript{23} However, if total modeling time is not an issue, these time restrictions can be relaxed and possibly better results might be obtained.

**RESULTS**

We tested our method on a large-scale benchmark set and compared the results with our previous method pro-s3-TASSER.\textsuperscript{22,31} For the CASP9 target set, we also compared our results to the top performing approaches of in CASP9.\textsuperscript{23}

**Large-scale benchmark**

A data set of 1192 targets was constructed from the SP\textsuperscript{3} threading library. These targets share a sequence identity <35% among themselves and to the threading template library used for template identification. All
library structures were released before the targets. We compare TASSER VMT with our previous developed method pro-sp3-TASSER using exactly the same template library. In our assessment, a target is classified as easy if the predicted first model has a TM-score to native > 0.5; otherwise, it is defined as a hard target. The benchmark set has 874 easy and 318 hard targets, respectively. In accordance with convention, the first model’s GDT-TS score is employed for model quality assessment. In Table II, we compare the performance of TASSER VMT and pro-sp3-TASSER as assessed by the cumulative GDT-TS scores of the first model as well as the foldability of hard targets. Foldability is defined as the number of targets having a first model TM-score to native ≥ 0.4. On average, TASSER VMT improves the GDT-TS scores by 3.5 and 4.3% for easy and hard targets, respectively. These improvements are statistically significant as shown by their respective P-values. For the 318 hard targets, foldability improves from 61 to 85 targets.

**Table II**
Comparison of TASSER VMT with Pro-sp3-TASSER on the 1192 Protein Benchmark Set

<table>
<thead>
<tr>
<th></th>
<th>318 Hard targets</th>
<th>874 Easy targets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cumulative</td>
<td>P-value</td>
</tr>
<tr>
<td></td>
<td>GDT-TS</td>
<td></td>
</tr>
<tr>
<td>TASSER VMT</td>
<td>615.27</td>
<td>–</td>
</tr>
<tr>
<td>Pro-sp3-TASSER</td>
<td>594.62</td>
<td>$7.7 \times 10^{-30}$</td>
</tr>
</tbody>
</table>

*a A target is defined as easy if the TM-score to native of the first model by at least one of the two methods is higher than 0.5; otherwise, it is classified as hard.

*b Two-sided P-value of the Student-t test between TASSER VMT and pro-sp3-TASSER. A P-value of ≤ 0.05 is considered statistically significant.

Numbers in brackets are the foldabilities defined as the number of targets having a first model TM-score to native ≥ 0.4.

**Figure 2**
Scatter plot comparison of the first model’s GDT-TS scores by pro-sp3-TASSER and TASSER VMT on the 1192 protein benchmark target set. The number of targets for which TASSER VMT is better/worse than pro-sp3-TASSER is 765/397.

**Figure 3**
Histogram comparison of first model GDT-TS score by pro-sp3-TASSER and TASSER VMT on the 1192 target set. (a) easy set; (b) hard set. The y-axis is the number of targets having first model GDT-TS score >x.
TASSER$^\text{VMT}$ and pro-sp3-TASSER for easy and hard targets, respectively. TASSER$^\text{VMT}$ improves over pro-sp3-TASSER models for all GDT-TS score cutoffs with the maximal improvements in GDT-TS score around 0.7 and 0.3 cutoffs for easy and hard targets, respectively.

We next analyze the alignment accuracy of our alternative alignment generation method in comparison with HHSEARCH and the structural alignment method TM-align$^4^1$ using native structures. We shall compare the model quality as measured by GDT-TS scores to native of models built from the alignments using MODELLER$^2^8$ as well as directly compare alignments with the structural alignment using TM-align.$^4^1$ In Table III, we show the average model GDT-TS score changes relative to using SP$^3$ alignments and the average percentage of exact, and $\pm1$, $\pm2$ shifted residue matches to the structural alignment as provided by TM-align for the top 10 SP$^3$ templates. Apparently, both the SP$^3$ alternative and HHSEARCH alignments are $1$–$3\%$ better than SP$^3$ alignments in terms of model GDT-TS scores to native for the easy targets. However, for hard targets, HHSEARCH is worse than SP$^3$, whereas SP$^4$ alternative alignments are still $\sim3\%$ better than SP$^3$. All methods are much worse than the best possible structural alignment generated by TM-align. For the exact match to the structural alignment given by TM-align, the SP$^3$ alternative alignment is around $2\%$ (absolute value) or $3\%$ (relative value) better than the SP$^3$ alignment for the easy targets. For the hard targets in the benchmark set, the average exact match to the structural alignment is quite low because many target-template alignments have zero matches. The relative improvement of SP$^3$ alternative alignment (13.09%) over SP$^3$ alignment (12.87%) for the hard targets is 1.7%, which is less than the GDT-TS score improvement of 2.6%. However, if the match is allowed to shift within $\pm1$ or $\pm2$ residues, then the relative improvement of SP$^3$ alternative alignment over that of SP$^3$ for the hard targets will be 3.0 or 3.5%, respectively. These improvements are consistent with the GDT-TS score result. Overall, the SP$^3$ alternative alignment is consistently better than the SP$^3$ alignment. A similar trend is also found for the CASP9 set described in the next subsection. Since the HHSEARCH provides an independent alignment for given target-template pair, we include it in our method to explore the alignment space more effectively.

