New Protein Structure Model Evaluation Methods That Include a Side-Chain Consensus Score for the Protein Modeling

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

[a] *School of Pharmacy, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan: and* [b] *RIKEN Systems and Structural Biology Center; 1–7–22 Suehiro-cho, Tsurumi-ku, Yokohama 230–0045, Japan.*
Received August 24, 2009; accepted November 24, 2009; published online November 27, 2009

Selecting the best quality model from a set of predicted structures is one of the most important aspects of protein structure prediction. We have developed model quality assessment programs that select high quality models which account for both the Cα backbone and side-chain atom positions. The new methods are based on the consensus method with consideration of the side-chain environment of a protein structure and the secondary structure agreement. This Side-chain Environment Consensus (SEC) method is compared with the conventional consensus method, 3D-Jury (Ginalski K. *et al.*, *Bioinformatics*, 19, 1015—1018 (2003)), which takes into account only the Cα backbone atoms of the protein model. As the result, it was found that the SEC method selects the models with more accurate positioning of the side-chain atoms than the 3D-Jury method. When the SEC method was used in combination with the 3D-Jury method (3DJ+SEC), models were selected with improved quality both in the Cα backbone and side-chain atom positions. Moreover, the CIRCLE (CCL) method (Terashi G. *et al.*, *Proteins*, 69 (Suppl. 8), 98—107 (2007)) based on the 3D–1D profile score has been shown to select the best possible models that are the closest to the native structures from candidate models. Accordingly, the 3DJ+SEC+CCL method, in which CIRCLE is used after reducing the number of candidates by the 3DJ+SEC consensus method, was found to be very effective in selecting high quality models. Thus, the best method (the 3DJ+SEC+CCL method) includes the consensus approaches of the Cα backbone and the side-chains, the secondary structure agreement and the 3D–1D profile score which corresponds to the free energy-like score in the residues of the protein model. In short, new algorithms are introduced in protein structure evaluation methods that are based on a side-chain consensus score. Additionally, in order to apply the 3DJ+SEC+CCL method and indicate the usefulness of this method, a model of human Cabin1, a protein associated with p53 function and cancer, is created using various internet modeling and alignment servers.

Key words protein structure prediction; model quality assessment; side-chain environment; consensus method; protein modeling; 3D–1D

A large number of genes from many genomes have been determined. Translation of the genes to amino acid sequences and the amino acid analysis of proteins have provided a wealth of protein sequence information. In order to elucidate the biological function of proteins, the three-dimensional (3D) structures of the proteins are essential. The determination of the 3D structure of proteins has been performed with the experimental methods such as X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. However, the number of protein structures determined using the experimental methods lags significantly behind the number of protein sequences. Therefore, computational approaches such as comparative or homology modeling for accurate protein structure prediction are urgently required.

Consensus methods have been used in the 3D modeling of protein structures.[1—3] For example, the reliability of the predicted 3D model is determined by rankings based on factors such as the similarity between Cα atoms of two protein backbones in the compared models. This representative consensus method is very useful if there are numerous accessible modeling or alignment servers. Here, an amino acid sequence of a target protein is used in the analysis of other excellent servers or websites and the output gives the 3D model or the sequence alignment between the target protein and a template protein 3D structure that has been determined experimentally. As shown in the Results and Discussion, we can submit an amino acid sequence under investigation to several web servers and obtain the 3D models or the sequence alignments. Thereupon, the consensus method is an available method, in spite of the complicated way in which we must collect 3D models and alignments created by each web server. Consequently, improving the consensus method is very important in creating a more accurate model based on many 3D models and alignments. In this paper, we have improved the consensus method in taking notice of the environment similarity of side-chains. We explain the algorithm of the new consensus method mentioned above and, actually, use our modeling method employing several server models and sequence alignments collected. As an example of a modeling target, we chose the human Cabin1 protein, a molecule involved in p53 function, which is a very important protein target involved in apoptosis and is combating with cancer in tissues.[4—6] Our 3D model for the N-terminal region consisting of 450 residues of human Cabin1 may be useful in the pharmaceutical, medicinal and biological fields.

In the meanwhile, 3D-Jury[1] method which is the representative consensus method was one of the most powerful methods to obtain a model with accurate Cα backbone atoms. In the 3D-Jury[1] method each amino acid is represented only by the Cα atom. As such, this method selects "good backbone" protein models that closely match with the experimental structure of the target protein. However, the quality of the side-chains of the selected models is not ensured by this method, because the 3D-Jury method does not refer to the

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

coordinates of the side-chains. Thus, we report a number of new consensus methods that consider the side-chain atoms as well as the backbone atoms. The purpose of this approach is to select protein models with correctly modeled side-chains and backbone atoms. In this paper, new consensus methods are shown to be effective and should be valuable to the protein structure modeling researchers. The scientifically appropriate reason for this consensus method is explained in the Methods.

To assess the performance of our new consensus methods, we performed the original assessment using the target proteins of the latest Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment as a test set. The CASP experiment is held once every two years with the aim of progressing technique in the modeling of protein structures.[7—13] In these CASP experiments, each participant receives over one hundred target protein sequences from the CASP organizers, and returns the predicted 3D models for each target. After the prediction period expires, the CASP assessors evaluate the accuracy of models predicted by each participating team using the Global Distance Test Total Score (GDT_TS)[14] as one of the evaluation criteria. The GDT_TS value represents the correctness of the Cα backbone geometry of the predicted model. A high GDT_TS value indicates that the predicted Cα backbone atoms match closely the position of the atoms in the native structure, which is the same meaning as the experimental structure in this paper. The use of the GDT_TS value for the assessment of the 3D structure of the protein model means, by a tacit consent, that the atomic coordinates of the side-chains will be near to the native structure if the main chain or the Cα backbone atoms of the target protein are modeled with high accuracy. In addition to the GDT_TS values, the accuracies of the side-chains were used as another criterion for the evaluation of the accuracy of the protein models.

Experimental

Methods. Side-Chain Environmental Consensus (SEC) Score We defined a new consensus score that considers the environment of the side-chains. The environment of the side-chain of a residue was assumed to be composed of two parameters: 'fraction buried' (fb) and 'fraction polar' (fp). The former means the fraction of the side-chain surface area buried by any other atoms and the latter term represents the fraction of the side-chain surface area that is exposed to water or covered by polar atoms.[15] The denominator of both parameters is the total surface area of the side-chain. To calculate the consensus score for the side-chain, we must compare the side-chain environment at the nth residue of a particular model with that of each model in a model set. We defined the environmental distance as a Euclidian distance between the two side-chain environments of the corresponding residues in two models as shown in Eq. 1 and Fig. 1A. In other papers, no researchers have reported this kind of Euclidian distance in relation to the environment of the side-chains.

$$env\_dis(\mathrm{M},\mathrm{M_i},n) = \sqrt{(\mathrm{fb}(\mathrm{M},n) - \mathrm{fb}(\mathrm{M_i},n))^2 + (\mathrm{fp}(\mathrm{M},n) - \mathrm{fp}(\mathrm{M_i},n))^2} \quad (1)$$

Here, $env\_dis(\mathrm{M}, \mathrm{M_i}, n)$ is the environmental distance between the nth residue of a particular model M and that of model $\mathrm{M_i}$ which is one of models in the model set. $\mathrm{fb}(\mathrm{M}, n)$ is the 'fraction buried' of the nth residue of model M. $\mathrm{fp}(\mathrm{M}, n)$ is the 'fraction polar' of the nth residue of model M. As the first approximation, in this paper, the weight of the term of fb was assumed to be equal to that of fp, though the weight value should be changed. The assignment of *env_dis* to each residue is shown in Fig. 1B. The environmental similarity score ($\mathrm{sim}(\mathrm{M}, \mathrm{M_i})$ in Fig. 1B) between two models is defined as the number of the corresponding residue pairs whose environmental distance is within 0.2. The threshold of 0.2 was determined by maximizing the total GDT_TS score using the CASP7 targets[12] as a training set. Based

