Method for Predicting Homology Modeling Accuracy from Amino Acid Sequence Alignment: the Power Function

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

[a] *Department of Biological Sciences, Faculty of Science and Engineering, Chuo University; 1–13–27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan: and* [b] *Department of Biomolecular Design, School of Pharmaceutical Sciences, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan.*
Received March 21, 2009; accepted October 29, 2009; published online November 2, 2009

We have devised a power function (PF) that can predict the accuracy of a three-dimensional (3D) structure model of a protein using only amino acid sequence alignments. This Power Function (PF) consists of three parts; (1) the length of a model, (2) a homology identity percent value and (3) the agreement rate between PSI-PRED secondary structure prediction and the secondary structure judgment of a reference protein. The PF value is mathematically computed from the execution process of homology search tools, such as FASTA or various BLAST programs, to obtain the amino acid sequence alignments. There is a high correlation between the global distance test-total score (GDT_TS) value of the protein model based upon the PF score and the GDT_TS_{MAX} value used as an index of protein modeling accuracy in the international contest Critical Assessment of Techniques for Protein Structure Prediction (CASP). Accordingly, the PF method is valuable for constructing a highly accurate model without wasteful calculations of homology modeling that is normally performed by an iterative method to move the main chain and side chains in the modeling process. Moreover, a model with higher accuracy can be obtained by combining the models ordered by the PF score with models sorted by the size of the CIRCLE score. The CIRCLE software is a 3D–1D program, in which energetic stabilization is estimated based upon the experimental environment of each amino acid residue in the protein solution or protein crystals.

Key words power function; comparative modeling; homology modeling; protein modeling accuracy

In various organisms, progress in sequencing technology has enabled the analysis of the amino acid sequences of all proteins encoded by the genome. Three-dimensional (3D) structure databases, such as Protein Data Bank (PDB),[1] have accumulated many structures through the development of international projects in structural genomics. PDB structures have been used as templates in homology or comparative modeling procedures. Subsequently, the number of amino acid sequences which can be used to build the tertiary structure of proteins by the homology modeling method is also increasing.[2] The initial step for building a 3D structural model by the homology modeling is to search for homology in the amino acid sequence on PDB. For the homology search, various alignment programs, including FASTA[3] and BLAST,[4] are used. The candidate for alignment is often chosen by referring to the various numerical values represented by the expected value (E-value) from the search result.[4] The E-value is determined by the number of amino acid residues for the query sequence, the length of sequence alignment obtained between the query protein and the template protein, and the sequence identity and similarity for the alignment. In the homology modeling method, a 3D structural model is computed by the data input as an alignment.[5] A 3D structure built by the homology modeling method is not an experimental structure; therefore, an assessment of the reliability of the model is required. Verification by a quality assessment program, such as a 3D–1D method, should be performed, because the experimental structure cannot be used before the modeling prediction. To obtain a 3D structural model that is physicochemically guaranteed, it is desirable to carry out each model construction for all the alignment candidates from the homology search results, and to apply the thus obtained 3D structure model to 3D–1D programs, such as Verify3D[6] and CIRCLE,[7] which estimate the energetic stability of 3D structures. However, it is not economically realistic to build all the models when there are too many alignment candidates. Thus, the model construction of all 3D structures with all the alignment candidates is impossible where computing resources are restricted. In this paper, a new method is introduced to selectively choose the alignment from which an accurate model can be computed. This method predicts the accuracy of the protein model before the modeling procedure. To construct the database for the 3D structural model represented by FAMSBASE,[5,8] it is necessary to build an accurate model of the amino acid sequence with many queries. Generally, the execution time of the homology modeling method is longer than that of the homology search, because the modeling calculation, including moving the main chain, is normally performed by an iterative method. Therefore, the information on the accuracy of the modeling, which is generally obtained after the homology modeling program execution, would help to obtain a correct structure more promptly, if the information was available prior to the execution of protein modeling. Usually, the degree of match of the character string of an alignment (homology percent value) or E-value is referred to after the execution of the homology search program represented by FASTA[3] or various kinds of BLAST(s).[4,9] Although the homology percent value and E-value are important as references for the accuracy of model construction, other structural parameters such as the agreement of the secondary structure between the target protein and the template protein is further required to restrict the model structure. Until now, there have been no useful parameters of accuracy estimation from sequence alignments except for the E-value or homology percent value. The Power Function (PF) score described in this paper prioritizes the re-

* To whom correspondence should be addressed. e-mail: shitakam@pharm.kitasato-u.ac.jp
# These authors contributed equally to this work.

sults when two or more homology search programs such as FASTA[3] or various kinds of BLAST(s)[4,9] are executed. This method is very useful where both high throughput and accuracy are required for modeling. The PF score is calculated using the model length, the percent of sequence homology, and the secondary structure agreement. Generally, structure modeling takes longer than homology searches or sequence alignments, meaning the PF saves on computer resources.

Methods

PDB Learning Set All 50226 sequences from PDB[1] May 4th 2004 were compared with one another in a pair-wise manner using BLAST. Sequences having percent homology above 95%, and that were at least 80% overlapping with each other, were clustered into 9224 families. The longest chain of amino acid sequences in each family was selected because they would lack fewer atomic coordinates in PDB, and the quality of the alignment from the longest chain is more guaranteed. We refer to the coordinate parts of PDB, because SEQRES does not necessarily include the coordinates. A non-redundant PDB dataset of 9224 chains, termed "template proteins", was then created. Although we used PDB from May 4th, 2004, which is from five years ago, we consider that the evaluation of the content described in this paper has not been affected by additions to the database in the past five years. As mentioned later, we constructed 240279 3D models for the 9224 PDB structures, and such a large number of models should provide a good benchmark test to develop a new method. Thus the number of sequences present on May 4th 2004 should be sufficient to provide a statistically significant number for our Power Function analysis. Nevertheless, we want to guarantee that the results do not change, when we recalculate the parameters for the PF using the latest PDB version. It is assumed that, to perform the benchmark test, we must construct about 1592000 ($=240276\times(23743/9224)^2$) 3D models for the 23743 PDB structures present on the latest PDB version of May 29th 2009. However, since we need too many computer resources even if we model only the Cα backbone, the above recalculation has not been performed until now. Accordingly, it should be noticed that the parameters for the PF is determined in the first approximation and still improved, though the PF is significantly superior to the E-value as the selection method of the sequence alignment giving the accurate 3D model as mentioned in Results and Discussion. The E-value is generally used for the estimation of the sequence alignment.

In this paper, 96 targets from the international contest Critical Assessment of Techniques for Protein Structure Prediction 7 (CASP7)[10] in 2006 were used as a test set. As a result, the above CASP7 test set targets were removed, or were not present, in the learning set of PDB from 2004.

The learning set includes not only X-ray structures, but also NMR structures and other experiments. The number of NMR structures added to the PDB database is not negligible, and we did not eliminate these structures for homology modeling. Although there are perturbations of structures in NMR structure or their complexes, suitable statistical or physicochemical filtering in the experiments should guarantee the quality of the structures. Threshold cut off using X-ray resolution was also not set for the same reason. Although the accuracy in the homology modeling naturally depends upon the height of X-ray resolution for the template protein, we thought much of the number of the amino acid residues or the sequence length of the template protein in this paper.

