HUMAN FAMSD-BASE: High Quality Protein Structure Model Database for the Human Genome Using the FAMSD Homology Modeling Method

Kazuhiko Kanou,[a] Tomoko Hirata,[a] Mitsuo Iwadate,[b,c] 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:* [b] *RIKEN Systems and Structural Biology Center; 1–7–22 Suehiro-cho, Tsurumi-ku, Yokohama 230–0045, Japan: and* [c] *Department of Biological Sciences, Faculty of Science and Engineering, Chuo University; 1–13–27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan.*

Received September 3, 2009; accepted October 29, 2009; published online November 2, 2009

Almost all proteins express their biological functions through the structural conformation of their specific amino acid sequences. Therefore, acquiring the three-dimensional structures of proteins is very important to elucidate the role of a particular protein. We had built protein structure model databases, which is called RIKEN FAMSBASE (http://famshelp.gsc.riken.jp/famsbase/). The RIKEN FAMSBASE is a genome-wide protein structure model database that contains a large number of protein models from many organisms. The HUMAN FAMSBASE that is one part of the RIKEN FAMSBASE contains many protein models for human genes, which are significant in the pharmaceutical and medicinal fields. We have now implemented an update of the human protein modeling database consisting of 242918 constructed models against the number of 20743 human protein sequences with an improved modeling method called Full Automatic protein Modeling System Developed (FAMSD). The results of our benchmark test of the FAMSD method indicated that it has an excellent capability to pack amino acid side-chains with correct torsion angles in addition to the main-chain, while avoiding the formation of atom-atom collisions that are not found in experimental structures. This new protein structure model database for human genes, which is named HUMAN FAMSD-BASE, is open to the public as a component part of the RIKEN FAMSBASE at http://mammalia.gsc.riken.jp/human_famsd/. A significant improvement of the HUMAN FAMSD-BASE in comparison with the preceding HUMAN FAMSBASE was verified in the benchmark test of this paper. The HUMAN FAMSD-BASE will have an important impact on the progress of biological science.

Key words protein structure prediction; protein model database; homology modeling; comparative modeling; sequence alignment; 3D–1D score

Genome sequencing projects have generated an enormous amount of deduced amino acid sequence information.[1] Almost all proteins express their biological functions through the structural conformation of their specific amino acid sequences. Thus, acquiring the three-dimensional (3D) structure of a protein is very important to elucidate its role. Recently, structural genomics projects,[2,3] *i.e.* post-genome sequencing projects, have increased the number of protein structures in the Protein Data Bank (PDB)[4,5] using experimental determination methods. Nevertheless, the number of structures lags behind the number of protein sequences in NCBI NR (/blast/db/FASTA directory on NCBI FTP site). Therefore, methods of accurate protein structure prediction are urgently required. One of the most effective approaches for protein structure prediction is the homology or comparative modeling method. Computational protein structure prediction methods, such as the homology modeling method, can provide valuable information for sequences whose structures have not been determined experimentally.

From a computer aided protein modeling point of view, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment has been performed every two years since 1994.[6—12] As a result, there has been good progress in structure prediction techniques. In the latest CASP experiment, CASP8 (2008),[12] we participated as an automatic server predictor using our Full Automatic protein Modeling System Developed (FAMSD) protein modeling method.[13] In the FAMSD method (see Methods), alignment programs such as a series of BLAST[14] programs, SP3,[15] and SPARKS2[16] programs, homology modeling program FAMS,[17] model quality estimation program CIRCLE,[18] and the molecular dynamics program APRICOT[19] were combined and used to construct reliable 3D protein models. In CASP8, the FAMSD team predicted 3D models for all 128 target proteins that were released by the CASP8 organizers. After the prediction period expired, we performed an original assessment of the FAMSD method in comparison with other server teams that participated in CASP8. The results of our assessment indicated that the FAMSD method has an excellent capability to pack amino acid side-chains with correct torsion angles, in addition to the correct Cα backbone, while avoiding the formation of atom-atom collisions that are not observed in native structures. Although the experimental structure is always not a native structure, we use the term of the native structure in place of the experimental structure in this paper.

On the other hand, Structure Based Drug Design (SBDD) has been developed to find bioactive compounds from a medicinal point of view. In order to perform the SBDD research, 3D structure of a target protein which is bound or docked by the bioactive compound is required. In the pharmaceutical and medical fields, again, it is important for us to obtain the 3D structure of human target protein. From the experiments such as the X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy, the 3D structure is obtained, and, however, there are many proteins, for which we are unable to analyze the 3D structure. In that case, the 3D model constructed with the homology modeling method is useful.

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

Then, if we have the modeling database of 3D structures for human genes, we can always use the model without consuming our time and computer resources. We had built a set of protein structure model databases,[20—23] which is called RIKEN FAMSBASE. It is open to the public at http://famshelp.gsc.riken.jp/famsbase/. The RIKEN FAMSBASE is a genome-wide protein structure model database that contains a large number of protein models from many organisms. The database is useful and valuable because structural models are easily obtained through the website. As one part of the RIKEN FAMSBASE, we had built HUMAN FAMSBASE, which contains many protein models for human genes that are significant in the pharmaceutical and medicinal fields. The models in the HUMAN FAMSBASE were constructed with POWER FUNCTION (PF) method mentioned later using the PDB version of September 2005, which was open to the public four years ago. It is natural that to update the PDB version in the homology modeling is very important as shown in 'Results and Discussion.' Moreover, the high accuracy of the model in the database is required, because the docking simulation of the bioactive compound with low molecular weight does not proceed properly without the side-chain geometry near to the native structure. Thus, we applied an improved modeling method, the FAMSD method, to human protein sequences using the PDB version of August 2008. The protein models of 20743 human genes were constructed with the FAMSD method and have been incorporated into a new relational database named HUMAN FAMSD-BASE. This database is open to the public as a component part of the RIKEN FAMSBASE at http://mammalia.gsc.riken.jp/human_famsd/, coincident with the publication of this paper.

The protein sequences of human genome in the HUMAN FAMSD-BASE were obtained from the sequence collection 'hsap2,' which contains 28954 protein sequences, in the Genomes TO Protein structures and functions (GTOP) database.[24,25] The sequences in 'hsap2' are collected from the National Center for Biotechnology Information (NCBI) database. The NCBI website provides genomic information infrastructure for medical researchers from around the world. The FAMSD method was applied to each sequence in 'hsap2' to construct reliable 3D structure models. However, when the FAMSD method assigned the target protein as 'very difficult' or impossible for protein modeling, 3D models of the targets were not included in HUMAN FAMSD-BASE, because these models are not sufficiently accurate or reliable. Consequently, for 20743 out of 28954 target proteins, one or more relatively reliable 3D models were constructed based on the 22297 non-redundant PDB chain set of August 2008. Thus, in this paper, 72% of proteins from the human genome were modeled with the homology method.

