Chem-Bio Informatics Journal, Vol. 1, No. 4, pp.136-148(2001)

Comparative Modeling of CAFASP2 Competition

Mitsuo Iwadate, Kazuyoshi Ebisawa and Hideaki Umeyama*

Department of Biomolecular Design
School of Pharmaceutical Sciences, Kitasato University
5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, JAPAN

**E-mail: umeyama@pharm.kitasato-u.ac.jp*

(Received October 25, 2001; accepted November 30, 2001; published online December 18, 2001)

Abstract

20 models were constructed for the comparative modeling section of the Critical Assessment of Fully Automated Structure Prediction-2 (CAFASP2) [1] [2]. Sequence identity between each target and the best possible parent(s) ranged between 6% and 52%. Searching for the reference proteins and sequence alignments between the targets and reference proteins was provided by PSI-BLAST[3]. The modeling protocol was executed by the automated computer software FAMS[4], consisting of a database search and simulated annealing. There was no human intervention in checking the process of sequence alignments and building models. Both our team and another team, which used 3D-JIGSAW[5], succeeded in solving eight target models. The accuracies of the modeled backbone and side chains were estimated by using torsion angles. In particular, our modeled side chains were significantly more accurate than the ones modeled by the JIGSAW team. Moreover, our backbone models were also better than those of the JIGSAW team.

Key Words: CAFASP2, Comparative Modeling, Protein Structure Prediction

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

Although protein function is best determined experimentally[6], since similar proteins sequences tend to have similar functions with a few exceptions[7], the prediction of function by matching the sequence of a given protein with those of proteins having known functions may be useful[8, 9, 6]. Recently the reports of this computational assignment of protein function have been

dramatically increased in many genome sequencing projects[10]. To determine the protein function computationally, it is useful to have comparative modeling for three-dimensional protein structure, which includes the functionally important active-site.
CAFASP means "Critical Assessment of Fully Automated Structure Prediction" [11]. CAFASP is a competition to determine the best software. Similarly, CASP[12-14]meaning "Critical Assessment of Techniques for Protein Structure Prediction" is the competition to determine the best researcher. CASP experiments were started in 1994 as CASP1 and were continued biennially to 2000 as CASP4. CAFASP was started in the CASP3 term (1998) as CAFASP1[15], and CAFASP2 was held in 2000[1]. Results from the comparative modeling section of CASP4 suggested that fully automated building procedures were less accurate than procedures with human intervention. Some human intervention effectively worked in CASP4, and the assessments have highlighted the algorithmic development of the improvement of sequence alignments.
Fully automated procedures are essential, however, and, indeed, have been used for large-scale genome modeling[16]. Therefore, CAFASP2 assessments judged not human investigation but only software performance. The use of typical alignment software such as FASTA[17], BLAST, or PSI-BLAST to determine which modeling software demonstrates the best performance is important for large-scale genome modeling.

2. MATERIALS AND METHODS

2.1 Selection of Parents and Sequence Alignments

To make position specific scoring matrix (PSSM) for PSI-BLAST[19], sequences of homologues to the target sequence were extracted from the PIR[18] sequence database, which was extracted on 19 June 2000 and contains 182,161 entries. The sequences were extracted and aligned by using the program BLASTP[19]. Homologues, for which there exist three-dimensional coordinates, were taken from and aligned the Protein Data Bank (PDB) [20, 21] using PSI-BLAST based on PSSM. A number of alignments between the parent(s) and target structure are produced automatically when the sequence identity is >20% and expected value (e-value) is <0.1 between the target and parent. The protein in the PDB database used as the comparative model is called the parent. The one of these alignments that had the lowest e-value was selected as the final alignment. Accordingly each of the twenty targets was constructed from single-parent templates.

2.2 Modeling

The modeling program used, FAMS, is also used by the Kitasato University team, but it is here called CAFASP. This program is available at the web site http://physchem.pharm.kitasato-u.ac.jp/. The input data for the program are only alignment and coordinates of PDB format.

