ELSEVIER

An automatic homology modeling method consisting of database searches and simulated annealing

Koji Ogata* and Hideaki Umeyama

School of Pharmaceutical Sciences, Kitasato University, Tokyo, Japan

We introduce a method of homology modeling consisting of database searches and simulated annealing. All processes involving searches for homologous proteins, alignment, the construction of Cα atoms, construction of main-chain atoms, and the construction of side-chain atoms are performed automatically. In this method, main-chain conformations are generated from the weighted average of main-chain coordinates in reference proteins. The weight is defined by the local space homology representing the similarity of environmental residues at topologically equivalent positions in reference proteins. Side-chain conformations are generated for constructed main-chain atoms by database searches, and main-chain atoms are optimized for the fixed side-chain conformations. These two processes, i.e., the side-chain generation and main-chain optimization, are repeated several times. This type of construction provides a structure similar to the x-ray structure, in particular, for main-chain and side-chain atoms in the residues belonging to structurally conserved regions (SCRs). The accuracy of our method was evaluated for 14 proteins whose structures are known. The average root mean square deviation between models and x-ray structures was 2.29 Å for all atoms, and the percentage of χ1 angles within 30° was 72.6% for SCRs residues. Some models were in good agreement with their respective x-ray structures. Our method, which has the advantage of being automated, gives results similar to, or better than, published results for three widely used test proteins. Our software, FAMS, is available on the World Wide Web. © 2000 by Elsevier Science Inc.

Color Plates for this article are on pages 305–306.

Corresponding author: H. Umeyama, School of Pharmaceutical Sciences, Kitasato University, 5-9-1, Shirokane, Minato-ku, Tokyo 108-8642, Japan. Tel.: (+81)-3-3446-9553; fax: (+81)-3-3446-9553. *E-mail address*: umeyamah@platinum.pharm.kitasato-u.ac.jp (H. Umeyama)

*Present address: Unite de Conformation de Macromolecules Biologique, University Libre de Bruxelles, Av. F.D. Roosevelt 50, B-1050 Bruxelles, Belgium.

Keywords: homology modeling, database searches, simulated annealing, conserved side-chain torsional angles, local space homology, protein structures

INTRODUCTION

The ability to predict tertiary structures of proteins has developed with computer technology and an increasing number of experimentally determined protein structures.[1–3] Homology modeling is an effective method for predicting tertiary structure, provided there exist homologous proteins whose three-dimensional structures are known. Models created by homology modeling often have been used to explain functions and characterization.[4]

Various methods for homology modeling have been developed, and it has been reported that these methods are successful in achieving accurate models.[5–10] The simple method of homology modeling, which constructs models from reference proteins and database searches, is divided into the following processes: searches for homologous proteins, alignment, main-chain construction, side-chain construction, and optimization. Grafting fragments of homologous proteins usually produces the main-chain construction. It is supposed that two tertiary structures within a homologous family of proteins are similar to each other. When the main-chain atoms for proteins whose structures are known are constructed with this method, the protein model created is similar to the native structure. Side-chain construction is performed for fixed main-chain atoms. Various side-chain modeling methods have been developed, and it is reported that these methods are able to obtain accurate models.[11–17] After side-chain construction, energy optimization using any of several molecular force fields[18] is performed to remove short contacts and strain in the structure.

Each process in homology modeling is dependent on the earlier processes. Therefore, errors may be inadvertently introduced and propagated. For example, when constructed main-chains are strained, side-chains frequently are selected in an unexpected conformation.

Methods for Monte Carlo and simulated annealing methods

Journal of Molecular Graphics and Modelling 18, 258–272, 2000

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00037-1

have been developed.[19–22] Conformations of the models created can be similar to the x-ray ones. Model structures were generated to satisfy an objective function derived from reference proteins. The objective function includes information of side-chain conformations. However, the conserved side-chain conformation within homologous proteins depends on amino acid type.[23] And, the side-chain conformation for the same amino acid type at topologically equivalent positions is sensitive to changes in spatially proximal amino acid residues.

In this article, we introduce a method of homology modeling in which database searches are alternated with the simulated annealing method. A tertiary structure of target protein was constructed using the simulated annealing method with information from homologous proteins. The process is as follows: searches for homologous proteins, alignment, construction of Cα atoms, main-chain construction, and side-chain construction with the optimization of main-chain atoms. These processes are carried out automatically and repeatedly. Each process uses topological information from homologous proteins for pairwise structural alignment between the target and reference proteins. The coordinates of main-chain atoms are generated and optimized with side-chain atoms consisting of conserved side-chain torsional angles within homologous proteins.

METHODS

Our modeling method was performed on Digital-UNIX based VT-alpha533 of Visual Technology Ltd., Japan, and all programs were written in ANSI C. The tertiary structures of proteins used were from the Protein Data Bank (PDB) released in October 1997.[24] The main-chain and side-chain constructions were performed as shown in Figure 1.

Searching for Homologous Proteins and Alignment

First, we selected reference proteins from a sequence database. The reference proteins are those having "amino acid identity" greater than 30% with respect to the target protein. We use the term "amino acid identity" based on the ratio of correct amino acids to the total number of residues for the length of the aligned sequences. The sequence database was generated from the PDB and consists of representative proteins having the highest resolution in highly similar proteins, which were defined as having the amino acid identity value greater than 90%. Next, to check the structural similarity, pairwise structural alignment based on the superposition of Cα atoms[25,26] was performed among reference proteins. If a protein did not have a structurally similar protein, i.e., if the ratio of structurally conserved regions (SCRs) residues compared to the total sequences was less than 0.25, it was excluded from the reference proteins. Here, we defined SCRs residues as those having the distance between corresponding Cα atoms less than 1.0 Å. Next, a multiple structural alignment based on the superposition of Cα atoms was performed among the reference proteins. For this alignment, the target sequence was put on by sequence alignment using the method of Smith and Waterman.[27] This alignment was evaluated to determine if inserted gaps were concentrated in loop and variable regions (VRs), which are defined by residues having the distance between Cα atoms

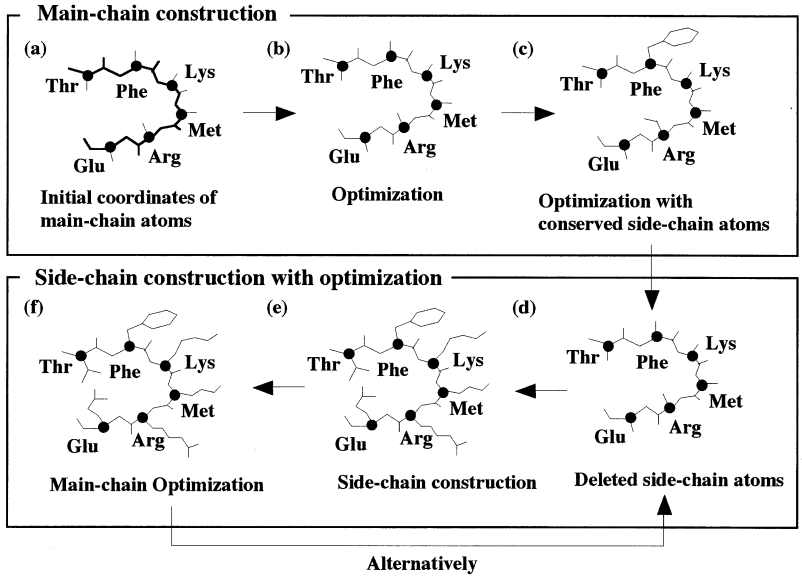


Figure 1. Main-chain and side-chain construction. Main-chain construction was performed as shown in panels a to c. (a) Initial coordinates of main-chain atoms were constructed by searching databases for optimized Cα atoms. (b) The initial main-chain atoms were optimized using the simulated annealing method. (c) For optimized main-chain atoms, the coordinates of side-chain atoms determined by the conserved side-chain torsional angles within homologous proteins were placed on fixed main-chain atoms. Main-chain atoms were optimized for partially generated side-chain atoms. Next, side-chain construction was performed as shown in panels d to f. (d) For the optimized main-chain atoms, side-chain atoms were deleted. (e) Side-chain conformations were generated for the fixed main-chain atoms. (f) Main-chain atoms were optimized for fixed side-chain atoms.

greater than 1.0 Å. We used a similarity (substitution) matrix derived from structural alignment.[28] Finally, we get a result of multiple alignment between a target and reference proteins.

Determination of Disulfide Bond and cis-Pro Residues

The information of Cys residues forming a disulfide bond and of cis-Pro residues is significant in homology modeling. The positions of these residues have been analyzed and predicted from results of alignment.[29] In this work, the positions of a disulfide bond and cis-Pro were determined from the results of sequence alignment on the reference proteins. In the case of Cys, first we pick the corresponding pair of Cys residues forming a disulfide bond in the reference proteins. If the number of the pairs in the aligned reference proteins is over 50% for the picked pair of Cys residues, such residues were regarded as forming a disulfide bond. Analogously, cis proline was determined.

Construction of Cα Atoms

Two separate processes were used to construct coordinates of Cα atoms. One process is the assignment of an initial set of coordinates from a reference protein and database searches. Another process is the optimization with an objective function. These two processes were performed alternatively.

(1) Construction of the initial Cα atoms Based on the results of alignment, the reference protein having the least number of inserted and deleted residues compared to the target sequence was chosen from the reference proteins. For convenience, we call this the base protein. If the target protein had many candidates for the base protein, the one having the highest amino acid identity value was chosen.

In aligned sequences of the target and base protein extracted from multiple alignment, we chose amino acid pairs belonging to a fragment continuing over three residue pairs. For these residue pairs, the coordinates of Cα atoms of the target protein were assigned based on those of the base protein. Fragments for which Cα positions were not determined were superimposed on basically assigned Cα atoms after searching a fragment database consisting of the number of residues from various proteins as shown in Figure 2.