"Power Function (PF)" Construction from Various Homology Search Results We used six different alignment tools [FASTA,[3] BLAST,[4] PSI-BLAST,[4] HMMER-Pfam,[11] RPS-BLAST and IMPALA[9]] to extract alignment features between all *vs.* all of the non-redundant PDB data set. For PSI-BLAST search, the profiles for queries were generated by searching the National Center for Biotechnology Information (NCBI) non-redundant (nr) database from April 25th, 2006, for amino acid sequences. For the HMMER-Pfam search, HMM sequence alignments from the Pfam database were searched using the hmmpfam tools of HMMER, and the profiles for queries were generated by searching the "Pfam-A.full 22.0" database. In both PSI-BLAST and HMMER-Pfam searches, the profile-sequence alignments were performed using the blastpgp tool of the PSI-BLAST package, which uses the matrices to solve the optimized sequence alignment. For sequence–profile alignments, we also used RPS-BLAST and IMPALA, in the PSI-BLAST package, to align the query sequence with the template profiles.

One difficulty in model construction is that it depends on handling the homology percent, which is also called sequence identity, to construct a model closer to the native structure for a query sequence or target protein. Normally, we can obtain many alignments between the sequence and a set of template proteins using the six alignment tools mentioned above. From each of the obtained alignments, the homology percent was calculated. It is easy to construct a 3D model for alignments having sequence identities of more than 50%, due to the rare insertion and deletion of amino acid residues on the sequence alignment. We thus treated the alignment having sequence identities below 50% as objects of homology modeling. We assumed that alignments with a homology percent value higher than a homology threshold (HomTh) should be abandoned to classify the difficulty of model construction.

Therefore, if we express the form mathematically, surviving alignments satisfy the following inequality,

$$\text{homology percent}(\ell) < \text{HomTh}(\ell) \tag{1}$$

The HomTh(ℓ) values were set to have five values of 50, 40, 30, 20, and 10%, which correspond to the variable number ℓ in the inequality (1). A sequence alignment of proteins normally becomes the base for model construction.

On the other hand, to assess the Cα coordinates of the backbone in the theoretical 3D structure model with the Cα coordinates of backbone in the experimentally obtained structure, Global Distance Test Total Score (GDT_TS) values have been used in the CASP experiments or contests until now.[12] The GDT_TS value is a range from zero to 100, the largest value of which is closest to the native structure.

An alignment having the highest GDT_TS value in comparison with the 3D coordinates obtained from the experimental analysis of a target protein, is called the alignment having $\text{GDT\_TS}_{\text{MAX}}$ in this paper, and is defined as the "best." Normally we cannot use this best alignment when we select some restricted alignments, which might give correct tertiary structures. Furthermore, the alignment, which we last selected as a representative among the restricted alignments, is called the "representative alignment." In relation to the best alignment, we term the representative alignment a "good alignment" if the GDT_TS of the model constructed using the representative alignment is equal to or more than 90% of the value of $\text{GDT\_TS}_{\text{MAX}}$. The value of 90% means that the modeled structure for a target protein is very close to the best modeled 3D structure based upon the GDT_TS evaluation. Under the condition that we must select a restricted number of alignments in the modeling process to save computer resources, we expect to select those belonging to the class of a "good alignment." Thus, it was assumed that "good alignment" satisfies the inequality,

$$\text{GDT\_TS}(\ell) \geq \text{GDT\_TS}_{\text{MAX}}(\ell)\times 0.9 \tag{2}$$

It is difficult to set the threshold of "good alignment" as a constant value in place of $\text{GDT\_TS}_{\text{MAX}}(\ell)$ due to diversity of difficulty in homology modeling in a set of targets. In other words, since we must deal with the alignments having various sequence identities, we need mobile standard values, such as the $\text{GDT\_TS}_{\text{MAX}}(\ell)$. Therefore, the threshold is set by considering the maximum GDT_TS values found in each of the five HomTh classes, after the modeling process according to the detected alignments of a set of template proteins for each queried target protein. The character ℓ of the $\text{GDT\_TS}_{\text{MAX}}(\ell)$ is the same as that in the inequality (1). To obtain good results for the CASP protein modeling contest,[13] it is very important to continuously submit the GDT_TS models satisfying the inequality (2) with whole targets during the contest period. Moreover, such good GDT_TS models, constructed from "good alignments," are also very important for high throughput modeling and in construction of a modeling database for all proteins. Using the method prepared in this paper, the ratio of judging an alignment as a good representative alignment in the sequence search of each target protein is calculated in the right side of Eq. 6 (described later) for each HomTh level of the five ℓ categories, using several parameters. The ratio, which is called "MGR" (Max GDT_TS Ratio), is defined by the following Eq. 3 of the conditional probability as shown in a vertical line ("|").

$$\begin{aligned}\text{MGR}(\ell) = \text{P}(&\text{GDT\_TS}(\ell) \geq \text{GDT\_TS}_{\text{MAX}}(\ell)\\ &\times 0.9 | \text{homology percent}(\ell) < \text{HomTh}(\ell)) \text{ in the } (\ell) \text{ category}\end{aligned} \tag{3}$$

It should be noticed that the inequality (1) is presented in relation to the estimation of the alignment, and that the inequality (2) is related to the 3D model constructed based upon the alignment. We use 3D models consisting of Cα backbone coordinates to develop the benchmark test in this paper. The modeling method of the Cα backbone is described later. It is questionable whether the representative alignment satisfies (1) or (2) or not. The states in which the representative alignments fit inequality (2) are counted as a number in the Eq. 3.

When the GDT_TS of the model for each representative alignment is larger than that of $GDT\_TS_{MAX} \times 0.9$, the alignment is estimated as the "good alignment." Thus, the ratio of "good alignment" number for the representative alignment number belonging to the ℓ category is estimated in the Eq. 3. In this paper, a alignment is significant as a representative when it survived due to having the maximum value from a set of values calculated from the power function (PF), mentioned later, for each query sequence. The protein model for this representative alignment is constructed only for the Cα backbone structure, using the FAMS program[14] to save the computer resources in the homology modeling process. The GDT_TS value for this representative alignment is substituted into the inequality (2). If the representative alignment is found to satisfy inequality (2), the query target is counted in addition to the query target number of the numerator in Eq. 3.

In the benchmark test, the numbers of query sequences or target proteins for the five ℓ categories were 8461, 8315, 7986, 7021, and 2911 in HomTh levels of 50, 40, 30, 20, and 10%, respectively. We considered that the numbers of query sequences are large enough to obtain statistically significant results. The ratio of the queries in a (ℓ) category for the all queries set is shown in Eq. 4

$$\mathrm{P(homology\ percent\ (\ell) < HomTh\ (\ell))} = \frac{\text{number of queries in the } (\ell) \text{ category}}{\text{9224 (all queries)}} \quad (4)$$

For example, the query ratio of HomTh level (50%) was 92% of the total 9224. Only the 8461 alignments having the maximum power function (PF) score for each of the 8461 queries survived in PF score calculations from among the alignments between the template protein sequences and each query sequence. In this paper, the number of the Cα backbone model structures constructed using Full Automatic protein Modeling System (FAMS) program[14] for the alignments of all 9224 queries with the PDB non-redundant template protein set is 240279. This number is huge and, because we wanted to save computer resources, we modeled only the Cα backbone structures when we used Eq. 3. In addition, even if the PDB database used in the benchmark test is the latest one, it is not thought that the results changed largely because of the largeness of the sample number in this benchmark test.

The GDT_TS value for the model evaluation is normally calculated from the comparison between the model and experimental structure. The best 3D model within the presumption, which might passively have the GDT_TS_{MAX}, is built based upon the alignment that has highest value of the following PF score. At this moment, the eight parameters of six k_i values, and m and n must have already been determined for the HomTh levels of 50, 40, 30, 20, and 10% using Eq. 6.

$$\text{power function} = \mathrm{PF} = k_i \times \text{model length} \times (\text{homology percent})^m \times (\text{secondary structure})^n \quad (5)$$

where k_i (i=1 to 6) are coefficients, and m and n are the power numbers for homology percent and the agreement of the predicted secondary structure of the target or subject protein and experimental secondary structure for the reference or experimental protein, respectively. The coefficients k_i (i=1 to 6) are introduced as weight factors for the purpose of grouping according to the alignment ability of various programs. The PF score was designed to have a linear correlation with the GDT_TS value.

