FAMSD: A Powerful Protein Modeling Platform that Combines Alignment Methods, Homology Modeling, 3D Structure Quality Estimation and Molecular Dynamics

Kazuhiko KANOU,[a] Mitsuo IWADATE,[b] Tomoko HIRATA,[a] Genki TERASHI,[a] Hideaki UMEYAMA,[a] and Mayuko TAKEDA-SHITAKA*,[a]

[a] *School of Pharmacy, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan: and* [b] *Department of Biological Sciences, Faculty of Science and Engineering, Chuo University; 1–13–27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan.* Received March 5, 2009; accepted September 8, 2009; published online September 30, 2009

The prediction of a protein three-dimensional (3D) structure is one of the most important challenges in computational structural biology. We have developed an automatic protein 3D structure prediction method called FAMSD. FAMSD is based on a comparative modeling method which consists of the following four steps: (1) generating and selecting sequence alignments between target and template proteins; (2) constructing 3D structure models based on each selected alignment; (3) selecting the best 3D structure model and (4) refining the selected model. In the FAMSD method, sequence alignment programs such as a series of BLAST programs, SP3 and SPARKS2 programs, the homology modeling program FAMS (Full Automatic Modeling System), the model quality estimation program CIRCLE and the molecular dynamics program APRICOT were used in combination to construct high quality protein models. To assess the FAMSD method we have participated in the 8th Critical Assessment of Techniques for Protein Structure Prediction (CASP8) experiment. The results of our original assessment indicate that the FAMSD method offers excellent capability in packing side-chains with the correct torsion angles while avoiding the formation of atom–atom collisions. Since side-chain packing plays a significant role in defining the biological function of proteins, this method is a valuable resource in biological, pharmaceutical and medicinal research efforts.

Key words protein structure prediction; homology modeling; comparative modeling; automatic protein modeling system; 3D-1D score; Critical Assessment of Techniques for Protein Structure Prediction

The number of three-dimensional (3D) structures of proteins solved by experimental methods is rapidly increasing. As of June 2009, more than 58000 structures were available in the Protein Data Bank (PDB).[1] However, the number of amino acid sequences whose 3D structures have not been determined remains more than 100 times greater. Therefore, some approaches for accurate protein structure prediction are urgently required. One of the most effective approaches for protein structure prediction is a comparative modeling (CM) method. This technique uses 3D template structures that have high sequence identities with the target protein. In this paper, we describe our comparative modeling platform which consists of the following four steps: (1) generating and filtering sequence alignments between the target and template proteins; (2) constructing 3D structure models based on each alignment; (3) selecting the best structural model among the candidates; and (4) refining the structure of the selected model. This automated protein modeling approach is called FAMSD. At each of the steps (1)—(4), sequence alignment programs such as SP3[2] and SPARKS2,[3] homology modeling program FAMS,[4,5] model quality estimation program CIRCLE[6] and the molecular dynamics program APRICOT[7] were primarily used and combined.

We have successively participated in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments[8—12] to assess our modeling methods. The CASP experiment is held once every two years with the aim of assessing technical progress in the prediction of protein structures. As a result, protein modeling techniques have progressed. In recent CASP experiments, more than one hundred protein sequences of unknown structures were released by the CASP organizers. The goal for each participating team is to correctly predict the 3D structures from the amino acid sequences. After the prediction period expires, the CASP assessors assess the quality of all models predicted by each participating team. From April to August 2008, the 8th CASP (CASP8) experiment was held and 128 target protein sequences were released.[13] We participated in CASP8 as an automatic predictor using the server for which our FAMSD method was internally included. In this paper, the basic algorithm of the FAMSD and the results of the protein modeling for CASP8 targets are described. We show that the modeling method using FAMSD generates reliable 3D protein structures. The protein structures predicted with the FAMSD method should be extremely valuable for experimental researchers.

Experimental

Methods. (1) Making the Sequence Alignments and the Selection To generate various sequence alignments between target and template proteins, eight types of alignment programs, BLAST,[14] PSI-BLAST,[15] PSF-BLAST,[16] RPS-BLAST, IMPALA,[17] Pfam[18]-BLAST, SP3[2] and SPARKS2[3] were executed. SPARKS2 and SP3 programs were shown to be excellent in the CASP6 experiment.[19] Various alignments were obtained and were filtered with our original alignment score value.

First, the alignment score value was calculated using the following Eq. 1 for six BLAST-related alignment methods. These methods were BLAST, PSI-BLAST, PSF-BLAST, RPS-BLAST, IMPALA and Pfam-BLAST.

$$Score_{ali}=k_i\times Len\times SEQid^{m}\times SS^{n} \quad (1)$$

Here, *Len* represents the number of residues of the region aligned to the template protein, *SEQid* represents the sequence identity percent, *SS* is the degree of match between the predicted secondary structure elements (SSE) from the target sequence and the SSE of the template protein. The predicted SSE from the amino acid sequence was obtained with PSI-PRED.[20] The SSE of the template protein was assigned by STRIDE.[21] The k_i value by which the significance weights are described in the six alignment methods is

* To whom correspondence should be addressed. e-mail: shitakam@pharm.kitasato-u.ac.jp

Table 1. Optimized m, n and k_i, (i=1: BLAST,[14] i=2: PSI-BLAST,[15] i=3: PSF-BLAST,[16] i=4: RPS-BLAST, i=5: IMPALA,[17] i=6: Pfam[18]-BLAST) Values

SEQid level	m	n	k_1	k_2	k_3	k_4	k_5	k_6
40—100	0.3	0.8	1	1.1	1.1	1.1	1.1	1.1
30—40	0.3	0.9	1	1.1	1.1	1.1	1.1	1.1
20—30	0.3	1.3	0.8	1	1	1	1	1
10—20	0.2	1.4	1	1	1	1	1	1
0—10	0	1.2	—	1	1	1	1	1

a coefficient for each alignment method. The k_i value and the parameters (m, n) are optimized for each sequence identity level as shown in Table 1. The details of this score will be reported elsewhere.