Furthermore, we describe the results of two types of benchmark tests ((1) and (2)) in relation to the HUMAN FAMSD-BASE. In benchmark test (1), using the query sequences in the CASP8, the FAMSD method was compared with the POWER FUNCTION (PF) method which was used to construct 3D models in the preceding HUMAN FAMSBASE of the RIKEN FAMSBASE. The HUMAN FAMSBASE is based on the PDB database (September 2005) in relation to the PDB version, which changes once in seven days. It is natural that the newer PDB version with the increased amount of the PDB data increases the number of the models constructed by FAMS because of the increased number of the templates available for sequence alignment with the targets. First, therefore, it is reasonable from a PDB version point of view that we insist on the superiority of the HUMAN FAMSD-BASE in comparison with the HUMAN FAMSBASE. As shown in 'Results and Discussion,' second, it was found that, from a modeling point of view, the FAMSD method is significantly superior to the PF method in benchmark test (1). We then implemented benchmark test (2) using target proteins whose structures have been determined by experimental methods in the human protein sequences of 'hsap2.' The experimental structures of those target proteins were hidden, and the constructed models were compared with solutions in the assessment of the modeling accuracy. Generally, biochemists and biologists can easily obtain the sequence identity % from the sequence alignment between target and template proteins used in the website of the HUMAN FAMSD-BASE. For homology based models, then, it is natural that they expect the assessment of the quality of a constructed model on the basis of the sequence identity between the target protein and the homologous template protein. In this paper, we show the relationship between the sequence identity and the quality of the model in the HUMAN FAMSD-BASE, based upon 'benchmark test (2),' to approximately assess the accuracy of the constructed model.

Methods

The FAMSD Modeling Method The FAMSD modeling method was applied to an amino acid sequence obtained from genes in 'hsap2' as follows: (1) many sequence alignments between a target protein and template proteins were generated; (2) the number of the alignment candidates was reduced based on our scoring function, which estimates the reliability of alignments and is called POWER FUCNTION (PF, see below), and based on the statistical scores about reliance for the sequence alignments obtained from two types of profile-profile alignment programs; (3) 3D models were constructed based on above selected alignments using the FAMS program[17]; and (4) constructed 3D models were evaluated based on our 3D–1D scoring function, which estimates the stability of a protein structure and is calculated in the CIRCLE program.[18] Thus, the FAMSD method was used for the 28954 query sequences in the 'hsap2' genome.

(1) Generating Sequence Alignments: To generate various sequence alignments between target and template proteins in the 22297 non-redundant PDB set, six types of BLAST-related alignment programs (BLAST,[14] PSI-BLAST,[26] PSF-BLAST,[27] RPS-BLAST, IMPALA,[28] and Pfam[29])-BLAST) and two types of profile-profile alignment programs (SP3[15] and SPARKS2[16]) were executed. The SPARKS2 and SP3 programs were shown to be excellent in relation to the sequence alignment in the CASP6 experiment.[30] As explained in the next step, various alignments were filtered with our alignment score, the PF score, and with the statistical score about reliance for the sequence alignment.

(2) Filtering Sequence Alignments: First, the alignment score value, $Score_{ali}$, which is also called the POWER FUNCTION (PF) score including the power numbers, was calculated using Eq. 1 for the six BLAST-related alignment methods. This PF score plays an important role in reducing the number of the alignments obtained from the step (1).

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

Here, *Len* represents the number of residues of a region in the target protein or query sequence aligned to the amino acid sequence of a template protein. *SEQid* represents the percent of sequence identity between the target and the template proteins. *SS* is the degree of match between the predicted secondary structure elements (SSE) from the target sequence and the SSE of the template protein. The predicted SSE from amino acid sequence was obtained with PSI-PRED.[31] The SSE of the template protein was assigned using 3D coordinates of the experimental structure by STRIDE.[32] The k_i value by which the significance weight is described in each of the six alignment methods is a coefficient for each alignment method. The k_i value and the parameters (m, n) were optimized for each sequence identity level of 50, 40, 30, 20, and 10%. The details of this scoring function or the PF score will be reported in another paper.

For the other two alignment methods, *i.e.* SP3[15] and SPARKS2,[16] the Z-scores of their output were used to filter the alignments. In a set of alignment score values, the Z-score is calculated in subtracting the average value of the set from a certain value in the set and dividing the standard deviation. Such a statistical Z-score is relatively reliable, especially when a target sequence has a high sequence identity with a template protein. We decided the following cut-off values to reduce the number of alignments, using the training set of CASP7 targets.[11] When the Z-score for an alignment was greater than or equal to the maximum Z-score×X, we adopted the alignment. Here, the X value is smaller than the value of 1, and the X depends on the modeling difficulty of the query sequence. The adopted alignments were used to construct the 3D structures in the next step (3). In other words, the parameter X is the cut-off value that was obtained by an optimization process in which we used the training set of the CASP7 targets.[11] The parameter X depending on the modeling difficulty of the target was decided on as shown in Table 1. In this paper, we used the Z-score and cut-off parameter X to select alignments for SP3 and SPARKS2 method. If we use the P-value of the alignment score instead of the Z-score and parameter X with confirming that alignment score distributes normally, the selection of the reliable alignments may be more successful. To estimate the difficulty of a target in the selection process of some alignments obtained from the SP3 and SPARKS2 programs, the support vector machine (SVM)[33] was used. Two values of score and sequence identity (%) of each top ranked alignment resulting from both PSI-BLAST[26] and SPARKS2[16] were used as vectors for SVM classification. Four classes of difficulty grade (‘CMeasy,’ ‘CMhard,’ ‘FR’ and ‘NF’) were obtained from each alignment program. To identify the difficulty grade of a target protein, the combination of two difficulty classes obtained from the two alignment programs was adopted, as shown in Table 1. The PSI-BLAST method is excellent for CMeasy and CMhard due to the base of sequence-profile alignment, and the SPARKS2 is excellent for CMhard and FR due to the base of profile–profile alignment. Then, both methods were used to take in the broad band in relation to the difficulty of the alignment.

(3) Constructing 3D Structure Models: We constructed 3D structure models using the homology modeling program FAMS[17] based on each selected alignment obtained in the steps (1) and (2). FAMS 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 protein, 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 the Critical Assessment of Fully Automated Structure Prediction (CAFASP-2) (2000) experiment, which is one category of CASP4 experiment, and CAFASP-3 (2002) experiment, which is one category of CASP5 experiment, FAMS was recognized as the good software for homology modeling.[34,35]

(4) Ranking Models According to Scoring Function for 3D Models: All the constructed models from the steps (1) and (2) were evaluated using the following scoring function (Eq. 2),

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

Here, *CCL* represents the CIRCLE score[18] which is based on a 3D–1D profile score (such as Verify3D[36]), and the *SSscore* represents the secondary structure agreement score that was calculated by comparing the secondary structure judged from the 3D model with the secondary structure predicted using the PSI-PRED[31] from the query sequence. The details of this score are mentioned by Terashi *et al.*[18] As shown in Table 1, the *w* value is the weighting factor for the *SSscore*, which was optimized using the training set based on the CASP7 targets.[11] The weight values of *w* were 0.3 and 1 for easy and difficult targets about the modeling process, respectively. It was shown that the agreement score in the secondary structure is also significant in addition to the CIRCLE score.

Table 1. Optimized Values of X and *w*

PSIB[a]	SPK2[b]	X[c]	*w*[d]
CMeasy	CMeasy	0.99	0.3
CMhard	CMeasy	0.9	0.3
CMeasy	CMhard	0.83	0.5
CMhard	CMhard	0.85	0.5
CMhard	FR	0.85	0.5
NF	CMhard	0.8	0.5
NF	FR	0.8	1
CMhard	NF	0.8	1
NF	NF	0.8	1

a) Predicted difficulty using alignment score and sequence identity of PSI-BLAST. *b*) Predicted difficulty using alignment score and sequence identity of SPARKS2. *c*) Cut-off parameter X changes from the 0.99 to the 0.8 as the difficulty increases. *d*) Parameter *w* is the weighting factor of *SSscore* as shown in the Eq. 2. We decided the parameters, X and *w*, using the training set of the CASP7 targets.