(2) Construction of Cα atoms by the simulated annealing method We refined the initial Cα coordinates using simulated annealing with an empirical objective function. The objective function was chosen to create a model similar to the native one (see Appendix). Our objective function does not have a physicochemical meaning. The weighted value of w_i^j in Equation A5 is an important factor in determining the frame of the target protein. It was defined as local space homology (LSH) represented by the homology value calculated for residues within 12 Å of a given residue (Figure 3). The relationship between LSH values and the ratio of residue pairs belonging to the SCRs within homologous families of proteins is shown in Figure 4. This graph shows that when two proteins have overlapping Cα atoms, residue pairs with high LSH value may have the distance between Cα atoms smaller than 1.0 Å. Therefore, it is thought that a homology modeling method, which takes LSH values into consideration, is significant. We

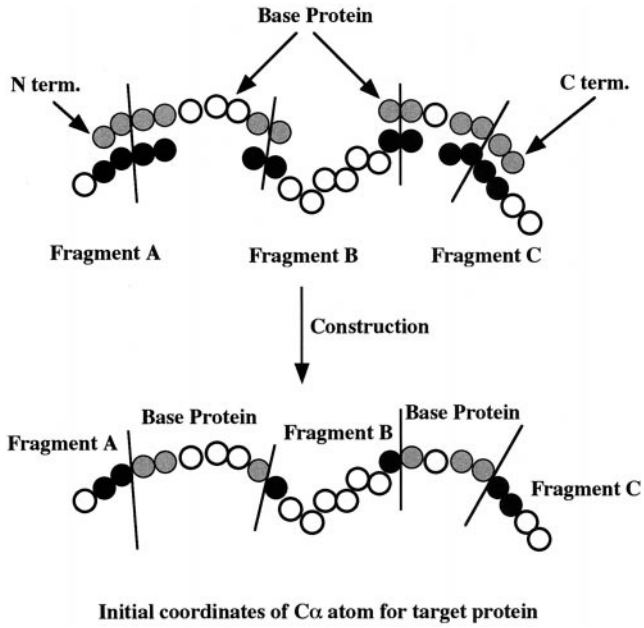
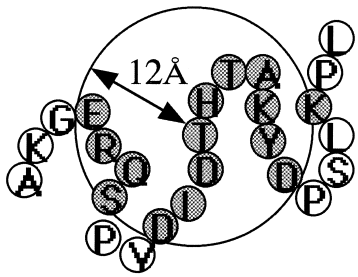


Figure 2. Construction of initial coordinates of Cα atoms. Cα atoms in the base protein provide Cα coordinates for the target protein. Fragment B having the lowest rmsd value calculated by the four overlapped residues was regarded as the optimal result of the database search. Fragment A involving the N terminal and fragment C involving the C terminal also were selected. Then, the four overlapped residues (colored gray in the target protein and colored black in the fragments) were used in the superimposing process. Finally, the target protein consisting of coordinates of Cα atoms was created.

defined M_i as the average distance between Cα atoms of the topologically equivalent positions in the reference proteins.

The processes of optimization are following. For the initial Cα coordinates, first, the weighted average of Cα coordinates $\langle x_i \mathbf{w}_i \rangle$ and the average distance M_i were obtained from the pairwise structural alignment based on the superposition of Cα atoms between the target and reference proteins. Next, simulated annealing with Equation A1 optimized the coordinates of Cα atoms. In the annealing procedure, perturbation of Cα atoms was performed within 1.0 Å. The equation of a random displacement for an atom, $x_n - x_{n-1} = 2.0[0.5 - \text{rand}\,(x_n - x_{n-1})]$, in which x_n and rand $(x_n - x_{n-1})$ are the coordinate vector in the n-th step and the random variable for the vector from (n−1)-th to n-th step, respectively, was used. The change in potential energy ΔV after displacement of the atom is calculated. If $\Delta V < 0$, the new configuration is accepted; on the other hand, if $\Delta V \geq 0$, a random number, i, in the interval (0,1) is selected for the vector change. Then, if exp (ΔV/kT), in which k and T are Boltzmann constant and temperature, respectively, is ≥i, the original configuration is kept, resuming at the preceding step (as in the Metropolis Monte Carlo procedure). Temperature is decreased by the equation, 0.5T, in each annealing step. The annealing step, which was calculated for all Cα atoms, was performed 100 times. The parameter corresponding to the temperature was started at 25 and decreased by




```
AKG ERQ-S PV DIDTHTAKYD PSL K PL
ARG NREAS SP NVDTHSARYD --L K PL
*   *   *   *   * *** *  **  * * **
```

Figure 3. Local space homology (LSH). LSH is calculated for a T residue in the figure. The circle has a radius of 12 Å on the central T residue. Residues within the circle for their Cα atoms are colored gray. The residue pairs on the structural alignment are enclosed by rectangles. The LSH value was calculated as the ratio of residues marked with asterisks. In this case, 9 residue pairs match in 16 residue pairs (including the gap). The LSH value is calcualted as (9/16) × 100 = 56.25%.

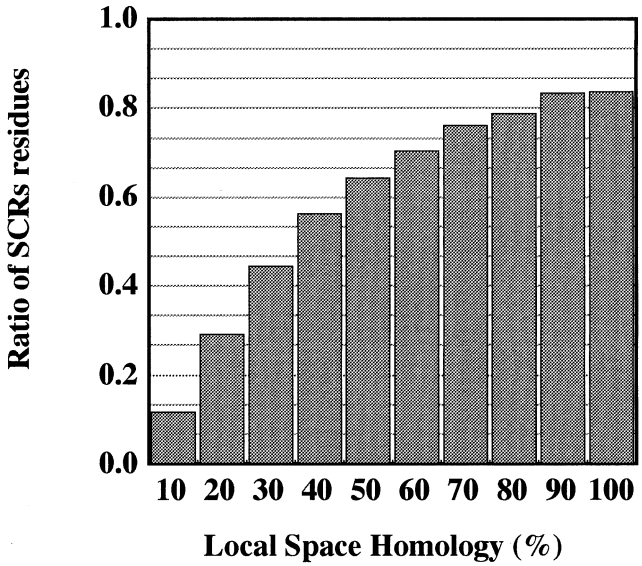


Figure 4. Relationship between local space homology (LSH) and the ratio of structurally conserved region (SCR) residues. The relationship between LSH values and the ratio of SCR residues is plotted for homologous families of proteins. The LSH values were obtained from the results of structural alignment based on the superposition of Cα atoms. The ratio of SCR is defined as the number of residues in the region compared to the total number of residues in the protein.

a factor of 0.5 and then by a fixed constant value after that step until 0.01 was reached.

These two procedures, obtaining structural information and the construction of Cα atoms, were repeated ten times, and the coordinates of Cα atoms having the least value of the objective function in the final annealing procedure were regarded as the optimal solution.

Main-Chain Construction

For the Cα coordinates, three separate processes were used to construct the main-chain atoms including Cβ atoms. One process is the assignment of main-chain atoms from the reference proteins and database searches. The second process is the optimization of the initial main-chain coordinates. The third process is also an optimization but different from the second; it considered the conserved side-chain atoms in the references.

(1) Construction of initial coordinates of main-chain atoms The other main-chain atoms, i.e., N, C, O, and Cβ (except Gly residue) atoms, were obtained from reference proteins and by searching a main-chain database.

First, structural alignment based on the superposition of Cα atoms was carried out between the target and reference proteins, and residues having a distance between Cα atoms less than 2.5 Å were picked up. Next, the coordinates of main-chain atoms except Cα atoms were assigned for the modeling structure from the reference residue having the least distance among Cα atoms.

Without considering the corresponding residue, searching a main-chain database consisting of four-residue fragments derived from the various proteins generated the coordinates of main-chain atoms. In this procedure, main-chain coordinates for residue i were assigned from a residue having the smallest root mean square distance (rmsd) value calculated between Cα atoms from $i-1$ to $i+2$. For residues at the N terminal, the superposition was performed for the Cα atoms from i to $i+3$; for the residue at the C terminal and one residue before the C terminal, the superposition was performed for Cα atoms from $i-3$ to i and from $i-2$ to $i+1$, respectively.

(2) Main-chain construction by simulated annealing Initial coordinates of main-chain atoms were refined by the simulated annealing method with an empirical objective function (see Appendix). It should be noted that our objective function does not have physicochemical meanings.

First, structural information for Equation A7, which consists of (1) the weighted average of the coordinates of main-chain atoms ($\langle x_i \mathbf{w}_i \rangle$), (2) the average of distance (M_i), and (3) the pair of N and O atoms forming the hydrogen bond, was obtained from the pairwise structural alignment between the target and reference proteins. Next, the refinement of main-chain coordinates including Cβ atoms was performed by the simulated annealing method. In the annealing procedure, perturbation of main-chain and Cβ atoms was performed within 1.0 Å of the starting position. The annealing step, which was the applied to all main-chain and Cβ atoms, was performed 200 times. The parameter corresponding to the temperature was started at 50 or 25 and decreased to 0.01 using a factor of 0.5 and then using a constant value for further decrements.

These procedures were performed six times, and the coordinates of main-chain atoms with the least value of objective function in final optimization procedure were regarded as an optimal solution. Then, the above parameter for temperature

was set at 50 for the first and second of six optimizations and was set at 25 for the rest.

