第３６回構造活性相関シンポジウム
講演要旨集

会 期　2008年11月2日(日)・3日(月)

会 場　神戸国際会議場

主 催　日本薬学会構造活性相関部会
共 催　日本化学会
　　日本農芸化学会
　　日本分析化学会
　　日本農薬学会
協 賛　日本薬学会医薬化学部会
　　日本薬学会薬学研究ビジョン部会

第 36 回構造活性相関シンポジウム　スケジュール

11 月 2 日	11:00-12:05	一般講演(K201-203) 座長：K201　竹田-志鷹真由子 K202-K203　合田 浩明
	12:05-13:30	昼休み（構造活性相関部会常任幹事会）
	13:30-15:30	ポスターセッション (KP201-KP233)
	15:30-15:45	休憩
	15:45-17:15	一般講演（K204-K207） 座長：K204-K205　川瀬 雅也 K206-K207　山下 富義
	18:00-20:00	懇親会・ポスター賞表彰式
11 月 3 日	9:50-11:30	一般講演（K301-304） 座長：K301-K302　仲西 功 K303-K304　飯島 洋
	11:30-12:45	昼休み（構造活性相関部会幹事会）
	12:45-13:45	特別講演（KS） 座長：高木 達也
	13:50-14:40	一般講演（K305-K306） 座長：加藤 博明
	14:40-15:00	休憩
	15:00-15:30	日中合同シンポジウム Opening Session
	15:30-16:10	Plenary Lecture（Joint Session）
	16:10-16:20	休憩
	16:20-17:30	Invited Lecture　(Joint Session; V01,V02)
	17:30-17:45	休憩
	17:45-19:30	Poster Session (Joint Session)

第 36 回構造活性相関シンポジウム

第 1 日目（11 月 2 日）

10:50-11:00　開会　（阪大院・薬、阪大・微研）高木達也

一般講演（会場：３F・国際会議室）

（座長）竹田—志鷹 真由子

11:00-11:25　**K201***　Scaffold Hopping Using Inductive Logic Programming
（アステラス製薬）○角山 和久，（Equinox Pharma Ltd）Ata Amini，（Imperial College London）Michael J. E. Sternberg，Stephen H. Muggleton.....5

（座長）合田 浩明

11:25-11:50　**K202***　Development of bioinformatics based ligand-docking and in-silico screening
（理化研）○高谷大輔，（北里大・薬）寺師玄記，加納和彦，（理化研、北里大・薬）竹田—志鷹 真由子，梅山秀明.....7

11:50-12:05　**K203**　Point-fluorination effect on molecular conformation in the aggregation pheromone, (3S,4S)-4-methyl-3-heptanol
（鳥取大院・工）○早瀬修一，田中絢子，川面 基，伊藤敏幸.....9

ポスターセッション（会場・３F レセプションホール）

13：30～15：30　**KP201-KP233**

13:30-14:30　奇数番号発表

14:30-15:30　偶数番号発表

一般講演（会場：３F・国際会議室）

（座長）川瀬 雅也

15:45-16:10　**K204***　Prediction of oral bioavailability of druglike chemical compounds
（阪大院・薬）○石塚賀彦，（田辺三菱製薬）大軽貴典，（阪大院薬）日高伸之介，山崎広之，高原淳一，岡本晃典，（阪大院薬、RCC-ERI）川下理日人，（阪大微研、RCC-ERI）安永照雄，（阪大院薬、阪大微研、RCC-ERI）高木達也.....11

第2日目（11月3日）

一般講演（会場：3F・国際会議室）

（座長）仲西 功

（座長）飯島 洋

特別講演（会場：3F・国際会議室）

（座長）高木 達也

一般講演（会場：3F・国際会議室）

（座長）加藤 博明

共同セッション（36th SAR and 8th Japan-China Joint Symposium Joint Session；会場・３F 国際会議室、レセプションホール）

15:30-16:10 **Plenary Lecture**

The assembly of pharmacophore and scaffold: some cases of their application

Zongru Guo

Institute of Materia Medica, Chinese Academy of Medical Sciences

Invited Lecture

16:20-16:55 **V01J** Reprofiling the Hansch-Fujita Type of Classical QSAR Using Modern Molecular Calculations

Tatsusada Yoshida, Kazuya Nagaoka, Toshio Fujita, and Hiroshi Chuman

Graduate School of Pharmaceutical Sciences, The University of Tokushima

16:55-17:30 **V02J** Prediction of ADME/Tox Properties with Machine Learning

Teruki Honma

Yokohama Institute, RIKEN

17:45-19:30 Poster Session (Reception Hall)

KP201 Development of application tools for protein-ligand interaction analyses in Protein Data Bank

（長浜バイオ大・生命情報）○齊藤美保子，白井 剛.....31

KP202 Development of a Software Tool for PharmacologicalActivity Prediction by using Collective SVM Models.

（豊橋技科大）○藤島悟志，（科研製薬）河合健太郎，（豊橋技科大）高橋由雅..33

KP203 Theoretical Study of Partial Agonist Ligands of Vitamin D Receptor (VDR): 19-nor Vitamin D3 Analogues

（立教大・理）○元吉 沙也加（立教大・極限生命情報研）山岸賢司，（日本大・医），槇島 誠，（日本大・医，立教大・極限生命情報研）山田幸子，（立教大・理）常盤広明.....35

P01 Application of the TFS-Based Support Vector Machines to in-silico Screening Strategy

Kentaro Kawai [1] *and Yoshimasa Takahashi*[2]

[1]Kaken Pharmaceutical Co., Ltd., [2]Toyohashi University of Technology

P02 Focused Library Design Based on Complex Crystal Structures and Application in FabZ Inhibitors Discovery

Hong Liu, Linyan He, Liang Zhang, Weiliang Zhu, Xu Shen, Hauliang Jiang

Drug Discovery and Design Center, Shanghai Institute of Materia Medica

P03 Three-Dimensional Structural Data Mining Based on Geometrical Fragment Spectra

Hiroaki Kato, Shigeru Yoshida, Yoshimasa Takahashi

Department of Knowledge-based Information Engineering, Toyohashi University of Technology