The details of the FAMSD method and the remarkable results of the benchmark test using the CASP8 target proteins[12] among the participating teams in the CASP8 experiment will be reported elsewhere. As a result of the benchmark test, we summarize that the FAMSD method has an excellent capability to pack amino acid side-chains with correct torsion angles, in addition to the correct main chain, while avoiding the formation of atom-atom collisions that are not observed in native protein structures.

Creation of the HUMAN FAMSD-BASE In the HUMAN FAMSD-BASE, the protein sequences of the human genome were obtained from the sequence collection ‘hsap2,’ which contains 28954 protein sequences, in the Genomes TO Protein structures and functions (GTOP) database.[24,25] The FAMSD method was applied to each sequence in ‘hsap2’ to construct reliable 3D structure models. However, if the FAMSD method assigned the target protein as ‘very difficult’ or impossible for protein modeling, 3D models of the targets were not constructed nor included in the HUMAN FAMSD-BASE, because these models were not sufficiently accurate or reliable. Consequently, for 72% (20743 out of 28954) of proteins from the human genome, one or more relatively reliable 3D models were constructed. This HUMAN FAMSD-BASE was created using the template structures of the 22297 non-redundant PDB chain set of August 2008. During a year, however, the number of the template 3D structures increased by 2197 in relation to the non-redundant PDB chains as at July 2009; therefore how the quality of the HUMAN FAMSD-BASE is affected by the change of the number of the non-redundant PDB chains is discussed in ‘Results and Discussion.’

Handling of Membrane Proteins In the ‘hsap2’ sequence collection of the GTOP database, there are many sequences of membrane proteins, such as G-Protein Coupled Receptor (GPCR) family proteins. According to the GTOP, 6470 out of 28954 sequences in the ‘hsap2’ were categorized as membrane proteins based on the prediction of transmembrane helices using the SOSUI program.[37]

In our FAMSD method, as described in ‘Methods,’ the 3D models constructed in the step (3) were ranked with the $Score_{str}$ including the CIRCLE score[18] in Eq. 2. The CIRCLE score is based on the sum of 3D–1D profile scores in the unit of the amino acid residue of the target protein, and estimates the stability in aqueous solution of a protein structure from a free energy point of view. In addition, the CIRCLE score represents the stability for soluble protein structures, and some parameters used in the CIRCLE program were determined using side-chains environments in experimental structures consisting of soluble proteins. Side-chains on the surface of a membrane protein are surrounded by lipids or hydrophobic molecules. The side-chain environments of membrane proteins are very different from those of soluble proteins. Thus the CIRCLE score should not be used as quality estimation for 3D models of membrane proteins. It should be noted that the $Score_{str}$ in Eq. 2 could not estimate the quality of 3D models for the membrane proteins appropriately. On the other hand, the $Score_{ali}$ in Eq. 1 is useful for both membrane proteins and soluble proteins, because the $Score_{ali}$ is obtained from an alignment that represents the evolutionary relationship between the target and template proteins.

Accordingly, for the 6470 membrane proteins defined based on the SOSUI in the HUMAN FAMSD-BASE, the 3D models were ranked using the $Score_{ali}$ in Eq. 1 for each query sequence. These ranking results are described when users select the ‘hsap2membrane’ as a species code on the top page of HUMAN FAMSD-BASE. Note that these 6470 membrane proteins are also included in the normal ‘hsap2’ collection in which the ranking for the 3D structures were executed using the $Score_{str}$ in Eq. 2, because false-positive judgments may be contained in the prediction of the transmembrane helices using the SOSUI program.[37]

Results and Discussion

The HUMAN FAMSD-BASE In the HUMAN FAMSD-

BASE, the total number of models and the average number of models per one target protein were 242918 and 11.7, respectively. The quality of the constructed models was estimated by the CIRCLE score,[18] which is a 3D–1D profile score based on Verify 3D.[36] According to our benchmark test using past CASP targets, in many cases the first ranked model by the CIRCLE score is statistically near to the native structure among the model candidates.[18] However, the model that is nearest to the native structure is infrequently a candidate model that is not ranked first. This means that the summation of isolated free energy-like score for each amino acid residue of the protein sequence is not absolutely enough to determine the order of the protein stability. Therefore, the HUMAN FAMSD-BASE includes all of the constructed models in the step (3) based on reliable alignments obtained in the steps (1) and (2), not only the first ranked models. Thus, the many models of total number 242918 were constructed, and the average number of constructed models per one target protein was 11.7. The models for each target were sorted by the CIRCLE score such that the first ranked model was the protein structure nearest to the native structure. Figure 1 shows the distribution of model number for the 10% band of sequence identity (%) in the alignment between the target proteins, which are the subjects of the structure prediction, and the templates, which are registered as the experimental structures in the PDB database. In Fig. 1, as only an alignment whose model was first ranked by the $Score_{str}$ was adopted in the modeling for each target protein, the total number of alignments used for describing this figure is 20743, which corresponds to the number of genes for which we could construct protein models. In the band of sequence identity from 0 to 10%, for example, there are 1,456 models corresponding to the 7% ratio of 20743. There are 6551 (32%) in the band of the sequence identity from 10 to 20%. Thus, 39% of the models of the 20743 proteins belong to the band of sequence identity from 0 to 20%. As the modeling of the target proteins having the sequence identity from 0 to 20% between the target and template proteins is very difficult technically, this HUMAN FAMSD-BASE includes many models for which we are generally unwilling to perform homology modeling due to the difficulty of performing the sequence alignment.

In the HUMAN FAMSD-BASE, if the template proteins have ligand binding sites, the sites may be significant for the function expression of the target protein. The HUMAN FAMSD-BASE provides a template chain file that contains coordinates of all ligands in PDB chains that have more than 95% sequence identity with the template protein. In other words, if the PDB chains in the 95% cluster have some ligands binding to them, then it suggests that the ligand-binding site is significant for the function of the subject protein.

Benchmark Test (1) in Comparison with the PF Method We wanted to perform the benchmark test for the quality of the models in the HUMAN FAMSD-BASE by comparing them with those in the previous version of the human protein structure model database called HUMAN FAMSBASE, which is part of the RIKEN FAMSBASE. However, the comparison of quality between both databases is very difficult, because the models in the above two human protein modeling databases have been constructed under two different conditions of the PDB version of August 2008 and September 2005, respectively, of when referring to the non-redundant PDB chain set. In the current HUMAN FAMSD-BASE, the FAMSD method was used to construct the models. On the other hand, in the preceding HUMAN FAMSBASE, the POWER FUNCTION (PF) method was used. The PF method employs BLAST-related programs, $Score_{ali}$ (that is, the PF score) in Eq. 1, and the FAMS modeling program. There are two major differences with the FAMSD method. In the preceding database, first, the SPARKS2 and SP3 programs were not used to generate alignments in the PF method. Second, the $Score_{ali}$ (PF score) in Eq. 1 was used in ranking models with the PF method, while the $Score_{str}$ in Eq. 2 was used in the FAMSD method. In the benchmark test of this paper, the models of the CASP8 targets in place of the models in the two human protein modeling databases compared were used. The quality of the models was assessed with the Global Distance Test Total Score (GDT_TS),[38] the %_correct_chi1 and the %_correct_chi2. The GDT_TS represents the accuracy of the Cα backbone geometry of the model, which is formally used in the CASP experiment. The GDT_TS value ranges from zero to 100. A high GDT_TS value indicates that the Cα backbone atoms of the model were predicted at near native positions. A GDT_TS value of 100 means that all the Cα coordinates of the model structure are within 1 Å in comparison with the experimental structure. The GDT_TS value is convenient and valuable to estimate the accuracy of the Cα backbone geometry of protein models. The %_correct_chi1 and the %_correct_chi2 are ratios of correct χ1 torsion angles and correct χ2 torsion angles, respectively. The χ1 torsion angle was considered "correct" if the value was within 40 degrees of the experimental value.[35] The χ2 torsion angle was considered "correct" if both the χ1 and χ2 values were within 40 and 60 degrees, respectively.