2.3 Evaluation of Models

Model Evaluations were basically constructed in three ways.
(1) Measuring Root Mean Square Distance (RMSD) of Cα atom co-ordinates from the superposition of all relevant homologues. The fitting program is written in C language.
(2) Measuring and comparing side-chain torsion angles (χ^1), then counting the number of

accurate amino acid residues,

(3) Counting accurate backbone torsion angles (ϕ, ψ).

3. Results

3.1 Teams in Comparative Modeling section of CAFASP2

In the CAFASP2 comparative modeling section, four teams participated: FAMS, 3D-JIGSAW, SDSC1, and msi-GeneAtlas (the team IDs used by CAFASP2 are 25, 26, 30, and 31), The web site is http://cafasp.bioinfo.pl/. According to this home page, the SDSC1 team (ID 30) produced only backbone co-ordinates for the targets, and the msi-GeneAtlas team (ID 31) produced their targets after the deadline, which is 48 hours after the target sequences were submitted. Assessor, Dr. Dunbrack in FCCC evaluated only two teams, FAMS and JIGSAW. His web site is http://www.fccc.edu/research/labs/dunbrack/cafasp-results.html[2]. In this paper we describe the two teams. It is unfortunate that the popular modeling software MODELLER[16], developed by Andres Sali of Rockefeller University, was used by many participants of the CASP4 comparative modeling section who participated in this competition. If their team participated, this competition would create more worth comparison.

3.2 Selecting Parents and Alignments

In CASP4, 44 targets, target codes T0086 ~ T0128 were submitted. PSI-BLAST searches with a 0.1 e-value found the parent PDB structures of about 20 targets. It means that if a set of many amino acid sequences is provided, we are able to get alignments for roughly half of them. In the assessor's evaluation, Dr. Dunbrack indicated that we mistakenly selected the parental assignments for T0091, T0094, T0095, T0096, T0110, T0116, T0120, T0121, and T0124 (as shown by the red numbers in Table 1). Table 1 shows the summarized RMSD fitting of Cα atoms between targets and parents.

Compared with JIGSAW, the FAMS program produced many answers (see Table1, FAMS 15, JIGSAW 13 and SDSC1 14 models), and the results were fairly good, especially regarding T0092, T0094, and T0110, for which only FAMS produced the models of 4 teams. The RMSD values are quite similar, because both teams used PSI-BLAST in the process of selecting the parents for the PDB structure. For target T0112, for example, FAMS selected 1DDA(B) (PDB code) as the parent, whereas, JIGSAW selected

Table. 1 Fitted RMSD value [Å] of Cα atoms.

	FAMS	JIGSAW	SDSC1
T0089	17.1773	21.8053	25.9982
T0092	4.22258		
T0094	14.2657		
T0099	4.90849	4.82491	4.76515
T0101			1.60513
T0103		20.3975	13.1962
T0110	8.70028		
T0111	1.73862	2.3315	2.73261
T0112	4.11556	4.11373	5.64019
T0114		13.3494	
T0115			34.9115
T0116	43.6884		27.372
T0120	23.6583	71.2137	37.306
T0121	3.30324	3.23358	3.45288
T0122	2.26866	2.52171	2.16785
T0123	3.59055	4.17979	3.59161
T0124	77.3679	104.299	
T0125		4.70701	
T0127	14.5661		30.0161
T0128	1.33571	1.20781	0.99166

1TEH(B) as the parent. Both proteins have quite similar amino acid sequences (identity is 63%) and structures (RMSD is 1.37 Å). Then the model structures of both teams have quite similar RMSD values, 4.116 Å and 4.114 Å, respectively, between the models and the parent structures (see Table 1). For most comparative modeling targets, similar parent PDB structures produce quite similar modeling structures, because they depend on the same alignment software of PSI-BLAST. There are no significant differences in the Cα superimposed comparisons. Biological discussions should therefore be performed for non-sequenced alignments or Cα fitting differences, as, for example, the comparison of side-chain torsion angles. By focusing on only the fitting at Cα atoms, some excellent models have lowest RMSD values were produced by the SDSC1 team. It is unfortunate, however, that the results are not useful for structure-based drug design (SBDD) or induced fitting, since only backbone residues were produced.