(3) Main-chain construction with conserved side-chain atoms For the generated main-chain atoms, conserved side-chain torsional angles were obtained from homologous proteins using the method of our previous work.[23] In this method, the probability of conserved side-chain torsional angles within homologous proteins was examined, and side-chain modeling based on this information was performed. The coordinates of side-chain atoms consisting of conserved side-chain torsional angles (Table 1) were placed in relation to the fixed main-chain atoms. For example, if an Arg residue had a conserved $\chi 1$ angle within the homologous proteins, the coordinates of the Cγ atom were placed; if a Phe residue had conserved $\chi 1$ and $\chi 2$ angles, all side-chain atoms of Phe were placed. Optimization using the simulated annealing method with Equation A7 was performed only for main-chain and Cβ atoms; atoms were perturbed within 1.0 Å. The annealing step, which was applied to all main-chain and Cβ atoms, was performed 200 times. The parameter corresponding to the temperature is started from 25 and decreased to 0.01 by a factor of 0.5 and then by a fixed constant value after that step. The term $U_{nonbond}$ in Equation A7 was calculated between main-chain atoms and partially generated side-chain atoms. At that time, the coordinates of side-chain atoms became relocated during every perturbation process of N, Cα, C, and Cβ atoms as the conserved side-chain torsional angles were retained through the optimization. The structural information, $\langle x_i \mathbf{w}_i \rangle$, M_i, and the pair of N and O atoms forming the hydrogen bond, was derived from homologous proteins, and this information was used in optimization procedure. To obtain various main-chain conformations, we performed the described procedures three times, and the coordinates of main-chain atoms with the lowest value of the objective function in the final optimization procedure were regarded as the optimal solution.

Side-Chain Construction

Side-chain construction was performed for the fixed main-chain and Cβ atoms. It was carried out using the method of our previous work, which provides an accurate model in a short time.[23] Next, main-chain atoms were optimized by the Monte Carlo method with low temperature, which was set at 0.001, using the objective function (Equation A7); the term of $U_{nonbond}$ in Equation A7 was calculated between main-chain and full side-chain atoms. Then, the coordinates of side-chain atoms in the perturbation of N, Cα, C, and Cβ atoms were relocated as side-chain torsional angles were kept fixed. The perturbation of atoms was done within 0.5 Å. Next, side-chain atoms were deleted, and side-chain modeling as mentioned ealier was performed again. The two processes were repeated until the constructed structure did not have a short contact of atoms less than 2.4 Å and did not have Cα atoms with virtual torsional angles N-Cα-Cβ-C over $-120 \pm 15°$.

Evaluation of Our Method

We tested our method for 14 known proteins, including homologous proteins and proteins of various lengths (Table 2). These proteins often are used to evaluate prediction methods. The calculation of rmsd, correct side-chain torsional angles, etc., was considered by averaging the five created models, because in the refinement procedure simulated annealing gives various solutions. Reference proteins having an amino acid

Table 1. Atoms corresponding to conserved side-chain torsional angles

Amino acids	Fixed atoms	χ^1	χ^2	Others[a]
Ala	Cβ			
Arg	Cβ	Cγ	Cδ	Nε,Cζ,Nη1,Nη2
Asn	Cβ	Cγ	Oδ1,Nδ2	
Asp	Cβ	Cγ	Oδ1,Oδ2	
Cys	Cβ	Sγ		
Gln	Cβ	Cγ	Cδ	Oε1,Nε2
Glu	Cβ	Cγ	Cδ	Oε1,Oε2
Gly				
His	Cβ	Cγ	Nδ1,Cδ2,Cε1,Nε2	
Ile	Cβ	Cγ1,Cγ2,Cδ1	Cδ1	
Leu	Cβ	Cγ	Cδ1,Cδ2	
Lys	Cβ	Cγ	Cδ	Cε,Nζ
Met	Cβ	Cγ	Sδ	Cε
Phe	Cβ	Cγ	Cδ1,Cδ2,Cε1,Cε2,Cζ	
Pro	Cβ,Cγ,Cδ			
Ser	Cβ	Oγ		
Thr	Cβ	Oγ1	Cγ2	
Trp	Cβ	Oγ	Cδ1,Cδ2,Nε1,Cε2,Cε3,Cζ2,Cζ3,Cη	
Tyr	Cβ	Cγ	Cδ1,Cδ2,Cε1,Cε2,Cζ,Oη	
Val	Cβ	Cγ1,Cγ2		

[a]Atoms that were not used as conserved values.

Table 2. Test proteins

Name of Proteins	ID code	No. of residues[a]	No. of homologous proteins	Identity of the closest homologue (%)[b]
Adipocyte lipid-binding protein	1ADL	131 (46)	6	42.2
FK506 binding protein	1FKB	107 (88)	1	40.5
HIV-1 protease	1HVI_A	99 (85)	3	48.5
Lignin peroxidase	1LLP	343 (192)	2	43.2
Lysozyme	1LZ1	130 (73)	4	37.7
Neocarzionstatin	1NOA	113 (52)	2	46.0
Phospholipase A2	1POA	118 (75)	7	50.0
Papain	1PPP	212 (90)	2	46.6
Beta trypsin	1TLD	223 (149)	25	43.6
Alpha-lytic protease	2ALP	198 (125)	4	36.6
Beta-lactamase	3BLM	257 (68)	2	42.3
Pepsin	4PEP	326 (146)	11	49.9
Bovine pancreatic trypsin inhibitor	5PTI	58 (37)	4	44.8
Aspartate aminotransferase	7AAT_A	401 (272)	3	46.8

[a]Values in parentheses are the number of SCR residues calculated from the pairwised structural alignment that maximizes SCRs.
[b]Amino acid identity of the closest homologue. The value was calculated from the sequence alignment.

identity value greater than 30% and less than 50% were selected from the sequence database.

A database for the fragment of Cα atoms was generated from 290 proteins with various folding for each number of residues from 3 to 20. A database for the main-chain atoms also was generated from the same 290 proteins. A side-chain library was generated from 72 proteins. The side-chain library involved 14,075 residues. When we modeled a target protein, we excluded its side-chain conformations from the side-chain library.

RESULTS

Table 3 shows our alignment accuracy, which is one of important factors for creating models. These were calculated for the structural alignment between target and base proteins. The number of incorrect residues is small in comparison with SCRs residues. 1PPP have the worst one. This protein was created from two reference proteins, which are 2ACT and 1THE_A. And 1THE_A has a different number of residues compared to the target protein. A fragment in SCRs was mistaken in the sequence correspondence.

Table 4 shows the average of rmsd values calculated for each of the five models and the native structure. The average rmsd for the models were 1.50, 1.51, and 2.29 Å for Cα, main-chain, and all atoms, respectively. 1TLD had the best rmsd value among all atoms, 1.55 Å. This protein had many homologous proteins for which their tertiary and secondary structures were very similar. And, as shown from rmsd for SCR residues, the active site in the created models for this protein showed conformations similar to the native one (Color Plate 1). Therefore, the models of this protein could take advantage of structural information from homologous proteins. On the other hand, 2ALP showed a very high rmsd value, although it also had four homologous proteins. In the alignment, 2ALP had a long insertion around Cys (170, sequential number; 220A, PDB-based number), and this residue formed a disulfide bond with Cys (137, sequential number; 189, PDB-based number) (Figure 5). As a result, the Cα atoms of the fragment around this residue could not be assigned from homologous proteins, suggesting this fragment separated from the native structure, resulting in a large rmsd value. In such cases, a long insertion in the target protein was involved, resulting in main-chain coordinates for the created models sometimes differing from the native one, even though it has a protein with similar sequence. Our reference proteins have sequence identity more than 30% for the target proteins. It was considered that this criterion of identity value might be enough to create an accurate model, but some created models are different from native

Table 3. Alignment accuracy

	No. of incorrect residues	
ID code	All	SCR
1ADL	4	0
1FKB	12	2
1HVI_A	0	0
1LLP	11	3
1LZ1	8	1
1NOA	8	0
1POA	11	9
1PPP	32	11
1TLD	27	1
2ALP	11	1
3BLM	3	2
4PEP	12	0
5PTI	0	0
7AAT_A	7	0

Table 4. Root mean square deviation between the five models and x-ray structure

	All residues (Å)									SCR residues (Å)								
	Cα			Main			All			Cα			Main			All		
ID code	Max[1]	Min.[2]	Ave.[3]	Max[1]	Min.[2]	Ave.[3]	Max[1]	Min.[2]	Ave.[3]	Max[1]	Min.[2]	Ave.[3]	Max[1]	Min.[2]	Ave.[3]	Max[1]	Min.[2]	Ave.[3]
1ADL	1.44	1.41	1.42	1.43	1.41	1.42	2.08	1.95	2.01	0.63	0.58	0.60	0.62	0.59	0.61	1.21	1.04	1.09
1FKB	1.54	1.43	1.49	1.45	1.35	1.42	2.80	2.58	2.67	0.74	0.68	0.72	0.79	0.72	0.75	1.49	1.34	1.42
1HVI_A	1.21	1.07	1.13	1.27	1.10	1.16	2.55	1.96	2.16	0.59	0.51	0.55	0.74	0.63	0.67	1.45	1.37	1.42
1LLP	1.55	1.21	1.37	1.54	1.21	1.38	2.01	1.72	1.87	0.59	0.56	0.57	0.62	0.59	0.60	1.36	1.27	1.30
1LZ1	2.07	1.92	1.98	2.15	2.02	2.07	3.21	2.73	2.94	0.73	0.65	0.70	0.72	0.65	0.70	1.58	1.46	1.57
1NOA	1.30	1.26	1.29	1.38	1.29	1.33	1.95	1.82	1.88	0.51	0.46	0.48	0.71	0.62	0.65	1.09	1.01	1.06
1POA	2.20	1.93	2.11	2.01	1.86	1.96	3.63	3.23	3.49	1.15	1.04	1.13	1.09	1.00	1.06	2.71	2.53	2.67
1PPP	1.19	1.09	1.14	1.23	1.12	1.15	2.31	2.14	2.14	0.72	0.52	0.85	0.77	0.57	0.89	1.71	1.31	1.74
1TLD	1.13	0.95	1.00	1.19	1.03	1.08	1.60	1.51	1.55	0.54	0.49	0.53	0.61	0.55	0.59	1.27	1.01	1.15
2ALP	2.46	2.15	2.44	2.36	2.04	2.36	2.96	2.67	2.94	0.69	0.60	0.66	0.74	0.64	0.71	1.72	1.46	1.62
3BLM	1.63	1.60	1.61	1.61	1.59	1.60	2.45	2.34	2.40	0.64	0.59	0.62	0.65	0.62	0.64	1.60	1.48	1.55
4PEP	1.52	1.40	1.45	1.54	1.41	1.47	2.04	1.88	1.94	1.14	0.83	0.97	1.16	0.88	1.01	1.75	1.26	1.45
5PTI	1.26	1.19	1.22	1.47	1.37	1.42	2.18	2.10	2.13	0.55	0.48	0.50	0.55	0.51	0.53	1.59	1.45	1.51
7AAT_A	1.38	1.36	1.37	1.34	1.31	1.33	2.02	1.93	1.98	0.65	0.63	0.63	0.68	0.66	0.67	1.40	1.33	1.36
Average			1.50			1.51			2.29			0.68			0.72			1.49

[1] Maximum of rmsd values in five structures.
[2] Minimum of rmsd values in five structures.
[3] Averaged of rmsd.