As shown in Table 2A, the FAMSD method provided significantly higher quality models than the FUNCTION method in the assessment of the Cα backbone geometry and side-chain conformation. The GDT_TS values of the FAMSD-based models were, on average, 9.6% higher than those of the FUNCTION-based models, as shown by diff % in Table 2A. In the assessment of side-chain conformation,

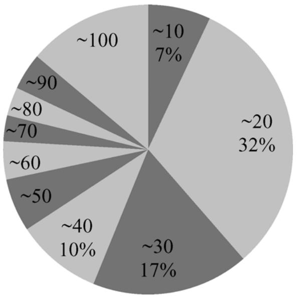

Fig. 1. Distribution of the Number Modeled against the Band of Sequence Identity (%) between Target and Template Proteins in HUMAN FAMSD-BASE

For each target protein, one alignment was used to calculate the distribution. The alignment whose model was first ranked by the CIRCLE score was adopted. Therefore, the total number of alignments in this distribution is 20743, which corresponds to the number of modeled genes. In this figure, for example, '~10' means that the sequence identity ranges from 0 to 10%, and '~20' means that the sequence identity ranges from 10 to 20%. The value written under the sequence identity range is the ratio of the number of alignments belonging to the sequence identity range for the number, 20743, of the total alignments.

Table 2. Benchmark Test (1) Results of the POWER FUNCTION and the FAMSD Methods

(A) Average Values of GDT_TS, %_correct_chi1 and %_correct_chi2

	PF[a]	FAMSD[b]	diff[c]	diff %[d]
GDT_TS[e]	54.4	59.6	+5.2	+9.6%
%_correct_chi1[f]	34.1%	39.0%	+5.0%	+14.6%
%_correct_chi2[g]	23.6%	26.9%	+3.3%	+13.9%

The 121 CASP8 target proteins were used as the benchmark set. *a*) POWER FUNCTION (PF) method that was used in the previous HUMAN FAMSBASE. *b*) FAMSD method that was used in the HUMAN FAMSD-BASE described in this paper. *c*) Difference calculated by subtracting PF from FAMSD. *d*) Ratio of increase from PF. *e*) Average of the GDT_TS value for the 121 CASP8 targets. *f*) Average of the %_correct_chi1 for the 121 CASP8 targets. *g*) Average of the %_correct_chi2 for the 121 CASP8 targets.

(B) The Number of Target Proteins in Which Either the PF Method or the FAMSD Method Was Superior to the Other Method

Sequence identity range[a]	PF[b]	FAMSD[c]	even[d]
0—10%	1	2	0
10—20%	2	28	24
20—30%	2	11	22
30—40%	0	3	7
40—50%	0	1	5
50—60%	0	0	7
60—70%	0	1	2
70—80%	0	1	1
80—90%	0	0	1
Total	5	47	69

a) Sequence identity which was obtained from the FAMSD method was used. *b*) The number of target proteins in which the GDT_TS value of the PF based model was higher by more than five points than that of the FAMSD based model. *c*) The number of target proteins in which the GDT_TS value of the FAMSD based model was higher by more than five points than that of the PF based model. *d*) The number of target proteins in which difference of the GDT_TS value between the PF based model and the FAMSD based model was less than 5.0.

the %_correct_chi1 and the %_correct_chi2 of the FAMSD-based models were higher by 14.6% and 13.9%, respectively, than those of the FUNCTION-based models. Table 2B shows the number of target proteins in which either the PF method or the FAMSD method was superior to the other method for each sequence identity range. Here, when the GDT_TS value of the model based on a method (Model A) was higher by more than five points than that of the model based on the other method (Model B), Model A was considered as superior to Model B. Five points of the GDT_TS value correlates to a 5% of full marks, which is estimated to be very small in relation to the difference in the 3D structure. As shown in Table 2B, for sequence identity range of "10—20%," the FAMSD method was superior to the PF method for the 28 target proteins. Conversely, the PF method was superior to the FAMSD method for only two target proteins. Both methods were approximately equal within the difference of 5 points for the 24 target proteins. Thus, the FAMSD method was superior to, or approximately as good as, the PF method for almost all the target proteins. Similarly, for the sequence identity range of "20—30%," the FAMSD method was superior to, or approximately as good as, the PF method for almost all the target proteins. We show an example, T0494, which was picked from the CASP8 target proteins and which belongs to the sequence identity range of "20—30%." For target protein T0494, The GDT_TS values of the FAMSD-based model and the PF-based model were 76 and 50, respectively, and the overlaid structures between the native structure and the two models based on the FAMSD and PF methods are shown in Figs. 2A and B, respectively. As shown in Fig. 2, it is apparent that the FAMSD-based model is nearer to the native structure than the PF-based model. In Fig. 1, the HUMAN FAMSD-BASE contains many models (56%) belonging to the band of the sequence identity from 0 to 30% in the sequence alignment between the target and template proteins; thus, the HUMAN FAMSD-BASE will provide many higher quality models than the preceding HUMAN FAMSBASE due to the superior modeling accuracy of the FAMSD method.

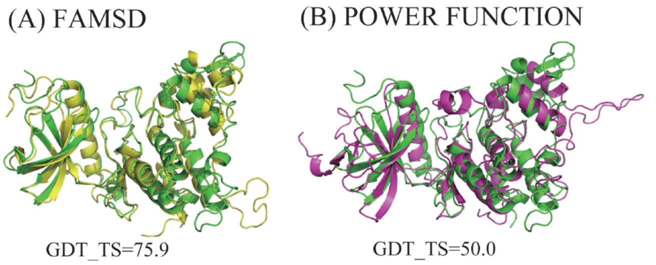

Fig. 2. Overlaid Structures between the Native Structure of T0494 and Two Models, the FAMSD Based Model and the PF Based Model

(A) Overlaid structure of the native and the FAMSD based model. Green and yellow colored ribbon model are the native structure and the FAMSD based model, respectively. The GDT_TS, %_correct_chi1 and %_correct_chi2 of the FAMSD based model were 75.9, 59.9% and 45.4%, respectively. (B) Overlaid structure of the native and the PF based model. Green and magenta colored ribbon model are the native structure and the POWER FUNCTION (PF) based model, respectively. The GDT_TS, %_correct_chi1 and %_correct_chi2 of the PF based model were 50.0, 33.4% and 25.9%, respectively.

Protein models are generally used to deduce the biological function where an experimental structure is not available, and, therefore, protein models are required to be highly accurate, not only in backbone geometry, but also in the side-chain conformation. Accordingly, the HUMAN FAMSD-BASE created in this paper is a useful and valuable tool for biological researchers, because the FAMSD method provides high quality models both in the Cα backbone geometry and in the side-chain conformation. The creation of the HUMAN FAMSBASE, based on the PF method, is economically effective because it does not consume vast computer resources due to the necessity of only performing the sequence analysis without many model constructions. However, because researchers need to access to useful models of target proteins belonging to the difficult modeling class, the creation of the HUMAN FAMSD-BASE is required, even if it is less economically advantageous in terms of computer resources due to the necessity of a higher number of model constructions for a given target protein. Therefore, it is useful and valuable that the HUMAN FAMSD-BASE becomes open to the public, along with the publication of our paper, in addition to the preceding HUMAN FAMSBASE.