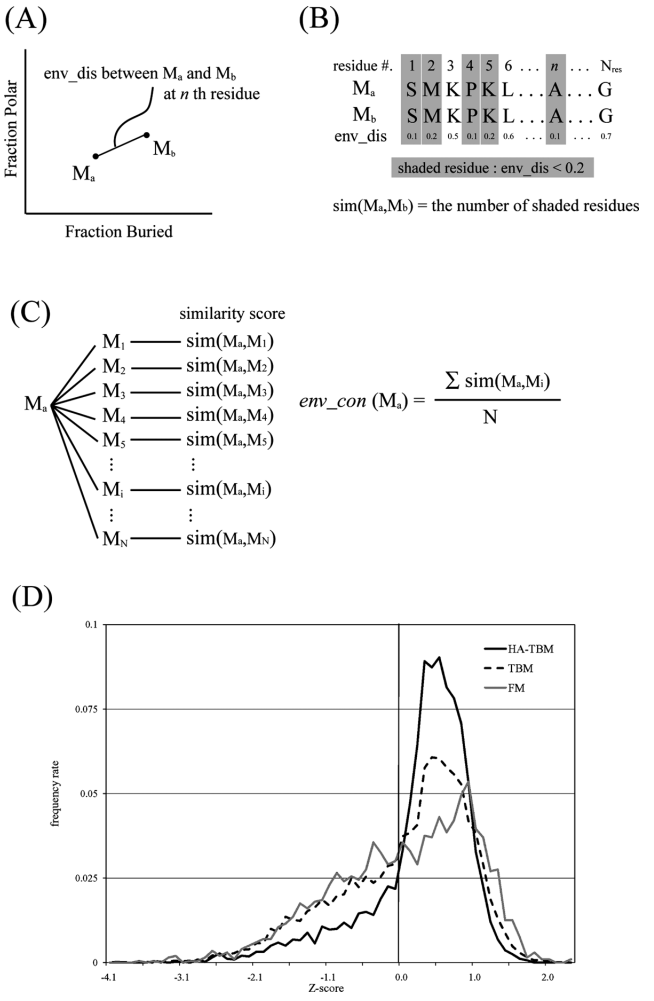

Fig. 1. Schematic Diagrams of the Calculation of the Environmental Consensus Score for a Particular Model M

(A) The environmental distance, *env_dis*, is defined as the Euclidian distance between the two side-chain environments of the corresponding amino acid residues of two models. The model $\mathrm{M_i}$ is a model in the model set. The $\mathrm{fb}(\mathrm{M}, n)$ and the $\mathrm{fp}(\mathrm{M}, n)$ are the 'fraction buried' for the nth residue of the model M and the 'fraction polar' for the nth residue of the model M, respectively. Since the Cα atom was included in the calculations of the fb and fp values, all amino acid residues including glycine residue have these values. (B) To calculate the environmental similarity score between the model M and the model $\mathrm{M_i}$, the *env_dis* mentioned in (A) is assigned to each residue. Residues with the $env\_dis <= 0.2$ are shaded. The environmental similarity score between model M and model $\mathrm{M_i}$ ($\mathrm{sim}(\mathrm{M}, \mathrm{M_i})$) is defined as the number of the residues with the $env\_dis <= 0.2$. (C) The environmental similarity scores between the model M assigned particularly and each model in the model set ($\mathrm{M_1}$—$\mathrm{M_N}$) were calculated. N means the number of models in the model set. In other words, N is the number of pairs between the model M and the model in the model set. $\mathrm{sim}(\mathrm{M}, \mathrm{M_i})$ is the environmental similarity score between the model M and the model $\mathrm{M_i}$ mentioned in (B). The environment consensus score, *env_con*, of the model M is the sum of the environmental similarity scores of the model M for each model in the model set divided by N. The model M is defined independently for the model set of $\mathrm{M_i}$ ($i = 1$ to N). (D) The distribution number of $env\_con_{native,n}$ of the native structures of the CASP7 targets against the horizontal axis. The horizontal axis is the Z-score of $env\_con_{native,n}$ for the set of the server models of each CASP7 target. The $env\_con_{native,n}$ is the *env_con* value at the nth residue of each native structure of the CASP7 targets. The Z-score of the $env\_con_{native,n}$ was calculated using the average and standard deviation values of the *env_con* for the server models of each CASP7 target. The vertical line at the Z-score = 0 represents the average of the *env_con* score among the CASP7 server models. Black, broken-black and gray lines are the Z-scores of env_con_{native} for the HA-TBM, TBM and FM targets, respectively. 75.6, 60.4 and 55.4% of the residues of the native structures for the HA-TBM, TBM and FM targets, respectively, are higher than the average score of the CASP7 server models. The higher ratio value (*i.e.* over 50%) shows the scientifically larger reliability of the consensus method. HA-TBM, TBM and FM, which represent high accuracy template based modeling, template based modeling and free modeling defined by CASP7 organizers, correspond to CMeasy, CMeasy+CMhard+FR and NF in our modeling difficulty presented in the Methods. Similarly, in the CASP8 targets, 74.6, 58.9 and 55.4% of the residues of the native structures for the HA-TBM, TBM and FM targets, respectively, are higher than the average score of the CASP8 server models. The data described above show that the *env_con* consensus method is suitable to select the model closest to the native structure.

on various values of *env_dis* including 0.2, model selection for each target of CASP7 was performed, and the summation of the GDT_TS value of each selected model was calculated as the total GDT_TS score. A particular model M is compared with each model in the model set, and the summation of the environmental similarity scores for each model in the model set was calculated as $\Sigma^{N} \sin(M, M_i)$ and is presented in Fig. 1C. N is the number of models in the model set for the specific target. As shown in Eq. 2, the environment consensus score, *env_con*, of model M is the summation of the environmental similarity scores for each model in the model set divided by the number of models in the model set:

$$env\_con(M) = \frac{\Sigma^{N} \mathrm{sim}(M, M_i)}{N} \quad (2)$$

For example, in the CASP8 experiment, the first (top) models from each server were used as the M_i model set. Since all the servers did not always submit models for all the target proteins, the value N was different for each target, and the average value and standard deviation of N were 65.4 and 2.9, respectively, in the CASP8 targets.

Figure 1C shows the explanation of *env_con*(M) in Eq. 2. Figure 1D shows the scientifically appropriate reason for the use of the *env_con* consensus score. The score of the native structures of the CASP7 targets are statistically higher than the average score, which is zero as the Z-score, for each of the CASP7 targets. Therefore, the investigation of the high *env_con* consensus score for each target is significantly connected with that of the corresponding native structure, and it has the scientific meaning as the method to select the structure near to the native.

The *env_con* score corresponding to the side-chain environments does not take into account the secondary structure agreement, although PSI-PRED,[16] which predicts secondary structures, achieved an average accuracy rate of about 80%. Consequently, a secondary structure agreement term was added to the *env_con* score term as a final Side-chain Environmental Consensus (SEC) score and is shown in Eq. 3.

$$SEC\ score(M) = Zscore(env\_con(M)) + w \times Zscore(SSscore(M)) \quad (3)$$

Here, *Zscore*(*env_con*(M)) and *Zscore*(*SSscore*(M)) represents the Z-score of the *env_con* score and the Z-score of *SSscore* for the model M, respectively. The *SSscore* represents the secondary structure agreement score which was calculated by comparison between the secondary structure of the 3D model and the secondary structure predicted from the sequence. The secondary structure prediction from the sequence was performed using PSI-PRED.[16] The details of this score were described in reference two.[15] The Z-score is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the standard deviation of the population. Before summing the two different measures, it was necessary to normalize these scores on a common scale. The symbol *w* in Eq. 3 is the weighting factor for the Z-score of *SSscore*. As shown in Table 1, the value of *w* is dependent on the target difficulty predicted by the Support Vector Machine (SVM)[17] as mentioned later. These *w* values were optimized using the training set based on CASP7 targets.[12] The optimization was performed by maximizing the sum GDT_TS, which is the summation of the GDT_TS value of the protein model with the highest SEC score for each target. The GDT_TS value was explained in the Introduction section, and, next, its calculation method is mentioned.

The GDT_TS value represents the correctness of the Cα backbone geometry of the model, and is defined as shown in Eq. 4[14]:

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

Here, *GDT_Pn* represents a percent of residues or Cα atoms separated by a distance shorter than *n* Å from a native structure. The *GDT_Pn* was calculated with a sequence-dependent superposition between a native and a model structure, where the number of residues which is separated by a distance shorter than *n* Å was maximized. The residue is represented by the Cα atom. The GDT_TS value is an average of GDT_P1, GDT_P2, GDT_P4 and GDT_P8, and ranges from zero to 100. A score of 100 for the GDT_TS indicates that all the Cα coordinates of the model structure are within 1 Å when compared with the native or experimental structure.

Target Difficulty Prediction The target difficulty is the difficulty of the protein modeling for a particular target. Generally, when the target protein has a template protein with high sequence identity, the difficulty of the protein modeling for the target protein is low. Inversely, when the target protein has no any template proteins with high sequence identity, the difficulty of the protein modeling for the target protein is high. In order to predict the target difficulty in Table 1, the SVM[17] program was used. The SVM program is based on a new type of learning machine, which is applied to pattern recognition estimations and other problems. Score and sequence identity (%) values of the best alignments resulting from PSI-BLAST[18] and SPARKS2[19] were used as vectors for difficulty classification in the SVM. Four difficulty classes were obtained from both alignment programs of PSI-BLAST and SPARKS2: CMeasy, CMhard, FR and NF in ascending order of difficulty. The two kinds of target difficulty classes resulting from the use of the two alignment programs were combined to identify the *w* value in the Eq. 3 as described in Table 1. The PSI-BLAST method is excellent for CMeasy and CMhard due to the base of sequence-profile alignments, whereas the SPARKS2 method is excellent for CMhard and FR due to the base of profile-profile alignments. Both methods were then used to accommodate the broad band in relation to the difficulty of the alignment. The *w* values in Table 1 were determined by maximizing the total GDT_TS value based on the training set consisting of the CASP7 target proteins. The PSI-BLAST and SPARKS2 programs were only used to classify the modeling difficulty of each target protein. The Z-score of *SSscore* was multiplied by a larger weight value in the case of difficult modeling targets than in the case of easy modeling targets. For the difficult targets, the *SSscore* term must be considered equally with the *env_con* term based on the Z-scores, whereas for the easy targets the *SSscore* term must be weighted by 30—50% to that of the *env_con* term. The *env_con* and *SSscore* terms are determined by the environmental states of the side-chains in the protein model and the agreement between the secondary structure of the predicted model and the secondary structure prediction for the query sequence, respectively.