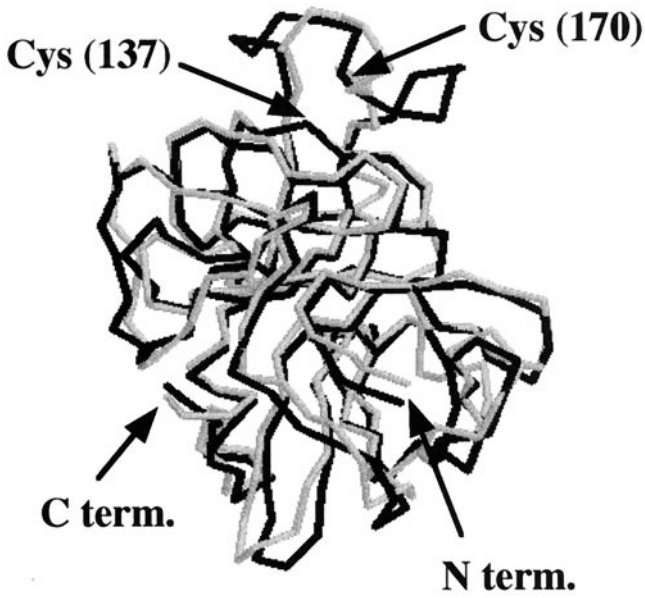


Figure 5. The Cα chain superimposed for native and modeled structures of alpha-lytic protease (2ALP). The native and modeled structures are shown in black and gray, respectively. The described model was selected from the five models.

Table 5. Percentage of correct side-chain torsional angles

	All (%)		SCRs (%)	
ID code	χ^1	χ^1 and χ^2	χ^1	χ^1 and χ^2
1ADL	71.1	57.7	82.7	67.3
1FKB	56.3	51.2	60.3	55.1
1HVI_A	57.7	46.5	65.4	55.4
1LLP	69.6	56.1	79.2	65.0
1LZ1	68.7	51.8	73.0	59.7
1NOA	66.6	57.6	70.8	63.1
1POA	51.2	42.6	61.6	53.2
1PPP	56.6	42.5	62.9	48.4
1TLD	72.4	64.5	78.5	71.7
2ALP	57.7	45.8	67.3	52.0
3BLM	59.9	43.8	79.6	60.0
4PEP	63.8	52.8	77.6	71.1
5PTI	75.7	62.4	82.1	69.0
7AAT_A	69.6	55.3	75.2	60.9
Average	64.1	52.2	72.6	60.8

structure owing to the long insertions and deletions on the alignment. This means that the number of inserted residues on the alignment is one factor affecting the accuracy of a model using homology modeling.

Table 5 shows the percentage of side-chain torsional angles within 30° of the x-ray structure values. These angles are one of the evaluations of the side-chain packing, and this criterion has been often used to evaluate the prediction of side-chain conformations. However, main-chain conformations of VR residues in the models frequently have different topologies from the native structure. Thus, we tabulated the percentage values for all atoms in addition to those for SCRs. Table 5 shows a high percentage of correct side-chain conformations in

the modeled structures. Our method took full advantage of the accuracy of the side-chain modeling method. The models of 1TLD had side-chain conformations similar to the native one. This protein could derive a large amount of information of conserved $\chi 1$ and $\chi 2$ angles from references. As shown in our previous article, the $\chi 1$ and $\chi 2$ angles of important residues not only determine the orientation of the side-chain structure of the residues, but also influence the surrounding side-chain conformations.[23] Thus, obtaining reliable information on $\chi 1$ and $\chi 2$ angles for conserved residues, we can generate accurate side-chain conformations surrounding the original residue. The strength of side-chain modeling is the high probability of conserved $\chi 1$ and $\chi 2$ angles within homologous proteins. The modeling of 1TLD is one such case, and using this model a small rmsd value was obtained.

The number of residues constructed from the conserved side-chain torsional angles of reference proteins also influences the accuracy of models. The difference in accuracy between 1FKB and 1ADL is an example. The models of 1FKB and 1ADL have similar rmsd values for Cα and main-chain atoms, although the number of reference proteins is different for 1FKB and 1ADL (Table 2). However, the rmsd value for all atoms of 1FKB was larger than that of 1ADL (Table 4). This difference was produced by the effect of the number of residues necessary to obtain conserved side-chain torsional angles within homologous proteins. The 1FKB model did not have a high percentage agreement for the conserved side-chains unlike the case of 1ADL (Table 5). Consequently, the rmsd for all atoms in 1FKB was worse than that of 1ADL.

Table 6 shows the percentage of reproduced hydrogen bonds in main-chain atoms for the native structure. Here, a hydrogen bond was defined by a distance between N and O atoms smaller than 2.9 $\pm$ 0.5 Å. The average percentage of reproduced hydrogen bonds is 85.6% for 14 tested proteins. From these values, it is seen that most of hydrogen bonds are reproduced in the created models. The 5PTI model had the best percentage. This protein had accurate main-chain atoms for a SCRs residue whose rmsd value was 0.53 Å. The values of 1LZ1, in which the main-chain atoms for models differed from native structure, were much lower than the average value. On the other hand, the percentage of hydrogen bonds for 2ALP was a little lower than the average value, although the rmsd value for main-chain atoms showed a worse value much like 1LZ1. This is due to the fact that the number of hydrogen bonds of SCRs was larger than those of 1LZ1.

DISCUSSION

Similarity of Models with Native Structures and Reference Proteins

We examined the structural similarity between our models, the native structure, and the reference proteins. Here, structural similarity was defined by the ratio of SCR residue for the length of target protein, and the SCR residues were obtained from a pairwise structural alignment based upon the superposition of Cα atoms. If the ratio of SCR residues between models and native structure (R_{m-n}) was larger than that of SCR residues between models and reference proteins (R_{m-r}), the created models could select similar fragments to the native structure. Oppositely, if R_{m-r} was larger than that of R_{m-n}, the created models were greatly affected by the structures of reference proteins. Moreover, the larger the ratio of SCR residues between native and reference proteins (R_{n-r}), the greater the need for accurate models with correct alignment. As shown in Table 7, the average values of R_{m-n} was smaller than that of R_{m-r} as a whole. In particular, the value of R_{m-n} for 1FKB, which had few homologous proteins, was smaller than those of R_{m-r}, and, moreover, this protein had larger R_{n-r} value. From these results, it seems that models similar to the reference structures were obtained.

To obtain an accurate model from a small number of homologous proteins, however, it is necessary that the reference proteins are topologically similar to the target protein, and that each process in the method provides accurate results. In particular, the alignment between the target and reference proteins is very important and influences the result of modeling. An alignment in which the amino acid correspondence in SCR residues is shifted provides an incorrect model from homology modeling. We performed sequence alignment with an amino acid similarity matrix obtained from structural alignment, and then the alignment was improved to insert gaps in the VRs. We think that our alignment would provide a correct amino acid correspondence of the target protein even when there is only a small number of homologous proteins. In addition, the other processes also influence the accuracy of modeling. Owing to these processes, we could obtain an accurate model. The modeling of 1NOA is such a case. The value of R_{m-n} for 1NOA is smaller than that of R_{m-r}, but the value of R_{n-r} is even smaller. Nevertheless, the accuracy of this protein for all residues is similar to those of 1FKB and 1PPP. This implies that each process in our method provides accurate results.

The values of R_{m-n} of 1ADL, 1TLD, and 4PEP were larger than those of R_{m-r}. These proteins have many homologous

Table 6. Percentage of reproduced hydrogen bonds in the main-chain[1]

ID code	Percentage (%)	ID code	Percentage (%)	ID code	Percentage (%)
1ADL	92.4	1NOA	73.5	3BLM	87.0
1FKB	88.7	1POA	89.2	4PEP	78.5
1HVI_A	79.4	1PPP	84.5	5PTI	98.9
1LLP	87.4	1TLD	87.0	7AAT_A	92.8
1LZ1	76.6	2ALP	82.3	Average	85.6

[1] Hydrogen bonds are defined by the distance between N and O atoms within 2.9 $\pm$ 0.5 Å.

Table 7. Ratio of the number of SCR residues[1]

ID code	Native vs references $(R_{n-r})^2$	Models vs references $(R_{m-r})^2$	Models vs native (R_{m-n})
1ADL	0.46 (0.61)	0.59 (0.65)	0.63
1FKB	0.82 (0.82)	0.95 (0.95)	0.78
1HVI_A	0.64 (0.86)	0.71 (0.83)	0.72
1LLP	0.72 (0.74)	0.85 (0.87)	0.79
1LZ1	0.58 (0.60)	0.74 (0.79)	0.65
1NOA	0.61 (0.65)	0.79 (0.85)	0.67
1POA	0.64 (0.71)	0.71 (0.72)	0.70
1PPP	0.66 (0.84)	0.78 (0.98)	0.79
1TLD	0.67 (0.76)	0.70 (0.75)	0.79
2ALP	0.67 (0.69)	0.80 (0.80)	0.66
3BLM	0.45 (0.56)	0.82 (0.87)	0.47
4PEP	0.50 (0.63)	0.56 (0.68)	0.62
5PTI	0.74 (0.88)	0.80 (0.86)	0.78
7AAT_A	0.65 (0.70)	0.89 (0.94)	0.67

[1] The ratio was obtained from the averaged ratio for reference protein. The SCR residues were calculated from the results of pairwised structural alignment based upon the superposition of Cα atoms.
[2] Values in parentheses are the ratio of the number of SCR residues for a reference protein having the largest ratio value.

proteins and a large amount of data was obtained from them. When our modeling is performed for a target protein having many homologous proteins, the models created were more similar to the native one rather than to the reference proteins. These results support the assumption that our modeling method is more likely to provide a native structure than a reference one.