Benchmark Test (2) in the Assessment of the Accuracy of the Constructed Model In addition to the above benchmark test (1) using CASP8 target proteins, we implemented another benchmark test (2) for the HUMAN FAMSD-BASE, using target proteins whose structures have been determined by experimental methods in the human protein sequences of 'hsap2.' Experimental structures that had missing residues were eliminated from this benchmark set. Consequently, 1370 target proteins were used as the first benchmark set. The FAMSD method was applied to each target in the first benchmark set. The native structures of the first benchmark set, which are answer structures, were omitted from the non-redundant PDB chain set for the purpose of no use. For 1227 (90%) out of 1370 target proteins, the FAMSD method could construct one or more 3D structure models. In the other 143 (10%) target proteins, no reliable alignments were detected; therefore no 3D models for this 10% were constructed. In the second benchmark set of 1227 target proteins, the total number of constructed models and the average number of models per one target protein were 13614 and 11.1, respectively. Figure 3 shows the distributed ratio of the number of the target proteins against the band of sequence identity (%) between the target and template proteins in the second benchmark set. We assessed the representative model, which is the first ranked model by the CIRCLE score, for each target protein in terms of RMSD_CA, GDT_TS, %_correct_chi1, and %_correct_chi2. The RMSD_CA is the Root Mean Square Deviation (RMSD) value between the Cα atoms of the experimental structure and those of the model. The %_correct_chi1 and the %_correct_chi2 are ratios of correct $\chi 1$ torsion angles and correct $\chi 2$ torsion angles, respectively. The $\chi 1$ torsion angle was considered “correct” if the value was within 40 degrees of the experimental value.[35] The $\chi 2$ torsion angle was considered “correct” if both the $\chi 1$ and $\chi 2$ values were within 40 and 60 degrees, respectively. Table 3 shows the average values of the RMSD_CA, the GDT_TS, the %_correct_chi1, and the %_correct_chi2 for each sequence identity region. These results indicate that the quality of a predicted model has a correlation with the sequence identity. As a whole, accordingly, the quality of the target protein in the HUMAN FAMSD-BASE for which a user such as biochemist or biologist wants to analyze might be inferred from the information in the band of the sequence identity between the target and template proteins in Table 3. The values of the RMSD_CA, the GDT_TS, the %_correct_chi1 and the %_correct_chi2 did not distribute normally for each band of the sequence identity between the target and template proteins; therefore, the box-and-whisker plots[39] for each band of the sequence identity are presented in Figs. 4A—D, which are corresponding to the RMSD_CA, the GDT_TS, the %_correct_chi1 and the %_correct_chi2, respectively. In each of many box-and-whisker plots[39] of Fig. 4, the horizontal line in the middle of the box represents the statistical median. The lower and the upper edges of the box are the 1st quartile (Q_1) and 3rd quartile (Q_3), respectively. For example, in the RMSD_CA assessment (Fig. 4A), the statistical median, the Q_1 and the Q_3 values are 3.6, 2.5 and 5.4 Å, respectively, in the sequence identity range of ‘∼30’ which represents the sequence identity from 20 to 30%. These values indicate that the number ratio of the models with the RMSD_CA <2.5 Å was 25 % in relation to the Q_1 value, the number ratio of the models with the RMSD_CA <3.6 Å was 50% in relation to the statistical median, and the number ratio of the models with the RMSD_CA <5.4 Å was 75% in relation to the Q_3. In other words, when a user obtains a model with 20—30% sequence identity, these values estimate the quality of the model with the RMSD_CA <2.5 Å, <3.6 Å and <5.4 Å with probability of 25%, 50% and 75%, respectively. The detailed data used to draw the box-and-whisker plots can be found in the Supplementary Data of the HUMAN FAMSD-BASE which is open to the public at http://mammalia.gsc.riken.jp/

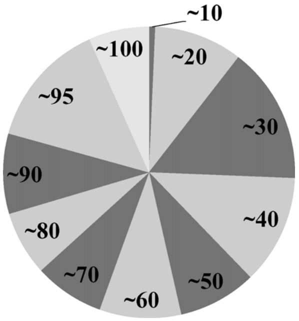

Fig. 3. Distribution of the Number Modeled against the Band of Sequence Identity (%) between Target and Template Proteins in the 1227 Benchmark Set in Benchmark Test (2)

For each target protein, one alignment was used to calculate the distribution. The alignment whose model was first ranked by the CIRCLE score was adopted. Therefore, the total number of alignments in this distribution is 1227. In this figure, for example, ‘∼10’ means that the sequence identity ranges from 0 to 10%, and ‘∼20’ means that the sequence identity ranges from 10 to 20%.

Table 3. Assessment of Predicted Models in the 1227 Benchmark Set against the Sequence Identity (%)

Sequence identity (%)	The number of targets	Average RMSD_CA[a)]	Average GDT_TS[b)]	Average %_correct_chi1[c)]	Average %_correct_chi2[d)]
0—10	9	12.08	34.80	11.16	5.62
10—20	129	9.83	46.68	25.83	16.01
20—30	199	5.02	66.12	45.55	31.57
30—40	160	3.59	73.15	51.30	35.61
40—50	114	3.35	78.15	56.78	40.89
50—60	119	2.53	81.68	61.59	45.10
60—70	102	3.02	83.65	63.63	49.61
70—80	94	1.75	87.99	71.72	55.81
80—90	117	2.40	85.91	69.53	55.72
90—95	184	1.96	89.57	75.61	61.98
95—100	88	0.60	96.34	90.57	83.99

a) Average value of the RMSD_CA for each sequence identity region. The RMSD_CA is Root Mean Square Deviation (RMSD) value between Cα atoms of the experimental structure and those of the model. *b*) Average value of the GDT_TS for each sequence identity region. The GDT_TS represents the accuracy of the Cα backbone geometry of the model, which is formally used in the CASP experiment as an alternative to the RMSD_CA. The GDT_TS value ranges from zero to 100. A high GDT_TS value indicates that the Cα backbone atoms of the model were predicted at near native positions. *c*) Average value of the %_correct_chi1 for each sequence identity region. The %_correct_chi1 is the ratio of correct $\chi 1$ torsion angles. The $\chi 1$ torsion angle was considered “correct” if the value was within 40 degrees of the experimental value. *d*) Average value of the %_correct_chi2 for each sequence identity region. The %_correct_chi2 is ratio of correct $\chi 2$ torsion angles. A $\chi 2$ torsion angle was considered “correct” if both the $\chi 1$ and $\chi 2$ values were within 40 and 60 degrees, respectively.