Table 1. Optimized *w* Value Based on the Training Set of the CASP7 Targets

PSIB[a]	SPK2[b]	*w*
CMeasy	CMeasy	0.3
CMhard	CMeasy	0.3
CMeasy	CMhard	0.5
CMhard	CMhard	0.5
CMhard	FR	0.5
NF	CMhard	0.5
NF	FR	1.0
CMhard	NF	1.0
NF	NF	1.0

a) Predicted difficulty obtained from the alignment score and the sequence identity of PSI-BLAST. *b*) Predicted difficulty obtained from the alignment score and the sequence identity of SPARKS2.

The SEC Method in the CASP8 Experiment (FAMSD_QA) In the Seventh CASP (CASP7),[12] a new prediction category called Quality Assessment (QA) was implemented.[20] This prediction category was introduced to develop a model quality estimation method without information of the experimental structure of target protein. In this category, predictors estimate the quality of the 3D models which were automatically predicted within 3 d after the receipt of the target sequence by server teams. After the prediction period expires, the CASP assessors of the QA category computed the correlation between the observed quality in comparison with the 3D coordinates of the target protein and predicted quality of the models achieved by the participating teams.[20] We participated in the latest CASP experiment (CASP8)[13] in 2008 using the SEC method as a QA predictor called 'FAMSD_QA.'[21] The team name was 'FAMSD'; however, the 'FAMSD' team was also a tertiary structure (TS) predictor, and therefore the inclusion of the '_QA' in the name of FAMSD avoids confusion. In this section, we explain the FAMSD_QA method including the SEC method in the CASP8 experiment.

Initially, we refined all the server models using the homology modeling program FAMS[22] to complement missing side-chain atoms and decrease collisions between side-chain atoms. The FAMS program was performed using each server model as a template structure. Since the tertiary structures of other server teams sometimes include geometric structures unrelated to the native structures of the proteins, the above reconstruction by FAMS should be performed. Without this reconstruction procedure, the side-chain environment could not be calculated correctly. Our 3D coordinates for all the target proteins were used in the QA predictions of the CASP8 experiment as the FAMSD_QA team. The SEC scores of every refined model were calculated and normalized into the range [0, 1]. For each target protein, the

value one was given for the 3D model having the maximum value of the SEC scores in all the server models, and the value zero was applied as the minimum value of the SEC scores. Consequently, we submitted QA predictions in the range of [0, 1] for all 128 targets. Our submission data are shown in website (CASP8 homepage: http://predictioncenter.org/download_area/CASP8/predictions/QA.tar.gz). The results of FAMSD_QA in the CASP8 experiment, such as the correlation coefficients for each target protein between the GDT_TS and the QA prediction for 3D models of participating server teams, will be published by the CASP8 assessors. In this paper, on the other hand, we focus on the protein model with the highest SEC score for each target protein to select the model closest in structure to the native conformer. Here, we assessed the quality of models ranked first by the SEC method and performed comparison with other CASP8 server models which were constructed by other server teams participating in the CASP8 experiment.

Combination of the SEC Method with Other Methods We combined the SEC method with two other methods, 3D-Jury (3DJ)[1] and CIRCLE (CCL),[15] as an approach to improve model selection. These combined methods were not used in the CASP8 experiment. After the CASP8 experiment, the following methods were developed.

Initially, we combined our SEC method with the 3D-Jury method to increase the accuracy of the Cα backbone. The 3D-Jury method gives the consensus of the Cα atoms for each model constructed with various modeling algorithms. Thus two types of consensus methods for the side-chains and the main chain are included in the combined method. The combined score, *Com_score*, was calculated as:

$$Com\_score(SEC + 3DJ) = Zscore(3DJ) + w' \times Zscore(SEC) \quad (5)$$

Zscore(3*DJ*) and *Zscore*(*SCE*) are the Z-scores of the 3D-Jury score and the SEC score, respectively. The symbol w' in Eq. 5 is the weighting factor for the Z-score of the SEC score and was set to 0.5. The value of 0.5 was optimized value using the training set of the CASP7 targets by maximizing the sum of the Z-score values of GDT_TS and the "correct χ1" for all target proteins. The sum of the Z-score values of GDT_TS and "correct χ1" for a target protein is explained in the Eq. 6 described later. In this case, in relation to the importance of the correctness for all the coordinates of the main chain and the side-chains, the training is executed to obtain a model that closely matches the native structure. As shown by the value of the weighting factor $w'=0.5$, the consensus of the main chain may be superior to that of the side-chains.

Furthermore, we re-ranked the top 3% models ranked by the *Com_score*(*SEC*+3*DJ*) in Eq. 5 using the 3D–1D profile score of the CIRCLE (CCL) program[15] to obtain models with higher quality in both the Cα backbone and the side-chain atoms from a free energy point of view. The value of 3% was determined based on the training set of the CASP7 targets in a manner similar to the determination of the weighting factor w'. The number of top 3% models varies for each target protein because the total number of models for each target protein varied. For example, in the CASP8 experiment, the average value of the number of top 3% models was 9.2 per one target protein. This model selection method including the CIRCLE program is termed the 3DJ+SEC+CCL method. As the CIRCLE program gives the 3D–1D profile score for the protein structure, the combined contents of the free energy for the protein structure, the secondary structure agreement and the consensus of the main chain and side-chains are included in the 3DJ+SEC+CCL method as the estimation of the constructed model.

Results and Discussion

Comparison of the SEC and 3D-Jury Methods Using the experimental structure or 3D coordinates of the target protein for each of the CASP8 targets, we evaluated the accuracy of the 3D model ranked first by the SEC score actually used by us in the CASP8 experiment. The results are discussed by comparison between the SEC and 3D-Jury methods. We calculated the GDT_TS[14] value to assess Cα backbone geometry. Furthermore, the number of residues which have the correct side-chain torsion angles, χ1 and χ2, was used to assess side-chain conformations. The χ1 torsion angle was considered "correct" if the value was within 40 degrees of the experimental value.[23,24] The χ2 torsion angle was considered "correct" if both the χ1 and χ2 values were within 40 and 60 degrees, respectively. The number of collect χ1 and χ2 torsion angles was counted when the Cα atom is positioned within 3.5 Å from the native structure following superposition onto the native structure using the MaxSub fitting method.[25] To compare our method with the 3D-Jury method, we also assessed the models ranked first by the 3D-Jury score. In CASP8, 121 of 128 targets were used for the assessment and they were divided into 162 domains by the CASP8 assessors.[13] We conveniently classified these domains into four categories based on the average GDT_TS values of the top 10% server models for each domain: "easy" (>70), "medium" (50—70), "hard" (30—50) and "very hard" (<30). We calculated the sum of GDT_TS, average GDT_TS, sum of the correct side-chains based on χ1 or χ2 values and the percentage of the correct side-chains for each category. However, in the absence of the correct protein backbone structure, the assessment of the side-chain quality of the models is meaningless. In other words, since the GDT_TS value is the index expressing the correctness of the Cα backbone or the folding conformation near to the native structure, two categories of "hard" and "very hard" defined above were slighted in discussing the assessment of both qualities of the main chain and the side-chains. Therefore, we paid special attention to "easy" and "medium" classes because the correctness of the protein model is not guaranteed when the GDT_TS value is below 50. We typically dealt with three groups based on domains, which were "easy" domains (above 70), "easy"+"medium" domains (above 50) and all domains including the "hard" and "very hard" domains (above 0). Table 2 shows the results of both the SEC and 3D-Jury methods for these three groups. As for the sum and the average of the GDT_TS values, which represent the quality of the backbone modeling, the SEC method was found to be slightly worse than the 3D-Jury method. However, the ratios of correctly predicted χ1 and χ2 values were higher using the SEC method. In particular, in the "easy" group, the ratios of correctly predicted χ1 and χ2 values using the SEC method were better than the 3D-Jury method by 4.7% and 4.3%, respectively. Therefore, the 3D-Jury method does not necessarily select models with good side-chain quality due to no consideration about the correctness of side-chain atoms. In contrast, since our SEC score considers the side-chain environment and the secondary structure agreement of the main chain, this method naturally ensures the selection of good side-chain models with well positioned Cα backbone geometries.