Because the model is created from homologous reference proteins, the ratio of the SCR residue number for the reference protein having the largest ratio value in comparison with the native structure or the model also were shown in parentheses (R_{n-r} and R_{m-r}) of Table 7. The difference of R_{n-r} between the averaged ratio and the largest one is larger for 1HVI_A, 3BLM, and 5PTI than for the others. It means that the reference structures distant from the native were used in the homology modeling. As with the results, the difference of R_{m-r} between the averaged ratio and the largest one increased. If the weight of the reference proteins, which is used to choose the best fragments in the reference proteins, is determined carefully, the modeling accuracy may be increased.

Modeling from the Reference Protein Having the Closest Native Structure

The LSH value was used as a weight of the reference proteins to choose the best fragments in the reference proteins. However, if we can choose the one best reference protein that has a structure closest to the native one, we may be able to create a more accurate model. Thus, we performed the modeling using the best template protein showing the largest superimposition for the target protein.

Table 8 shows the SCR ratios for the models created from the closest reference protein. Those are shown in the columns labeled "The closest." It is natural that the values of R_{n-r} and R_{m-n} are smaller than that of R_{m-r} in the case without other reference proteins. R_{m-n} from "The closest " were larger than R_{m-n} from "Family." It means that if we find the best reference structure near the native, we can get a more accurate model. However, it is difficult to find the best reference protein among some references before the modeling. Therefore, we used the LSH value to construct main-chain atoms. Nevertheless, we acknowledge the fact that, in principle, it is possible to find a better reference structure for any given model. This statement might benefit potential researchers of human intervention in which the protein model is created from the reference having the largest sequence similarity.

As mentioned in the discussion for Table 7, changing the weight of the reference protein obtained from the LSH value may improve the homology modeling in the case where the ratio of SCR residue number is largely different among homologous proteins. It follows that if the selected reference proteins show the sequence similarity largely different from each other, the reference protein having the largest sequence similarity should be the only one used as the template protein. In order to create a more accurate model for the main-chain, on the other hand, if the reference protein having the largest SCR residue number of R_{m-r} is used as the only reference protein in repeated modeling, the variable loop regions of the target protein may not be created correctly. Those loops often are obtainable from other homologous proteins, and SCR structures of N-terminal and C-terminal sequences may be similar to those of other homologous proteins. In this case, including the LSH value using some reference proteins is appropriate.

Table 8. Comparison of structural similarity for models created from the closest reference proteins[12]

	From family			From the closest		
PDB ID	Native vs references $(R_{n-r})^2$	Models vs references $(R_{m-r})^2$	Models vs native (R_{m-n})	Native vs references (R_{n-r})	Models vs references (R_{m-r})	Models vs native (R_{m-n})
1HVI_A	0.64 (0.86)	0.71 (0.83)	0.72	0.86	1.00	0.83
3BLM	0.45 (0.56)	0.82 (0.87)	0.47	0.56	0.99	0.55
5PTI	0.74 (0.88)	0.80 (0.86)	0.78	0.88	0.95	0.83

[1] The ratio was obtained from the averaged ratio for reference proteins. The SCR residues were calculated from the results of pairwised structural alignment based upon the superpoison of Cα atoms.
[2] Values in parentheses are the ratio of the number of SCR residues for a reference protein having the largest ratio value.

Variability of Highly Similar Proteins and Models

To obtain an evaluation index for homology modeling in our tested proteins, we calculated rmsd values between highly similar proteins having amino acid identity value greater than 95%, having the same length and lacking no coordinates of Cα atoms. To compare the variability of models and highly similar proteins, we calculated the rmsd values among models. Generally, the optimization procedure using a probabilistic method, e.g., Monte Carlo, simulated annealing or genetic algorithm, tends to give various solutions, even if these are performed under the same conditions. In addition, main-chain and side-chain constructions were performed using different methods, and simulated annealing and database search were used in the main-chain construction and the side-chain construction, respectively. Therefore, attention should be paid to variability in the models.

From Table 9, the average rmsd among highly similar proteins were 0.51, 0.53, and 0.99 Å for Cα, main-chain, and all atoms, respectively. These values imply a smaller limitation of variability among models. The averages of rmsd values calculated between models are 0.57, 0.64, and 1.23 Å for Cα, main-chain, and all atoms, respectively. These values were similar to the above limitation values of variability. The models of 2ALP had larger rmsd value among the seventeen proteins. This protein had a large insertion, which resulted in variability in the models, and so the created models differed from each other. The models of 7AAT_A, on the other hand, showed very small rmsd values. This protein had the largest number of residues among tested proteins. This implies that our method provides a unique model that is not influenced by the number of residues. The models of 1ADL and 1TLD were described in Color Plate 2 as examples of variability.

The rmsd between the highly similar proteins implies a limitation in the accuracy of created models. If the rmsd value of the models is near to that of highly similar proteins, the method of homology modeling is very reliable. The difference of rmsd values for all atoms in comparison with Tables 4 and 9 is about 1.3 Å. It should be noted that the differences for 1TLD and 4PEP, which have many homologous proteins, are about 0.8 Å. Thus, it is expected that if we can obtain a large number of homologous proteins, our results would improve in comparison with this work.

Table 9. Comparison of variability in highly similar proteins and models

ID code	No. of highly similar proteins	Average of rmsd among highly similar proteins			Average of rmsd among models[1]		
		Cα	Main-chain	All	Cα	Main-chain	All
1ADL	8	0.33	0.35	0.80	0.30	0.40	0.96
1FKB	14	0.49	0.51	1.07	0.31	0.40	1.07
1HVI_A	83	0.53	0.56	1.11	0.52	0.61	1.44
1LLP	3	0.38	0.39	0.70	0.94	0.98	1.28
1LZ1	46	0.42	0.44	0.92	1.09	1.10	1.92
1NOA	3	0.60	0.61	1.09	0.37	0.47	1.03
1POA	3	0.58	0.58	1.25	0.71	0.75	1.41
1PPP	13	0.41	0.44	0.91	0.39	0.44	1.29
1TLD	23	0.30	0.33	0.73	0.47	0.55	0.92
2ALP	23	0.14	0.14	0.34	1.26	1.27	1.87
3BLM	8	0.47	0.49	1.03	0.40	0.47	1.12
4PEP	5	0.79	0.81	1.15	0.58	0.65	1.15
5PTI	23	0.76	0.83	1.39	0.31	0.49	0.90
7AAT_A	23	0.96	0.96	1.33	0.29	0.35	0.87
Average		0.51	0.53	0.99	0.57	0.64	1.23

[1]The rmsd values were calculated for five models.

Comparison with Other Homology Modeling Methods

To compare with other homology modeling methods, we performed the modeling of ten proteins used in the challenge of comparative modeling, Critical Assessment of Structure Prediction: Round 2 (CASP2; *http://predictioncenter.llnl.gov/casp2/Casp2.html*) and 3 (CASP3; *http://predictioncenter.llnl.gov/casp3/Casp3.html*). The codes of the predicted models were T0001, T0003, T0008, T0009, T0012, T0017, T0024, T0028, T0058, and T0060; the protein data must be contained in the PDB files in order to calculate the rmsd in comparison with the x-ray data. The created models were compared with the masked PDB data for all the residues and SCR residues. Only three proteins, T0001, T0003, and T0024, were evaluated in the following conditions. The compared residues in evaluations were T0001: 2–156 residues, T0003: 10–154, and T0024: 9–165.[30] The models of T0001 (1VDR_A), T0003 (2GPR), and T0024 (1U9A_A) were constructed from (3DFR, 7DFR, 8DFR), (1GPR, 1F3G), and (1AAK, 2UCE), respectively. Table 10 shows the rmsd for the predicted models of ten proteins. These average values were slightly larger than the average ones in Table 4. It was difficult for our system to create models for these proteins. In particular, T0009 (1JER) shows the largest rmsd for all the residues. The alignment between 1JER and its

Table 10. rmsd for CASP2 and CASP3 proteins[1]

CASP code	PDB ID	All residues (Å)			SCR residues (Å)		
		Cα	Main	All	Cα	Main	All
T0001	1VDR_A	1.87	1.86	2.64	0.84	0.96	1.60
T0003	2GPR	1.57	1.58	2.32	0.68	0.71	1.61
T0008	1COI	2.03	2.03	2.97	1.38	1.28	2.15
T0009	1JER	3.80	3.66	4.67	1.69	1.68	2.72
T0012	1PCI_A	2.86	2.84	3.58	1.64	1.65	2.38
T0017	6GSV_A	1.89	1.86	2.55	0.79	0.82	1.58
T0024	1U9A_A	2.34	2.23	3.26	1.55	1.42	2.60
T0028	1EG1_A	2.38	2.38	2.80	0.99	0.97	1.39
T0058	1UUG_A	1.74	1.75	2.28	0.60	0.64	1.55
T0060	1DPT_A	1.75	1.68	2.51	1.01	1.10	1.91
Average		2.22	2.19	2.96	1.12	1.12	1.95

[1]The rmsd values for our results were calculated as the averaged rmsd for five created models. The included residue in the rmsd calculations of T0001, T0003, and T0024 is the same as the criterion of the report by Martin et al.[30]

references had many incorrect residues in comparison with its structural alignment, and such a wrong alignment provided a model distant from the native structure. We emphasize that our algorithm is 100% automatic, which is an advantage, and our additive emphasis is the optimization of knowledge-based constraints. However, some of models generated to compare against the CASP models show inferior results compared to the best, for example, for T0017: best 0.41, ours 1.89; for T0009: best 2.28, ours 3.80. The details of implementation and optimization of our constraints may offer little in the way of added insight beyond the excellent work of human intervention done by other researchers at the beginning of this decade.