Figure 1. Superimposed structure of CASP4 target ID T0112 (DHSO, Ketose Reductase / Sorbitol Dehydrogenase) with the X-ray structure. White is X-ray diffraction structure. FAMS is shown in Brown and RMSD is 4.116Å; JIGSAW is shown in Green and RMSD is 4.114Å.

3.3 Side-Chain Accuracy

Side-Chains were assessed by comparing the side-chain dihedral angles between X-ray structures, except for T0099, for which the NMR structure and models were used. Table 2 shows the number and rate of correct side-chain dihedral angles of each model produced by FAMS and JIGSAW.

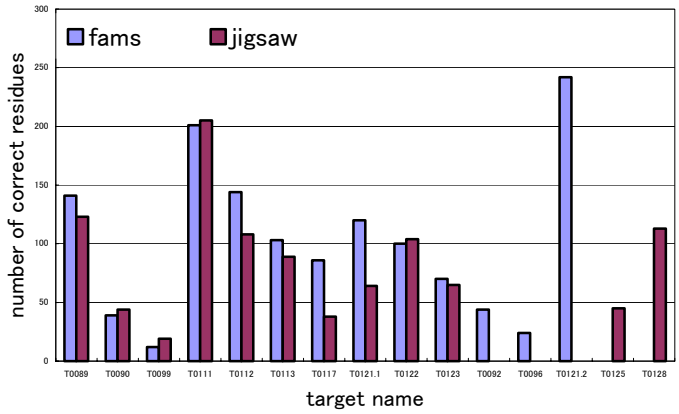

Figure 2. Number of correct side-chains.

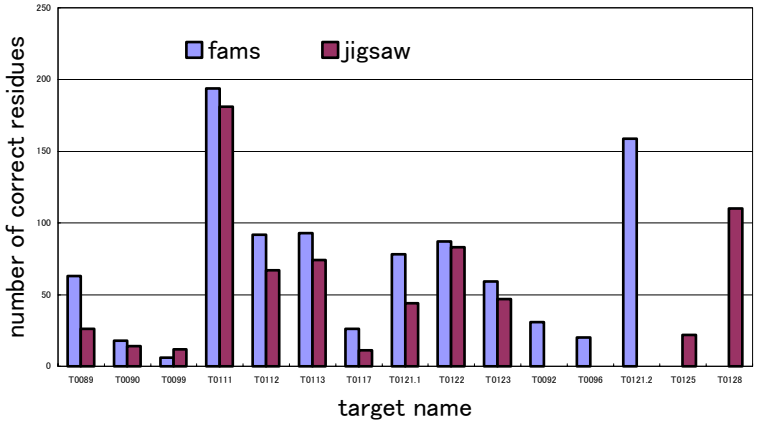

Figure 3. Number of correct side-chains of correct alignment regions.

Figure 2 and Figure 3, respectively, show the numbers of correct(*ncor*) side-chains and *ncor* of correct alignment regions defined as the distance between Cα atoms < 3.5 Å by using program MaxSub, which assesses the quality of a predicted protein structure (http://www.cs.bgu.ac.il/~dfischer/MaxSub/)[22]. Total numbers of ncor are 1206 and 1017 and it appears that FAMS does a significantly better job with side-chains than does JIGSAW. It also appears that FAMS predicts the backbone within 3.5 Å for a higher proportion of the target backbone Cα atoms. The number of side-chains (*nsc*) of FAMS is usually greater than *nsc* of JIGSAW for correctly aligned residues. This means that FAMS usually produced more reasonable model structures from similar parental structures and alignments than did JIGSAW. Moreover *ncor* of FAMS is usually greater than *ncor* of JIGSAW for correctly aligned residues (see Figure 3), except for the target T0099 (NMR structure). This is a very important point to determine the biological meaning from the modeling results, since most functions of enzymes or receptors are given from the side-chain conformations of the amino acids.