Performance of the Three SEC-Related Methods We examined the performance of three SEC-related methods using the CASP8 targets which were not used in the training set to determine various parameters such as the threshold value, 0.2, of the environmental distance in the Eq. 1, the weighting factor w (0.3, 0,5, 1.0) of the term of secondary structure agreement in the Eq. 3, the weighting factor, 0.5, of the term of the Side-chain Environmental Consensus (SEC) score in the Eq. 5 and the cut-off value, 3%, of the 3DJ+SEC+CCL method. A comparison between the SEC method and the combined method (3DJ+SEC) showed that the average GDT_TS improved slightly from 73.7 to 74.3 without reducing the χ1 and χ2 quality values of the side-chains (Table 3). Thus, the 3DJ+SEC method which has not been reported in other papers was successful. Furthermore, the quality val-

Table 2. Comparison between SEC and 3D-Jury

		"easy"[a]			"easy"+"medium"[b]			ALL[c]		
		3DJ[d]	SEC[e]	diff[f]	3DJ	SEC	diff	3DJ	SEC	diff
GDT_TS	Sum	7916.07	7828.33	−87.74 (−1.12%)	10079.06	9945.39	−133.67 (−1.34%)	11020.08	10896.86	−123.22 (−1.13%)
GDT_TS	Average	81.61	80.7	−0.91 (−1.13%)	74.66	73.67	−0.99 (−1.34%)	68.03	67.26	−0.77 (−1.14%)
$\chi 1$	Sum	5919	6464	**+545 (+8.43%)**	7493	8116	**+623 (+7.68%)**	7977	8632	**+655 (+7.59%)**
	%[g]	51.20%	55.91%	**+4.71% (+8.43%)**	45.53%	49.32%	**+3.79% (+7.68%)**	40.66%	44.00%	**+3.34% (+7.59%)**
$\chi 2$	Sum	3257	3637	**+380 (+10.45%)**	4079	4474	**+395 (+8.83%)**	4331	4743	**+412 (+8.69%)**
	%[h]	36.80%	41.10%	**+4.29% (+10.45%)**	32.10%	35.20%	**+3.11% (+8.83%)**	28.38%	31.08%	**+2.70% (+8.69%)**

a) "easy" in which the average GDT_TS of the top 10% servers is >70 consists of 97 domains. *b*) "easy"+"medium" in which the average GDT_TS of the top 10% servers is >50 consists of 135 domains. *c*) ALL consists of 162 domains. *d*) First-ranked models by 3D-Jury. *e*) First-ranked models by SEC. *f*) The difference calculated by subtracting 3DJ from SEC. *g*) Percentage of the correct $\chi 1$ values. *h*) Percentage of the correct $\chi 2$ values. The text is bold when the SEC method was better than the 3D-Jury method. The value in parentheses (as percentages) is the increase in the rate that was calculated using the value obtained from the 3D-Jury method as the standard.

Table 3. Comparison between 3D-Jury and the SEC-Related Methods

		"easy"+"medium" targets						
		3DJ[a]	SEC[b]	diff[e]	3DJ+SEC[c]	diff[f]	3DJ+SEC +CCL[d]	diff[g]
GDT_TS	sum[h]	10079.06	9945.39	−133.67 (−1.34%)	10027.21	−51.85 (−0.51%)	10076.81	−2.25 (−0.02%)
GDT_TS	average[i]	74.66	73.67	−0.99 (−1.34%)	74.28	−0.38 (−0.51%)	74.66	0 (0%)
$\chi 1$	sum[j]	7493	8116	**+623 (+7.68%)**	8208	**+715 (+9.54%)**	8346	**+853 (+11.38%)**
	%[k]	45.53%	49.32%	**+3.79% (+7.68%)**	49.88%	**+4.35% (+9.54%)**	50.71%	**+5.18% (+11.39%)**
$\chi 2$	sum[l]	4079	4474	**+395 (+8.83%)**	4499	**+420 (+10.30%)**	4652	**+573 (+14.05%)**
	%[m]	32.1%	35.20%	**+3.11% (+8.83%)**	35.40%	**+3.30% (+10.30%)**	36.60%	**+4.50% (+14.03%)**

a) The 3D-Jury method. *b*) The SEC method. *c*) Combined 3D-Jury and SEC method. *d*) Combined 3D-Jury, SEC and CIRCLE method. *e*) Difference calculated by subtracting 3DJ from SEC. The value in parentheses (as percentages) is the increase in the rate that was calculated using the value obtained from the 3D-Jury method as the standard. *f*) Difference calculated by subtracting 3DJ from 3DJ+SEC. The value in parentheses (as percentages) is the increase rate which was calculated using the value obtained from the 3D-Jury method as standard. *g*) Difference calculated by subtracting 3DJ from 3DJ+SEC+CCL. The value in parentheses (as percentages) is the increase rate which was calculated using the value obtained from the 3D-Jury method as standard. *h*) Summation of the GDT_TS value for 135 domains in the "easy"+"medium." *i*) Average the GDT_TS value for the 135 domains in the "easy"+"medium." *j*) Summation of the number of the correct $\chi 1$ for the 135 domains in the "easy"+"medium." *k*) Percentage of the correct $\chi 1$ for the 135 domains in the "easy"+"medium." *l*) Summation of the number of the correct $\chi 2$ for the 135 domains in the "easy"+"medium." *m*) Percentage of the correct $\chi 2$ for the 135 domains in the "easy"+"medium." The shaded text indicates the difference from 3DJ for each SEC-related method. The text is bold when the SEC-related method was superior to the 3D-Jury method.

Table 4. Ratios of the Correctly Predicted $\chi 1$ and $\chi 2$ Values of the Side-Chains for the 3D-Jury Method and the 3DJ+SEC+CCL Method against the 0.1 Fraction Buried (fb) Bands Determined between 0.0 and 1.0

fb[a]	$\chi 1$[b]			$\chi 2$[c]		
	3DJ[d]	3DJ+SEC+CCL[e]	diff[f]	3DJ[d]	3DJ+SEC+CCL[e]	diff[f]
0.0—0.1	18.9%	22.4%	+3.5%	13.2%	16.5%	+3.3%
0.1—0.2	29.4%	32.2%	+2.8%	20.0%	21.2%	+1.2%
0.2—0.3	29.5%	34.1%	+4.5%	20.2%	23.9%	+3.7%
0.3—0.4	34.3%	38.6%	+4.2%	23.2%	27.1%	+3.9%
0.4—0.5	37.3%	40.0%	+2.8%	26.8%	29.2%	+2.4%
0.5—0.6	36.1%	41.0%	+4.9%	26.0%	29.3%	+3.3%
0.6—0.7	37.8%	43.4%	+5.6%	25.9%	30.3%	+4.4%
0.7—0.8	40.6%	46.9%	+6.4%	27.6%	34.1%	+6.5%
0.8—0.9	43.4%	49.0%	+5.5%	30.1%	34.6%	+4.5%
0.9—1.0	49.8%	56.1%	+6.4%	36.1%	42.1%	+5.9%

a) Fraction of the native structure buried. A value of 1.0 means the state which is completely buried in the protein. *b*) Percentage of the correct $\chi 1$. *c*) Percentage of the correct $\chi 2$. *d*) The 3D-Jury method. *e*) Combined 3D-Jury, SEC and CIRCLE method. *f*) Difference calculated by subtracting the 3DJ from the 3DJ+SEC+CCL.

ues of the side-chains improved by re-ranking with the CIRCLE score (3DJ+SEC+CCL). Compared with the 3D-Jury method, the ratios of the correctly predicted $\chi1$ and $\chi2$ values improved from 45.5 to 50.7% and from 32.1 to 36.6%, respectively.

In order to analyze the increases in the ratios of the correctly predicted $\chi1$ and $\chi2$ values, the $\chi1$ and $\chi2$ quality values of the side-chains for the 3D-Jury and the 3DJ+SEC+CCL methods and the given difference between the two $\chi1$ and $\chi2$ quality values of each method were calculated against the 0.1 fraction buried (fb) band presented in Table 4. In the fb bands of 0.6—0.7, 0.7—0.8, 0.8—0.9 and 0.9—1.0, the $\chi1$ and $\chi2$ quality values in the 3DJ+SEC+CCL increased by 5.5 to 6.4% and 4.4 to 6.5%, respectively, in comparison with 3DJ. In the fb bands of 0.0—0.1, 0.1—0.2, 0.2—0.3, 0.3—0.4, 0.4—0.5 and 0.5—0.6, the $\chi1$ and $\chi2$ quality values in the 3DJ+SEC+CCL increased by 2.8 to 4.9% and 1.2 to 3.9%, respectively, in comparison with 3DJ. Therefore, it was shown that the 3DJ+SEC+CCL method gives larger quality ratios of the $\chi1$ and $\chi2$ values not only in the buried regions of the protein but also for residues located on the surface of the protein. The functional importance of buried or solvent exposed side-chains depends on the function of the protein under investigation. The 3DJ+SEC+CCL method gave more correctly predicted $\chi1$ and $\chi2$ values as a whole and represents a better modeling method for examining the side-chain structure-function relationship in proteins.