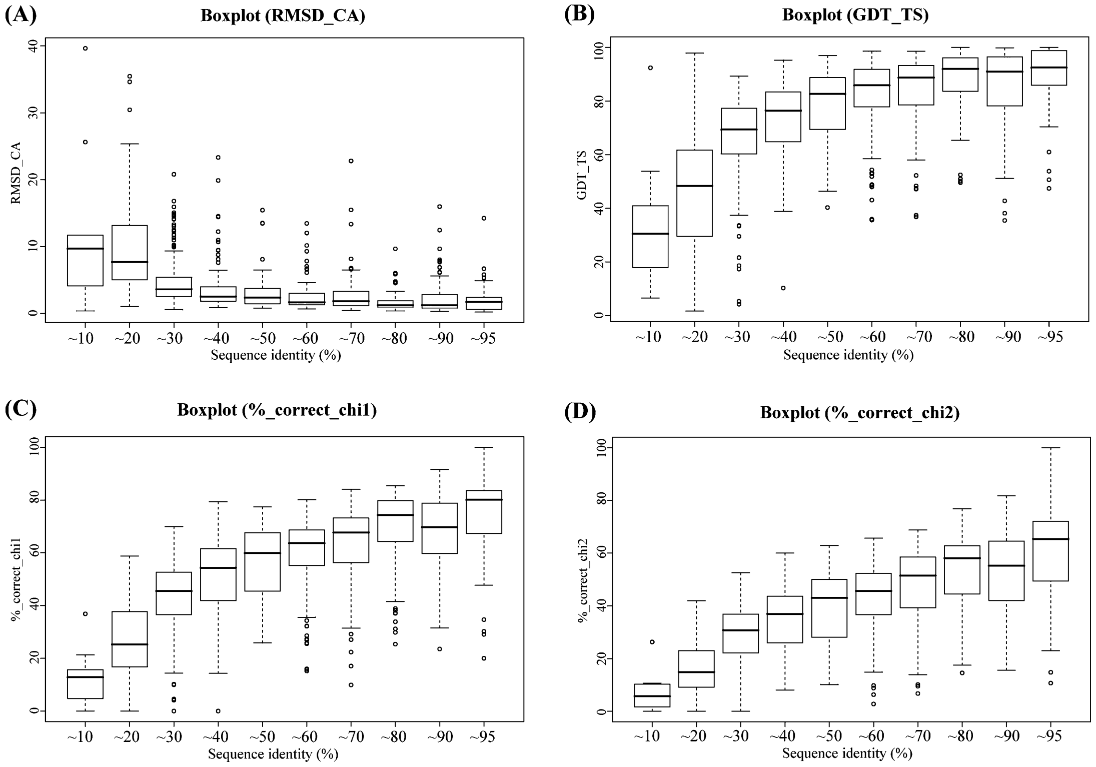

Fig. 4. Box-and-Whisker Plots against the Band of Sequence Identity (%) between Target and Template Proteins from the Assessment Results for the 1227 Benchmark Set in Benchmark Test (2)

In each of the box-and-whisker plots, the horizontal line in the middle of the box represents the statistical median. The lower and the upper edges of the box are the 1st quartile (Q_1) and 3rd quartile (Q_3), respectively. The upper and lower "whiskers" represents the farthest points that are not outliers (*i.e.*, that are within 3/2 times the interquartile range of Q_1 and Q_3). Open circles represent outliers that are over 3/2 times the interquartile range of Q_1 and Q_3. The detailed data used to draw the box-and-whisker plot can be found in the Supplementary Data. The statistical difference of the RMSD_CA values between each sequence identity ranges was determined by two-sided Wilcoxon's rank sum test, since the values of the RMSD_CA did not distribute normally. Difference with $p<0.05$ was considered significant. In most of the pairs, except for those detailed below, the difference was significant. The difference between '~50' and '~60' of the sequence identity range ($p=0.145$), the difference between '~70' and '~80' of the sequence identity range ($p=0.566$), and the difference between '~70' and '~90' of the sequence identity range ($p=0.157$) were not significant. (A) Box-and-whisker plot for RMSD_CA. (B) Box-and-whisker plot for GDT_TS. (C) Box-and-whisker plot for %_correct_chi1. (D) Box-and-whisker plot for %_correct_chi2.

human_famsd/. Thus, the HUMAN FAMSD-BASE indicates the approximate accuracy of the model, when a user of our database wants to know the accuracy of the referred model.

Example: Aspartyl tRNA Synthetase Aspartyl tRNA synthetase[40] consists of 501 amino acids. The 3D structure of aspartyl tRNA synthetase has not been determined experimentally, such as by X-ray diffraction. The HUMAN FAMSD-BASE provides the 3D structure models for Aspartyl tRNA synthetase as shown in Fig. 5. On the top page of the HUMAN FAMSD-BASE, when users input 'Aspartyl tRNA synthetase' in the Keyword search box, the result obtained is shown in Fig. 5A. The search results provide some information; (1) information for the query protein, such as the gene ID, the amino acid length and the annotation; and (2) information derived from alignment between target and template proteins, such as the pdb code, the sequence identity, and the alignment method for each constructed model. The first ranked model for aspartyl-tRNA synthetase was constructed using the structure of pdb code 1ASY as a template. The sequence alignment between the query and template proteins is shown in Fig. 5A. In this case, the sequence identity between the query sequence and sequence of the template protein was 57%. According to the benchmark test (2), the accuracy of the model in the sequence identity range of "50—60%" is estimated as the RMSD_CA <1.7 Å, the GDT_TS >86, the %_correct_chi1 >64% and the %_correct_chi2 >46%, with a probability of 50% in relation to the statistical median (Table 4). Thus, the quality of the model in the HUMAN FAMSD-BASE could be estimated roughly using the results of the benchmark test (2). For biochemist and biologist, thus, the model of aspartyl-tRNA synthetase is useful and valuable, because they cannot obtain the experimental 3-dimensional structure of this synthetase.

On the other hand, moreover, we picked another target protein different from aspartyl-tRNA synthetase as a standard-type of the comparison to estimate the 3D structure visually on the same sequence identity range ("50—60%") from the target proteins of the benchmark test (2), whose structures have already been experimentally determined. Protein of NP_001019195.1 (GTP cyclohydrolase I) was selected as such a representative protein or a standard-type. Figure 6 shows the overlaid structures of the predicted and experimental structures of protein NP_001019195.1, which is one of the target proteins in benchmark test (2). The sequence identity between NP_001019195.1 and template (pdb code 1FB1) sequences was 58.4%. The sequence identity range belonged to "50—60%," just as in the case of aspartyl-tRNA synthetase. The superimposed structures of the predicted 3D model for NP_001019195.1 are almost overlapping to each other from a visual point of view, and, therefore, Fig. 6 visually represents the typical statistical-medians of GDT_TS and RMSD_CA for the sequence identity range of 50—60%.

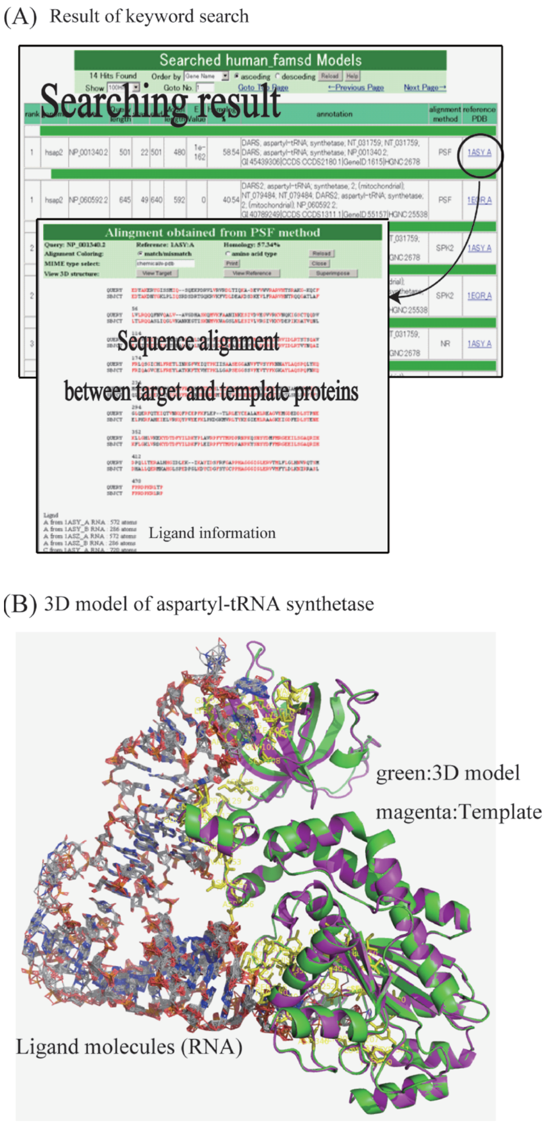
(A) Result of keyword search
Searched human_famsd Models
Searching result
Alignment obtained from PSF method
Sequence alignment
between target and template proteins
Ligand information
(B) 3D model of aspartyl-tRNA synthetase
green:3D model
magenta:Template
Ligand molecules (RNA)