Table 2. Comparison of FAMS and JIGSAW models for the accuracy of side-chain modeling. *ncor* is the number of residues with $\Delta\chi1 < 40°$where $\Delta\chi1$ is the dihedral difference in $\chi1$ between the model and target of the experimental structure. Correctly aligned residues were determined by MaxSub alignments within the distance between $C\alpha < 3.5$ Å[22].

	target	ncor	nsc	ncor/nsc	ncor	nsc	ncor/nsc
fams	T0089	141	318	44.340	63	125	50.400
Jigsaw	T0089	123	297	41.414	26	53	49.057
fams	T0090	39	91	42.857	18	35	51.429
Jigsaw	T0090	44	136	32.353	14	44	31.818
fams	T0092	44	107	41.121	31	69	44.928
fams	T0096	24	54	44.444	20	39	51.282
fams	T0099	12	40	30.000	6	25	24.000
Jigsaw	T0099	19	41	46.341	12	28	42.857
fams	T0111	201	321	62.617	194	305	63.607
Jigsaw	T0111	205	315	65.079	181	273	66.300
fams	T0112	144	282	51.064	92	168	54.762
Jigsaw	T0112	108	238	45.378	67	139	48.201
fams	T0113	103	192	53.646	93	164	56.707
Jigsaw	T0113	89	186	47.849	74	147	50.340
fams	T0117	86	180	47.778	26	51	50.980
Jigsaw	T0117	38	110	34.545	11	27	40.741
fams	T0121.1	120	202	59.406	78	128	60.938
fams	T0121.2	242	404	59.901	159	259	61.390
Jigsaw	T0121.1	64	156	41.026	44	88	50.000
fams	T0122	100	197	50.761	87	160	54.375
Jigsaw	T0122	104	199	52.261	83	158	52.532
fams	T0123	70	139	50.360	59	101	58.416
Jigsaw	T0123	65	142	45.775	47	94	50.000
Jigsaw	T0125	45	119	37.815	22	49	44.898
Jigsaw	T0128	113	168	67.262	110	157	70.064

3.4 Backbone Accuracy

Backbones were assessed by comparing the backbone dihedral angles between X-ray structures, except for target T0099 (NMR) and models. Table 3 shows the number and rate of the correct set of backbone dihedral angles, $\mathrm{sqrt}(\Delta\phi^2 + \Delta\psi^2) < 60°$ where $\Delta\phi$ is the dihedral difference in

ϕ between the model and target of the experimental structure, and Δψ is the dihedral difference in ψ between the model and target. For the FAMS and JIGSAW models, Figures 4 and 5 respectively show the numbers of the correct(*ncor*) backbone and *ncor* of correct alignment regions, defined as the distance between Cα atoms < 3.5 Å by using the program MaxSub. Here again, FAMS came out ahead, giving that it predicted for a larger fraction of each structure. The "*ncor*" of aligned residues is higher for FAMS than JIGSAW in almost all cases, although the percentages are sometimes lower as shown in Table 3. Thus, it appears that FAMS also does a significantly better job with the backbone than does JIGSAW.