Second, for the other two alignment methods, *i.e.* SPARKS2 and SP3, the Z-score of their output was used to filter alignments. The Z-score is relatively reliable, especially when the target sequence has a high sequence identity with the template protein. We decided cut-off values for filtering alignments using the training set of CASP7 targets.[12] The alignment was adopted when the Z-score was greater than or equal to the maximum Z-score×X. The adopted alignments were used to construct the 3D structures in the next step. In other words, the parameter X is the cut-off value which was obtained by the optimization process in which we used the training set of CASP7 targets.[12] As shown in Table 2, the value of the parameter X was decided according to the difficulty of the target.

To decide the difficulty of the target, the support vector machine (SVM)[22] was used. Score and sequence identity (%) values of the top ranked alignments resulting from both PSI-BLAST and SPARKS2 were used as vectors for SVM classification. Four classes of difficulty ('CMeasy', 'CMhard', 'FR' and 'NF') were obtained from each alignment program. To identify the difficulty of a particular target, the combination of two difficulty classes which were obtained from two alignment programs was used.

Figure 1 shows the schematic diagram of the alignment selection. In this figure, when the target difficulty class was 'CMeasy-CMeasy' (*i.e.* two classes obtained from BLAST and SPARKS2 were both 'CMeasy'), this indicated that the target protein has high sequence identity with the template protein, and thus, the top two alignments were selected to construct the 3D structure. In contrast, when the target difficulty class was 'NF-CMhard' this indicated that the difficulty level was higher and therefore four more alignments were selected.

(2) Constructing the Three-Dimensional Structures Based on each selected alignment obtained in step (1), 3D structure models were constructed by using the FAMS program.[4] We repeated the modeling six times for each alignment, because the position of the side-chains varies from one model to the next as the FAMS program employs a Monte Carlo method.[23]

(3) Selecting a Structure from the Model Candidates All constructed models in the step (2) were evaluated using the following scoring function for a structural model:

$$Score_{str} = CCL + w \times SSscore \quad (2)$$

Here, *CCL* represents the CIRCLE score[6] which was based on a 3D-1D profile score (like Verify3D[24]), and the *SSscore* represents the secondary structure similarity score which was calculated by comparison between the secondary structure of the 3D model and the secondary structure predicted from the sequence. The details of this score were mentioned in reference six.[6] It is confirmed in reference six that the score consisting of the CIRCLE score and the *SSscore* is very useful to select a good quality model from many candidates.[6] As shown in Table 2, the w value is the weighting factor for the *SSscore* which was optimized using the training set based on CASP7 targets. The weight values of w were 0.3 and 1 for easy and difficult targets, respectively. It was shown that the secondary similarity score is also significant in addition to the CIRCLE score, especially for difficult targets.

In the optimization process we maximized the summation of the Global Distance Test Total Score (GDT_TS)[25] for all Template Based Modeling (TBM) targets.[26] 'TBM' is explained later. The GDT_TS represents the correctness of the Cα backbone geometry of the model, which is formally used in the CASP experiment. The GDT_TS value was calculated as shown in Eq. 3.

$$\text{GDT\_TS} = \frac{GDT\_P1 + GDT\_P2 + GDT\_P4 + GDT\_P8}{4} \quad (3)$$

Table 2. Optimized Values of X and w

PSIB[a]	SPK2[b]	X[c]	w[d]
CMeasy	CMeasy	0.99	0.3
CMhard	CMeasy	0.9	0.3
CMeasy	CMhard	0.85	0.5
CMhard	CMhard	0.85	0.5
CMhard	FR	0.85	0.5
NF	CMhard	0.8	0.5
NF	FR	0.8	1
CMhard	NF	0.8	1
NF	NF	0.8	1

a) Predicted difficulty using the alignment score and sequence identity of PSI-BLAST. *b*) Predicted difficulty using the alignment score and sequence identity of SPARKS2. *c*) Cut-off parameter X is explained in Fig. 1. *d*) Parameter w is the weighting factor of the *SSscore* as shown in the Eq. 2. We decided the parameters using the training set based on the CASP7 targets.

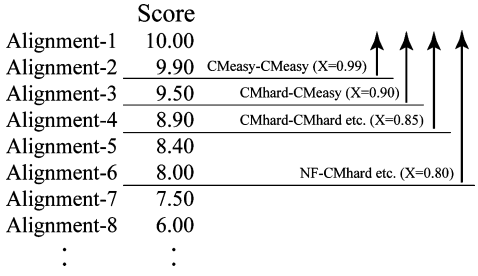


Fig. 1. Schematic Diagram of the Alignment Selection

Alignments are sorted by the alignment score. When the Z-score of the alignment-n is greater than or equal to the maximum Z-score×X, we adopted the alignment-n. In this figure, the maximum alignment Z-score is 10. The number of alignments used to construct the three-dimensional structures varies according to the combination of two predicted difficulty classes resulting from the alignments using PSI-BLAST[15] and SPARKS2.[3] If the predicted difficulty classes of a particular target are CMeasy and CMeasy, parameter X is 0.99. So alignments whose scores are more than or equal to 9.9, in this case two alignments, are used to construct the three-dimensional structures. If the predicted difficulty classes of this target are NF and CMhard, parameter X is 0.80, and therefor, six alignments are used.