As shown in Fig. 2, the quality of models ranked first by our three SEC-related methods was compared with other CASP8 server models. In the assessment for the total Z-score of the GDT_TS in the CASP8 experiment, the Zhang-Server and RAPTOR were ranked in the top 2 of the all 71 servers that participated in the CASP8.[14] Therefore, the two servers, the Zhang-Server and RAPTOR, were included in this figure. In Fig. 2, "easy"+"medium" domains (135 domains) were used to calculate the accuracies of the models. The broken line with cross symbols represents the average GDT_TS. The three SEC-related methods were comparable in level to the top-ranked Zhang-Server. The broken lines with triangle and square symbols represent the ratios of correct $\chi1$ and $\chi2$ values, respectively, which are graduated in the right vertical axis. The re-ranking method using the CIRCLE program (*i.e.* the 3DJ+SEC+CCL method) performed better than the Zhang-Server in both the $\chi1$ and $\chi2$ accuracies. The bar represents the summation of the combined Z-score for the same 135 domains. The combined Z-score ($Z_{combined}$) was calculated as the summation of the Z-scores for the GDT_TS and the number of correct $\chi1$ torsion angles:

$$Z_{combined} = (Z_{GDT\_TS} + Z_{\chi1}) \quad (6)$$

Here, Z_{GDT_TS} and $Z_{\chi1}$ represent the Z-scores for GDT_TS and the number of correct $\chi1$ torsion angles, respectively. Higher $Z_{combined}$ values indicate that the main chain and the side-chains were both structurally similar to the native structure in relation to Z_{GDT_TS} and $Z_{\chi1}$, respectively. In Table 2, the six teams (the 3D-Jury method, the three SEC-related methods and the two CASP8 servers (Zhang-Server and RAPTOR)) are sorted in descending order of the $Z_{combined}$ value, which are graduated in the left vertical axis. In the assessment of this $Z_{combined}$ value, the 3DJ+SEC+CCL method did better than the Zhang-Server which is the best server among the CASP8 servers, although the SEC method and the 3DJ+SEC method followed the Zhang-Server. However, it should be remembered in the above comparison that models created by the Zhang-Server are used in the SEC-related methods as shown in the explanation of Table 2.

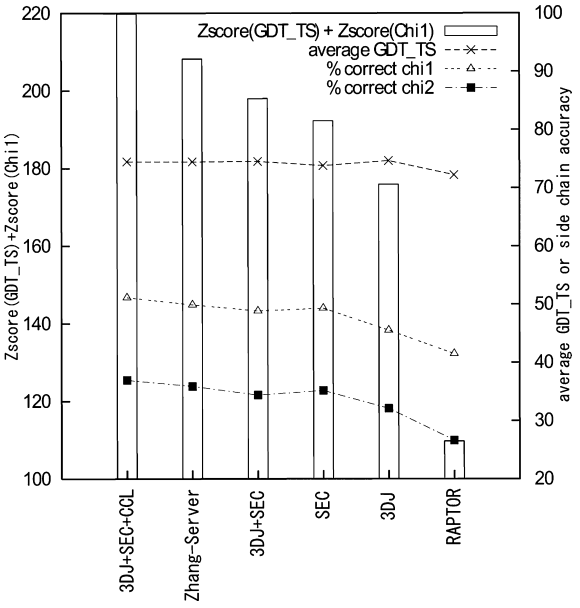

Fig. 2. Comparisons between 3DJ, Our Three SEC-Related Methods (SEC, 3DJ+SEC and 3DJ+SEC+CCL) and the Top Two CASP8 Server Teams: the Zhang-Server and RAPTOR

The bar represents the sum of the combined Z-score ($Z_{combined}$) in Eq. 6 (the left vertical axis). The $Z_{combined}$ was calculated as the summation of the Z-scores for the GDT_TS and the number of correct $\chi1$ torsion angles. The broken lines with the cross, triangle and square symbols represent the average GDT_TS, percentage of correct $\chi1$ values and percentage of correct $\chi2$ values, respectively (right axis). The average values of all 71 CASP8 servers for the sum of the $Z_{combined}$, the average GDT_TS, the correct $\chi1$ % and correct $\chi2$ % are −0.04, 60.3, 41.7 and 21.9, respectively. The values of sum $Z_{combined}$, average GDT_TS, percentage of correct $\chi1$ values and percentage of correct $\chi2$ values for the Zhang-Server model structures are 208.2, 74.5, 49.9 and 35.9 before the remodeling by our FAMS homology modeling program, and they are 207.0, 74.0, 49.9 and 35.9 after the remodeling by the FAMS. Although the remodeling process improves short-contacts of atom–atom with no appearance in the native structures, it does not increase the model accuracy indicated by the GDT_TS value. About the ratio % of the number which were selected as the top team for each target by using the 3DJ method, the Zhang-Server, the HHpred4, the MULTICON-CLUST, the METATASSER, the FALCON-CONSENSUS and the MULTICOM-REFINE selected above 5% were 29.7, 7.8, 7.0, 6.3, 6.3 and 5.5%, respectively. About the ratio % of the number which were selected as the top team for each target by using the SEC method, the Zhang-Server, the LEE-SERVER, the MUProt and the BAKER-ROBETTA selected above 5% were 13.3, 10.9, 9.4 and 5.5%, respectively. About the ratio % of the number which were selected as the top team for each target by using the 3DJ+SEC method, the Zhang-Server, the MUProt and the LEE-SERVER selected above 5% were 20.3, 10.2 and 7.0%, respectively. About the ratio % of the number which were selected as the top team for each target by using the 3DJ+SEC+CCL method, the Zhang-Server, the BAKER-ROBETTA, the MULTICOM-REFINE, the MUProt and the LEE-SERVER selected above 5% were 26.6, 10.9, 8.6, 6.3 and 6.3%, respectively. When the Zhang-Server models were excluded from server models, the values of $Z_{combined}$ are 196.1, 176.8 and 162.8 for the 3DJ+SEC+CCL, the 3DJ+SEC and the SEC, respectively, and, the other hand, the values of $Z_{combined}$ are 206.4 and 107.9 for the top Zhang-Server and the second RAPTOR, respectively; the three values, 196.1, 176.8 and 162.8, excluding the Zhang-Server are inserted between both values of the top two server teams.

Good Examples of the SEC-Related Methods for the CASP8 Targets T0447 is one of the CASP8 targets, the putative Formyltetrahydrofolate Synthetase (TM1766) from *Thermotoga maritima* (pdb code 3DO6) which consists of 542 residues. Many CASP8 servers predicted this target well, and the average GDT_TS of all server models and that of the top 10% server models were 64.50 and 89.02, respectively. The 3D-Jury method which is the Cα backbone consensus method selected the HHpred5_TS1 model whose GDT_TS was 88.24. The best GDT_TS among all server models of

Table 5. Correctness of Our Three Models Obtained from the Three Methods, SEC, 3DJ+SEC and 3DJ+SEC+CCL in Comparison with Two Models Obtained from the Two Methods, Best GDT_TS and 3DJ, Tested on One of the CASP8 Targets, T0447

Model	GDT_TS	$\chi1$[a)]	$\chi2$[b)]
Best GDT_TS[c)]	89.30	312	188
3DJ[d)]	88.24	243	132
SEC[e)]	87.92	319	181
3DJ+SEC[f)]	87.96	319	190
3DJ+SEC+CCL[g)]	87.59	319	195

a) $\chi1$ is the number of correct $\chi1$ torsion angles. *b*) $\chi2$ is the number of correct $\chi2$ torsion angles. *c*) Best GDT_TS is the highest GDT_TS model among all CASP8 server models for the T0447 target. *d*) 3DJ indicates the model selected by the 3D-Jury method. *e*) SEC indicates the model selected by the side-chain environment consensus (SEC) method. *f*) 3DJ+SEC indicates the model selected by the combined method of the 3D-Jury and the SEC score. *g*) 3DJ+SEC+CCL indicates the model selected by the combined method of the 3D-Jury, the SEC and the CIRCLE score. Our SEC-related methods were shaded. This table indicates that the selections of a protein model using the SEC-related methods were achieved with very high quality in both the Cα backbone and side-chains.

this target was 89.30, so the difference in the two GDT_TS values was 1.06, which is very small in comparison with 100 of full marks (Table 5). The selection of a high quality model was successful in terms of the GDT_TS value, *i.e.* accuracy of the Cα backbone geometries. However, the number of "correct $\chi1$" values of the model selected by the 3D-Jury method was 243, and significantly lower than the highest value of 319 obtained from a different server model. As such, the 3D-Jury model showed very high quality selection based on the backbone prediction but showed weak model selection for side-chains positions. Conversely, the SEC method selected the SAM-T08-server_TS3 model whose GDT_TS and the number of "correct $\chi1$" were 87.92 and 319, respectively. Thus our SEC methods could select a protein model with very high quality in both the Cα backbone atom positions and side-chain atoms positions. Although the 3DJ+SEC+CCL method was not used in the CASP8 experiment, it also selected a high quality model (Table 5). This may become a powerful method as a participating team in the CASP9 (2010) experiment which will greatly contribute to the progress of the protein modeling techniques.