In Figure 4, we present some examples of TASSER$^\text{VMT}$ improvements over pro-sp3-TASSER. Target 2qsv_A domain 1 is a 101 residue hard target for threading method SP$^3$. The predicted first model’s GDT-TS score by pro-sp3-TASSER is only 0.19, whereas for TASSER$^\text{VMT}$, it is 0.61. The best model in the structure pool of pro-sp3-TASSER generated by short TASSER runs has a GDT-TS score of 0.60 while that of TASSER$^\text{VMT}$ is 0.70 due to better alignments and more template combinations. The final prediction is dictated by the selected model quality for the TASSER refinement procedure. The GDT-TS score of the top model selected by FTCOM$^3^8$ in TASSER$^\text{VMT}$ is 0.62, while that of the model selected by TASSER-QA$^3^6$ in pro-sp3-TASSER is 0.24. Therefore, the main reason for improvement is due to better model selection for this target. Another example is 3bci_A, a 165 residue easy target for SP$^3$, whose GDT-TS improves from 0.62 to 0.72. Since it is an easy target, the model selection process alone does not make a big difference. The difference comes from the fact that the best model in the structure pool in pro-sp3-TASSER has a GDT-TS score of 0.63 while that from TASSER$^\text{VMT}$ has a GDT-TS score of 0.73 due to better alignments employed. For comparison, TASSER models are also shown in Figure 4. Although on average, the pro-sp3-TASSER is better than TASSER$^2^2$ for the target 2qsv_A, it is slightly worse.

Table III
Alignment Accuracy on the Top 10 SP$^3$ Templates for the 1192 Benchmark Set

<table>
<thead>
<tr>
<th></th>
<th>Benchmark set</th>
<th>CASP9 set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>874 Easy</td>
<td>318 Hard</td>
</tr>
<tr>
<td>GDT-TS score to native$^a$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change relative to SP$^3$b</td>
<td>Change relative to SP$^3$</td>
</tr>
<tr>
<td>SP$^3$ alternative alignment</td>
<td>2.9%</td>
<td>2.6%</td>
</tr>
<tr>
<td>HHSEARCH alignment</td>
<td>0.7%</td>
<td>-2.9%</td>
</tr>
<tr>
<td>TM-align alignment</td>
<td>15.3%</td>
<td>40.7%</td>
</tr>
<tr>
<td>HHSEARCH alignments</td>
<td>58.95/68.09/72.17</td>
<td>12.87/19.04/23.16</td>
</tr>
<tr>
<td>SP$^3$ alternative alignment</td>
<td>60.77/70.62/74.47</td>
<td>13.09/19.61/23.96</td>
</tr>
<tr>
<td>HHSEARCH alignment</td>
<td>59.70/68.67/72.56</td>
<td>12.01/17.97/21.88</td>
</tr>
</tbody>
</table>

$^a$GDT-TS scores are calculated using full length models built from alignments with MODELLER$^2^8$.
$^b$Defined as (model GDT-TS of this method–model GDT-TS of SP$^3$)/(model GDT-TS of SP$^3$).
Correct match is defined as that when the threading and structural alignments of the target residue to the template residue are identical, within $\pm1$, or $\pm2$ (third number) residues apart. The number of matches in a given target is normalized by alignment length of TM-align.$^4^1$ Presented numbers are the average per target per template and in percentages.
CASP9 targets

To compare TASSERVMT with other state-of-the-art template-based methods, we tested TASSER VMT on all 112 CASP9 targets. A target is defined as easy if the average TM-score\(^4\) of the first models generated by the best 50% of all CASP9 servers is higher than 0.5; otherwise, it is defined as hard. The list of easy/hard targets can be found at http://zhanglab.ccmb.med.umich.edu/casp9/. This definition is equivalent to that in the above benchmark test where two methods are involved: TASSER VMT and pro-sp3-TASSER. In what follows, CASP servers with small variations will be represented by one server. For example, HHpredA represents all three HHpred servers, RaptorX represents all RaptorX servers, and MULTICOM_CLUSTER represents all MULTICOM\(^4\) servers.