Fig. 5. Top Page of the HUMAN FAMSD-BASE and 3D Protein Model for Aspartyl-tRNA Synthetase in the HUMAN FAMSD-BASE

(A) Searching results when users input the key word "aspartyl tRNA synthetase" at the top page of the HUMAN FAMSD-BASE. The search results provide some information; (1) information for the query protein such as the gene ID, the amino acid length and the annotation, (2) information derived from alignment between the target and template proteins, such as the pdb code, the sequence identity and the alignment method for each constructed model. The constructed 3D models are sorted by the $Score_{str}$ score for constructed models. When a user chooses "reference PDB" such as "1ASY_A," the sequence alignment between the query and template proteins will appear. If the template protein contains some ligand molecules, information for ligand molecules will appear below the alignment. (B) 3D protein model for aspartyl-tRNA synthetase. The green and magenta colored ribbon models are the predicted 3D model of aspartyl-tRNA synthetase and the template structure (pdb code 1ASY), respectively. The stick models are ligand molecules included in 95% sequence identity cluster proteins for the template structure. Yellow colored stick residues on the 3D model of aspartyl-tRNA synthetase are residues within 4 Å from the RNA molecules.

If the 95% sequence identity cluster proteins for the template protein contain some ligand molecules, information for ligand molecules will appear below the sequence alignment. In Fig. 5A, the "Superimpose" button provides the coordinates of the 3D structure model of the query protein and the template protein with superimposition. The coordinates of the ligand molecules are also included. Figure 5B shows the superposition between the model for aspartyl-tRNA synthetase and the template protein (pdb code 1ASY) with the ligand molecules. In this case, the RNA molecules that were obtained from 95% sequence identity cluster proteins for the template protein are included. Thus, a user might identify the RNA binding amino acid residues in the aspartyl-tRNA synthetase protein from the 3D structure model in the HUMAN FAMSD-BASE.

Table 4. Results of the Benchmark Test (2) for the Sequence Identity Range of 50—60%

	RMSD_CA		GDT_TS	%_correct_chi1	%_correct_chi2
Min[a)]	0.68 Å	Max[e)]	98.58	80.1%	65.7%
1st Qu[b)]	1.27 Å	3rd Qu[d)]	91.88	68.7%	52.3%
Median[c)]	1.67 Å	Median[c)]	85.86	63.7%	45.7%
3rd Qu[d)]	3.02 Å	1st Qu[b)]	77.82	55.2%	36.6%
Max[e)]	13.46 Å	Min[a)]	35.62	15.2%	2.8%

a) Minimum value in the benchmark test. *b*) 1st quartile (Q_1) value. *c*) Statistical median value. *d*) 3rd quartile (Q_3) value. *e*) Maximum value in the benchmark test. The values for the other sequence identity ranges can be found at the Supplementary Data.

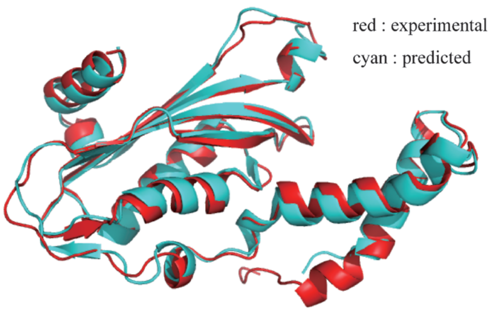

Fig. 6. Overlaid Structures of the Predicted and Experimental Structures of the Representative Protein (NP_001019195.1) for the Sequence Identity Range of 50—60%

Cyan and red colored ribbon models are predicted 3D model and experimental structure for NP_001019195.1, respectively. NP_001019195.1 (GTP cyclohydrolase I, pdb code is 1FB1) is one of the target proteins in the benchmark test (2). The predicted 3D model was constructed using pdb code 1WUR as a template, and the sequence identity between NP_001019195.1 and pdb code 1WUR was 58.4%. The values of the RMSD_CA, GDT_TS, %_correct_chi1 and %_correct_chi2 for the predicted 3D model were 1.3 Å, 87, 62% and 46%, respectively. Since the value for each assessment criterion is near by the value of the corresponding statistical median for the sequence identity range of 50—60%, NP_001019195.1 was selected as a representative protein or a standard-type protein in the sequence identity range.

Updating the Database The first version of the HUMAN FAMSD-BASE was created with the non-redundant PDB chain set of August 2008. The non-redundant PDB chain set is created based on the clustering with 95% sequence identity. At August 2008, the number of the representative PDB chains was 22297. However, by July 2009, the number of the representative PDB chains had increased by 2197. We implemented an update of the HUMAN FAMSD-BASE using the newly added 2197 PDB chains, as shown in Table 5.

As the result of the update, the number of models that were newly constructed in this update was 15489 (Table 5). The total number of more targets modeled newly or increasingly was 4979 in this update. Of these, 330 targets had no models at August 2008. The other 4649 targets had one or more models at August 2008. Consequently, in the latest HUMAN FAMSD-BASE, the numbers of total models and targets modeled totally were 258407 and 21073, respectively. The 22.4% ratio of the modeled targets (20743 in August 2008) changed in relation to the constructed model number, and the number of newly modeled targets increased by 1.6% during a year from August 2008 to July 2009. From the result

Table 5. Statistics for the July 2009 Update

	August 2008	Increased number by July 2009	July 2009
PDB[a)]	22297	2197	24494
Models[b)]	242918	15489	258407
Modeled targets[c)]	20743	4979 (330)[d)]	21073

a) The number of PDB chains of the non-redundant PDB chain set. *b*) The number of the constructed models. *c*) The number of targets that have at least one model. *d*) The number of the modeled targets increased by 4979 in the update of July 2009. Of these, 330 targets had no models at August 2008. The other 4649 targets had one or more models at August 2008, *i.e.* the number of models for the 4649 targets increased from equal to or more than one.

of the large change of the 22.4% ratio mentioned above, thus, the HUMAN FAMSD-BASE should be updated periodically with the latest version of the PDB database, at least once in a year.

Conclusion

We created a new protein structure model database called HUMAN FAMSD-BASE for human protein sequences of 'hsap2.' This model database includes protein models constructed with the FAMSD protein modeling method. Assessing the FAMSD method using the CASP8 targets showed that it has an excellent capability to pack amino acid side-chains with correct torsion angles, while avoiding the formation of atom-atom collisions that are not observed in native protein structures. As shown in Table 2, furthermore, in comparison with the POWER FUNCTION method that was used to construct models in the previous HUMAN FAMSBASE, the FAMSD method provides significantly higher quality models than the FUNCTION method in the assessment of Cα backbone geometry and side-chain conformation. Thus, due to the high quality models, both for Cα backbone geometry and side-chain conformation, provided by the FAMSD method, the HUMAN FAMSD-BASE is a useful and valuable tool for biological, pharmaceutical and medicinal researchers. As shown in Fig. 1, the 39% ratio of human protein models in the HUMAN FAMSD-BASE have the alignments of low sequence identity (under 20%), for which we are generally unwilling to construct homology models due to the technical difficulty for having the statistically significant alignment between the target and template proteins. Then, the HUMAN FAMSD-BASE should provide many useful models especially in the band of low sequence identity in addition to that of the high sequence identity for researchers.