Here *GDT_Pn* represents a percent of the residues separated by a distance shorter than nÅ. The GDT_TS value is an average of the GDT_P1, GDT_P2, GDT_P4 and GDT_P8 values, which ranges from zero to 100. A high GDT_TS value indicates that the positions of the Cα backbone atoms of the model matched closely the positions of the Cα backbone atoms in the native or experimental structure. 'TBM' mentioned above is a category that was assigned to the CASP7 target proteins by the CASP7 assessors. The CASP7 assessors divided each target into domains and assigned a category to each domain. Target domains for which at least one structurally similar template was available were categorized as 'TBM'.[26]

The average $Score_{str}$ of the six models repeatedly constructed from each alignment was calculated to select the alignment with an average value that was the highest. For the selected alignment, next, the highest score model among the six models based on the selected alignment was chosen as the final 3D structure. Figure 2 shows the distribution ratio of the alignment method of finally ranked first models using the above scoring function (Eq. 2) in CASP8. Accordingly, the alignment methods such as SP3 and SPARKS2 that showed the larger ratio values primarily contributed to the creation of the model with the highest GDT_TS value.

As shown in Fig. 1, in cases where the modeling difficulty of the target

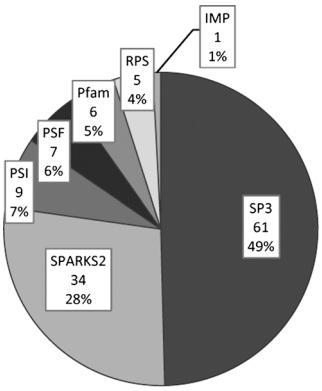


Fig. 2. Percent Distribution of the Alignment Method Which Was Used to Construct the Final FAMSD Models

The percent distribution of the alignment method which was used to construct the final FAMSD model in CASP8. PSI, PSF, Pfam, RPS and IMP represents PSI-BLAST,[15] PSF-BLAST,[16] Pfam[18]-BLAST, RPS-BLAST and IMPALA,[17] respectively.

was easy, it was better to decrease the modeling number determined from the alignment score. As such, the alignment score was reliable in the case of easy targets. In cases where the target difficulty was higher, that is, the modeling would be difficult, it was better to increase the modeling number for the ranking by $Score_{str}$. Here, it is important to indicate that for the difficult targets, the quality estimation score such as the CIRCLE score for the 3D structures is equally as important as the score derived from the alignments.

The alignment score is always an important estimate of the quality of a homology-based model, because an alignment represents an evolutionary relationship between the target and template proteins. For easy targets, where there are some template proteins with high sequence identities and the alignment scores are relatively high, the alignment score provides an appropriate estimate of the quality of a model. However, for difficult targets which have no template proteins with adequate high sequence identities, the alignment scores may be evolutionally meaningless, and the alignment scores do not appropriately estimate the quality of a model. Conversely, the CIRCLE score, which is the major part of the $Score_{str}$, estimates the stability of a protein structure from a free energy point of view.[6] For difficult targets in which the alignment scores are not reliable, the CIRCLE score estimates a model quality more appropriately than the alignment score. In some cases, however, the CIRCLE score provides a high score to a model which has many stable regions locally in the model even if the global structure was not close in structure to the native fold. Therefore, to select a good model for difficult targets, both the alignment score and the structural score which includes the CIRCLE score are required.

(4) Performing the Refinement of the Finally Selected Models The selected model was refined using molecular mechanics and molecular dynamics (MM-MD). The MD program APRICOT[7] was used in this process. The parameters including temperature used in this process were determined with the training set based on the CASP5 target proteins[10] (Table 3). In the same training set, the refined models which deviated from the initial models with the MM-MD refinement were discarded as follows. Root Mean Square Deviation (RMSD) values of the structure deviations were more than 0.5, 0.7 and 1.2 Å, for Cα atoms, main chain atoms and all atoms including side-chains, respectively. Using this procedure, hydrogen bonds and side-chain torsion angles were refined slightly, and unfavorable collisions between hydrophobic atoms were reduced.

Results and Discussion

The Effect of Combining Eight Alignment Methods The FAMSD method used eight alignment methods, and only one model based on one alignment was finally selected as described in the Methods section. In order to confirm our FAMSD method, we tested whether the final models of the FAMSD method were better than the models based on each eight individual alignment methods. These referenced models were constructed based on an alignment which was ranked as number one in each alignment method. The FAMS

Table 3. Parameters Used in the MM-MD Optimization Process

Run time	5 ps
Temperature	100 K
Constraints[a)]	Position constraint for Cα atoms $k(\mathrm{Ca}-\mathrm{Ca}_{\mathrm{init}})^2$, $k=100$
	Torsion angle constraints for main chains and side chains $k'(\theta-\theta_{\mathrm{init}})^2$, $k'=600$
Solvent	Box of water molecule 7.5 Å from protein surface, Periodic boundary condition[27)]
Force field	AMBER united atom force field[28)]

a) The Ca and the $\mathrm{Ca}_{\mathrm{init}}$ indicate the Cα coordinates of the MM-MD refinement model and the Cα coordinates of the initial model, respectively. The θ and θ_{init} indicate the torsion angles of the MM-MD refinement model and that of the initial model, respectively.

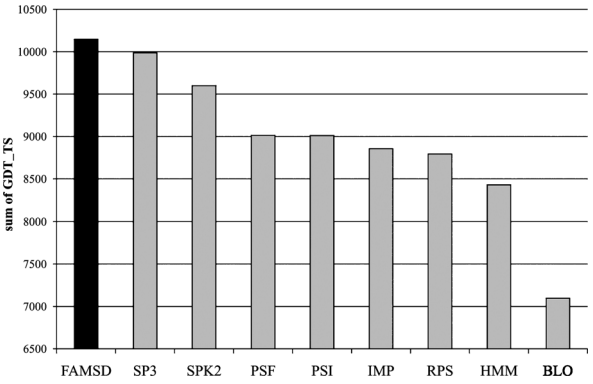


Fig. 3. Comparison between the FAMSD Routine and Each of the Eight Individual Alignment Methods

FAMSD and each of the eight alignment methods are sorted by the sum GDT_TS for all CASP8 targets, (SP3:SP3,[2] SPK2:SPARKS2,[3] PSI:PSI-BLAST,[15] PSF:PSF-BLAST,[16] IMP:IMPALA,[17] RPS:RPS-BLAST, HMM:Pfam[18]-BLAST, BLO:BLAST[14]).

program was also used to construct these models in the same manner as the FAMSD method. In this case, the difference in the GDT_TS values was primarily due to the difference in the amino acid sequence alignments. The summation of the GDT_TS value of all CASP8 targets was calculated for each individual alignment method (Fig. 3). Thus, as shown in Fig. 3, the sequence alignment obtained from the FAMSD method contributed to the better GDT_TS score. The FAMSD method was confirmed to be an excellent method when compared with the eight methods.