Table 11 shows the rmsd for the predicted models of three proteins by using various methods. From this table, results of the rmsd of 1VDR_A and 1U9A_A were similar to the best ones in the challenge groups for CASP2[30] because our alignments for two proteins were similar to those of the best results in CASP2 groups. Accuracy of created models usually depends on the alignment for their reference proteins. On the other hand, our rmsd values for 2GPR was slightly poorer than those of the Sali and MSI groups.[30,31] This difference might be due to the wrong alignment shown in Figure 6. Such misalignments were also found in the modeling of the Sternberg[32] and Weber[33] groups. The rmsd from these groups, therefore, were similar to the ones from our modeling. In the modeling of 2GPR, when we used the alignment around the C-terminal as shown for the aligned X-ray sequence, the rmsd were similar to those of the Sali and MSI groups (data not shown). Therefore, the alignment between target and reference proteins should be performed including more structural information. Nevertheless, our rmsd were slightly better than those by the Sternberg and Weber groups. This reason may be that we used LSH to determine the position of atoms. This 2GPR protein is an example for the effect of our LSH weighting. For 1VDR_A and 1U9A_A, moreover, similar calculations as shown in Table 8 were performed. R_{m-n} from "Family" and "The closest" are 0.44 and 0.51 SCR number ratio, respectively, for 1VDR_A; 0.49 and 0.54 for 1U9A_A. Again it was shown that if we find a reference protein closest to the native structure, we can get a more accurate model.

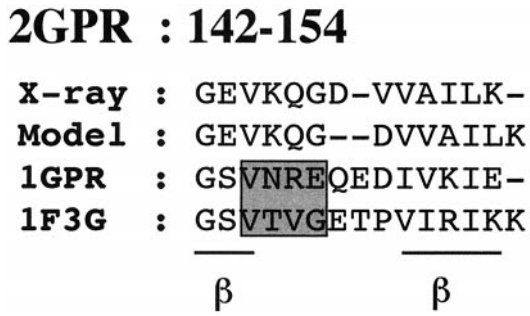


Figure 6. Alignment with 2GPR and homologous proteins. The residues from 142 to 154 for 2GPR and model are shown with the reference proteins of 1GPR and 1F3G. The bottom lines indicate secondary structures for 1GPR and 1F3G. SCR residues are shaded gray.

Again, three protein models (T0001, T0003, and T024) were created to compare with the results of MODELLER, which was able to create more accurate models in the trial of CASP2. The modeling coordinates for MODELLER were obtained from the structures given in CASP2. In this case, our modeling was performed using an alignment identical to that generated by MODELLER, and the reference proteins were the same as those used by MODELLER. Table 12 shows the results from our method and MODELLER. The rmsd for all the atoms in our models were better than those by using MODELLER, although the rmsd for Cα are a little worse. Moreover, our percentage values for correct side-chain torsional angles of $\chi 1$ and ($\chi 1$,$\chi 2$)were a little better than those using MODELLER. The better rmsd for all the atoms in our models may be due to the better prediction of side-chain torsional angles.[17,23]

Although it was shown in Table 12 that in the use of the same given alignment the method presented in this article is

Table 11. Comparison with other methods based on root mean square deviation[1]

1VDR_A (T0001)			2GPR (T0003)			1U9A_A (T0024)		
Cα	All	Group	Cα	All	Group	Cα	All	Group
1.87	2.64	This work	1.34	2.12	Šali	2.34	3.26	This work
1.89	2.70	Šali	1.46	2.25	MSI	2.39	3.24	Moult
1.97	2.66	Fidelis	1.57	2.32	This work	2.60	3.45	Glaxo-Wellcome
2.02	3.11	Glaxo-Wellcome	1.67	2.51	Sternberg	2.69	3.44	Abagyan
2.03	2.77	Abagyan	1.75	2.54	Weber	2.69	3.60	Sutcliffe
2.04	2.91	Schering AG	1.90	2.46	Wolynes	2.71	3.56	Šali
2.23	2.92	Wolynes	2.00	2.69	Abagyan	3.48	4.42	Taylor
2.26	3.13	Sternberg	2.10	2.93	Bruccoleri			
2.93	4.26	Taylor	2.45	3.47	Taylor			
3.59	4.28	MSI	3.06	3.58	Wolynes			
3.74	4.52	Bruccoleri						
3.75	4.48	Weber						

[1]The rmsd values for the other groups are taken from the reported paper by Martin et al.[30]. The rmsd values for our results were calculated as the averaged rmsd for five created models. The included residue in the rmsd calculations is the same as the criterion of the report by Martin et al.

Table 12. Comparison of accuracy between our method and MODELLER[1]

	rmsd				Correct of side-chain torsional angles (SCRs)			
	Cα		All		χ^1		(χ^1,χ^2)	
Code	Our method	MODELLER	Our method	MODELLER	Our method	MODELLER	Our method	MODELLER
T0001	1.83	1.89	2.58	2.70	49.60	48.00	38.40	28.00
T0003	1.37	1.34	2.04	2.12	60.91	52.27	34.55	29.55
T0024	2.76	2.71	3.55	3.56	50.00	60.00	40.80	44.00

[1] Alignment in our method was the same as that in the CASP 2 by Sánchez and Šali.[31]

comparable to the widely accepted MODELLER approach, the differences in our annealing protocols may not be statistically significant. Our rmsd values for the Cα atoms are slightly worse than those of the Sali group. We may need to include more simulated annealing steps. It may be of interest that different annealing protocols give similar main-chain structures to those of the Sali group, where the same alignments and the same reference structures are used. On the other hand, if a skillful changing of the weight of the LSH value is devised, our rmsd for the Cα atoms may be equal to those of the Sali group. Thus, although our method may not be a great or groundbreaking method, our work is fairly solid and as robust as those of the Sali group. We have shown possible results on CASP targets and, therefore, believe that experts in homology modeling will be better able to put the capabilities of our method in perspective.

CONCLUSION

We constructed an automated homology modeling system based on an LSH and including the probability of conserved side-chain torsional angles within homologous proteins. Our method is based on iterative cycles of side-chain and main-chain optimization. The initial backbone model is derived as a weighted average of a set of similar reference proteins. The side-chains are selected in a database search, followed by reoptimization of the backbone via simulated annealing, further side-chain placement, etc. Optimization uses a potential function that resembles standard force fields with addition of disulfide and proline terms. The final procedure is automated. Our method yields results similar to those of MODELLER, which uses a procedure different from our FAMS program. We examined the accuracy of our method on 14 proteins whose structures are known. The accuracy of rmsd values was 2.29 Å and 1.49 Å for all residues and SCR residues, respectively. The percentage of $\chi 1$ angles within 30° of the native structure was 72.6% for SCRs residues. For a few proteins having a large number of homologous proteins, modeled structures obtained were similar to native structure rather than references. This means that our modeling method is very useful for homology modeling. Moreover, it was shown that our method provides a very appropriate model for comparing the variability of models with that of the highly similar proteins. Accordingly, this modeling system is useful as a new tool for structural biology studies.

This modeling method is available on the World Wide Web. The URL is *http://physchem.pharm.kitasato-u.ac.jp/FAMS/fams.html*. The system will create a model from a provided amino acid sequence.

ACKNOWLEDGMENTS

We thank Dr. Teruyo Yoneda and Dr. Mayuko Takeda-Shitaka for providing helpful discussions. This work was supported by a grant-in-aid for special project research from the Ministry of Education, Science, Sports and Culture of Japan.

REFERENCES

1 Covell, D.G., and Jernigan, R.L. Conformations of folded proteins in restricted spaces. *Biochemistry* 1990, **29,** 3287–3294
2 Bowie, J.U., Luthy, R., and Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991, **253,** 164–170.
3 Ota, M., and Nishikawa, K. Assessment of pseudo-energy potentials by the best-five test: A new use of the three-dimensional profiles of proteins. *Prot. Eng.* 1997, **10,** 339–351
4 Takeda-Shitaka, M., and Umeyama, H. Elucidation of the cause for reduced activity of abnormal human plasmin containing Ala^{55}-Thr mutation: Importance of highly conserved Ala^{55} in serine proteases. *FEBS Lett.* 1997, **425,** 448–452
5 Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., and Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature (Lond)* 1987, **323,** 347–352
6 Greer, J. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins* 1990, **7,** 317–334
7 Holm, L., and Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a Cα trace. Application to model building and detection of co-ordinate errors. *J. Mol. Biol.* 1991, **218,** 183–194
8 Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 1992, **226,** 507–533
9 Yoneda, T., Komooka, H., and Umeyama, H. A computer modeling study of the interaction between tissue factor pathway inhibitor and blood coagulation factor Xa. *J. Prot. Chem.* 1997, 16, 597–605
10 Samudrala, R., and Moult, J. A graph-theoretic algorithm for comparative modeling of protein structure. *J. Mol. Biol.* 1998, **279,** 287–302