To find the optimized six k_i values, and m and n values in Eq. 5, the MGR from Eq. 3 was used. Here, the MGR must be maximized as a function of six k_i values, and m and n values. Thus, following Eq. 6 was supposed to put the Eq. 3 into definite shape:

$$\mathrm{MGR}\ (\ell)\ (k_i, m, n) \text{ for the set of } (k_i\ (i = 1 \text{ to } 6), m, n) = \frac{\text{number of "good alignments" selected by largeness of formula (5)}}{\text{number of representative alignments satisfying the inequality formula (1) in the } (\ell) \text{category.}} \quad (6)$$

The numerator is calculated by transferring into Eq. 3 for all target proteins after selection of the representative alignment by sorting by largest PF score from Eq. 5

Here, three values of model length, homology percent, and secondary structure agreement for the set of (k_i (i=1 to 6), m, n) are substituted for the PF in Eq. 5 in the many alignments for each target protein, and, after sorting by largest PF score, the representative alignment is determined.

By using this representative alignment, the Cα 3D model is constructed with the FAMS homology modeling program,[14] and the GDT_TS value of this model is calculated in fitting to the experimental structure of the template protein. This GDT_TS value is compared with the GDT_TS_{MAX} by transferring it into Eq. 3. This transferring process is repeated for all the target proteins or query sequences. Thus, the MGR is calculated for the set of (k_i (i=1 to 6), m, n). Equation 6 shows that the MGR depends on the parameters of the six k_i values, and m and n used in Eq. 5. The six k_i values, and m and n were changed with an interval of 0.1.

MGR (ℓ) is calculated for the right term having any k_i (i=1 to 6), and m and n in relation to number intervals of k_i (i=1 to 6) (0.0—2.0), m (0.0—1.0) and n (0.0—1.5), respectively. Contour plots for changing the values of m and n of the MGR at homology thresholds of 50, 40, 30, 20, and 10% are shown in Fig. 1. The effect on the MGR value of changing each i of k_i (i=1, 2, 3, 4, 5 and 6) is shown in Fig. 2. Thus $\mathrm{MGR_{MAX}}$ (ℓ) (k_i (i=1 to 6)=a_i (i=1 to 6), m=b, n=c) is determined, and the six a_i (i=1 to 6) set values, b and c are optimized for the five ℓ values of HomTh (ℓ). Once again, it is assumed that the order of predicted GDT_TS value is determined by the size of the scalar value of PF (a_i (i=1 to 6), b, c), considering various alignment methods described in this paper. Accordingly, the propriety for determining the order of the modeling accuracy using the scalar of the right side in the Eq. 5 is investigated in this paper. In Eq. 5, in addition, the PF is a supposed function to calculate a score that corresponds to the MGR, and another revised function, in place of the PF, might be reported by other researchers. Then, using local alignment tools such as FASTA or various kinds of BLAST, generate alignments that usually do not cover the whole query length. The length of the amino acid sequence of the obtained local alignment is approximately proportional to the GDT_TS value, under conditions where the folding between a target protein model and template X-ray structure is the same. In this paper, the power number of the model length in Eq. 5 is set to 1 in the first approximation. The homology percent value generally correlates with the accuracy of homology-based modeling structures. The agreement regarding the secondary structure is determined by the identity percent between the PSI-PRED[15] prediction of a query and the STRIDE program.[16] Since it was reported in the ref. 16 that the improvement of the STRIDE over the DSSP relative to PDB assignment has been objectively demonstrated, we used the STRIDE in place of the DSSP, even if both programs return the similar results.

Results and Discussion

To determine the parameter values of m, n, and k_i (i=1 to 6), MGR values were calculated for all combinations of m and n within the range of 0.0—3.0 (with a step of 0.1), and the ranges for the k_i coefficients that correspond to six kinds of alignments were 0—2.0 (with a step of 0.1). As mentioned in Methods, the value of k_i corresponds to each weight of various alignment methods, which include the sequence-sequence alignment and the sequence-profile alignment. Therefore, the weight value of k_i is required. On the other hand, if we want to determine the m and n power numbers for each of the various sequence alignments, more computer resources are needed. In this paper, we determined the m and n values as a first approximation without reference to the sequence alignment method. Figure 1 shows contour plots of MGR defined in Eq. 6 for m and n with the above ranges. In the optimization process of m and n, the k_i (i=1 to 6) values were fixed to be one. The contour plots of MGR are shown in Figs. 1a to e for the five categories of HomTh. The maximum MGR values for each category were determined using the sets of m and n as shown in Table 1. In Fig. 2 the k_i (i=6) were determined. One value for each of the six k_i was set in the five categories of HomTh in the initial step; these values are shown as the maximum positions in the results of Figs. 2a to e. As a first approximation, the changing of the six parameters was performed in the order of k_1, k_2, k_3, k_4, k_5, and k_6, since the optimization process of eight parameters is very complicated.

As shown in Fig. 1c, for example, the maximum value was found in the HomTh=30% condition, and the optimized m

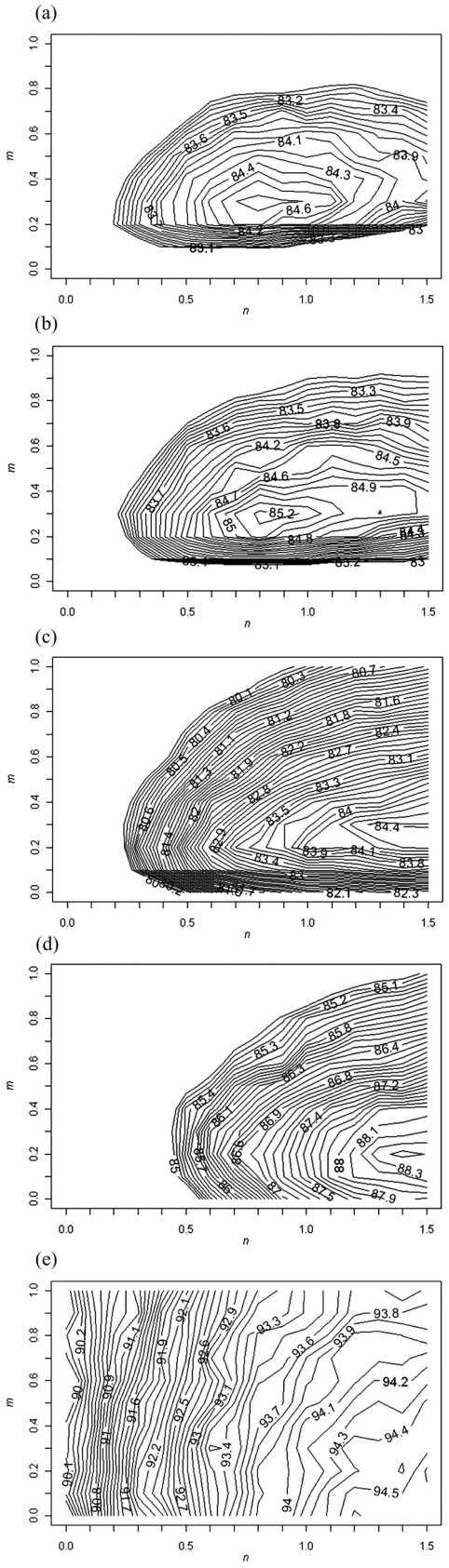

Fig. 1. Contour Plots of MGR from Eq. 3 at a Homology Threshold of HomTh=(a) 50%, (b) 40%, (c) 30%, (d) 20%, and (e) 10% for the Power Numbers of m and n in the Eq. 5