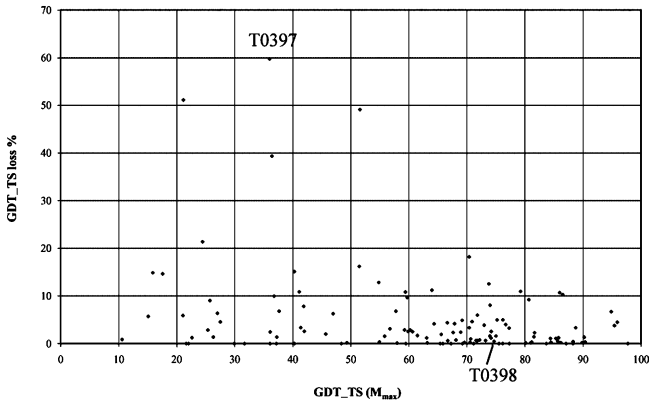
In this paper, the highest GDT_TS score model among all the models given by the eight different alignment methods employed was termed as the max GDT_TS model. The GDT_TS loss and GDT_TS loss % of the FAMSD model (M_a) was defined in the following Eqs. 4 and 5, respectively, and refers to the max GDT_TS model.

$$\mathrm{GDT\_TS\ loss}(M_a)=\mathrm{GDT\_TS}(M_{max})-\mathrm{GDT\_TS}(M_a) \quad (4)$$

$$\mathrm{GDT\_TS\ loss\ \%}(M_a)=\frac{\mathrm{GDT\_TS\ loss}(M_a)}{\mathrm{GDT\_TS}(M_{max})} \quad (5)$$

Here, M_{max} indicates the max GDT_TS model, and GDT_TS(M_{max}) represents the GDT_TS value for the model M_{max}. The GDT_TS loss % of the model M_a was plotted

(A)



(B)

	Average GDT_TS loss	Average GDT_TS loss %
GDT_TS(Mmax) >= 60	2.07	2.71
GDT_TS(Mmax) < 60	3.26	8.43
All	2.54	4.12

Fig. 4. GDT_TS Loss

(A) GDT_TS loss % value is plotted against the max GDT_TS value for each target protein. T0398 used as a good example in Fig. 7 later is GDT_TS(M_{max})=74.7 and GDT_TS loss %=0.4, and T0397 used as a poor example in Table 4 later is GDT_TS(M_{max})=36 and GDT_TS loss %=59.7. (B) Average value of the GDT_TS loss and the GDT_TS loss % for GDT_TS(M_{max})>=60, GDT_TS(M_{max})<60 and all of target proteins.

against the max GDT_TS value for each target protein as shown in Fig. 4A. As shown in Fig. 4B, the average loss % of the FAMSD method was only 2.7% in comparison with the max GDT_TS model when the maximum value of GDT_TS was more than or equal to 60. Conversely, when the max GDT_TS value was less than 60, the average loss % increased to 8.4%. The average loss % for all targets was 4.1%. Accordingly, the easy targets with max GDT_TS values greater than 60 will be modeled by the FAMSD method with a modeling percentage loss of only *ca.* 3%. In this section, as a note, it is provisionally assumed that target proteins with max GDT_TS values greater than 60 are grouped into the easy category.

For example, in the case of T0396 which is one of the CASP8 targets, the GDT_TS value, the GDT_TS loss and the GDT_TS loss % of the FAMSD model were 86.5, 3.0 and 3.3%, respectively. This 3.3% of the GDT_TS loss % value is near 2.7% of the average value for easy targets. In this case, the RMSD_CA (Root Mean Square Deviation for Cα atoms between an experimental structure and a model) of the FAMSD model and that of the max GDT_TS model were 1.88 Å and 1.75 Å, respectively. Consequently, the difference in the RMSD_CA between the FAMSD model and the max GDT_TS model was only 0.13 Å. This indicated that a *ca.* 3% value of the GDT_TS loss % represents a very successful selection of the best model from the model candidates. Accordingly, the FAMSD method provides the high accuracy model for easy target proteins.

Effects of the MM-MD Refinement We compared the quality of the models with and without MM-MD refinement to elucidate the influence of the MM-MD refinement described in the step (4). In CASP8, models in which the refined models deviated from the initial models during the

(A)

	without refinement	MM-MD	difference
GDT_TS	5872.43	5878.15	+5.72 (+0.10%)
χ1	6409	6453	+44 (+0.69%)
χ2	3472	3497	+25 (+0.72%)
H-bond	8838	9961	+1123 (+12.71%)

(B)

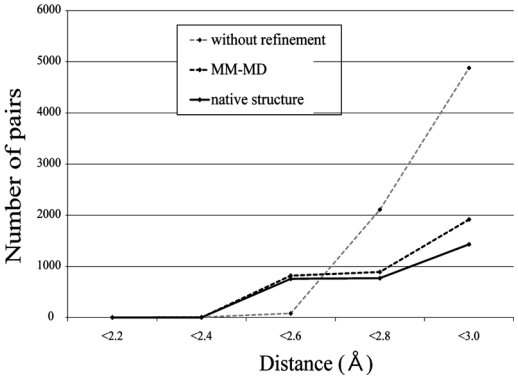


Fig. 5. Comparison between the Models with and without MM-MD Refinement

(A) Comparison between the quality of the models with and without the MM-MD refinement against GDT_TS, the number of correct χ1 torsion angles ('χ1'), the number of correct χ2 torsion angles ('χ2') and hydrogen bonds ('H-bond'). (B) The number of collisions of C–C pairs for the initial models without MM-MD refinement, the MM-MD refined models and the native structures from the experiments. Black and gray colored broken lines represent the collision numbers included in the models with and without MM-MD refinement, respectively. The solid line represents the collision numbers included in the native structures.