11 Ponder, J.A., and Richards, F.M. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequence for different structural classes. *J. Mol. Biol.* 1987, **193,** 775–791
12 Lee, C., and Subbiah, S. Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* 1991, **217,** 373–388
13 Holm, L., and Sander, C. Fast and simple Monte Carlo algorithm for side-chain optimization in proteins: Application to model building by homology. *Prot. Struct. Funct. Genet.* 1992, **14,** 213–223
14 Desmet, J., Maeyer, M.D., Hazes, B., and Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature (Lond)* 1992, **356,** 539–542
15 Dunbrack R.L. Jr., and Karplus, M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* 1993, **230,** 543–574
16 Tanimura, R., Kidera, A., and Nakamura, H. Determinants of protein side-chain packing. *Prot. Sci.* 1994, **3,** 2358–2365
17 Ogata, K., and Umeyama, H. Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms. *Prot. Eng.* 1997, **10,** 353–359
18 Weiner, S.J., Kollman, P.A., Nguyen, D.T., and Case D.A. An all-atom force field for simulation for proteins and nucleic acids. *J. Comput. Chem.* 1986, **7,** 230–252
19 Sali, A., and Blundell, T.L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* 1993, **234,** 779–815
20 Abagyan, R.A., Totrov, M.M., and Kuznetsov, D.A. ICM: A new method for structure modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* 1994, **15,** 488–506
21 Evans, J.S., Chan, S.I., and Goddard W.A. Prediction of polyelectroyte polypeptide structures using Monte Carlo conformational search methods with implicit solvation modeling. *Prot. Sci.* 1995, **4,** 2019–2031
22 Li, H., Tejero, R., Monleon, D., Bassolino-Klimas, D., Abate-Shen, C., Bruccoleri, R., and Montelione, G.T. Homology modeling using simulated annealing of restrained molecular dynamics and conformational search calculations with CONGEN: Application in predicting the three-dimensional structure of murine homeodomain Msx-1. *Prot. Sci.* 1997, **6,** 956–970
23 Ogata, K., and Umeyama, H The role played by environmental residues on side-chain torsional angles within homologous families of proteins: A new method of side-chain modeling. *Prot. Struct. Funct. Genet.* 1998, **31,** 355–369
24 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer based archival file for micromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542
25 Russell, R.B., and Barton, G.J. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Prot. Struct. Funct. Genet.* 1992, **14,** 309–323
26 Holm, L., and Sander, C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* 1998, **26,** 316–319
27 Smith, T.F., and Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* 1981, **147,** 195–197
28 Johnson, M.S., and Overington, J.P. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.* 1993, **233,** 716–738
29 Sali, A., and Overington, J.P. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Prot. Sci.* 1994, **3,** 1582–1596
30 Martin, A.C.R., MacArthur, M.W., and Thornton, J.M. Assessment of comparative modeling in CASP2 . *Proteins Suppl.* 1997, **1,** 14–28
31 Sánchez, R., and Sali, A. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Suppl.*1997, **1,** 50–58
32 Bates, P.A., Jackson, R.M., and Sternberg, J.E. Model building by comparison: A combination of expert knowledge and computer automation. *Proteins Suppl.* 1997, **1,** 59–67
33 Harrison, R.W., Reed, C.C., and Weber, I.T. Analysis of comparative modeling predictions for CASP2 targets 1.3.9.17. *Proteins Suppl.* 1997, **1,** 68–73

APPENDIX

Objective Function for the Coordinates of Cα Atom

The objective function in the construction of Cα atoms was defined by following sum:

$$U_{C\alpha} = U_{len} + U_{ang} + U_{pos} + U_{vdw}, \qquad \text{(A1)}$$

where U_{len}, U_{ang}, U_{pos}, and U_{vdw} are explained as follows.

The U_{len} term in Equation A1 is the function of distance between Cα atoms located side by side, and the pair of Cys residues forming a disulfide bond is defined by

$$U_{len} = K_l \sum_i (D_{i,i+1} - 3.8)^2 + K_{SS} \sum_i (D_i^{SS} - 5.4)^2, \qquad \text{(A2)}$$

where $D^{i,i+1}$ is the distance between residues i and $i+1$, and D_i^{SS} is the distance between a pair of Cys residues forming a disulfide bond. K_l and K_{SS} are constants set at 2 and 5, respectively.

U_{ang} is a function for the virtual angle for Cα atoms, and is defined by

$$U_{ang} = K_a \sum_i (\theta_i - \theta_0)^2, \qquad \text{(A3)}$$

where θ_i(rad) is the angle for the residue i, $i+1$, and $i+2$. θ_0 was set at $(100/180)\pi$ (rad), which was determined from the structure of x-ray studies in PDB. K_a is a constant value and was set at 1.

U_{pos} is a function of the positions of Cα atoms, and is defined by

$$U_{pos} = K_{pos} \sum_i \frac{1}{M_i} \| \mathbf{x}_i - \langle w_i \mathbf{x}_i \rangle \|^2, \qquad \text{(A4)}$$

where $|¥|$ means norm and M_i is the average distance between Cα atoms at topologically equivalent positions on the structural alignment. For residue i, if the value of M_i was not taken from reference proteins because none of the amino acid residues match its position, then $M_i$was set at 10. K_{pos} is a

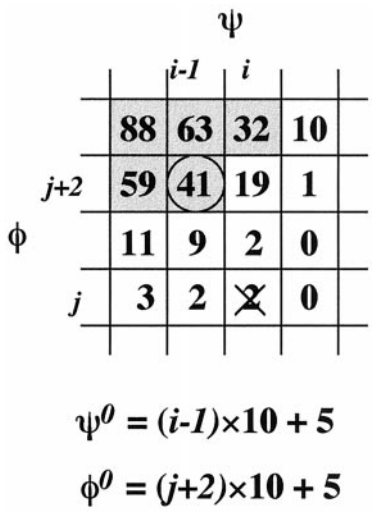


Figure A1. Selection of main-chain torsional angles. The distribution of the number of residues for every 10° was obtained for Gly, Pro, and the other residues. For each element in the distribution of Gly and Pro residues, the number was multiplied by 17 to balance with the others. These distribution figures are the same as those of the Ramachandran plot. Let us consider a residue having the main-chain torsional angles ϕ and ψ which are located at (i,j) on the lattice graph. If the number of residues at (i,j) is greater than 25, the indicating torsion angles ϕ^0 and ψ^0 in Equation A14 are given by angles ϕ and ψ, respectively. If the number of residues at (i,j) is less than 25, the nearest point having the number of residues more than 25 is selected to determine the angles ϕ^0 and ψ^0. As an example, the number of residues at (i,j) is 2, and the nearest point satisfying the above condition is at (i−1 ,j+2). Therefore, the values of ϕ_0 and ψ_0 are given by {(i−1) × 10} + 5° and {(j+2) × 10} + 5°.

constant value set at 10. Here, $\langle x_i \mathbf{w}_i \rangle$ is the weighted average of the coordinates of Cα atoms and is defined by

$$\langle w_i \mathbf{x}_i \rangle = \frac{1}{W} \sum_i w_i^j \mathbf{x}_i^j, \tag{A5}$$

where $\mathbf{x}_i^j$ is the coordinates of Cα atoms on the residue i in reference j. w_i^j is the weight of the coordinate of the i th Cα atom in reference j, and W is sum of w_i^j for all j. This w_i^j is important in determining the frame of the created models. For example, when we obtain the average coordinates of Cα atoms, w_i^j was set at 1 for all i and j. In this work, this value was defined by the value of LSH (see Methods).

U_{vdw} is defined by

$$U_{vdw} = K_{vdw} \sum_{i,j(>1+2)} \left\{ \left(\frac{3.8}{D_{i,j}} \right)^{12} - \left(\frac{3.8}{D_{i,j}} \right)^{6} \right\}, \tag{A6}$$

where K_{vdw} is a constant set at 0.01 ($D_{i,j}$ = 3.2 Å) and at 0.001 ($D_{i,j}$ 3.2 Å), and the cutoff value was 6 Å.

Objective Function for Main-Chain Atoms

The objective function was defined by the following sum:

Table A1. Parameters for bond lengths and angles

Bond lengths		Bond angles	
Bond	Value (Å)	Angle	Value (degree)
C–N	1.32	Cα–C–N	116.6
N–Cα	1.45	Cα–C–O	120.4
Cα–C	1.54	O–C–N	122.9
C–O	1.23	C–N–Cα	121.9
Cα–Cβ	1.53	N–Cα–C	110.1
		N–Cα–Cβ	109.7
		Cβ–Cα–C	111.1
		O–C–OXT	126.0
		S–S–Cβ	103.7

Table A2. Parameters for nonbonded interaction

Atom types	r_i^{*1}	ε_i^{*2}
–NH–, $-NH_2$, $-NH_3$	2.25	0.16
>CH–, $-CH_{2-}$, $-CH_3$	2.30	0.06
>C=	1.85	0.12
=O	1.60	0.20
>N–	1.75	0.16
=CH–	2.35	0.12
–OH	2.15	0.15
–SH	2.50	0.20
–S–	2.00	0.20

[1] $r_{i,j}$* in Equation A10 was defined by $(r_i^* + r_j^*)/2$.
[2] $\varepsilon_{i,j}$* in Equation A10 was defined by $(\varepsilon_i^* + \varepsilon_j^*)/2$.

$$U_{main} = U_{bond} + U_{ang} + U_{nonbond} + U_{SS} + U_{pos} + U_{tor} + U_{chi} + U_{hydr}. \tag{A7}$$

U_{bond} term in Equation A7 is a function of bond length and is defined by

$$U_{bond} = K_b \sum_i (b_i - b_i^0)^2, \tag{A8}$$

where b_i^0 is the standard bond length and is given in Table A1 for different types of bonds. K_b is a constant set at 225.

U_{ang} is a function of bond angle and is defined by

$$U_{ang} = K_a \sum_i (\theta_i - \theta_i^0)^2, \tag{A9}$$

where θ_i is the ith bond angle and is given in Table A1 for different types of angles. K_a is a constant set at 45.

$U_{nonbond}$ is a function of interaction of nonbonded atoms and is defined by

$$U_{nonbond} = K_{non} \sum_{i,j} \varepsilon_{i,j} \left\{ \left(\frac{r_{i,j}^*}{r_{i,j}} \right)^{12} - 2 \left(\frac{r_{i,j}^*}{r_{i,j}} \right)^{6} \right\}, \tag{A10}$$

where ε_{ij} and r_{ij}* are constant values that were defined between atomic types (Table A2). K_{non} is a constant set at 0.25. The cutoff value was 8 Å.