P04 Synthesis and Evaluation of N_4-Arylsulfonylquinoxalinones as HIV-1 Reverse Transcriptase Inhibitors

Bailing Xu[1]*, *Yan Sun*[1], *Tao Yu*[1], *Ying Guo*[1], *Yingli Cao*[1], *Decai Fu*[2]

[1] Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College; [2]Hebei University of Science and Technology

P05 Theoretical Consideration on Enzymatic QSAR; Energy Decomposition Analysis of the Hammett σ Constant and QSAR of MMP-9 Inhibitors Using Ab Iinitio MO Calculations

Kazuya Nagaoka, Miku Oonishi, Tatsusada Yoshida, Hiroshi Chuman

Graduate School of Pharmaceutical Sciences, The University of Tokushima

P06 Design, synthesis of Phosphodiesterase IV Inhibitors based on Pharmacophore and Scaffold Hopping

Yanshen Guo, Xueshi Mao, Fengming Chu, Zongru Guo

Institute of Materia Medica Chinese Academy of Medical Sciences

P07 Docking-pose prediction by receptor-based tailor-made scoring function

*Shinya Nakamura** [1], *Kazuo Kitaura* [2] *and Isao Nakanishi* [1]

[1] Department of Pharmaceutical Sciences, School of Pharmacy, Kinki University

[2] Graduate School of Pharmaceutical Sciences, Kyoto University

K202

Development of bioinformatics based ligand-docking and in-silico screening.

Daisuke Takaya[1,2], Genki Terashi[2], Kazuhiko Kanou[2], Mayuko Takeda-Shitaka[1,2], Hideaki Umeyama[1,2]

1) RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

2) Biomolecular Design, School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

takayad@gsc.riken.jp

1. Introduction

Many protein targets implicated in diseases have been discovered through biochemical experiments. As a result, the competition between pharmaceutical companies and other organizations to discover drug-like compounds which inhibit or activate those protein targets is fierce. Many pharmaceutical companies are using in-silico screening programs such as DOCK, AutoDock and GOLD. These programs use classical mechanical potentials. Here, we report a novel method, ChooseLD[1] (CHOOse biological information Semi-Empirically on the Ligand Docking), which uses simulated annealing (SA) based on bioinformatics for protein-ligand flexible docking.

2. Method

First, we placed a target protein of interest and aim to select one or more ligands with low molecular weight. The query amino acid sequence of the target protein is aligned in the filter of CE[2] Z-Score (more than 3.7) of Protein Data Bank (PDB) database, which includes ligand molecules, termed the family ligand set. Alignment methods such as PSI-BLAST are then applied. A fingerprint (FP) of a chemical descriptor is determined on the basis of bioinformatics. The FP includes the information about the atom-type such as used in SYBYL and bond-type; single, double, triple or aromatic. The FPs of the same type including different Cartesian coordinates of the docked ligand composes another Cartesian coordinates vector of the redundant FP set. FPAScore (FP Alignment Score) to calculate alignment between target ligand and library ligands is defined by equation (1). The FPAScore is used to determine the most stable docking conformation of the ligand. The value of

Fig 1. Defined Equation

$$\begin{aligned} FPAScore &= F(aligned\_fp, fp\_rmsd, molecule) \\ &= BaseScore(aligned\_fp, fp\_rmsd) \\ &\times fp\_volume(molecule) \\ &\times fp\_contact\_surface(molecule) \end{aligned} \quad (1)$$

$$BaseScore(aligned\_fp, fp\_rmsd) = \frac{RawScore(aligned\_fp)}{1+\ln(fp\_rmsd^{k1}+1)} \quad (2)$$

$$fp\_volume(molecule) = \ln\frac{1.0+nafp^{k2}}{1.0+nap^{k3}} \quad (3)$$

$$fp\_contact\_surface(molecule) = \frac{\sum_{i=1}^{n} density\_of\_atom(atom(i))}{total\_density\_of\_atom(molecule)} \quad (4)$$

"*RawScore(aliged_fp)*"is maximized in the SA calculation process accompanied by variation of the value of "$ln(fp\_rmsd^{k1}+1.0)$ ". The *fp_rmsd* is the root mean square deviation (rmsd) value that is the result of the least-square fitting using the FP alignment. The *nafp* is the number of docking ligand atoms covering the FPs region. The *nap* is the number of docking ligand atoms covering the target protein region. In the calculation of the benchmark sets, both *k2* and *k3* are equal to one. "*atom(i)*" is the sequential atom number of a ligand. "*density_of_atom(atom(i))*" shows the ligand atom number in the region of significance in the area of direct interaction. "*total_density_of_atom(molecule)*" is the denominator for the ligand and is used to standardize the numerator

3. Results and Discussion

In order to test the protein-ligand docking method based upon bioinformatics, two benchmark tests were performed by using two database sets composed of either 85[3] or 133[4] PDB structures respectively. After the ligand molecule was docked to the target protein in the two benchmark sets, rmsd value of Cartesian coordinates between the docked ligand and the X-ray analyzed ligand was calculated. If the docking state is within 2.0 Å, docking is considered successful.

Tc range	Success rate %
0.08~0.16	12.6
0.08~0.24	20.8
0.08~0.36	29.2
0.08~0.56	40.1
0.08~0.76	44.8
0.08~0.96	46.4

	Success rate %		
Docking soft	Corina	MINI[4]	average
DOCK	21.6	20.6	21.1
AutoDock	26.2	27.0	26.6
GOLD ChemScoreSTD	45.5	45.3	45.4
GOLD ChemScoreLib	44.1	44.9	44.5
GOLD GOLDScoreSTD	45.2	46.7	46.0

Table 1: Success rates of ChooseLD, DOCK, AutoDock and GOLD.