MM-MD refinement were discarded. This is because structures naturally break if initial models have incorrectly folded regions that are beyond the capacity of the MM-MD refinement. For these particular models, we submitted the initial models without MM-MD refinement. In 92 targets out of 128, MM-MD refined models were submitted.

As shown in Fig. 5A, we compared the quality of the models with and without MM-MD refinement in terms of the sum GDT_TS value, the number of correct χ1 and χ2 values and hydrogen bonds. A χ1 torsion angle was considered "correct" if the value was within 40 degrees of the experimental structure value.[29,30] A χ2 torsion angle was considered "correct" if χ1 and χ2 were within 40 and 60 degrees, respectively.[29,30] As the result of the comparison, we found that the sum GDT_TS score and the number of correct χ1 and χ2 values of the MM-MD refined models remained almost the same as those of the initial models. In contrast, the number of correctly placed hydrogen bonds was found to increase by 12.7%. Furthermore, we calculated the number of collisions of C–C pairs for the MM-MD refined models, initial models and their native structures (Fig. 5B). For example, the numbers of C–C pairs whose distances were below 2.8 Å for the initial models and the MM-MD refined models were 2109 and 818, respectively. The latter value was comparable to that of 768 observed in the native structures. In addition, at the other threshold of distances between C–C pairs in Fig. 5B, the number of collisions observed in the MM-MD

models was similar to the number observed in the native structures. Consequently, using the MM-MD refinement, hydrogen bonds were refined, and the number of collisions in the models was reduced and became similar to the number observed in the native structures. Nevertheless, this refinement method did not lead to noticeable degradation in the quality of the Cα backbone geometry and side-chain torsion angles. Accordingly, the MM-MD refinement method proposed in this paper is very useful in refining the models to closely match the native state.

Ranking of the FAMSD Method among the CASP8 Server Predictors As shown in Fig. 6 A, the team using the FAMSD method was ranked 13th among the 71 CASP8 servers for the sum GDT_TS of 154 domains which were categorized into the TBM target by the CASP8 assessors.[31] The FAMSD team was ranked 10th for the summation of the number of correct $\chi 1$ torsion angles for the same 154 domains. Furthermore, the CASP8 servers including the FAMSD team were ranked by calculating a summation of a combined mixed Z-score. The combined Z-score ($Z_{combined}$) was calculated as the average of the Z-scores for GDT_TS and the number of correct $\chi 1$ torsion angles:

$$Z_{combined}=(Z_{GDT\_TS}+Z_{\chi 1})/2 \quad (6)$$

The Z-scores for GDT_TS and the number of correct $\chi 1$ torsion angles were calculated on each model using the average and standard deviation values from all models. Higher $Z_{combined}$ values indicated that the main chain and side-chains were both structurally similar to the native structure in relation to Z_{GDT_TS} and $Z_{\chi 1}$, respectively. In this assessment the FAMSD team was ranked 10th.