The vertical axis and horizontal axes mean m and n values, respectively. The k_i (i=1 to 6) values were fixed to be one in the contour plots of Fig. 1.

and n values were 0.3 and 1.3, respectively (Table 1). For each HomTh level, optimized m, and n values were similarly determined from the highest MGR values. Using the above set of (m, n)=(0.3, 1.3), as shown in Fig. 2c, the MGR value was plotted against the k_i (i=1 to 6) value. In the six cases, the MGR values were maximum around about k_i (i=1, 3, 4, 5, 6)=1.0 as shown in Table 1. The n value is larger than the m value for all HomTh levels in inequality (1). A higher n value means that consideration of the secondary structure agreement is more important in comparison with the homology percent. Alignment outputs of FASTA[3] and various kinds of BLAST programs[4] already involve the effect of the homology percent ratio, as detected by model length, in their calculations; therefore, its contribution, observed as the m value, is relatively low in comparison with secondary structure agreement. As shown in Eq. 5, the power number for the sequence length for a model is constantly defined to be one in this paper. In other words, the contribution of homology percent might be small, correlating internally with the model length. Remarkably, the n value tends to become higher when the HomTh levels become lower. Therefore, when the modeling of the target is very difficult, consideration of the secondary structure agreement becomes more important under these circumstances. With HomTh=10%, however, the n value of the power number becomes smaller with the decrease of the power number m. The behavior of number n indicates the contribution of the machine learning algorithm that is used for secondary structure prediction of the PSI-PRED program. As shown by the change of the n value, except for the case of HomTh=10%, consideration of the secondary structure might be working stably, even for low homology targets. On the other hand, the values for coefficient k_i (i=1 to 6) were around 1.0 for all HomTh levels. In BLAST,[4] which includes no profile, and FASTA[3] alignments, low k values appear to be shown for HomTh=20—30%, and no alignment results were detected with default threshold of E-value (10) for BLAST or FASTA at HomTh=10% level. If the contribution of some k_i values to the PF score is thought to be low enough, the PF score could work without k values.

Application of the PF Score to CASP7 Targets The correctness of the order sorted by largest PF score in Eq. 5 should be checked for whether the PF is useful to determine the GDT_TS order or not. Thus, the PF score was applied to the all 104 targets in the CASP7 contest[10] in 2006 as a test set. Six kinds of alignment were performed for the amino acid sequences of all the 104 targets against the PDB data of November, 2007. The alignments with experimental structures registered after the CASP7 contest were removed to perform a fair assessment. Moreover, no experimental structures were available for 12 targets, T0284, T0285, T0286, T0287, T0320, T0333, T0334, T0343, T0344, T0352, T0355 and T0386, in this version of PDB database. In addition, three targets, T0336, T0337 and T0377, had no significant alignments in the PDB sequence database. Thus, 89 targets were evaluated. All the alignments in the obtained results were used for model construction, including the main chain and the side-chains, using the homology modeling program FAMS.[14] FAMS, which is the homology or comparative modeling program, constructs the 3D model of the target protein based on the sequence alignment between the query sequence and the amino acid sequence of the template pro-

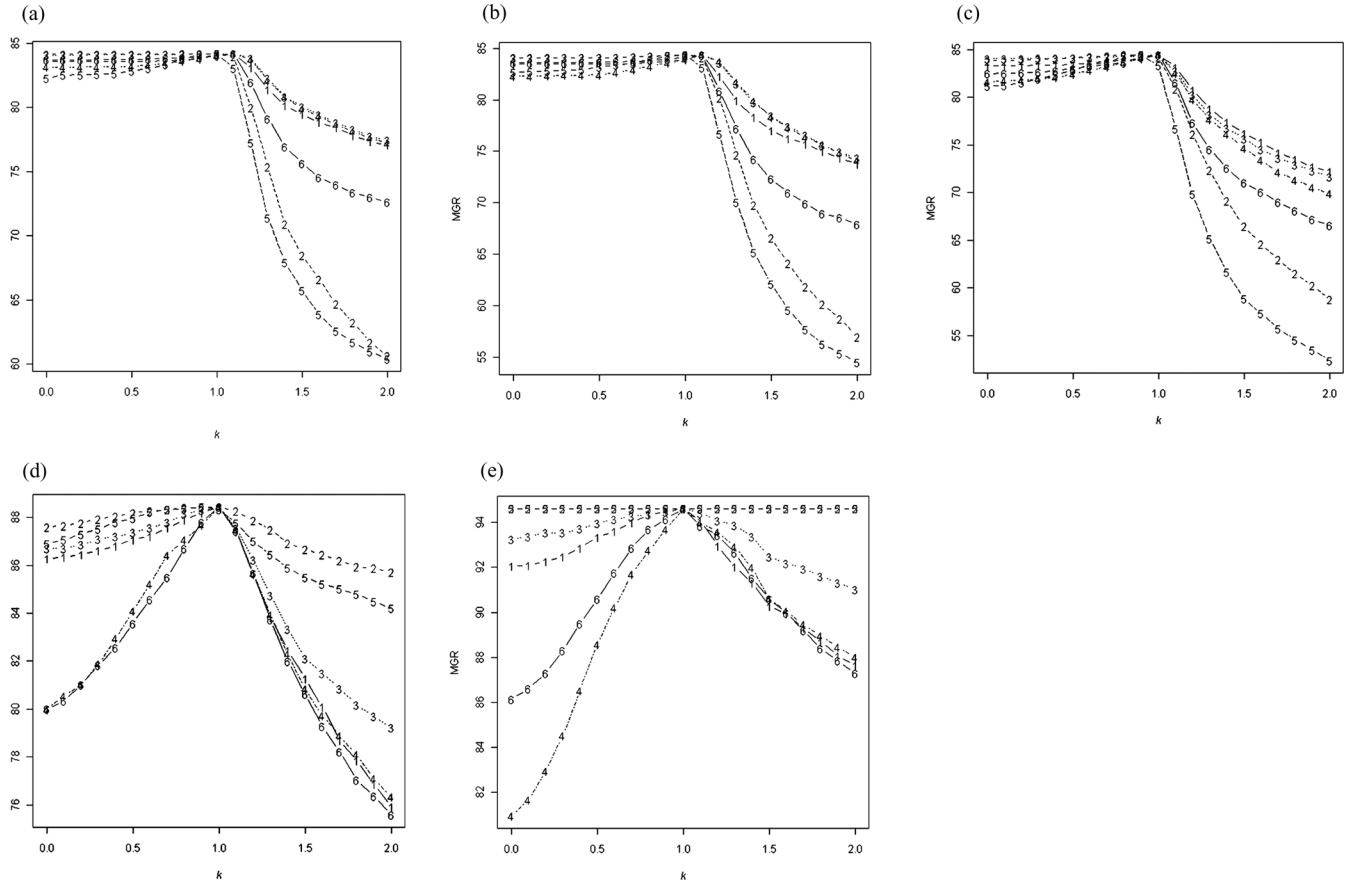

Fig. 2. MGR Value Plots Verifying k_i (i=1 to 6), (i=1: PSI-BLAST, i=2: BLAST, i=3: RPS-BLAST, i=4: IMPALA, i=5: FASTA, i=6: Pfam-BLAST) of HomTh=(a) 50%, (b) 40%, (c) 30%, (d) 20%, and (e) 10% Using the Fixing Condition of Parameters m and n in Table 1

When the k_i (i=p) value is changed from 0.0 to 2.0 in the 0.1 step, other five k_i (i≠p) (i=1 to 6) values were fixed to about one. Arabic numerals of 1 to 6 on the curves in Figs. 2a to e indicate the above alignment methods in the changes of k_i (i=1 to 6) values. As the first approximation, changing the six parameters was carried out in the order of k_1, k_2, k_3, k_4, k_5, and k_6. In the optimization process of k_i (i=1 to 6), the set values of m=b and n=c giving the maximum values in the contour plots of MGR in Figs. 1a to e were used as the two parameters of m and n.