For target T0447, Fig. 3 shows the superposition of the native structure (white), the model selected by the 3D-Jury method (gray) and the model selected by the 3DJ+SEC+CCL method (black) for particular side-chains which have been labeled. In Fig. 3A, residues L142, I144, V149 and V221 (gray) are positioned in different orientations when compared to the corresponding position of the residues in the other two structures. The side-chains for these residues from the native structure and 3DJ+SEC+CCL model are in good agreement. In Figs. 3B, C and D, similarly, residues F150, I211, I220, R222 (Fig. 3B), T154, I183, T185 (Fig. 3C), T60, S85 and I255 (Fig. 3D) are in different spatial positions for the 3D-Jury model when compared to positions of the same side-chains in the native and 3DJ+SEC+CCL structures. Higher accuracies of the torsion angles of the side-chains in the 3DJ+SEC+CCL model than in the 3DJ model (Table 5) were explained from the comparisons of the deviations of the side-chains from the native conformation. The environment of fb and fp for a residue is determined in the free energy state of the residue of the protein. Therefore, it is reasonable that the modeling accuracy is improved as a whole in the inclusion of the side-chain environment.

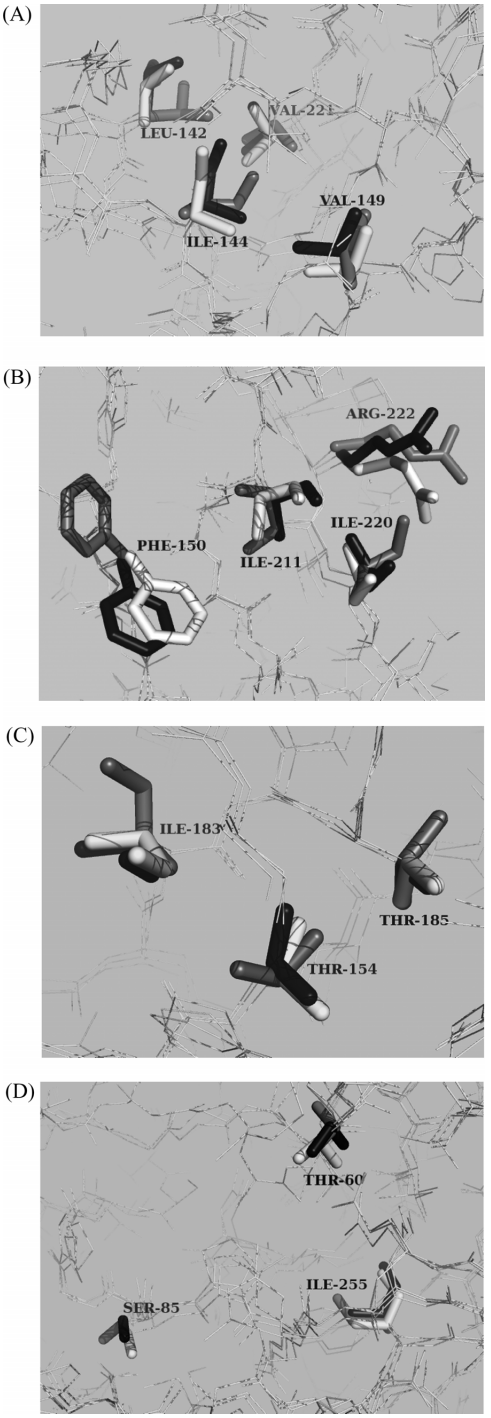

Fig. 3. Comparisons between the Native Structure, Model Selected by 3D-Jury Method and Model Selected by the 3DJ+SEC+CCL Method for Target T0447

Superposition of native structure, model selected by 3D-Jury method and model selected by the 3DJ+SEC+CCL method for target T0447. White, gray and black show native structure, model selected by the 3D-Jury method and model selected by the 3DJ+SEC+CCL method, respectively. It is shown that white side-chains of native structure overlap to black side-chains of 3DJ+SEC+CCL model more than gray side-chains of 3D-Jury model. Since the conformation of the side-chain is intrinsically related to the function of the protein (*e.g.* the catalytic triad consisting of histidine, serine and aspartic acid residues in the serine-proteases is a good example[35]), the conformations of the black colored side-chains of the 3DJ+SEC+CCL model may be useful in explaining the biological function of the target T0447. In this paper, the PyMol program[36] was used to present the protein structures.

Cα-atom fluctuations in the protein for the target T0447 were calculated with Normal Mode Analysis (NMA),[26—30]

which is an analysis method that uses harmonic molecular dynamics. The Cα-atom fluctuations of the native structure, the model selected by the 3D-Jury and the model selected by the 3DJ+SEC+CCL were near identical as a whole. As shown in Fig. 4, however, when we compared the Cα-atom fluctuations of the three structures in the peptide regions of 206—230, 239—267 and 449—469, the superposition of the 3DJ+SEC+CCL model with the native structure was better than the superpositioning between the native structure and the 3DJ model. Therefore, it was shown that the higher accuracy of the side-chains of the 3DJ+SEC+CCL model brings the harmonic molecular dynamics of the main chain closer to the native structure. Since protein dynamics is generally related to biological functions such as protein–protein interactions[27,28,30] and protein–ligand docking,[31] the 3DJ+SEC+CCL method may provide biologically valuable 3D models.

For the target T0412, in order to show better quality of the side-chains in the 3DJ+SEC+CCL method, we described the positions of L47 and F49 in Fig. 5A and N31, D32 and I33 in Fig. 5B. The side-chains of these residues in the 3DJ+SEC+CCL model are in good agreement with side-chain positions in the native structure. However, the side-chains of the best GDT_TS model are positioned in different orientations with respect to the spatial orientations of these side-chains in the native structure. The target T0412 is a good example, in which the side-chain conformations in the 3DJ+SEC+CCL method are much closer to the native structure in comparison with the best GDT_TS model. Again, the 3DJ+SEC+CCL method giving the structure nearer to native means that the methods considering the free energy state such as SEC and CCL are important in the protein modeling.

For the target T0511, in order to show better quality of the side-chains in the 3DJ+SEC+CCL method, we described the positions of I70 and N75 in Fig. 6A and N96 and L146 in Fig. 6B. Side-chain positions in the 3DJ+SEC+CCL method are in good agreement with side-chains positions in the native structure. In contrast, for the 3DJ+SEC model, the position of the side-chains differs to the position of the same side-chains in the native structure. The target T0511 is a good example, in which the side-chain conformations in the 3DJ+SEC+CCL method are much closer to the native structure in comparison with the 3DJ+SEC method. This fact means that both methods of SEC and CCL considering the free energy state are significant in the protein modeling.

Poor Example of SEC-Related Methods for the CASP8 Targets T0498 is one of the CASP8 targets and consists of 56 residues. In this target, many template proteins with >50% sequence identity were obtained, and these templates were classified into two folds which were dissimilar. One fold is an all-alpha protein such as pdb code 2FS1 (Fig. 7B),

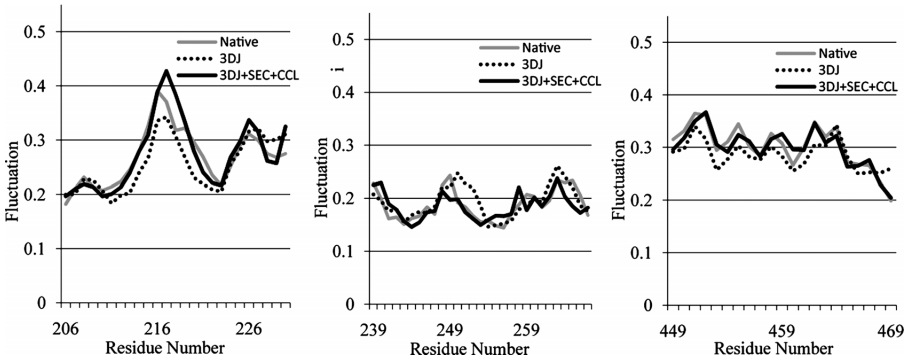

Fig. 4. Cα-Atom Fluctuations Calculated with Normal Mode Analysis (NMA)[26] for Target T0447

The gray, broken and black lines denote Cα-atom fluctuations of the native structure (pdb code 3DO6), the model selected by the 3D-Jury method and the model selected by the 3DJ+SEC+CCL method, respectively. In three regions of target sequence, 206—230, 239—267 and 449—469, comparison of Cα-atom fluctuations is shown. In these regions, the Cα-atom fluctuations of the 3DJ+SEC+CCL based model were nearer to that of the native structure than the 3D-Jury based model, although the Cα-atom fluctuations of both the 3D-Jury based model and the 3DJ+SEC+CCL based model were similar to that of the native structure as a whole.