Interestingly, the benchmark test (2) showed that the accuracy of the constructed model can be estimated from referring to the various values of the average RMSD_CA, the average GDT_TS, the average %_correct_chi1, and the average %_correct_chi2 in Table 3, if researchers using the HUMAN FAMSD-BASE check the sequence identity percent between the target and template proteins. From another box-and-whisker plot point of view, moreover, the supplementary data of this paper gives information about the model quality from the sequence identity between the target protein and template proteins. We took up aspartyl-tRNA synthetase protein as an example. The RMSD value for minimum, first quartile (Q_1), statistical median, third quartile (Q_3) and maximum, and the GDT_TS, %_correct_chi1 and %_correct_chi2 values for maximum, Q_3, statistical median, Q_1 and minimum are shown in Table 4.

Moreover, it is important that newly published experimental structures are reflected in the model database. We implemented an update of the HUMAN FAMSD-BASE using newly published PDB chains added between August 2008 and July 2009. The database should be periodically updated with the latest version of the PDB database, at least once in a year, as the number of available template proteins increases. By including the latest modeling structures in the update, the HUMAN FAMSD-BASE will have an important impact on the progress of biological science. Researchers will be provided with useful models of target proteins belonging to the difficult modeling class; therefore, the successive creation of the HUMAN FAMSD-BASE is required, even if it is less economically advantageous in terms of computer resources than previous database. Furthermore, it is useful and valuable that the version of the HUMAN FAMSD-BASE based on the PDB version of August 2008 becomes open to the public, in addition to the preceding HUMAN FAMSBASE, which was based on the PDB version of September 2005.

Supplementary Data Supplementary Data are available at the Help page of the HUMAN FAMSD-BASE, http://mammalia.gsc.riken.jp/human_famsd/docs/manual.pdf.

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. Authors thank Dr. Fukuchi at National institute of Genetics for providing data of the membrane proteins in the 'hsap2.'

References

1) Nierman W. C., Eisen J. A., Fleischmann R. D., Fraser C. M., *Curr. Opin. Struct. Biol.*, **10**, 343—348 (2000).

2) Nakayama T., Fujii M., Yokoyama S., *Tanpakushitsu Kakusan Koso*, **47**, 982—986 (2002).

3) Yokoyama S., Hirota H., Kigawa T., Yabuk, T., Shirouzu M., Terada T., Ito Y., Matsuo Y., Kuroda Y., Nishimura Y., *Nat. Struct. Biol.*, **7** (Suppl.), 943—945 (2000).

4) Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E., *Nucleic Acids Res.*, **28**, 235—242 (2000).

5) Berman H., Henrick K., Nakamura H., Markley J. L., *Nucleic Acids Res.*, **35**, D301—D303 (2007).

6) Moult J., Hubbard T., Bryant S. H., Fidelis K., Pedersen J. T., *Proteins*, **29** (Suppl. 1), 2—6 (1997).

7) Moult J., Hubbard T., Fidelis K., Pedersen J. T., *Proteins*, **37** (Suppl. 3), 2—6 (1999).

8) Moult J., Fidelis K., Zemla A., Hubbard T., *Proteins*, **45** (Suppl. 5), 2—7 (2001).

9) Moult J., Fidelis K., Zemla A., Hubbard T., *Proteins*, **53** (Suppl. 6), 334—339 (2003).

10) Moult J., Fidelis K., Rost B., Hubbard T., Tramontano A., *Proteins*, **61** (Suppl. 7), 3—7 (2005).

11) Moult J., Fidelis K., Kryshtafovych A., Rost B., Hubbard T., Tramontano A., *Proteins*, **69** (Suppl. 8), 3—9 (2007).

12) CASP8 home page ⟨http://www.predictioncenter.org/casp8/index.cgi⟩ 2008.

13) CASP8 abstracts ⟨http://www.predictioncenter.org/casp8/doc/CASP8_book.pdf⟩ 2008.

14) Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J., *J. Mol. Biol.*, **215**, 403—410 (1990).

15) Zhou H., Zhou Y., *Proteins*, **58**, 321—328 (2005).

16) Zhou H., Zhou Y., *Proteins*, **55**, 1005—1013 (2004).

17) Ogata K., Umeyama H., *J. Mol. Graph. Model.*, **18**, 258—272, 305—306 (2000).

18) Terashi G., Takeda-Shitaka M., Kanou K., Iwadate M., Takaya D., Hosoi A., Ohta K., Umeyama H., *Proteins*, **69** (Suppl. 8), 98—107 (2007).

19) Yoneda S., Yoneda T., Kurihara Y., Umeyama H., *J. Mol. Graph. Model.*, **21**, 19—27 (2002).

20) Umeyama H., *Nippon Yakurigaku Zasshi*, **120**, 43—46 (2002).

21) Yamaguchi A., Iwadate M., Suzuki E., Yura K., Kawakita S., Umeyama H. and Go M., *Nucleic Acids Res.*, **31**, 463—468 (2003).

22) Umeyama H., Iwadate M., "Current Protocols in Bioinformatics," Chapter 5:Unit5.2, John Wiley & Sons, U.S.A., 2004.

23) Yura K., Yamaguchi A., Go M., *J. Struct. Funct. Genomics*, **7**, 65—76 (2006).

24) Kawabata T., Fukuchi S., Homma K., Ota M., Araki J., Ito T., Ichiyoshi N. and Nishikawa K., *Nucleic Acids Res.*, **30**, 294—298 (2002).

25) Fukuchi S., Homma K., Sakamoto S., Sugawara H., Tateno Y., Gojobori T., Nishikawa K., *Nucleic Acids Res.*, **37** (Database issue), D333—D337 (2009).

26) 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).

27) Nanatani K., Fujiki T., Kanou K., Takeda-Shitaka M., Umeyama H., Ye L., Wang X., Nakajima T., Uchida T., Maloney P. C., Abe K., *J. Bacteriol.*, **189**, 7089—7097 (2007).

28) Schäffer A. A., Wolf Y. I., Ponting C. P., Koonin E. V., Aravind L., Altschul S. F., *Bioinformatics*, **15**, 1000—1011 (1999).

29) Sonnhammer E. L., Eddy S. R., Durbin R., *Proteins*, **28**, 405—420 (1997).

30) Tress M., Ezkurdia I., Graña O., López G., Valencia A., *Proteins*, **61** (Suppl. 7), 27—45 (2005).

31) Jones D. T., *J. Mol. Biol.*, **292**, 195—202 (1999).

32) Frishman D., Argos P., *Proteins*, **23**, 566—579 (1995).

33) Vapnik V., "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995.

34) 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).

35) Fischer D., Rychlewski L., Dunbrack R. L. Jr., Ortiz A. R., Elofsson A., *Proteins*, **53** (Suppl. 6), 503—516 (2003).

36) Eisenberg D., Lüthy R., Bowie J. U., *Methods Enzymol.*, **277**, 396—404 (1997).

37) Hirokawa T., Boon-Chieng S., Mitaku S., *Bioinformatics*, **14**, 378—379 (1998).

38) Zemla A., *Nucleic Acids Res.*, **31**, 3370—3374 (2003).

39) Chambers J., Cleveland W., Kleiner B., Tukey P., "Graphical Methods for Data Analysis," 1983.

40) Scheper G. C., van der Klok T., van Andel R. J., van Berkel C. G., Sissler M., Smet J., Muravina T. I., Serkov S. V., Uziel G., Bugiani M., Schiffmann R., Krägeloh-Mann I., Smeitink J. A., Florentz C., Van Coster R., Pronk J. C., van der Knaap M. S., *Nat. Genet.*, **39**, 534—539 (2007).