Table 1. Optimized m, n and k_i, (i=1: PSI-BLAST, i=2: BLAST, i=3: RPS-BLAST, i=4: IMPALA, i=5: FASTA, i=6: Pfam-BLAST) Values with Maximizing MGR in Eq. 6

HomTh	m	n	k_1	k_2	k_3	k_4	k_5	k_6
50	0.3	0.8	1.1	1.0	1.1	1.1	1.0	1.1
40	0.3	0.9	1.1	1.0	1.1	1.1	1.0	1.1
30	0.3	1.3	1.0	0.8	1.0	1.0	0.9	1.0
20	0.2	1.4	1.0	1.0	1.0	1.0	0.9	1.0
10	0.0	1.2	1.0	—	1.0	1.0	—	1.0

Using the Eq. 6, MGR values maximized for the set of m and n in the five categories were 84.7, 85.2, 84.4, 88.4 and 94.6 in HomTh levels 50, 40, 30, 20 and 10%, respectively.

tein, or between the former and several template proteins. In the modeling process, FAMS moves the main chain and the side-chain atoms of the target protein alternatively in maintaining the conformational space between the model and the template 3D structure, and performs the conformational search iteratively as close as possible to the native structure in the packing state of the main chain and the side-chains. In addition, FAMS can iteratively construct the 3D models of both the Cα backbone and the main chain without including the side-chains.

After the 3D structures, including the main chain and the side-chains, were constructed by FAMS,[14] two important scores were computed to compare the capability of the PF score. One is the CIRCLE[7] value, which is calculated based upon solvent exposure and polar state of each amino acid residue of the protein. The CIRCLE program is a 3D structure evaluation program[7] that needs all the modeling structures, including the side-chains, to evaluate the quality of various models. Another is the GDT_TS value obtained from the fitting of each model and the experimental or natural structure for the target protein. Thus all the GDT_TS values were calculated for all the alignments through modeling. The correlations between each of the two score sets and the PF score set were analyzed. The PF score correlated well with both the CIRCLE score set and the GDT_TS values set (Table 2). Averages and standard deviations of the correlation coefficients between the PF score, the CIRCLE score, and the GDT_TS values are shown in Table 2. The statistical significance of the differences of correlation coefficients was evaluated using the T-test at the 1% significance level, and higher correlation coefficients of the PF score than those of the CIRCLE score against the GDT_TS values were confirmed for the 59 CM, 30 non-CM, and all 89 targets. In Table 2, the distinction of Comparative Modeling (CM) target and non-CM target was judged from the support vector machine trained with E-value and the homology percent value of PSI-BLAST.[4] The treatment of the SVM with E-value and the homology percent value of PSI-BLAST were exercised using the set of the CASP6 targets. Then, we per-

Table 2. Averages and Standard Deviations of Correlation Coefficients between Power Function (PF Score), CIRCLE Score, and GDT_TS Value

	PF score *vs.* CIRCLE		PF score *vs.* GDT_TS		CIRCLE *vs.* GDT_TS	
	r	ρ	r	ρ	r	ρ
Average of CM targets (59 targets)	0.68±0.25	0.56±0.29	0.95±0.05	0.89±0.06	0.76±0.23	0.64±0.28
Average of non-CM targets (30 targets)	−0.01±0.39	0.02±0.31	0.69±0.13	0.70±0.13	0.01±0.30	−0.05±0.30
Average of all	0.45±0.45	0.38±0.39	0.86±0.15	0.83±0.13	0.51±0.44	0.41±0.44

Three correlation coefficients were calculated for each target protein in the CASP7 contest. r indicates Pearson product-moment correlation coefficient. ρ indicates Spearman's rank correlation coefficient.

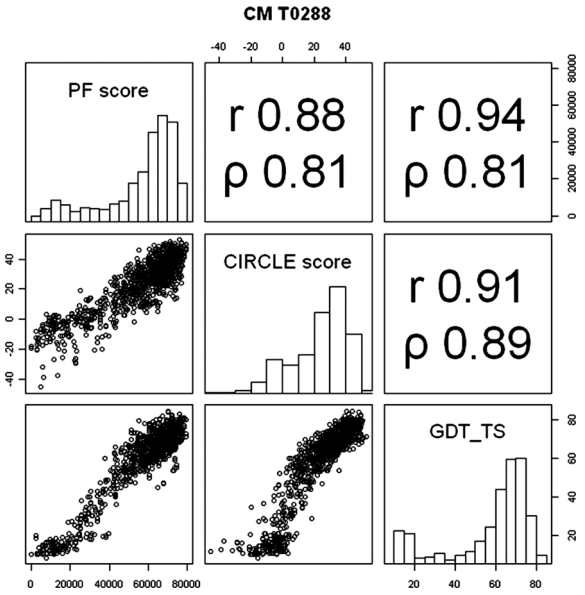

Fig. 3a. Matrix Correlation Map for Three Variables, PF Score, CIRCLE Score, and GDT_TS Value, of the T0288 Target in CASP7, as a Typical Example of the 58 CM Targets

The diagonal three panels show histograms of each variable's distribution. The upper right three panels show two correlation coefficients of Pearson's method (r) and Spearman's method (ρ). PF score, CIRCLE score, and GDT_TS value range from 0 to 80000, from −45 to 60, and from 0 to 100, respectively. For example, two correlation coefficients of the PF score against the GDT_TS value are r=0.94 and ρ=0.81. Those of the PF score against the CIRCLE score are r=0.88 and ρ=0.81.

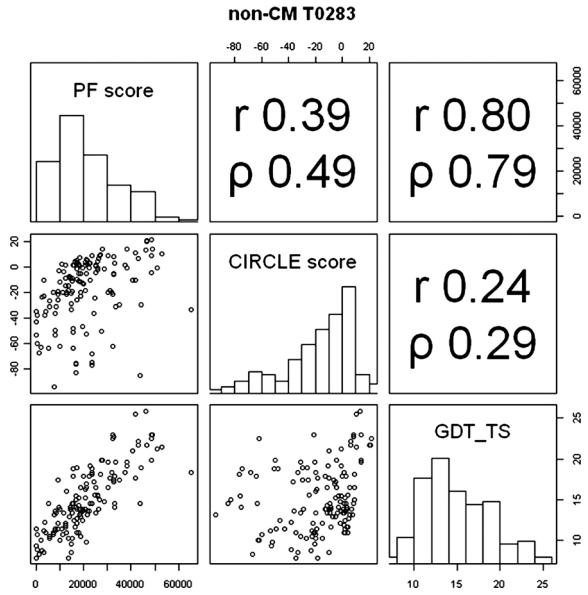

Fig. 3b. Matrix Correlation Map for Three Variables, PF Score, CIRCLE Score, and GDT_TS Value, of the T0283 Target in CASP7, as a Typical Example of the 30 Non-CM Targets

The diagonal three panels show histograms of each variable's distribution. The upper right three panels show two correlation coefficients of the Pearson's method (r) and Spearman's method (ρ). PF score, CIRCLE score and GDT_TS value range from 0 to 70000, from −100 to 20 and from 8 to 26, respectively. For example, two correlation coefficients of the PF score against the GDT_TS value are r=0.80 and ρ=0.79. Those of the CIRCLE score against the GDT_TS value are r=0.24 and ρ=0.79.