To have a fair comparison, all necessary inputs (profiles, templates, fragments, secondary structure predictions, etc.) were generated using information that was available at the time of CASP9. The results are compiled in Table IV. TASSERVMT outperforms all other methods for the 80 easy targets and the differences between TASSERVMT and other methods are all statistically significant. TASSERVMT is only slightly worse than Zhang-Server and QUARK for the 32 hard targets, although the differences are not statistically significant. This could be because the Zhang-Server and QUARK used template-free modeling for some hard targets. Compared to our own method, pro-sp3-TASSER, TASSERVMT improves the results by 8.2 and 9.3% for easy and hard targets, respectively, and all improvements are statistically significant. The percentages of improvement are larger than those for the large benchmark set but the corresponding \(P\)-values are less significant. The difference is mainly due to different sample size (112 vs. 1192). For a small size sample set, for example, for the CASP9 set, a small number of outliers can dramatically influence the results.

Table IV
Comparison of TASSERVMT with Top CASP9 Servers on the 112 Targets

<table>
<thead>
<tr>
<th></th>
<th>Cumulative GDT-TS</th>
<th>Cumulative GDT-TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TASSERVMT</td>
<td>54.28</td>
<td>–</td>
</tr>
<tr>
<td>Zhang-Server</td>
<td>53.50</td>
<td>0.04</td>
</tr>
<tr>
<td>QUARK</td>
<td>53.45</td>
<td>0.03</td>
</tr>
<tr>
<td>HHpredA</td>
<td>52.63</td>
<td>0.007</td>
</tr>
<tr>
<td>RaptorX</td>
<td>52.97</td>
<td>0.007</td>
</tr>
<tr>
<td>Seek-server</td>
<td>51.95</td>
<td>2.9 \times 10^{-5}</td>
</tr>
<tr>
<td>MULTICOM_CLUSTER</td>
<td>51.72</td>
<td>9.63 \times 10^{-4}</td>
</tr>
<tr>
<td>BAKER-ROSETTASERVER(^\text{c})</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>chunk-TASSER</td>
<td>50.43</td>
<td>2.0 \times 10^{-11}</td>
</tr>
<tr>
<td>pro-sp3-TASSER</td>
<td>50.16</td>
<td>1.2 \times 10^{-13}</td>
</tr>
<tr>
<td>Zhang-Server</td>
<td>53.50</td>
<td>0.04</td>
</tr>
<tr>
<td>QUARK</td>
<td>53.45</td>
<td>0.03</td>
</tr>
<tr>
<td>HHpredA</td>
<td>52.63</td>
<td>0.007</td>
</tr>
<tr>
<td>RaptorX</td>
<td>52.97</td>
<td>0.007</td>
</tr>
<tr>
<td>Seek-server</td>
<td>51.95</td>
<td>2.9 \times 10^{-5}</td>
</tr>
<tr>
<td>MULTICOM_CLUSTER</td>
<td>51.72</td>
<td>9.63 \times 10^{-4}</td>
</tr>
<tr>
<td>BAKER-ROSETTASERVER(^\text{c})</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>chunk-TASSER</td>
<td>50.43</td>
<td>2.0 \times 10^{-11}</td>
</tr>
<tr>
<td>pro-sp3-TASSER</td>
<td>50.16</td>
<td>1.2 \times 10^{-13}</td>
</tr>
</tbody>
</table>