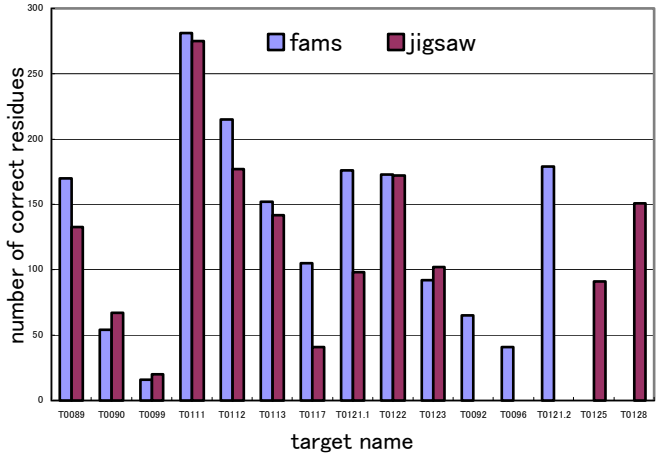

Figure 4. Number of correct backbone residues for each of modeled proteins.

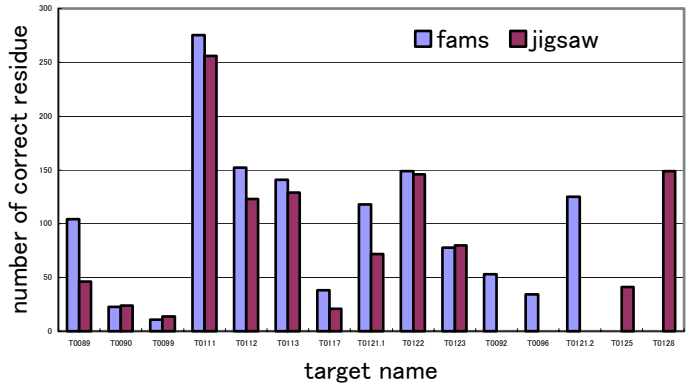

Figure 5. Number of correct backbone residues of correct alignment regions for each of modeled proteins.

Table 3. Comparison of FAMS and JIGSAW models for target of the experimental structure. *ncor* is the number of residues with sqrt($\Delta\phi^2 + \Delta\psi^2$) <60° where $\Delta\phi$ is the dihedral difference in ϕ between the model and target, and $\Delta\psi$ is the dihedral difference in ψ between the model and target. Correctly aligned residues were determined by MaxSub alignments within the distance between Cα< 3.5 Å [22].

		all residues in model			correctly aligned residues (maxsub 3.5)		
	target *)	ncor	nres	rate %	ncor	nres	rate %
fams	T0089	170	318	53.459	104	125	83.200
jigsaw	T0089	133	297	44.781	46	53	86.792
fams	T0090	54	91	59.341	23	35	65.714
jigsaw	T0090	67	136	49.265	24	44	54.545
fams	T0092	65	107	60.748	53	69	76.812
fams	T0096	41	54	75.926	34	39	87.179
fams	T0099	16	40	40.000	11	25	44.000
jigsaw	T0099	20	41	48.780	14	28	50.000
fams	T0111	281	321	87.539	275	305	90.164
jigsaw	T0111	275	315	87.302	256	273	93.773
fams	T0112	215	282	76.241	152	168	90.476
jigsaw	T0112	177	238	74.370	123	139	88.489
fams	T0113	152	192	79.167	141	164	85.976
jigsaw	T0113	142	186	76.344	129	147	87.755
fams	T0117	105	180	58.333	38	51	74.510
jigsaw	T0117	41	110	37.273	21	27	77.778
fams	T0121.1	176	202	87.129	118	128	92.188
fams	T0121.2	179	202	88.614	125	131	95.420
jigsaw	T0121.1	98	156	62.821	72	88	81.818
fams	T0122	173	197	87.817	149	160	93.125
jigsaw	T0122	172	199	86.432	146	158	92.405
fams	T0123	92	139	66.187	78	101	77.228
jigsaw	T0123	102	142	71.831	80	94	85.106
jigsaw	T0125	91	119	76.471	41	49	83.673
jigsaw	T0128	151	168	89.881	149	157	94.904

*) Target T0121 was separately evaluated as N- and C- terminal domains, T0121.1 and T0121.2, respectively.