Using the 85 benchmark set, the constant *k1* value in equation (2) was optimized to be 4.0 by varying it from 1.0 to 6.0. In the docking calculations carried out for the 133 benchmark set using the *k1* value of 4.0, the success rates of Tanimoto coefficient (Tc) ranges 0.08~0.16, 0.08~0.24, 0.08~0.36, 0.08~0.56, 0.08~0.76 and 0.08~0.96 were 12.6, 20.8, 29.2, 40.1, 44.8 and 46.4%, respectively. (Table 1)

Compared with the success rates using programs DOCK, AutoDock and GOLD, the success rates of ChooseLD are almost equivalent to those of the docking program. Our program is mainly based upon a bioinformatics basis set called FP. Thus, direct comparison of the success rates of those may be meaningless. Nevertheless, our program is comparably powerful when the researchers want to carry out protein-ligand docking and in-silico screening of a target protein. In the future, the number of PDB codes with the interacting ligand will be increased. It is anticipated that such an increase will improve the success rate of our ChooseLD method.

References

1. Takaya, D., et al *Chem Pharm Bull.* 56(5) 742-744. (2008)
2. Shindyalov, I. N., et al *Protein Engineering* 11, 739-747 (1998)
3. Hartshorn, M. J., et al *J. Med. Chem.* 50, 726-741 (2007).
4. Onodera, K., et al. *Chem. Inf. Model.* 47, 1609-1618 (2007)

KP229

Fams-ace2: Structure evaluation program using the combination of local consensus method and circle quality assessment method

Genki Terashi[1], Hiroko Sakai[1], Kazuhiko Kanou[1], Tomoko Hirata[1], Mayuko Takeda-Shitaka[1], Hideaki Umeyama[1]

1) School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN
terashig@pharm.kitasato-u.ac.jp

1. Introduction

The Critical Assessment of Protein Structure Prediction (CASP) is carried out in order to survey the capabilities and limitations of current methods of modeling protein structure from sequence. The methods are assessed on the basis of the analysis of a large number of blind predictions of protein structures. In the eighth round (CASP8 [1]), our fams-ace2 server participated in the 3D coordinate prediction category as a human expert group. We applied two different scoring functions for the fully automated model prediction server, fams-ace2: (1) the local consensus score; and (2) the model quality score based on classification of the side-chain environment for each residue. The local consensus score was used as a filter to select the models which have locally similar structures comparing with the set of models. The model quality score was then used for the final selection of the best model. This model quality score was calculated by our model quality assessment program CIRCLE [2] (see KP230).

2. Method

The procedure of fams-ace2 can be summarized as the following 4 steps:
(1) Obviously incorrect models which have serious physical clashes or broken main-chain structures were removed. (2) The top 10% (an optimized parameter of fams-ace2) of server models were selected in the order of the local consensus score; the local consensus score is calculated as the equation (1).

$$LocatCons_m = \frac{\sum_{n}^{N}\sum_{i}^{R_m} MAXSUB\left(LOC_{m,i}, LOC_{n,i}\right)}{N}$$

Equation 1

N is the number of server models. $LOC_{m,i}$ is a set of C-alpha coordinates which exist within 10Å from the i th residue of model m. *MAXSUB(a,b)* is a maximum number of C-alpha coordinates (subset a) which superimpose well (within 3Å) upon their corresponding C-alpha coordinates in subset b. The values of 10 and 3 Å are optimized parameters of fams-ace2. (3) All of the server models, selected in step (2), were refined and rebuilt utilizing our homology modeling program FAMS[3]. (4) The top 5 structures were selected, according to a model quality evaluation based on their CIRCLE score. The coefficients of *SSscore* in the KP230 which do not use the consensus method were changed in the fams-ace2 from 0.35 and 0.75 to 0.30 and 0.30, respectively. The fams-ace2 is a fully automated server and does not require human intervention.

The parameters of fams-ace2 were optimized by the data set of previous CASP7. We used the Global Distance Test Total Score (GDT_TS) as the quality of model compared to native. When we applied optimized fams-ace2 to CASP7 targets, fams-ace2 obtained the best results over all server groups (Fig.1). Moreover, in Template Based Modeling Targets, fams-ace2 also achieved best results over all groups including human groups.

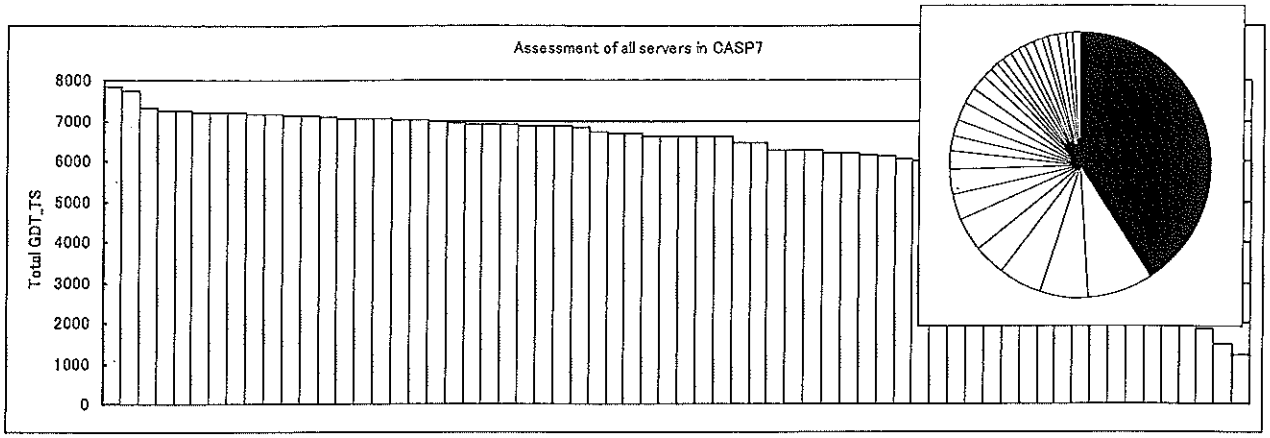

Fig. 1 Results of fams-ace2 (gray bar) and distribution of selected servers (pie graph) in CASP7

3. Results and Discussion

The 103 native protein structures of CASP8 128 targets were published in CASP8 web site (Sep 03 2008). We calculated GDT_TS of all models submitted by servers and fams-ace2 (Fig. 2). The total GDT_TS of fams-ace2 (gray in Fig.2) were obviously better than almost all of the other servers. The fams-ace2 selected models of the best server (Zhang-server) among 40% and 34% of targets in CASP7 and CASP8, respectively (The black area in pie graph of Fig.1 and Fig.2).