The side-chain accuracies were further analyzed, as we

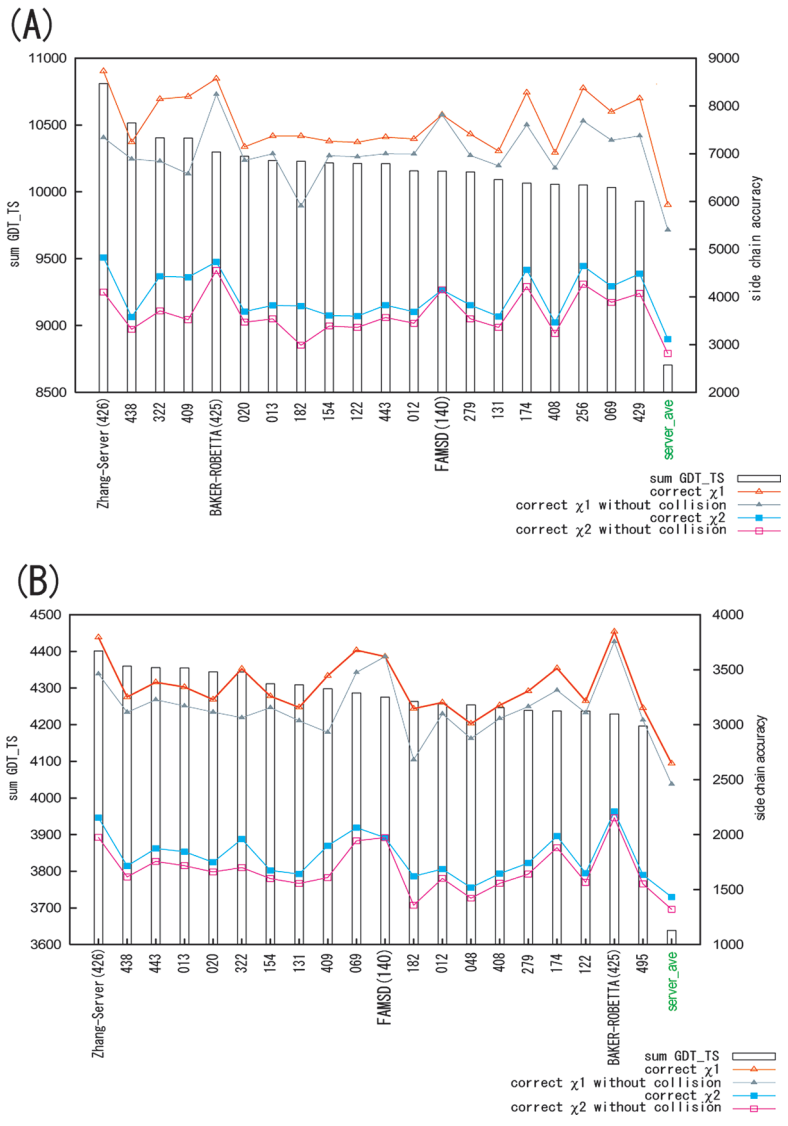


Fig. 6. CASP8 Server Rankings

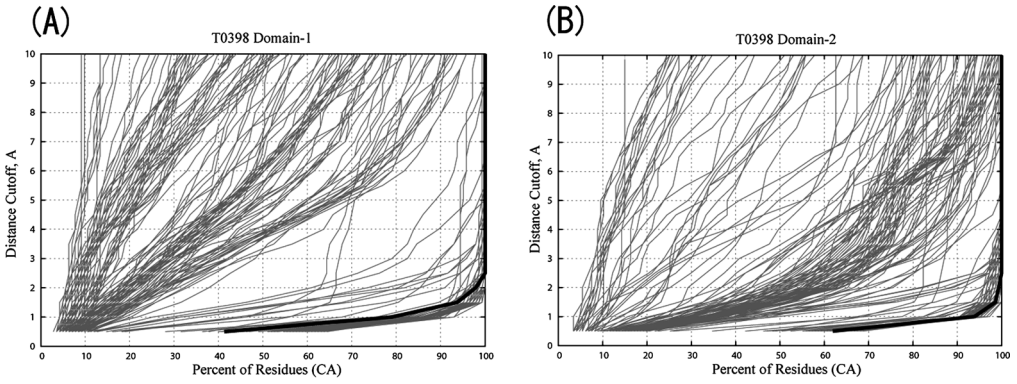
The summation of GDT_TS values ('sum GDT_TS'), the number of correct $\chi 1$ torsion angles ('correct $\chi 1$'), the number of correct $\chi 2$ torsion angles ('correct $\chi 2$'), and the number of correct $\chi 1$ and $\chi 2$ torsion angles without unfavorable collisions ('correct $\chi 1$ without collision' and 'correct $\chi 2$ without collision', respectively) are plotted against the top 20 of the 71 CASP8 servers. (A) The bar represents the sum of the GDT_TS values for 154 TBM (Template Based Modeling) target domains (left axis). Solid lines with open triangles and closed squares represent the number of correct $\chi 1$ values and the number of correct $\chi 2$ values, respectively (right axis). Solid lines with filled triangles and open squares represent the number of correct $\chi 1$ values and the number of correct $\chi 2$ values without collisions, respectively (right axis). Group names corresponding to the group numbers will be found on the CASP8 website shown below. http://predictioncenter.org/casp8/docs.cgi?view=groupsbynumber 'server_ave' imaginary team which is not participating as the CASP8 server represents the average of all 71 CASP8 servers. (B) 50 TBM-HA (Template Based Modeling-High Accuracy) target domains were used instead of above 154 TBM targets. TBM-HA was a category that was decided by the CASP8 assessors. These target domains have at least one model with a GDT_TS value of 80.

wanted to identify excellent aspects of the FAMSD method in comparison with other CASP8 servers. The protein models should not contain short contacts in which atom–atom pair interactions are unfavorable. This is because such unfavorable atom–atom pair interactions are not observed in native structures. As such, unfavorable side-chain interactions were excluded in relation to the success of the side-chain modeling. We used the criteria of unacceptable distances as 2.4 and 2.0 Å for C–C pairs and non C–C pairs, respectively, because atom–atom pairs within these distances are extremely rare in native structures. As the results, the ranking of the FAMSD team was 2nd and 5th for the correct $\chi 1$ estimations and the $Z_{combined}$, respectively. The protein model is generally used to explain the biological function in cases where no experimental structure exists. Consequently, it is important that the model has a high assessment result in the conformation which might be observed in the native structure. Therefore, as the FAMSD method was ranked 2nd and 5th in the above mentioned tests, this platform represents a valuable modeling method.

The results for 50 domains which were categorized into TBM-HA (TBM-High Accuracy) are shown in Fig. 6B. Target domains in which the best prediction had at least a GDT_TS of 80 were categorized as TBM-HA. These proteins represent the easy targets to model. The performance of the FAMSD team in this category was better than the performance in the TBM category (Fig. 6A). Consequently, the FAMSD method compared well with the methods of other modeling servers; FAMSD is world-leading in providing physically meaningful models that have good side-chain conformations without unfavorable collisions. Accordingly, the FAMSD modeling method as presented in the Methods section represents a valuable protein modeling platform.

The FAMSD method is not the best among the CASP8 servers, and, in terms of GDT_TS, the best server among the CASP8 servers was the Zhang–Server. However, there are some cases in which the FAMSD models have higher GDT_TS values than the corresponding Zhang–Server models. The superior ratio of the FAMSD team was 18% in 147 domains. Both teams tied with 7 domains out of a possible 154 domains. Although the FAMSD team ranked 2nd for the correct $\chi 1$ estimation, taking into consideration the atom–atom collisions, the FAMSD team was superior to the BAKER-ROBETTA team (ranked 1st) in the 35% ratio of 146 domains in the comparison of the correct $\chi 1$ numbers of both models. Both teams tied with 8 domains out of a possible 154 domains. Moreover, although the FAMSD team was 5th for the ranking of the $Z_{combined}$, the 5th ranked FAMSD team was superior to the BAKER-ROBETTA team (ranked 1st) in the 38% ratio of 154 domains in the comparison of the $Z_{combined}$ scores of both models. This means that the best method such as Zhang–Server or BAKER-ROBETTA does not necessarily provide the best model for every target proteins. In other words, if the best model among many models constructed by many powerful protein modeling approaches could be always selected for each target, by using many approaches we could obtain better models than by using each individual approach. Therefore, not only the Zhang–Server or the BAKER-ROBETTA, various protein structure modeling approaches including the FAMSD method are needed. Additionally, it is important that methods are developed for selecting the best model from the many models constructed with the various modeling approaches.