4. Discussion

In the "Results" section, we described selecting parent structures, getting alignments, and estimating side-chain and backbone accuracy. FAMS produced more accurate and precise structures, and its alignments in CAFASP2 are fully dependent upon PSI-BLAST software. It means that PSI-BLAST was working well in CAFASP2 in the 2000 fully automatic protein modeling contest. But focusing on the "Fold Recognition" competition (accurate comparison of alignments), there are many excellent teams, such as Baker, Murzin, and Blundell. In CASP4 regarding the target of low homology with a known structure, many participants searched for an algorism that would produce results closer to the experimental structures at the alignment level.
In 2 or 3 years, the CAFASP-type competition in the Comparative Modeling section will become the most important event in bio-informatics, since comparative modeling may be expected to become economically valuable with the determination of many genome sequences.

Acknowledgement

We thank Dr. Roland L. Dunbrack for his fair evaluation in CAFASP2.

References

[1] J. M. Bujnicki, A. Elofsson, D. Fischer, and L. Rychlewski, *Bioinformatics*, **17**, 750-751 (2001).
[2] D. Fischer, *Proteins*, **Suppl**, in press (2001).
[3] A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul, *Nucleic Acids Res.*, **29**, 2994-3005 (2001).
[4] K. Ogata and H. Umeyama, *J. Mol. Graph. Model.*, **18**, 258-272, 305-306 (2000).
[5] P. A. Bates and M. J. Sternberg, *Proteins*, **37**, 47-54 (1999).
[6] S. G. Oliver, *Nature*, **379**, 597-600 (1996).
[7] C. A. Orengo, D. T. Jones, and J. M. Thornton, *Nature*, **372**, 631-634 (1994).
[8] B. Dujon, *Trends Genet.*, **12**, 263-270 (1996).
[9] E. V. Koonin and A. R. Mushegian, *Curr. Opin. Genet. Dev.*, **6**, 757-762 (1996).
[10] G. L. Miklos and G. M. Rubin, *Cell*, **86**, 521-529 (1996).
[11] D. Fischer, C. Barret, K. Bryson, A. Elofsson, A. Godzik, D. Jones, K. J. Karplus, L. A. Kelley, R. M. Maccallum, K. Pawowski, B. Rost, L. Rychlewski, and M. Sternberg, *Proteins*, **Suppl**, 209-217 (1999).
[12] J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen, *Proteins*, **Suppl**, 2-6 (1997).
[13] J. Moult, T. Hubbard, K. Fidelis, and J. T. Pedersen, *Proteins*, **Suppl**, 2-6 (1999).
[14] E. E. Lattman, *Proteins*, **44**, 399 (2001).
[15] J. De Vlieg, R. M. Scheek, W. F. Van Gunsteren, H. J. Berendsen, R. Kaptein, and J. Thomason, *Proteins*, **3**, 209-218 (1988).
[16] R. Sanchez and A. Sali, *Proteins*, **Suppl**, 50-58 (1997).
[17] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 2444-2448 (1988).

[18] W. C. Barker, J. S. Garavelli, Z. Hou, H. Huang, R. S. Ledley, P. B. Mcgarvey, H. W. Mewes, B. C. Orcutt, F. Pfeiffer, A. Tsugita, C. R. Vinayaka, C. Xiao, L. S. Yeh, and C. Wu, *Nucleic Acids Res.*, **29**, 29-32 (2001).
[19] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.*, **215**, 403-410 (1990).
[20] H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook, *Nat. Struct. Biol.*, **7 Suppl**, 957-959 (2000).
[21] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.*, **28**, 235-242 (2000).
[22] N. Siew, A. Elofsson, L. Rychlewski, and D. Fischer, *Bioinformatics*, **16**, 776-785 (2000).