Although the advantage of fams-ace2 over other servers is slightly smaller than the results applying for CASP7 (123 domains, Fig.1), the intended results are accomplished. This small difference between CASP7 and CASP8 might be caused by the change of the distribution of target difficulty and performance of servers. When we calculate GDT_TS of CASP8 models, we did not consider the domain regions. Therefore the results of some targets will be changed. The advantages of fams-ace2 are the fully automated process, the low calculation costs and a high accuracy similar to the top of human groups. We are planning to optimize fams-ace2 according to the target difficulty and performance of each server by using much huger data set.

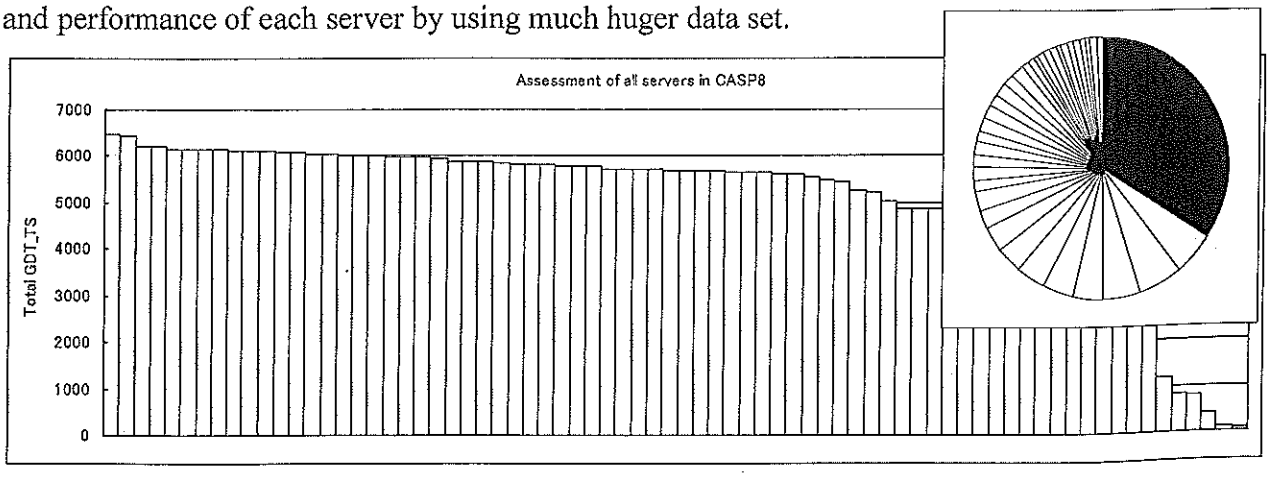

Fig. 2 Results of fams-ace2 (gray bar) and distribution of selected servers (pie graph) in CASP8

References

1. http://predictioncenter.org/casp8/index.cgi
2. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Fams-ace: a combined method to select the best model after remodeling all server models. *Proteins*. 2007;69 Suppl 8:98-107.
3. Ogata, K. and Umeyama, H. An automatic homology modeling method consisting of database searches and simulated annealing. *J. Mol. Graphics Mod.* 2000 Jun;18(3):258-72, 305-6.

KP230 CIRCLE: Development of 3D Protein Model Quality Assessment program using secondary structure prediction method and side-chain environment

Genki Terashi[1], Hiroko Sakai[1], Kazuhiko Kanou[1], Tomoko Hirata[1], Mayuko Takeda-Shitaka[1], Hideaki Umeyama[1]

1) School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN
terashig@pharm.kitasato-u.ac.jp

1. Introduction

The accurate prediction of protein structure is one of the major challenges in the field of bioinformatics. The Model Quality (MQ) assessment technique for distinguishing the near native models (high quality models) from decoys which are inferior models is one of the most important factors to achieve the accurate protein structure prediction. Many of the scoring functions for evaluating protein structures are founded on knowledge-based potentials, clustering methods, structural energies using molecular mechanics force fields, and the profile of sequence or structure (e.g. Verify3D, Inbgu, 3D-PSSM, ProQ). These scoring functions are used to assess the model quality and ultimately select the best model among a set of models. In this work, we have developed MQ Assessment Programs CIRCLE [1] and participated in Quality Assessment (QA) category of CASP8 [2] (The 8th Critical Assessment of Protein Structure Prediction, May-Aug 2008). CIRCLE aims at identifying the near native models and incorrect models without using consensus methods.

2. Method

CIRCLE considers two terms for the model quality: (1) model quality calculated from the side-chain environment of each residue (SideChainScore in equation(1)); and (2) similarity between the secondary structure propensities predicted for an amino acid sequence by PSI-PRED and the secondary structure of the three-dimensional model (SSscore in equation (1)). The side-chain environment for each residue is determined from the fraction of the molecular surface area of the side-chain covered by the polar atoms, the fraction of the side-chain area buried by any other atoms, and the secondary structure. According to the target difficulty, a total score is calculated as:

$$TotalScore = \begin{cases} \sum_{n}^{length} \left(0.35 \times SSscore + SideChainScore_{CM}\right)_n & CM \\ \sum_{n}^{length} \left(0.75 \times SSscore + SideChainScore_{FRNF}\right)_n & FRorNF \end{cases} \quad (1)$$

As shown in equation (1), the similarity score of the secondary structures (SSscore) is emphasized in difficult targets (FR: Fold Recognition, NF: New Fold) than easy targets (CM: Comparative Modeling).

In the QA category of CASP8, predictor groups provide quality estimates comprising scores between 0.0 and 1.0 for each protein structure model produced by server groups participating CASP8. Therefore, for each target, we convert estimated score of models into the values from 0.0 to 1.0 by scaling circle score of models which has minimum and maximum values.

3. Results and Discussion

The 103/128 (80%) native protein structures of CASP8 targets were published in CASP8 web site (Sep 2008). We calculated Pearson's correlation coefficient between converted CIRCLE score and the quality of models. We used the Global Distance Test Total Score (GDT_TS) as the quality of model compared to native.