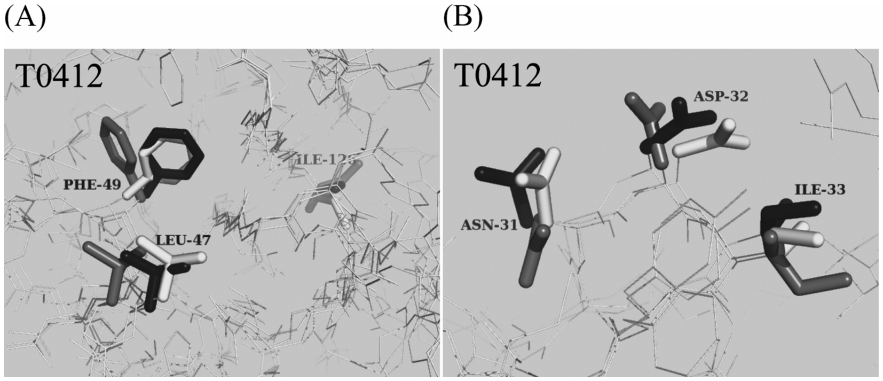

Fig. 5. Comparisons between the Native Structure, the Beset GDT_TS Model and Model Selected by the 3DJ+SEC+CCL Method for Target T0412

White, gray and black show native structure, the best GDT_TS model and model selected by the 3DJ+SEC+CCL method (3DJ+SEC+CCL based model), respectively. The GDT_TS, correctly predicted χ1 and correctly predicted χ2 values for the best GDT_TS model were 79.7, 54 and 33, respectively. The GDT_TS, correctly predicted χ1 and correctly predicted χ2 values for the 3DJ+SEC+CCL based model were 78.2, 67 and 39, respectively. The target T0412 is a good example, in which the side-chain conformations in the 3DJ+SEC+CCL method are near to the native structure in the comparison with the best GDT_TS model.

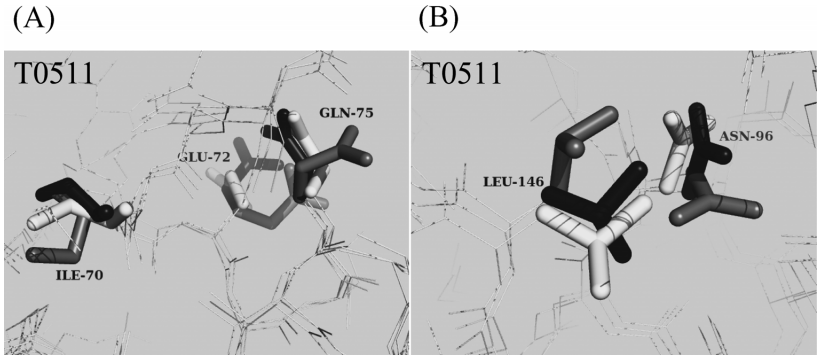

Fig. 6. Comparisons between the Native Structure, the Model Selected by the 3DJ+SEC Method and the Model Selected by the 3DJ+SEC+CCL Method for Target T0511

White, gray and black represent the native structure, the model selected by the 3DJ+SEC method (3DJ+SEC based model) and the model selected by the 3DJ+SEC+CCL method (3DJ+SEC+CCL based model), respectively. The GDT_TS, correctly predicted $\chi 1$ and correctly predicted $\chi 2$ values for the 3DJ+SEC based model were 81.5, 95 and 55, respectively. The GDT_TS, correctly predicted $\chi 1$ and correctly predicted $\chi 2$ values for the 3DJ+SEC+CCL based model were 83.7, 107 and 63, respectively. The target T0511 is a good example, in which the side-chain conformations in the 3DJ+SEC+CCL method match closely the native structure in comparison with the model selected using the 3DJ+SEC method.

Table 6. Modeling or Alignment Servers Located on the Internet That Were Used for the 3D Structure Prediction of Human Cabin1

No.	Server name	URL in the internet	Number of candidate models
1	3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html	2
2	FFAS03	http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl	5
3	FUGUE	http://tardis.nibio.go.jp/fugue/prfsearch.html	2
4	genThreader	http://bioinf.cs.ucl.ac.uk/psipred/psiform.html	5
5	I-TASSER	http://zhang.bioinformatics.ku.edu/I-TASSER/	5
6	PHYRE	http://www.sbg.bio.ic.ac.uk/~phyre/	1
7	SAM-T02	http://compbio.soe.ucsc.edu/HMM-apps/T02-query.html	5
8	SP3	http://sparks.informatics.iupui.edu/hzhou/anonymous-fold-sp3.html	5
9	SPARKS2	http://sparks.informatics.iupui.edu/hzhou/sparks2.html	5
10	FAMSD	Our modeling method[32]	5

Forty candidate models corresponding to the model M in Fig. 1C were obtained from the 10 servers. The numbers of 2, 5, 2, 5, 5, 1, 5, 5, 5 and 5 are the number of the candidate models obtained from No. 1 to No. 10 servers, respectively. The total number of candidate models from the ten servers was 40. The set of 10 models corresponding to the model set, M_1 to M_N, in Fig. 1C were determined from 10 models that represented each of 10 internet servers.

whereas the other fold is an alpha–beta protein such as pdb code 2IGD (Fig. 7C). The experimental structure of this target was similar to pdb code 2FS1 like fold (Fig. 7A). However, most of the CASP8 server models were constructed based on the pdb code 2IGD like proteins. The models which have a pdb code 2FS1 like fold (correct fold) were a minority in the CASP8 server models. In this case, the consensus methods such as the 3D-Jury or our SEC-related method selected a model which was the incorrect fold, because the majority of the server models were incorrect fold. Thus, the consensus methods such as the 3DJ method and the SEC-related methods will fail to select the correct models if the majority of the servers make incorrect predictions, even if there are some correct models present in all the server models.

Application of the 3DJ+SEC+CCL Method for Human Cabin1 Protein Human Cabin1 (also known as Cain) is a ubiquitously expressed 2220 residue protein that regulates protein phosphatase activity of calcineurin and the transcriptional activity of myocyte enhancer factor 2 (Mef2).[4—6] Jang *et al.* have reported that Cabin1 regulates expression of a subset of p53 target genes in both human and mouse cells in the absence of genotoxic stress.[4—6] Furthermore, they had reported that Cabin1 physically interacts with p53 and negatively regulates p53 on specific p53 target promoters by regulating chromatin structure.[4] We implemented the 3D structure prediction of human Cabin1 using the 3DJ+SEC+CCL method. We found modeling or alignment servers on the internet as shown in Table 6. After we determined the sequence region to create a model by our FAMSD method,[32] we submitted 450 residues of the N-terminal sequence of Cabin1 to servers in Table 6. Alignments obtained from the alignment servers were used to construct the 3D models with the FAMS program. For models obtained from the modeling servers, the reconstruction procedure with the FAMS program was performed to decrease collisions between side-chain atoms. Consequently, forty candidate models were obtained. The average value and standard deviation of the sequence identity between the target and template proteins were 12.4 and 4.3%, respectively. The modeling for the 12.4% value of the sequence identity is generally thought to be very difficult. The 3DJ+SEC+CCL method was used to select the best model among the forty candidate models. Ten models that were ranked first in each of the ten servers were then used as the model set presented in Fig. 1C. As a result, the model based on the SP3 method was selected from the forty candidate models. Figures 8A and B show the 3D models for human Cabin1, which were predicted using the modeling or alignment servers, and Figs. 8C and D show the model obtained using the 3DJ+SEC+CCL method. The 3D coordinates of this model are obtained at http://mammalia.gsc.riken.jp/human_famsd/SEC/human_cabin1.pdf. This model may provide insight into the structure-function rela-

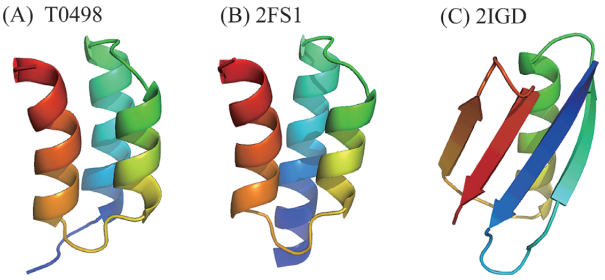

Fig. 7. A Comparison between the Native Structure of T0498, 2FS1 Structure and 2IGD Structure

(A) The native structure of T0498. N- and C-terminal regions of the protein are colored red and blue, respectively. (B) The NMR structure of pdb code 2FS1. This structure is similar to the experimental structure of T0498. The sequence identity with the target protein (T0498) was 54%. (C) The X-ray structure of pdb code 2IGD. This structure is not similar to the experimental structure of T0498. The sequence identity with the target protein (T0498) was 57%.