Appendix

Table S1. Comparison of FAMS and JIGSAW By residue type, *ncor* is the number of side-chains with $\Delta\chi^1<40°$ from the experimental structure for all residues, except Pro; $\Delta\chi^1<20°$ for Pro. *nsc* is the number of side-chains. Correctly aligned residues were determined by Maxsub alignments, distance between Cα< 3.5 Å [22].

		all residues in model			correctly aligned residues (maxsub 3.5)		
	residue type	Ncor	Nsc	rate %	ncor	Nsc	rate %
fams	ALL	1206	2325	51.871	848	1501	56.496
jigsaw	ALL	1017	2107	48.268	691	1257	54.972
fams	ARG	87	155	56.129	65	100	65.000
jigsaw	ARG	56	162	34.568	35	73	47.945
fams	ASN	57	107	53.271	39	66	59.091
jigsaw	ASN	41	88	46.591	30	54	55.556
fams	ASP	78	144	54.167	52	80	65.000
jigsaw	ASP	51	121	42.149	26	59	44.068
fams	CYS	16	32	50.000	6	17	35.294
jigsaw	CYS	6	24	25.000	3	12	25.000
fams	GLN	41	87	47.126	29	52	55.769
jigsaw	GLN	40	79	50.633	25	44	56.818
fams	GLU	99	221	44.796	56	131	42.748
jigsaw	GLU	89	193	46.114	52	112	46.429
fams	HIS	22	44	50.000	14	25	56.000
jigsaw	HIS	26	46	56.522	20	28	71.429
fams	ILE	112	194	57.732	86	138	62.319
jigsaw	ILE	93	174	53.448	70	119	58.824
fams	LEU	146	266	54.887	111	191	58.115
jigsaw	LEU	135	234	57.692	100	156	64.103
fams	LYS	95	184	51.630	64	121	52.893
jigsaw	LYS	96	207	46.377	61	115	53.043
fams	MET	27	62	43.548	20	47	42.553
jigsaw	MET	32	62	51.613	22	40	55.000
fams	PHE	55	90	61.111	41	53	77.358
jigsaw	PHE	55	93	59.140	38	54	70.370
fams	PRO	40	118	33.898	21	67	31.343
jigsaw	PRO	43	89	48.315	24	51	47.059
fams	SER	52	143	36.364	34	85	40.000
jigsaw	SER	48	122	39.344	27	66	40.909

fams	THR	76	144	52.778	58	96	60.417
jigsaw	THR	47	113	41.593	37	73	50.685
fams	TRP	9	15	60.000	5	8	62.500
jigsaw	TRP	12	19	63.158	9	13	69.231
fams	TYR	51	79	64.557	32	48	66.667
jigsaw	TYR	47	76	61.842	34	48	70.833
fams	VAL	143	240	59.583	115	176	65.341
jigsaw	VAL	100	205	48.780	78	140	55.714

Chem-Bio Informatics Journal, Vol. 1, No. 4, pp.136-148(2001)

CAFASP2 における Comparative Modeling

岩舘満雄　海老沢計慶　梅山秀明*

北里大学薬学部生物分子設計学教室

**E-mail: umeyamah@pharm.kitasato-u.ac.jp*

要旨

完全自動タンパク質構造予測を行うプログラムの性能を完比較・検討するCAFASP2(Critical Assessment of Fully Automated Structure Prediction-2)において、FAMSをエントリし、20個のモデル座標を提出した。参照タンパク質の選定及び配列アライメントは、PSI-BLASTを用いた。モデリングはFAMSを用いた。FAMSはデータベース探索とシミュレーティッドアニーリングを組み合わせたアルゴリズムを持つ。配列アライメントからモデリングを通じて全ての過程は完全に自動的に行われた。北里大学チーム「FAMS」と英国「JIGSAW」の両者は8個の目的タンパク質について構造を提出した。モデルの主鎖及び側鎖の正確さはねじれ角によって推定された。特にFAMSのモデルはJIGSAWに比べて側鎖の正確度が高いことが有意に示された。主鎖においても同様の傾向があった。

キーワード：CAFASP2, 比較モデリング, タンパク質構造予測

領域区分：分子生物学における情報計算技術