The average of GDT_TS (x-axis) and correlation coefficient (y-axis) are shown in Fig.1. These results show that QA performance of CIRCLE depends on the quality of set of models which are evaluated (Table 1). The good correlation coefficients were obtained above 0.9 for the targets having the high average value of GDT_TS (above 50).

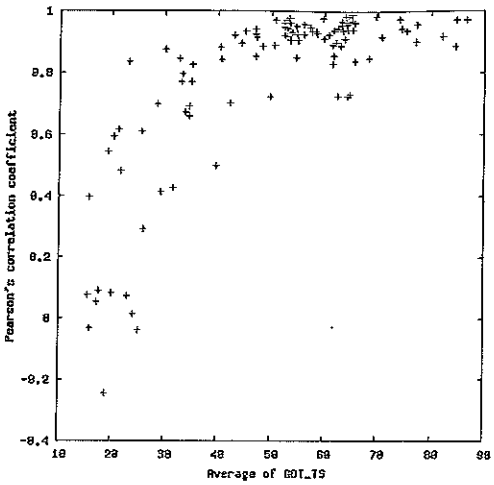

Fig. 1

Average of GDT_TS	Average of Pearson's correlation coefficient
0-25	0.24
25-50	0.75
50-75	0.92
75-100	0.93
ALL	0.78

Table 1

Additionally the best (T0423) and worst (T0460) examples of CIRCLE results are shown in Fig.2 and Fig.3. The x-axis and y-axis represents the circle score and GDT-TS of each model, respectively. In T0423 (Fig.2), CIRCLE score has high value of correlation coefficient (0.98), because high quality models (GDT_TS > 50) has high proportion of set of models. In contrast, in the case that no good models existed in the set of models (T0460 of Fig.3), CIECLE could not perform well (correlation coefficient = -0.24). These results indicate that CIRCLE still has a room to improve especially in difficult targets. We are planning to add other kind of scoring function calculated from evolutional information such as a sequence alignment score and consensus method.

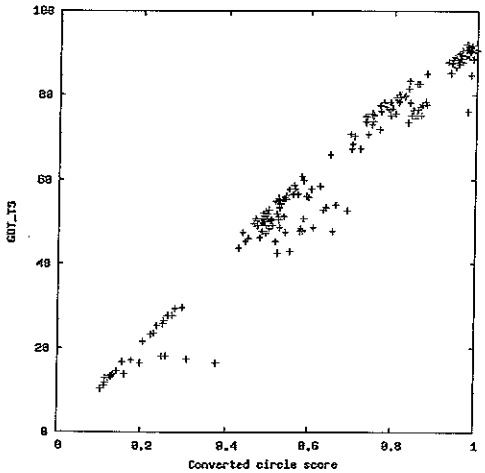

Fig. 2 T0423

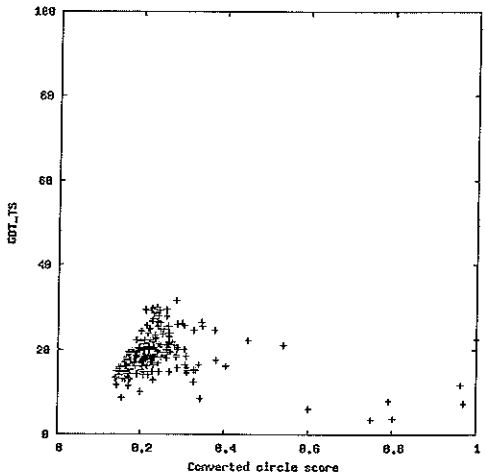

Fig. 3 T0460

References

1. http://predictioncenter.org/casp8/index.cgi
2. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Fams-ace: a combined method to select the best model after remodeling all server models. Proteins. 2007;69 Suppl 8:98-107.

KP231

FAMS_multi: Automated homology modeling based upon multiple reference proteins using better pairwise alignments in CASP8

Kazuhiko Kanou[1], Tomoko Hirata[1], Genki Terashi[1], Hiroko Sakai[1], Mayuko Takeda-Shitaka[1] and Hideaki Umeyama[1]

1) School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN

INTRODUCTION

We developed an automated method of protein structure prediction called FAMS (Full Automatic Modeling System) [1,2]. FAMS is a homology modeling program consisting of database search and simulated annealing, and can construct high accuracy model when appropriate reference protein was detected. For predicting more accurate model, especially of loop structure and side chain torsion angles, we developed a new version of FAMS, called FAMS-multi, which uses multiple reference proteins.

For the purpose of assessing this method, we participated in CASP8 (8th Critical Assessment of Techniques for Protein Structure Prediction) experiment (our team name is '*FAMS_multi*'). CASP is a world-wide experiment for protein structure prediction held every two years since 1994. CASP provides participants with more than 100 protein sequences, and the each of participants must submit the predicted structures within 76 hours and more 2 weeks as an automatic predictor and a non-automatic one, respectively. Non-automatic predictors can use models which have been predicted by automatic predictors. We participated as a non-automatic predictor for the purpose of using automatically predicted models, but all processes were performed automatically. Models which were predicted by other automatic predictors were used to generate better alignments, and we rebuilt models by using FAMS-multi program which uses multiple reference proteins. In the following, we describe the scheme of this method and our results for CASP8.

METHODS

1. Generation of better pairwise alignments

We used the predicted models by other teams to generate better pairwise alignments between the target and its template in the PDB. First, we rebuilt these models by using FAMS program for the purpose of removing collisions. These rebuilt models were used to generate pairwise sequence alignments between the target and its template. The pairwise alignments were generated by structural superposition between each refined model and the its template using CE program [3]. When the superposition of the model and its template was not performed with the criteria of Z-score > 3.7, the alignment was not used.

Next, we constructed Cα models from these alignments using FAMS-multi program, and calculated 3D-jury scores of these Cα models which is Cα consensus score. Some alignments whose Cα model has a high 3D-jury score were used to construct full atom models using FAMS-multi program, and these models were evaluated using fams-ace2 method. Figure 1 shows the distribution of teams whose alignment was used to construct models that were finally selected by fams-ace2 method.