\(^a\) A target is defined as easy, if the averaged TM-score of the first models by the best 50% of all CASP9 servers is higher than 0.5; otherwise, it is classified as hard. The list of targets can be found at http://zhanglab.ccmb.med.umich.edu/casp9/.
\(^b\) Two-sided \(P\)-value of the Student-t test between TASSERVMT and the given method. A \(P\)-value of \(\leq 0.05\) is considered statistically significant.
\(^c\) Because of missing targets, we did not put results for easy targets here.
CASP targets, we conducted several different modeling protocols–A: Running TASSER refinements using a single set of top (10 for easy, 20 for hard targets, respectively) template models from SP3 without employing the variable multiple template protocol; B–E: protocols employing the Variable Multiple Template protocol with different combinations of SP3 default, SP3 alternative and HHSEARCH alignments. In protocol A, there is no intermediate stage of generating a structure pool. The results are compiled in Table V. Obviously, all other protocols are significantly better than protocol A due to the usage of a variable multiple template protocol. Thus, the variable multiple template protocol contributes a lot to TASSERVMT performance. By comparing TASSERVMT without SP3 alternative alignments (D in Table V, replacing SP3 alternative with SP3 default in TASSERVMT) with the full TASSERVMT protocol, we see the effect of the alternative alignments (in combination with the HHSEARCH alignment). For easy targets, the SP3 alternative alignments contribute around 2% of the increase in GDT-TS score; whereas for hard targets, this contribution is around 6%, which is higher than the average alignment accuracy increase given in Table III. This could be due to the fact that HHSEARCH is independent of SP3 alternative alignments and provides some complementary alignments, whereas SP3 default alignments and SP3 alternative alignments share common alignment profiles.

The direct effect of HHSEARCH can be obtained by comparing protocol E with TASSERVMT. Protocol E is slightly worse than TASSERVMT by 1.5/1.1% for easy/hard targets. Therefore, the HHSEARCH alignment improves easy targets more than hard targets. HHSEARCH still improves hard targets even though on average the HHSEARCH alignment is worse than SP3 default for hard targets (see Table III). This could be due to the fact that HHSEARCH is independent of SP3 alternative alignments and provides some complementary alignments, whereas SP3 default alignments and SP3 alternative alignments share common alignment profiles.

We now examine how good the SP3 parametric alignment for alternative alignment generation is. We compare the actual model quality of models built by MODELLER from the SP3 default alignment and from SP3 parametric alignments for the top 10 SP3 templates of the 112 CASP9 targets in Table VI. For each target-template pair, the average, standard deviation, best GDT-TS score of the models, and percentage of models having better GDT-TS scores than that of SP3 default alignment are obtained. The average model quality of the parametric alignments is slightly worse than that of the SP3 default alignment, which is understandable because the parameters of SP3 default alignment is optimized against structure alignments. The variance (standard deviation) of the parametric alignments is 11/14% of the default alignment.
for easy/hard targets. Around 30% of the alignments are better than the default ones. The best alignment is on average 9/18% better than the SP³ default alignment for easy/hard targets. Thus, SP³ parametric alignment can generate models much better than the SP³ default alignment. Even though the best models are still far from the models generated by TM-align (see Table III; 16.7/38.8% better than SP³ default), they are much better than the SP³ alternative alignment (see Table III; 2.6/3.2% better than SP³ default). If a better model selection method could be used for alternative alignment generation, then the improvement would be more significant.

**DISCUSSION**

We have improved our automated template-based modeling approach pro-sp³-TASSER method significantly by improving alignment accuracy, more sampling of multiple template combinations and a better model selection method, FTCOM. The resulting method, TASSERVTMT performs better than the best servers in CASP9 for easy targets. We note that in contrast to pro-sp³-TASSER and many other CASP servers, TASSERVTMT, does not separately consider any unaligned domains. However, based on our analysis of the two main components of our method, alignment and model selection, there is still much room for further improvement. As shown in Table III, current state-of-the-art alignment methods like HHSEARCH and the method developed in this work are far from optimal. Moreover, model selection could also be improved. For example, on average, the best model as measured by GDT-TS score in the structure pool generated by short TASSER simulations is 4.6 and 28% better than our predictions for easy and hard targets, respectively. A better model quality assessment prediction method could further improve these results, especially for hard targets. In summary, for easy/hard targets, improving the template alignment to coincide with the structural alignment would result in a ~15/40% improvement while improving model selection would provide another ~5/30% improvement. We are currently investigating the possibility of improving alignment and model selection further towards their upper limits. In this implementation of TASSERVTMT, we have explored only a relatively small number of multiple templates and generated alternative alignments by a simple parametric approach. It is possible that a more effective alternative alignment generation method that samples more near native alignments will result in further improvement. We are currently exploring a variety of approaches designed to achieve this goal.

**ACKNOWLEDGMENTS**

The authors thank Dr. Bartosz Ilkowski for managing the cluster on which this work was conducted.

**REFERENCES**