U_{SS} is a function of Cys residues forming an SS-bond and is defined by

$$U_{SS} = \sum_i \{K_{C\alpha}^{SS}(D_i^{C\alpha} - 5.4)^2 + K_{C\beta}^{SS}(D_i^{C\beta} - 3.8)^2\} \tag{A11}$$

where the constant values $K^{SS}{}_{C\alpha}$, and $K^{SS}{}_{C\beta}$ were both set to 7.5.

U_{pos} is a function of atomic positions and is defined by

$$U_{pos} = K_{pos} \sum_i \frac{1}{M_i} \|\mathbf{x}_i - \langle w_i \mathbf{x}_i \rangle\|^2, \tag{A12}$$

where $w_i\mathbf{x}_i^j$ is given by

$$\langle w_i \mathbf{x}_i \rangle = \frac{1}{W} \sum_j w_i^j \mathbf{x}_i^j. \tag{A13}$$

These functions are the same as Equations A4 and A5. K_{pos} is a constant set at 0.3.

U_{tor} is a function of main-chain torsional angles and is defined by

$$U_{tor} = K_t \sum_i \sqrt{(\phi_i - \phi_i^0)^2 + (\Psi_i - \Psi_i^0)^2} + K_\omega \sum_i (\omega_i - \omega_i^0)^2, \tag{A14}$$

where ϕ_i^0 and ψ_i^0 are the nearest torsional angles ϕ_i and ψ_I in the Ramachandran plot (Figure A1). ω_i^0 has a torsional angle value 0 (radian) for cis-Pro or π (radian). K_t and K_ω are constants set at 10 and 50, respectively.

U_{chi} is a function of Cα chirality and is defined by

$$U_{chi} = K_{chi} \sum_i \left(\tau_i + \frac{2}{3} \pi \right)^2 \tag{A15}$$

where τ_i is a virtual torsional angle defined by N-Cα-Cβ-C, and K_{chi} was set to 50.

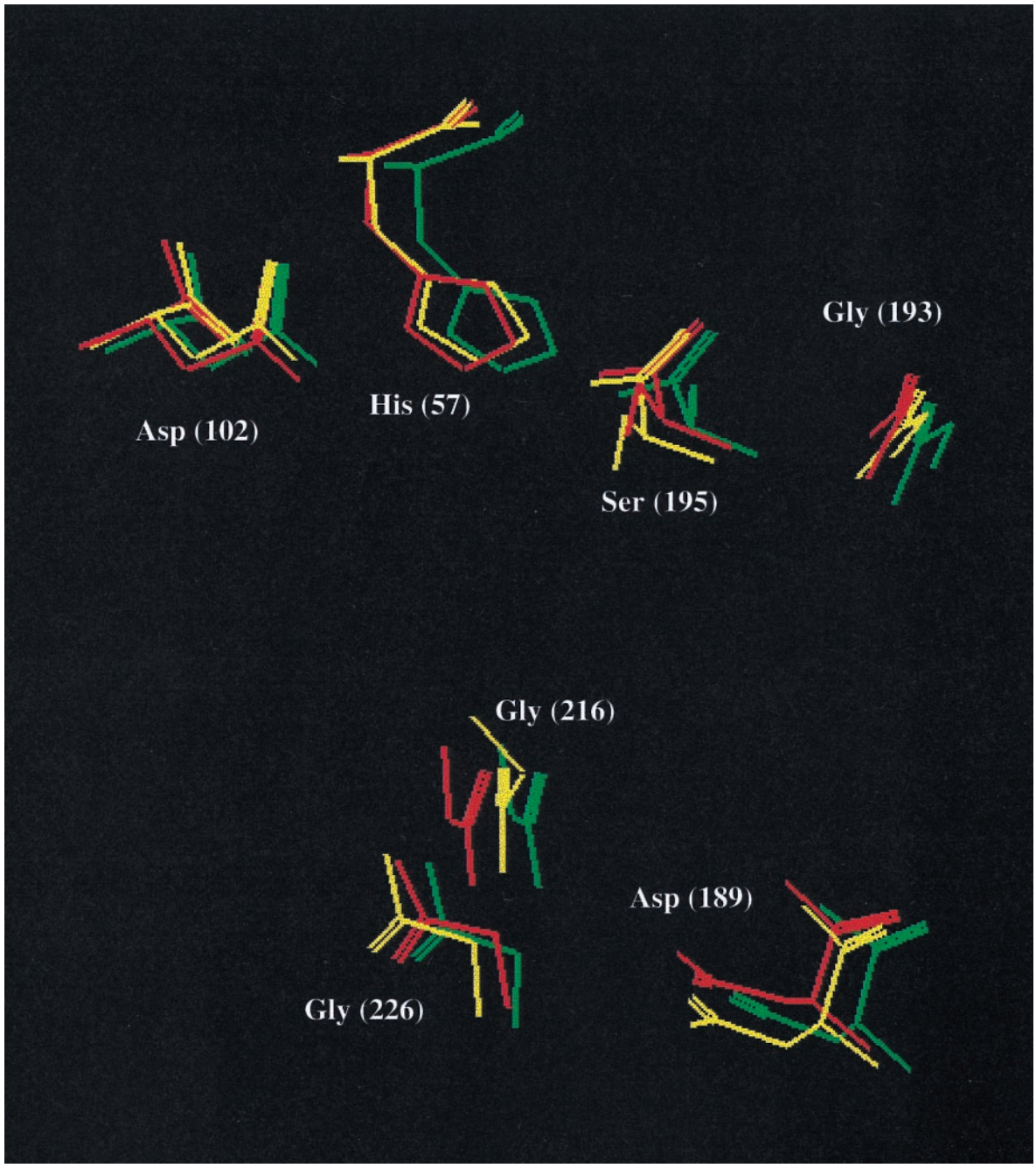
U_{hydr} is a function of the conservation of hydrogen bonds in main-chain atoms within homologous proteins and is defined by

$$U_{hydr} = K_{hydr} \sum_{i,j} (D_{i,j}^{N-O} - 2.9)^2. \tag{A16}$$

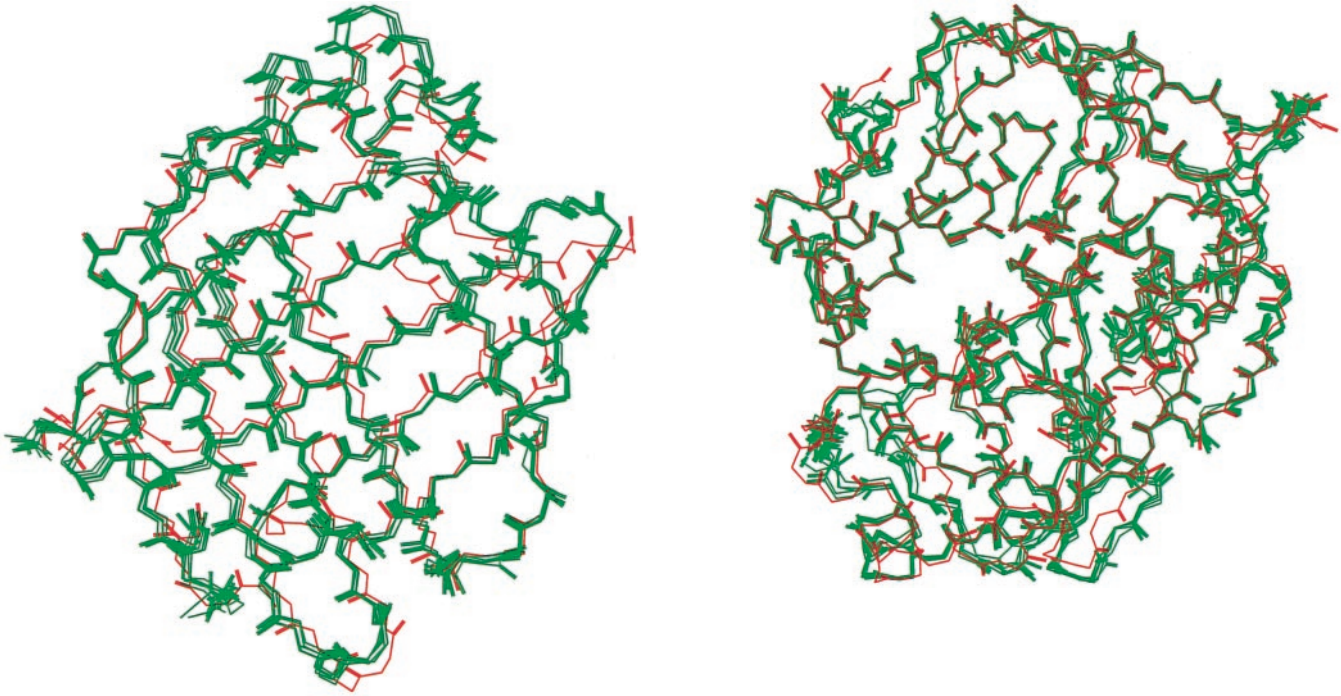
The hydrogen bond was defined by the distance between N and O atoms within a range of 2.9 $\pm$ 0.5 Å. The conservation of hydrogen bonds was regarded as the ratio of the pair of N and O atoms forming hydrogen bonds at topologically equivalent positions greater than 0.75 in the reference proteins. K_{hydr} is a constant set at 0.6.

Koji Ogata and Hideaki Umeyama

An automatic homology modeling method consisting of database searches and simulated annealing



Color Plate 1. Comparison of the conformation of the active site in beta-trypsin (1TLD). The native structure is shown in red. The structures closest to and farthest from native structure in five models are shown in yellow and green, respectively.



Color Plate 2. Variability of constructed models. Color Plate 2a and Color Plate 2b show the variability of models for 1ADL and 1TLD, respectively. In addition, the native structure of these proteins is described for each of these figures. The native and five modeled structures are shown in red and green, respectively.