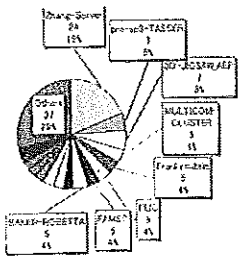

Figure 1

2. Construction of models by FAMS-multi

Some reference proteins were chosen based on the sequence and structural similarity with template. Next, a multiple structural alignment based on the superposition of Cα atoms was performed among the reference proteins. The target sequence was put on for this alignment based on the piarwise alignment between target and template mentioned in the preceding section. Thus, we get a result of multiple alignment between a target protein and reference proteins.

Using this alignment, tertiary structures were constructed mainly with next three steps, Cα construction, main chain construction, side chain construction. In each step, optimization was executed by the simulated annealing method.

Cα construction step: For the initial Cα coordinates,

kanouk@pharm.kitasato-u.ac.jp

first, the weighted average of Cα coordinates and the average distance were obtained from pairwise structural alignment based on the superposition of Cα atoms of the target and reference proteins. The weight factor of Cα coordinates for each reference proteins was decided based on Local Space Homology (LSH) calculated for each secondary structure segment. Next, the coordinates of Cα atoms were optimized by simulated annealing.

Main chain construction step: Initial coordinates of main chain atoms were constructed with the same method as FAMS. In the simulated annealing step, the potential function, which is consisting of (1) the weighted average of the coordinates of main chain atoms, (2) the average of distance and (3) the pair of N and O atoms forming the hydrogen bond as structural information, was used.

Side chain construction step: For the generated main chain atoms, conserved side chain torsion angles were obtained from homologous proteins. The coordinates of side chain atoms consisting of conserved side chain torsion angles were placed in relation to the fixed main chain atoms. The structural information such as the weighted average of the coordinates, average of distance, 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.

3. Evaluate models (fams-ace2 method)

Thus, some full atom models were constructed. These models were evaluated using fams-ace2 selecting method (combined Cα consensus and Circle score [4]). Consequently top five models were selected.

4. Refine models

Five selected models were refined using Energy minimize & Molecular dynamics. With this procedure, hydrogen bonds, main chain torsion angles and side chain torsion angles were refined slightly and collisions of hydrophobic atoms were decreased.

All the procedures were implemented automatically.

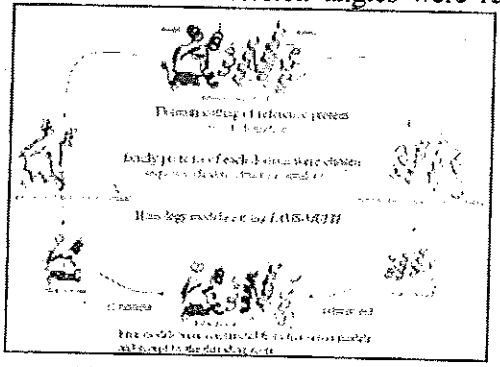

Figure 2. The scheme of *FAMS_multi*.

RESULTS & DISCUSSION

103 experimental structures of 128 CASP8 targets became available by September 3, 2008. We evaluated the accuracy of *FAMS_multi* models and that of the other server models, and compared them. The accuracy of backbone geometry was assessed by GDT_TS score, and the accuracy of side chain was assessed by the number of residues which have a sufficiently accurate side chain (chi1 torsion angle within 30 degrees from native structures or chi2 torsion angle within 60 degrees from native). Figure 3 shows the server ranking with the cumulative GDT_TS score of 103 targets (bar graph). Line graphs of square and triangle point is the cumulative number of accurate Chi1 torsion angles and Chi2 torsion angle, respectively.

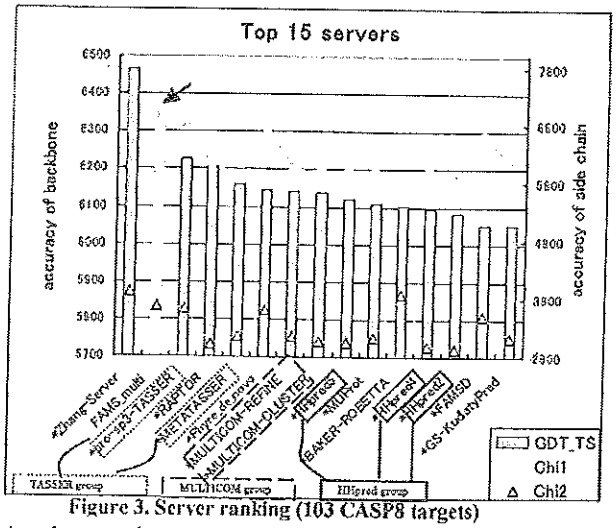

Figure 3. Server ranking (103 CASP8 targets)

As the results, *FAMS_multi* ranked second following Zhang-Server with GDT_TS score. *FAMS_multi* also ranked second following Zhang-Server with side chain accuracy. *FAMS_multi* could construct good models in terms of backbone geometry and side chain conformation.

REFERENCE

[1] Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.

[2] Ogata K, Umeyama H. Proteins. 1998; 31(4):355-69.

[3] Shindyalov IN, Bourne PE. Protein Engineering 1998; 11(9) 739-747.

[4] Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.

KP232 FAMSD: Individual Comparative Modeling server using SP3, FAMS and CIRCLE.

Kazuhiko Kanou[1], Tomoko Hirata[1], Genki Terashi[1], Hiroko Sakai[1], Mayuko Takeda-Shitaka[1] and Hideaki Umeyama[1]

1) School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641 JAPAN

INTRODUCTION

Recently, the number of proteins whose three-dimensional structures are already solved is increasing more and more. In September 2008, more than 52,000 structures are available on the Protein Data Bank (PDB) website. But the sequences whose structure has not been solved yet are more than 100 times as many as the sequence which were solved structurally. Therefore some approaches for protein structure prediction are required for implementing the structure based drug design and so on. Some effective approaches have been developed all over the world. Among those approaches, the most effective one is the Comparative Modeling when suitable template structures which have high sequence identities are detected. Our comparative modeling consists of following four steps: (1) making sequence alignments between target protein and template structures, (2) constructing three-dimensional structures based upon each alignment, (3) selecting the best structure model and (4) refinement of the selected model. We have developed an automatic protein structure prediction server called FASMD. Programs such as SP3 [1], FAMS (Full Automatic Modeling System) [2], CIRCLE [3] and Molecular dynamics were used at the each step (1) ~ (4), respectively.