formed the distinction of CM category and non-CM category for each of the CASP7 targets. We must distinguish between CM category and non-CM category before the selection process of the best 3D model among many constructed models by the CIRCLE, because the CIRCLE scoring equation of the model selection for the CM target is different from that for the non-CM target. In this paper we use "homology modeling" to mean the same as "comparative modeling." The CM targets and the non-CM targets represent the proper query sequence for comparative modeling and free modeling, respectively. For the CM targets, the PF score using the homology search tools of FASTA[3] and various BLAST[4] programs showed a high correlation against the GDT_TS value. Figure 3 shows the correlation maps for T0288 and T0283 as typical examples of CM and non-CM targets, respectively, in CASP7. The correlation maps for other targets are supplied on the web site (http://www.bio.chuo-u.ac.jp/iwadate/PF score/). Thus, the method allowed us to choose the alignments giving higher GDT_TS values from the homology search results, using the PF score. In other words, we can remove the alignments giving lower GDT_TS values before 3D modeling from among the alignment candidates. In the case of the non-CM targets, the correlation coefficient of the PF score was higher against the GDT_TS, although that of the CIRCLE score was almost zero against the GDT_TS. The PF method showed a correlation value of 0.69, which is a meaningful value for the non-CM targets. Additionally, the results indicated that it was relatively difficult for the non-CM targets to provide meaningful information about GDT_TS values using only the CIRCLE score, which is said to be useful in selecting the model closest to the native structure.[7] This indicated that the supposition of Eq. 6 in the methods was appropriate. Equation 6 dictates that the GDT_TS value will be predicted based upon the model length, the homology percent or the sequence similarity, and the secondary structure agreement.

On the other hand, because the PF score correlated with the CIRCLE score with a value of only 0.68 for the CM targets, the propensity of the model estimation is different between the PF and CIRLCE methods. The PF method does not have a strong internal correlation for selecting models with larger GDT_TS values in comparison with the CIRCLE method. This fact is significant for the development of a selection method for the model closest to the native structure.

Table 3. Averages and Standard Deviations of the Number of Models for Each Target, GDT_TS$_{MAX}$ Value, GDT_TS Value of a Top Ranking Model Selected by CIRCLE Score, and GDT_TS Value of a Top Ranking Model Selected by PF Score

	Numbers of models	GDT_TS$_{MAX}$ [a)]	GDT_TS of top ranking model selected by CIRCLE score	GDT_TS[a)] of top ranking model selected by PF score
Average of CM targets (59 targets)	672.44	68.52±15.79	62.24±20.01 (0.91)	62.73±16.39 (0.92)
Average of non-CM targets (30 targets)	127.80	29.05±14.93	18.04±16.13 (0.62)	23.94±14.34 (0.82)
All	488.85	55.22±24.29	47.34±28.13 (0.86)	49.65±24.19 (0.90)

a) Correlation coefficients between GDT_TS$_{MAX}$ and GDT_TS of models selected by PF score as a top ranking score are 0.92 (for 59 CM targets), 0.93 (for 30 non-CM targets), and 0.97 (for all targets).

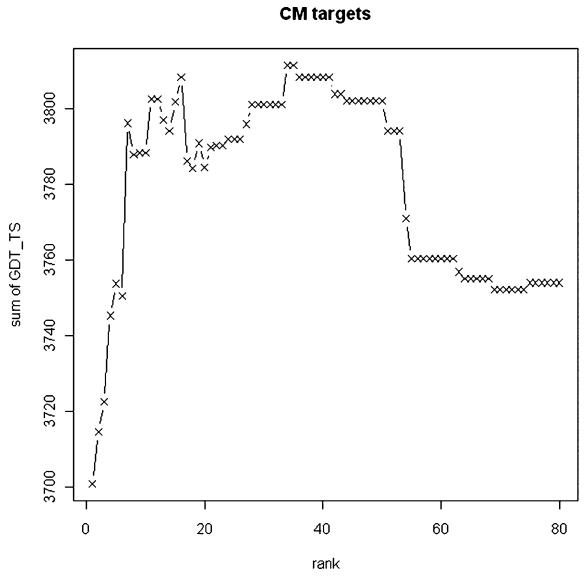

Fig. 4a. Vertical and Horizontal Axes Show the Sum of Each GDT_TS Value of Selected Models, Sorted about the Size by CIRCLE Score from 1st to Xth Ranking and Xth Ranking, Respectively, with the PF Score

The larger values of the sum of the GDT_TS values of 59 CM targets show that the combination method of PF score and CIRCLE score is effective.

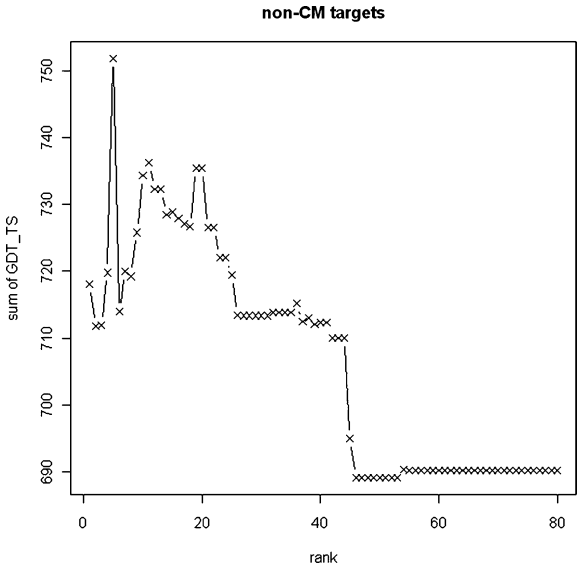

Fig. 4b. Vertical and Horizontal Axes Show the Sum of Each GDT_TS Value of Selected Models, Sorted about the Size by CIRCLE Score from 1st to Xth Ranking and Xth Ranking, Respectively, with the PF Score

The larger values of the sum of the GDT_TS values of 30 non-CM targets show that the combination method of PF score and CIRCLE score is effective.

In Table 3, the maximum GDT_TS values after modeling for all the alignments detected in six homology search tools are called the GDT_TS$_{MAX}$ values. Moreover, the averages of GDT_TS$_{MAX}$ for CM and non-CM targets show the averages for the modeling of CM and non-CM targets of all the CASP7 targets. The average GDT_TS$_{MAX}$ values differ greatly by 39.5 between the CM and the non-CM targets. The difference of the technical difficulty of the modeling is clear. The average GDT_TS value, 62.2, in the models chosen with the CIRCLE score is almost same as that, 62.7, obtained from the PF score for the CM targets, though it is lower by 6.3 than the average GDT_TS$_{MAX}$. For the non-CM targets, the PF score had a little more selection capability, because the average GDT_TS values were 18.0 and 23.9 for the CIRCLE and PF scores, respectively. Thus, we can select the models having the ratios of 92 and 82% of GDT_TS$_{MAX}$ for CM and non-CM targets, respectively, using the PF score. The circle score selected models having ratios of 91 and 62 % of GDT_TS$_{MAX}$ for CM and non-CM targets, respectively. Moreover, the GDT_TS of models selected as the top ranking model by the PF score had high correlation coefficients of 0.92, 0.93 and 0.97, against the GDT_TS$_{MAX}$ for CM, non-CM and all targets, respectively. This indicated that the PF score proposed in Eq. 5 was very useful in selecting the sequence alignments having higher GDT_TS values. Naturally, we cannot obtain the actual GDT_TS value in the process of comparative or homology modeling, since the 3D structure of the query sequence has not been experimentally determined yet.

Combined Method of the PF Score and the CIRCLE Value Although the PF method showed a good correlation with the CIRCLE method, as shown by the 0.68 score for the CM targets in Table 2, the differences indicated that a combined method might be effective.