Good Example: T0398 Target T0398 (pdb code: 3D4O) is one of the CASP8 targets, dipicolinate synthase subunit A (NP_243269.1) from *Bacillus halodurans*. This protein consists of two domains. In this target, the GDT_TS loss % of the FAMSD model was 0.35%. This value showed that the model selection mentioned in steps (1), (2) and (3) of the Methods was successful. The GDT_TS values of domains 1 and 2 (D1 and D2) were 94.23 and 98.30, respectively. Fig. 7A and B are the GDT plots[32] for T0398 domain-1 and do-



(C)

	GDT_TS	RMS_CA[a]	RMS_ALL[b]	correct χ1 %[c]	correct χ2 %[d]
Domain 1	94.23	0.90	1.46	68.10	51.22
Domain 2	98.30	0.65	1.22	72.07	59.76

[a]RMSD for Cα atoms
[b]RMSD for all atoms
[c]ratio % of the number of correct χ1 torsion angle
[d]ratio % of the number of correct χ2 torsion angle

Fig. 7. Assessment Result for T0398

(A) GDT plot[32] for T0398 domain-1. The thick black line represents the FAMSD model. The other lines indicate models submitted by other groups. The GDT_Pn represents a percent of residues separated by a distance shorter than nÅ. Here, nÅ is called the 'Distance Cutoff, Å'. A percent of residues separated by a distance shorter than nÅ is called the 'Percent of Residues (CA)'. The 'Distance Cutoff, Å' is plotted against the 'Percent of Residues (CA)' in relation to all the server models. (B) GDT plot[32] for T0398 domain-2. The thick black line represents the FAMSD model. The other lines indicate models submitted by other groups. (C) The GDT_TS value, RMSD for Cα atoms, RMSD for all atoms, the ratio % of the number of correct $\chi 1$ and ratio % of the number of correct $\chi 2$ values for the T0398 domain-1 and domain-2.

Table 4. Comparisons between the FAMSD Model and the Max GDT_TS Model for the Whole Chain, the N-Terminal Domain and C-Terminal Domain of T0397

(A) whole chain

	FAMSD model	max GDT_TS model
$Score_{str}$ [a)]	23.476	12.825
GDT_TS	14.50	36.00

(B) N-terminal domain (T0397_D1)

	FAMSD model	max GDT_TS model
$Score_{str}$ [a)]	2.665	−16.845
GDT_TS	20.43	19.82

(C) C-terminal domain (T0397_D2)

	FAMSD model	max GDT_TS model
$Score_{str}$ [a)]	20.811	29.669
GDT_TS	27.21	72.79

a) $Score_{str}$ was calculated with Eq. 2. In this target T0397, as the target level was 'CMhard, CMhard', the parameter *w* was set to 0.5 as shown in Table 2.

main-2, respectively. This figure shows that the accuracy of the Cα backbone of the FAMSD model was one of the best predictions. Furthermore, the percentage ratios of correctly predicted χ1 and χ2 torsion angles of the FAMSD model were 70.0% and 55.5%, respectively. The domains of T0398-D1 and T0398-D2 of the FAMSD model were both ranked 3rd out of 71 CASP8 servers for the number of the correct χ1 and χ2 torsion angles. No unfavorable atom–atom pair interactions were observed in the FAMSD models.

Poor Example: T0397 Target T0397 (pdb code 3D4R) is a domain of unknown function from the Pfam-B_34464 family from *Methanococcus maripaludis*. The FAMSD method selected a model based on the alignment with pdb code 1H2W. However, the max GDT_TS model of this target was constructed based on the alignment with pdb code 2QJ8. In Table 4A, the GDT_TS, the GDT_TS loss and the GDT_TS loss % of the FAMSD model were 14.5, 21.5 (=36.0—14.5) and 59.7% (=21.5/36.0), respectively. From the 59.7 of GDT_TS loss % value, the FAMSD method failed to select the correct template or alignment from many candidates.

The CASP8 assessor divided this target protein into two domains; T0397-D1 (1—82) and T0397-D2 (83—150) as shown in Tables 4B and 4C, respectively. By comparing the quality of the FAMSD model and the max GDT_TS model, it was found in Table 4C that there were significant differences in the quality of the C-terminal domain. The T0397-D2 of the max GDT_TS model was predicted well; however, the T0397-D2 of the FAMSD model was poor. On the other hand, there was only a small difference in quality between the max GDT_TS model and the FAMSD model for the N-terminal domain (Table 4B). Table 4 shows the $Score_{str}$ values of the FAMSD model and the max GDT_TS model for whole chain and the two domains. The $Score_{str}$ of the C-terminal domain of the max GDT_TS model was higher than that of the FAMSD model (Table 4C). FAMSD failed to predict the domain boundary for this T0397 target. If the correct domain boundary between D1 and D2 for this target was found, the selection of the max GDT_TS model for the C-terminal domain from many candidates using the $Score_{str}$ would have been possible (data is not shown). However, the domain boundary prediction for the target to model was not easy and remains a critical issue for us.