We had participated in the past Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments. CASP is held once in 2 years, and each participant receives more than 100 protein sequences whose structure was unknown, and returns the predicted three-dimensional structures. After prediction season, the organizers of CASP assess the quality of all predicted models using its experimental structures. From April to August 2008, the 8th CASP (CASP8) experiment was held and 128 protein sequences were released totally. We participated in CASP8 as an automatic predictor using FAMSD. We describe the algorithm of FAMSD and our results for CASP8.

METHODS

(1) Making sequence alignments

8 kinds of alignment programs, BLAST, PSI-BLAST [4], PSF-BLAST, RPS-BLAST, IMPALA, Pfam-BLAST, SPARKS2 and SP3 were executed for each target protein sequence. Various alignments were generated and were filtered with its alignment score. The alignment scores for 6 kinds of methods except SPARKS2 and SP3 were calculated with following equation,

$$score = f(k_i, Hom, Len, SS) \qquad (1)$$

Here *Len* is the number of residues of a predicted model. *Hom* indicates sequence identity % value, *SS* is the degree of secondary structure agreement between the secondary structures predicted one from sequence using PSI-PRED [5] and one calculated from model using STRIDE. k_i is a coefficients for each alignment method.

And as the alignment score for SPARKS2 and SP3, Z-score of their output was used.

When the alignment score was more than the maximum score of all alignments * X, these alignments were used to construct model. A parameter X is a cut-off value which was decided using CASP7 targets as a training set.

(2) Constructing three-dimensional structures

We constructed three-dimensional structures using FAMS program based on each selected alignment which was mentioned in the preceding section.

kanouk@pharm.kitasato-u.ac.jp

(3) Selecting the best structure

All constructed models were evaluated using following scoring function,

score = *CIRCLE* + *w* * *SSscore*

Here, *Circle* represents the 3D1D score which was improved based on verify3D and *SSscore* represents the degree of secondary structure agreement. *w* is the weighting factor for *SSscore* which was optimized using CASP7 models as a training set.

Figure 1 shows the distribution of alignment method of finally ranked first models by above scoring function.

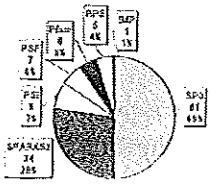

Figure 1.

(4) Refinement of the selected models

Five selected models were refined using Energy minimize & Molecular dynamics. With this procedure, hydrogen bonds, main chain torsion angles and side chain torsion angles were refined slightly and collisions of hydrophobic atoms were decreased.

RESULTS & DISCUSSION

103 experimental structures of 128 CASP8 targets became available by September 3, 2008. We evaluated the quality of all server models, and compared GDT_TS of FAMSD model and the average of all server models (Figure 2). As a result, in almost of all targets GDT_TS of FAMSD model is higher than the average of all server models. Figure 2 shows that two targets were failed to prediction and these targets are in the difficult category.

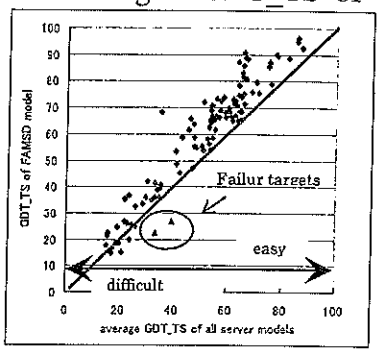

Figure 2.

Figure 3 shows top 15 of 126 server predictors sorted by the cumulative GDT_TS score of all 103 targets. Then FAMSD ranked at 13th. The accuracy of side chain was also assessed with the number of residues in the case that each model have a sufficiently accurate side chain, i.e., chi1 and chi2 torsion angle which is within 30 and 60 degrees, respectively, from native structure. In Figure 3, line graphs of square and triangle point is the cumulative number of accurate chi1 torsion angles and chi2 torsion angle, respectively.

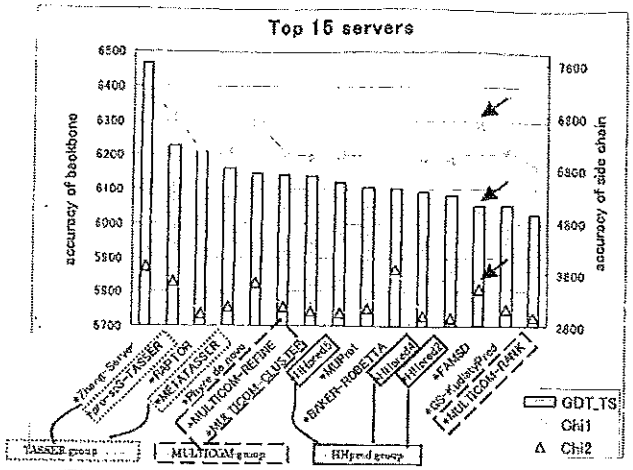

Figure 3 Top 15 servers for all 103 targets

Furthermore we calculated the cumulative score of GDT_TS, chi1 and chi2 for only 75 targets in the relatively easy category. Target classification is referred to on Robetta evaluation page [6]. As the results, the rank of FAMSD with GDT_TS, chi1 and chi2 were 11th, 7th and 11th, respectively. The six servers (Zhang-Server, Phyre_de_novo, pro-sp3-TASSER, FAMSD, BAKER-ROBETTA and COMA-M) predicted high quality models in terms of not only backbone geometry but also side chain conformation.