We made the combined method between the PF score and the CIRCLE score for the model constructed. First, for a query sequence or a target protein, the PF scores of all alignments detected by FASTA and various kinds of BLAST were calculated and ranked by size. Second, the model construction by homology modeling using FAMS program[14] was performed for the alignments until a certain (Xth) number of the PF score. Finally, the CIRCLE scores of the constructed models were calculated, and we selected the model that had the highest CIRLCE score until the (Xth) number of the PF score. Concretely, for target proteins in the class of CM or non-CM in the contest, the calculations were executed to find the models having the highest CIRCLE score in the limitation (Xth) number of the PF score for each query sequence. As shown in Fig. 4a, the combination of the CIRCLE score and the PF score was superior to the CIRCLE score alone for

Table 4. Comparison of GDT_TS Values between Two Models Selected Using E-Value and PF Score for (A) CM Targets and (B) Non-CM Targets in the CASP7 Contest

(A) CM targets

Target	Min E-value	GDT_TS selected by		
		E-value		PF score
T0288	1.00E-29	72.8	<	80.22
T0289	1.00E-81	40.15	<	43.89
T0297	1.00E-61	61.02	<	64.22
T0302	5.00E-50	74.81	>	53.98
T0303	7.00E-62	68.3	>	59.26
T0311	2.00E-15	51.86	<	55.05
T0313	1.00E-140	67.48	<	70.17
T0316	5.00E-72	20.74	<	28.09
T0318	1.00E-135	47.39	<	56.85
T0322	2.00E-25	53.7	<	57.75
T0329	1.00E-45	65.48	>	48.74
T0330	5.00E-47	60.22	>	43.67
T0331	9.00E-38	49.46	<	64.93
T0338	2.00E-62	50.29	<	57.91
T0339	1.00E-125	55.02	<	67.49
T0357	1.00E-15	23.67	<	39.2
T0364	3.00E-31	67.97	>	62.4
T0366	5.00E-33	82.61	>	75
T0367	3.00E-42	69.4	<	77.2
T0373	5.00E-31	64.93	>	60.61
T0374	1.00E-32	62.5	>	55.47
T0376	2.00E-90	62.9	>	58.15
T0378	5.00E-66	58.5	<	69.5
T0379	4.00E-44	50.49	>	46.81
T0380	2.00E-32	65.96	>	57.45
T0384	1.00E-106	57.79	<	64.67
Total of all 59 targets		3681.31		3701.01

(B) Non-CM targets

Target	Min E-value	GDT_TS selected by		
		E-value		PF score
T0299	1.20E-02	10.2	<	13.83
T0300	5.50E-02	22.19	>	19.1
T0304	3.00E-02	12.38	<	23.6
T0306	4.80E-02	11.58	<	20.26
T0314	1.60E-01	13.92	<	17.22
T0319	4.00E-14	14.44	<	17.22
T0325	8.30E-04	36.69	>	10.82
T0327	5.00E-03	33.65	<	50.96
T0335	7.40E-02	17.86	<	51.19
T0347	2.50E-02	15.56	>	8.67
T0348	3.00E-04	10.66	<	45.9
T0350	8.00E-04	10.68	<	26.71
T0351	5.40E-02	20	<	28.75
T0353	2.60E-01	32.06	>	25.59
T0354	1.00E-03	11.04	<	13.12
T0358	8.00E-03	14.77	<	37.88
T0361	7.00E-04	14.48	>	10.67
T0369	4.00E-13	60.37	>	56.12
T0372	1.00E-02	5.45	<	7.8
T0382	1.60E-02	18.6	<	21.49
T0383	4.20E-01	8.6	<	20.2
Total of all 30 targets			584.96	718.07

The targets having differences of less than 2.0 between GDT_TS values of two models selected by E-value and PF score are not listed. In (A), the numbers of '<' and '>' are 15 and 11, respectively, and, in the (B), they are 15 and 6, respectively. In this table, the superior ratios of the PF method to the E-value method are 1.4 and 2.5 times for CM and non-CM targets, respectively.

CM targets. The total GDT_TS value for the CM targets gave lower values over the 55th rank of the PF score. If the CIRCLE method were superior to the combined PF and CIRCLE method, the sum of GDT_TS values would increase against the rank of the PF score in Fig. 4a. The 3D structures from homology modeling for the 1st to Xth PF scores need to be the target protein model set having a very high total GST_TS value. For each query sequence, the 3D models obtained from the selection of the PF score were reassessed with the CIRCLE score, and, again, the results were sorted by the size. Thus, the protein model for each target protein in the class of CM or non-CM was selected using the CIRCLE score among the 3D structures constructed after being selected by the size of PF score. As already noted, the CIRCLE score can select a model near to the native or experimental structure from the models constructed for the sequence alignments of the 1st to Xth PF score. The CIRCLE score helps the PF score in selecting the model near to the native structure from a free energy point of view, using the 3D–1D score. It should be noted that the combined PF and CIRCLE method is available within the limited ranking of the PF score.

In Fig. 4a, the CIRCLE score functioned until the ranking of the PF score reached about 50. The high estimation of the selection of the 3D model by the CIRCLE program is shown against the ranking order of PF score, X, of 10 to 50, as shown in the higher total GDT_TS values of the vertical axis for the CM targets. For the non-CM targets of Fig. 4b, the sum of the GDT_TS or the total GDT_TS value showed the highest value at PF rank of X=5, and the value decreased rapidly at around X=10. In the ranking value of from 10 to 40, the level of the total GDT_TS value was relatively higher. Thus, for using the combination of the CIRCLE score for the PF score, the rank of the PF score at which models should be included are 1st to 40th and 1st to 10th for CM and non-CM targets, respectively. These results indicated that, for the CM targets and the non-CM targets, the combined method of the PF and CIRCLE methods was useful for obtaining a model having a high GDT_TS value.

Comparison with E-Value The E-value of various homology search programs has been used as a statistical index for considering the degree of agreement of the character string of an alignment. Comparison between the alignment chosen by E-value and that chosen by the PF score is shown in Table 4. The results of the comparison for all the 89 targets are available on the web site shown at the end of this section. In Table 4, the targets having a difference of less than 2.0 between the GDT_TS values of two models selected by the E-value and the PF score are not listed. Using the selections performed by the E-value and by the PF score, we chose the same alignment in 12 targets among the 59 CM targets. Selection of the same alignment by the two methods indicates that, because the same experimental structure was used in a limited fashion in the 20% ratio of the CM targets,

the two methods select the template proteins independently. The PF score selected alignments that showed higher GDT_TS values for 28 targets, and the E-value selected alignments that showed higher GDT_TS values for 19 targets. In non-CM targets, we chose the same alignment in 23% (7 targets) of the 30 targets. The two methods also select the template proteins independently for the non-CM targets. The PF score selected alignments that showed a higher GDT_TS values in 17 targets, and the E-value selected six targets with higher GDT_TS values. Thus the superiority of the PF method in comparison with the E-value method was demonstrated. The T-test confirmed that this alignment selection superiority of the PF method over the E-value method was confirmed at the 5% (p-value=0.03624) significance level.

Our PF method consists of three elements; (1) the length of a model; (2) the degree of homology percent or sequence identity of characters composed of amino acid sequence; and (3) agreement ratio of the secondary structure. Terms of (1) and (2) are also contained in the E-value method. We showed that the agreement of the secondary structure might contribute a little to CM targets and lot to non-CM targets. The superiority ratios of the PF method against the E-value method are 1.5 (=28/19) and 2.8 (=17/6) for CM and non-CM targets, respectively. For CM targets, the total GDT_TS value, 3701, of the models selected using the PF method was larger by 37 than the 3681 selected using the E-value method. For non-CM targets, the total GDT_TS value, 718, of the models selected using the PF method was larger by 133 than the 585 selected using the E-value method.