Conclusions

We have developed an automatic protein 3D structure prediction method called FAMSD. This method is based on comparative modeling which is an effective protein structure prediction method. Alignment programs such as a series of BLAST programs, the SP3 and SPARKS2 programs, the homology modeling program FAMS, the 3D structure quality estimation program CIRCLE and the molecular dynamics program APRICOT were combined to construct high quality protein structure models. In our original assessment, we mainly used the combined Z-score ($Z_{combined}$) as an assessment criterion. A higher $Z_{combined}$ value indicates that the main chain and side-chains are similar in conformation (*i.e.* geometric spatial positioning of atoms within the 3D fold of the protein) to the native structure. In this assessment, the FAMSD team was ranked 10th out of 71 CASP8 servers. On the other hand, taking into consideration the residue–residue collisions in the assessment of the conformations of the side-chains, it was shown that the rankings of the FAMSD team were 2nd and 5th for the correct χ1 estimation and the $Z_{combined}$ value, respectively. This result indicates that the FAMSD method offers excellent capability in packing side-chains with the correct torsion angles while avoiding atom–atom collisions. Since protein modeling is generally used to help defining the biological function of proteins when no experimental structures are available, the FAMSD method, with excellent performance in positioning side-chains, should be a valuable platform for use in the biological, pharmaceutical and medicinal research efforts. Finally, the FAMSD method represents a very valuable tool for modeling the structures of a large number of proteins arising from all human genes or the genes of other species. This is because the FAMSD method is a fully automated protein structure prediction approach.

Acknowledgement This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research(B), 08021917, 2007.

References

1) Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E., *Nucleic Acids Res.*, **28**, 235—242 (2000).
2) Zhou H., Zhou Y., *Proteins*, **58**, 321—328 (2005).
3) Zhou H., Zhou Y., *Proteins*, **55**, 1005—1013 (2004).
4) Ogata K., Umeyama H., *J. Mol. Graph. Model.*, **18**, 258—72, 305–6 (2000).
5) Takeda-Shitaka M., Terashi G., Takaya D., Kanou K., Iwadate M., Umeyama H., *Proteins*, **61** (Suppl. 7), 122—127 (2005).
6) Terashi G., Takeda-Shitaka M., Kanou K., Iwadate M., Takaya D., Hosoi A., Ohta K., Umeyama H., *Proteins*, **69** (Suppl. 8), 98—107 (2007).
7) Yoneda S., Yoneda T., Kurihara Y., Umeyama H., *J. Mol. Graph. Model.*, **21**, 19—27 (2002).
8) Moult J., Hubbard T., Fidelis K., Pedersen J. T., *Proteins*, Suppl. **3**, 2—6 (1999).
9) Moult J., Fidelis K., Zemla A., Hubbard T., *Proteins*, Suppl. **5**, 2—7 (2001).
10) Moult J., Fidelis K., Zemla A., Hubbard T., *Proteins*, **53** (Suppl. 6),

334–339 (2003).
11) Moult J., Fidelis K., Rost B., Hubbard T., Tramontano A., *Proteins*, **61** (Suppl. 7), 3—7 (2005).
12) Moult J., Fidelis K., Kryshtafovych A., Rost B., Hubbard T., Tramontano A., *Proteins*, **69** (Suppl. 8), 3—9 (2007).
13) CASP8 home page ⟨http://www.predictioncenter.org/casp8/index.cgi⟩, 2008.
14) Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J., *J. Mol. Biol.*, **215**, 403—410 (1990).
15) Altschul S. F., Madden T. L., Schäffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J., *Nucleic Acids Res.*, **25**, 3389—3402 (1997).
16) Nanatani K., Fujiki T., Kanou K., Takeda-Shitaka M., Umeyama H., Ye L., Wang X., Nakajima T., Uchida T., Maloney P. C., Abe K., *J. Bacteriol.*, **189**, 7089—7097 (2007).
17) Schäffer A. A., Wolf Y. I., Ponting C. P., Koonin E. V., Aravind L., Altschul S. F., *Bioinformatics*, **15**, 1000—1011 (1999).
18) Sonnhammer E. L., Eddy S. R., Durbin R., *Proteins*, **28**, 405—420 (1997).
19) Tress M., Ezkurdia I., Graña O., López G., Valencia A., *Proteins*, **61** (Suppl. 7), 27—45 (2005).
20) Jones D. T., *J. Mol. Biol.*, **292**, 195—202 (1999).
21) Frishman D., Argos P., *Proteins*, **23**, 566—579 (1995).
22) Vapnik V., "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995.
23) Metropolis N., Ulam S., *J. Am. Stat. Assoc.*, **44**, 335—341 (1949).
24) Eisenberg D., Lüthy R., Bowie J. U., *Methods Enzymol.*, **277**, 396—404 (1997).
25) Zemla A., *Nucleic Acids Res.*, **31**, 3370—3374 (2003).
26) Clarke N. D., Ezkurdia I., Kopp J., Read R. J., Schwede T., Tress M., *Proteins*, **69** (Suppl. 8), 10—18 (2007).
27) Cheatham T. E., Miller J. H., Fox T., Darden P. A., Kollman P. A., *J. Am. Chem. Soc.*, **117**, 4193 (1995).
28) Case D. A., Cheatham T. E. 3rd, Darden T., Gohlke H., Luo R., Merz K. M. Jr, Onufriev A., Simmerling C., Wang B., Woods R. J., *J. Comput. Chem.*, **26**, 1668—1688 (2005).
29) Fischer D., Elofsson A., Rychlewski L., Pazos F., Valencia A., Rost B., Ortiz A. R., Dunbrack R. L. Jr., *Proteins*, Suppl. **5**, 171—183 (2001).
30) Fischer D., Rychlewski L., Dunbrack R. L. Jr, Ortiz A. R., Elofsson A., *Proteins*, **53** (Suppl. 6), 503—516 (2003).
31) CASP8 domain classification ⟨http://www.predictioncenter.org/casp8/doc/Target_classification_1.html⟩, 2008.
32) Oldfield T. J., Hubbard R. E., *Proteins*, **18**, 324—337 (1994).