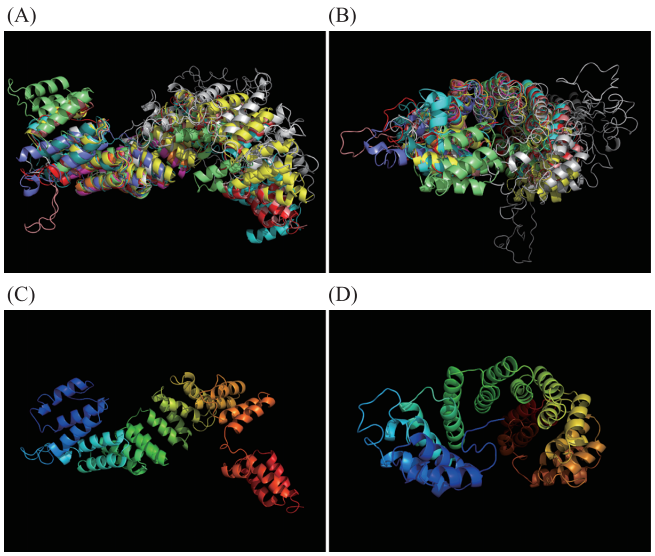

Fig. 8. Predicted 3D Models for Human Cabin1 Using the 3DJ+SEC+CCL Method

(A) and (B) show the 3D models for human Cabin1 which were predicted using the modeling or alignment servers, and (C) and (D) show the models derived from the 3DJ+SEC+CCL method. Models (A) and (C) are presented in a different orientation to models (B) and (D). The 3D coordinates of this model are available at http://mammalia.gsc.riken.jp/human_famsd/SEC/human_cabin1.pdf. This model should be useful in the pharmaceutical, medicinal and biological fields interested in understanding the relationship between the structure and function of human Cabin1. In (A) and (B), models obtained from various servers are colored to be unit colors for each server model, and the SP3 based model which was selected by the 3DJ+SEC+CCL method is colored red. In (C) and (D), N- and C-terminal regions of the protein are colored red and blue, respectively.

tionship of the protein and therefore may be useful in the pharmaceutical, medicinal and biological fields.

Conclusion

We have developed new consensus methods for the purpose of selecting high quality models which account for both the Cα backbone and side-chain atom positions. The new consensus methods are based upon the consideration of the side-chain environment, which is determined in the free energy state of the amino acid residue of the protein. As shown in Fig. 1, this SEC method employs a very simple algorithm, and it is scientifically appropriate consensus method in the set of all the server models for the CASP7 or CASP8 targets. The SEC method reinforces the traditional consensus method for the main chain, 3D-Jury, in terms of selecting models with side-chains of high quality. Thus, models were selected with improved quality both in the Cα backbone and side-chain positions when the SEC method was used in combination with the 3D-Jury method (3DJ+SEC). Accordingly, the 3DJ+SEC method including both consensus of the main chain and the side-chains in addition to the secondary structure agreement is useful. Moreover, in the calculations of the combined score, *Com_score*(3DJ+SEC) in Eq. 5, contribution of the Z-score term of the 3D-Jury score was greater than that of the SEC score as shown by the weighing factor w' in Eq. 5. As such, the consensus of the backbone is more important than the consensus of the side-chains during selection process of a 3D protein model.

The 3D–1D profile score CIRCLE has been shown to select the model that is closest to the native structure. We have combined the CIRCLE score (CCL) with the 3DJ+SEC consensus method. The 3DJ+SEC+CCL method includes the 3D–1D profile score based upon the residue unit of the amino acid, the score of the secondary structure agreement and the consensus methods for both the main chain and the side-chains of a protein model. The consensus method of the side-chain environment is based on the consensus of the free energy state of the residue of the protein. The 3DJ+SEC+CCL method, in which we use the CIRCLE score after reducing the number of candidates using the consensus method, may be very effective in the biological, pharmaceutical and medicinal fields. The example of a modeling application was presented in the Results and Discussion using the human Cabin1 protein, which has functional connections with p53 activity and cancer.

Recently, top-ranked servers in the CASP8 experiment such as the Zhang-server[33] and ROBETTA[34] became available to the science community *via* websites. If more servers that participated in CASP8 become publicly available, the SEC-related methods may be very useful in providing higher quality models in terms of both main chain and side-chain atom position when compared to other individual servers.

Acknowledgment 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 and Notes

1) Ginalski K., Elofsson A., Fischer D., Rychlewski L., *Bioinformatics*, **19**, 1015—1018 (2003).
2) Wallner B., Elofsson A., *Proteins*, **69** (Suppl. 8), 184—193 (2007).
3) Wallner B., Elofsson A., *Protein Sci.*, **15**, 900—913 (2006).
4) Jang H., Choi S. Y., Cho E. J., Youn H. D., *Nat. Struct. Mol. Biol.*, **16**, 910—915 (2009).
5) Tolstonog G. V., Deppert W., *Nat. Struct. Mol. Biol.*, **16**, 900—901 (2009).
6) Vousden K. H., *Cell*, **103**, 691—694 (2000).
7) Moult J., Hubbard T., Bryant S. H., Fidelis K., Pedersen J. T., *Proteins*, **29**, (Suppl. 1), 2—6 (1997).
8) Moult J., Hubbard T., Fidelis K., Pedersen J. T., *Proteins*, **37**, (Suppl. 3), 2—6 (1999).
9) Moult J., Fidelis K., Zemla A., Hubbard T., *Proteins*, **45**, (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 homepage 〈>http://www.predictioncenter.org/casp8/index.cgi〉,

2008.
14) Zemla A., *Nucleic Acids Res.*, **31**, 3370—3374 (2003).
15) Terashi G., Takeda-Shitaka M., Kanou K., Iwadate M., Takaya D., Hosoi A., Ohta K., Umeyama H., *Proteins*, **69** (Suppl. 8), 98—107 (2007).
16) Jones D. T., *J. Mol. Biol.*, **292**, 195—202 (1999).
17) Vapnik V., "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995.
18) 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).
19) Zhou H., Zhou Y., *Proteins*, **55**, 1005—1013 (2004).
20) Cozzetto D., Kryshtafovych A., Ceriani M., Tramontano A., *Proteins*, **69** (Suppl 8), 175—183 (2007).
21) CASP8 abstracts ⟨http://www.predictioncenter.org/casp8/doc/CASP8_book.pdf⟩, 2008.
22) Ogata K., Umeyama H., *J. Mol. Graph. Model.*, **18**, 258—272, 305—306 (2000).
23) Fischer D., Elofsson A., Rychlewski L., Pazos F., Valencia A., Rost B., Ortiz A. R., Dunbrack R. L. Jr., *Proteins*, **45** (Suppl. 5), 171—183 (2001).
24) Fischer D., Rychlewski L., Dunbrack R. L. Jr., Ortiz A. R., Elofsson A., *Proteins*, **53** (Suppl. 6), 503—516 (2003).
25) Siew N., Elofsson A., Rychlewski L., Fischer D., *Bioinformatics*, **16**, 776—785 (2000).
26) Kamiya K., Sugawara Y., Umeyama H., *J. Comput. Chem.*, **24**, 826—841 (2003).
27) Nojima H., Takeda-Shitaka M., Kanou K., Kamiya K., Umeyama H., *Chem. Pharm. Bull.*, **56**, 635—641(2008).
28) Nojima H., Takeda-Shitaka M., Kurihara Y., Kamiya K., Umeyama H., *Chem. Pharm. Bull.*, **51**, 923—928 (2003).
29) Kurihara Y., Watanabe T., Nojima H., Takeda-Shitaka M., Sumikawa H., Kamiya K., Umeyama H., *Chem. Pharm. Bull.*, **51**, 754—758 (2003).
30) Nojima H., Takeda-Shitaka M., Kurihara Y., Adachi M., Yoneda S., Kamiya K., Umeyama H., *Chem. Pharm. Bull.*, **50**, 1209—1214 (2002).
31) Okimoto N., Futatsugi N., Fuji H., Suenaga A., Morimoto G., Yanai R., Ohno Y., Narumi T., Taiji M., *PLoS Comput Biol.*, **5**, Epub (2009).
32) Kanou K., Iwadate M., Hirata T., Terashi G., Umeyama H., Takeda-Shitaka M., *Chem. Pharm. Bull.*, **57**, 1335—1342 (2009).
33) http://zhang.bioinformatics.ku.edu/I-TASSER/.
34) http://robetta.bakerlab.org/.
35) Umeyama H, Hirono S, Nakagawa S., *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 6266—6270 (1984).
36) http://pymol.sourceforge.net/.