Superimposed views of T0331 and T0327 are shown in Fig. 5 as modeling examples of CM and non-CM targets, respectively. We have shown that the PF score is more effective than the E-value score in the comparison between the GDT_TS values for both scores. Superimposed views of another 87 targets are supplied on the web site (http://www.bio.chuo-u.ac.jp/iwadate/PFscore/).

Nevertheless, as the E-value method shows superiority to the PF method by 42% and 29% for the CM and non-CM target, respectively, in Table 4, a good researcher should properly use both methods as the case requires.

We have discussed the results shown in the Tables and Figures of this paper, and deduce that the PF score is very useful. However, we did not analyze in detail the reason why this score is powerful. To clarify this point, investigation of the results of several more benchmark tests is required. We used six different alignment tools [FASTA,[3] BLAST,[4] PSI-BLAST, HMMER-Pfam,[10] RPS-BLAST and IMPALA[9]] to extract alignment features between all *vs.* all of the non-redundant PDB data set. However, these local alignment tools occasionally provide only a short alignment, especially for the non-CM targets. According to these local alignment programs, only a protein segment would be modeled instead of whole protein domain, which is the state of the native or experimental structure. In this case, core residues might be exposed to water, which would probably lead to a bad CIRCLE score based upon 3D–1D score, even if the protein segment has a good GDT_TS value. When we obtain such a protein segment from a difficult alignment between the target protein and the template protein, how should a benchmark test be performed that includes the core residues exposed to water in relation to the Power Function? We could consider the PF as including the term of the total areas of the core residues exposed to water when we refer to the experimental structure becoming the template protein. Next, we mixed NMR structures, and high-resolution and low-resolution X-ray structures in the dataset. Naively speaking, NMR structures and low-resolution X-ray structures would be bad templates. We incorporated high-resolution X-ray structures composed of long sequences into this paper in a first approximation, but we have not performed the training about the appropriateness of this approximation. In future, we should check which templates of NMR structures, low-resolution X-ray structures and high-resolution X-ray structures become good templates, as judged by their GDT_TS score, which shows the accuracy of the Cα backbone of the main chain with no consideration of the accuracy of the side-chains. Moreover, if we use the PDB learning set combining the templates of NMR structures, low-resolution X-ray structures and high-resolution X-ray structures, in addition to the PDB sequences set having homology percent values above 95%, and overlapping at least 80% with each other, how the selection using the PF score changes becomes very interesting. Next, we employed

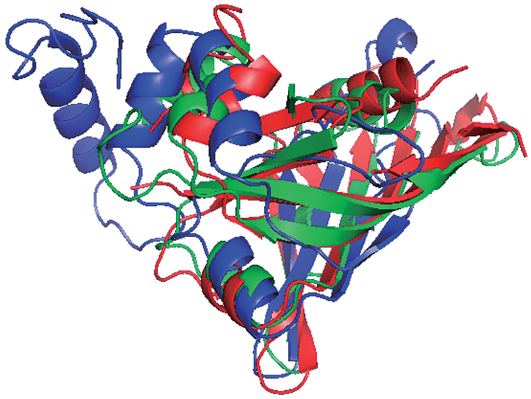

Fig. 5a. Superimposed View of Model Structures Selected by PF Score (Green) and E-Value (Blue), against Native Structure (Red) in the T0331 Target as an Example of the 59 CM Targets in CASP7

The green structure is closer than the blue one against the red experimental structure. As the GDT_TS value, 64.93, of the model selected by PF score is larger, by 15.47, than that of the model selected by the E-value (49.46), the green structure from the PF score is closer to the red native structure than the blue structure from the E-value.

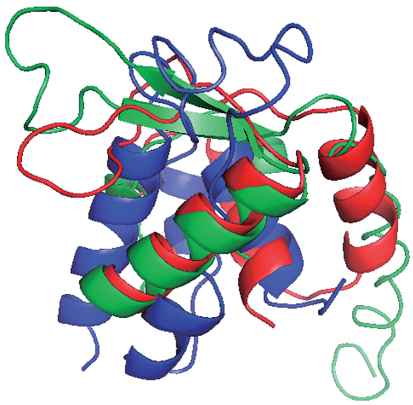

Fig. 5b. Superimposed View of Model Structures Selected by PF Score (Green) and E-Value (Blue) against Native Structure (Red) in the T0327 Target as an Example of the 30 Non-CM Targets in CASP7

As the GDT_TS value, 50.96, of the model selected by PF score is larger, by 17.31, than that of the model selected by the E-value (10.68), the green structure from the PF score is closer to the red native structure than the blue structure from the E-value.

PSI-PRED prediction for the target protein in the agreement of the secondary structure between the target protein and the template protein, but we have not checked the relationship between the prediction accuracy and the PF score performance. Because we have thought that the agreement ratio between the prediction accuracy of PSI-PRED and the secondary structure of the template protein almost depends upon the prediction accuracy of PSI-PRED from the results shown in the ref. 15. The prediction accuracy described in the paper of the PSI-PRED method was about 76% in the average, and the prediction accuracy of PSI-PRED varies depending on the target protein. Thus, it was assumed that the relationship between the prediction accuracy and the PF score performance is naturally positive correlation. Lastly, the correlations of the PF score with the alignment coverage of the domain, the X-ray resolution of templates and the prediction accuracy of PSI-PRED should be evaluated in a future paper, because those positive or negative correlations might affect the performance of the PF score.

Acknowledgments The authors wish to thank Prof. Masayuki Ohmori of Chuo University for assistance in preparing this manuscript. 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., Henrick K., Nakamura H., Markley J. L., *Nucleic Acids Res.*, **35**, 301—303 (2007).
2) Fukuchi S., Homma K., Sakamoto S., Sugawara H., Tateno Y., Gojobori T., Nishikawa K., *Nucleic Acids Res.*, **37**, 333—337 (2009).
3) Pearson W., "Current Protocols in Bioinformatics," Chapter 3, Unit 3.9, John Wiley & Sons, Inc., 2004.
4) Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J., *Nucleic Acids Res.*, **25**, 3389—3402 (1997).
5) Umeyama H., Iwadate M., "Current Protocols in Bioinformatics," Chapter 5, Unit 5.2, John Wiley & Sons, Inc., 2004.
6) Bowie J. U., Lüthy R., Eisenberg D., *Science*, **253**, 164—170 (1991).
7) Terashi G., Takeda-Shitaka M., Kanou K., Iwadate M., Takaya D., Hosoi A., Ohta K., Umeyama H., *Proteins*, **69** (Suppl. 8), 98—107 (2007).
8) Yamaguchi A., Iwadate M., Suzuki E., Yura K., Kawakita S., Umeyama H., Go M., *Nucleic Acids Res.*, **31**, 463—468 (2003).
9) Schaffer A. A., Wolf Y. I., Ponting C. P., Koonin E. V., Aravind L., Altschul S. F., *Bioinformatics*, **15**, 1000—1011 (1999).
10) Kryshtafovych A., Fidelis K., Moult J., *Proteins*, **69** (Suppl. 8), 194—207 (2007).
11) Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R., Griffiths-Jones S., Howe K. L., Marshall M., Sonnhammer E. L., *Nucleic Acids Res.*, **30**, 276—280 (2002).
12) Venclovas C., Zemla A., Fidelis K., Moult J., *Proteins*, **53** (Suppl. 6), 585—595 (2003).
13) Moult J., *Curr. Opin. Struct. Biol.*, **15**, 285—289 (2005).
14) Ogata K., Umeyama H., *J. Mol. Graph. Model.*, **18**, 258—272, 305—256 (2000).
15) Jones D. T., *J. Mol. Biol.*, **292**, 195—202 (1999).
16) Frishman D., Argos P., *Proteins*, **23**, 566—579 (1995).