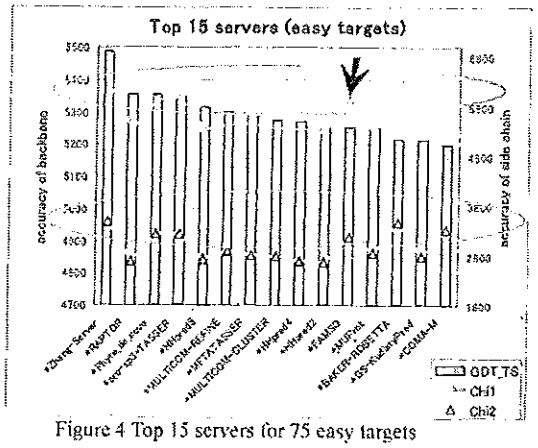

Figure 4 Top 15 servers for 75 easy targets

REFERENCE

[1] Zhou H, Zhou Y. Proteins. 2005;61 Suppl 7:152-6.

[2] Ogata, K. and Umeyama, H. J Mol Graph Model 2000; 18, 258-272.

[3] Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Proteins. 2007;69 Suppl 8:98-107.

[4] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. Nucleic Acids Res. 1997; 25, 3389-3402.

[5] McGuffin LJ, Bryson K, Jones DT. Bioinformatics. 2000; Apr;16(4):404-5.

[6] http://robetta.bakerlab.org/CASP8_eval/index.html

KP233 FAMSD_QA: Model quality assessment using the side chain environment consensus score

Kazuhiko Kanou[1], Tomoko Hirata[1], Genki Terashi[1], Hiroko Sakai, Mayuko Takeda-Shitaka[1] and Hideaki Umeyama[1]

[1]School of Pharmaceutical Sciences, Kitasato University

INTRODUCTION

Selecting the best quality model from a set of predicted structures is one of the most important parts of protein structure prediction. In the Seventh Critical Assessment of Techniques for Protein Structure Prediction (CASP7) experiment, the new prediction category called "QA" (Quality Assessment) was implemented. In this category participants assess the quality of the models which were predicted automatically by other teams. Each QA participant gives a "reliability score" to these models from 0 (indicates 'bad model') to 1 (indicates 'good model'). After the experimental structures of CASP7 targets became available, the GDT_TS score was calculated for all predicted models and was compared with the "reliability score". Several values such as the Pearson correlation coefficient between GDT_TS and "reliability score" of each QA method were calculated to measure the quality of each QA method.

As a result of QA category in CASP7, we found that the most powerful method was consensus method like 3D-Jury [1]. 3D-Jury score is the summation of the number of residues (Cα atom) within 3.5Å from each predicted model. This method can select "good backbone" models but the quality of the side chain of selected models is not so good.

Thus we developed a new consensus method which considers side chain environment for the purpose of selecting good side chain models, and participated in the latest CASP experiment (CASP8) [2] using this method. We describe the algorithm of this method and the results of CASP8.

METHODS

First, we calculated the side chain environment composed of 'fraction buried' and 'fraction polar' for each residue of predicted model. 'Fraction buried' is the fraction of buried area within the surrounding side chain atoms, and 'fraction polar' is the fraction of buried area within the surrounding polar atoms. These values range from 0 to 1.0 per residue. When the model A was assessed, for each residue of model A, the side chain environment was calculated and is compared with the other models. If the Euclidian distance between the side chain environment ('fraction buried' and 'fraction polar') of one residue of model A and that of corresponding residue of another model was within 0.2, we considered that the two residues were in the same environment. For each model, we counted the number of residues in the same environment and the side chain environment score is the summation of those numbers. The threshold of 0.2 was determined using CASP7 models as a training set.

In CASP8, we participated in QA category as a team 'FAMSD_QA'. We had refined all predicted models by FAMS [3] and had assessed quality of these models using following combined score.

$$\text{score} = env\_con + w * SSscore$$

Here, *env_con* represents the side chain environment consensus score and *SSscore* represents the degree of match between the secondary structure of a predicted model and the secondary structure predicted from the given sequence with PSIPRED [4]. w is the weighting factor for *SSscore* and ranges from 0 to 1. In the case of difficult targets, more weight is given to *SSscore* than easy targets. This value was optimized using CASP7 models.

RESULTS and DISCCUSSION

Correlation coefficients

103 experimental structures of 128 CASP8 targets became available by September 2008. We calculated GDT_TS (accuracy score of backbone geometry) of all predicted models for 103 structure available targets, and calculated Pearson and

kanouk@pharm.kitasato-u.ac.jp

Spearman correlation coefficients between GDT_TS and FAMSD_QA score. As a result, average Pearson and Spearman correlation coefficients for all targets were 0.85 and 0.75, respectively. Furthermore the averages for 75 relatively easy targets were 0.91 and 0.79, and for 28 relatively difficult targets were 0.69 and 0.67, respectively. (Target classification is referred to on Robetta evaluation page [5].) Given this, it can be considered that FAMSD_QA scoring is more effective for easy targets than for difficult targets. The reason for the difference between Pearson and Spearman correlation coefficients for easy targets is that some targets of in the easy category have the bipolar distribution (there are both moderately good models and extremely bad models), that is, non normal distribution. The target that has the biggest difference between Pearson and Spearman correlation coefficients was T0444 (PDB code is 2VUX), these coefficients were 0.857 and 0.289, respectively. Figure 1 shows the scatter plot of FAMSD_QA score versus GDT_TS. In this case FASMD_QA scoring could judge the moderately good models (GDT_TS > 50) as "good model" and could judge the extremely bad models (GDT_TS < 30) as "bad model". Therefore Pearson correlation coefficient was very high (0.857). But among the moderately good models, FASMD_QA scoring couldn't distinguish relatively good models from relatively bad models, so Pearson correlation coefficient calculated with only these models was 0.416.

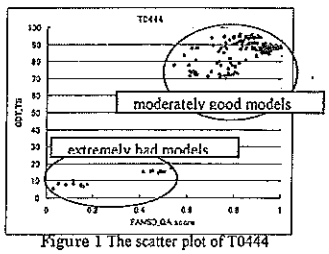

Figure 1 The scatter plot of T0444

This is not so good, but the GDT_TS of the first ranked model by FASMD_QA score is 88.6 and the highest GDT_TS among all models is 96.0. The ratio of the GDT_TS of the first ranked model to the highest GDT_TS, we call MGR (Max GDT_TS Ratio), is 92.3 (88.6/96.0) %.

The average MGR value for all targets, easy targets and difficult targets were 89.6, 93.8 and 79.0 %, respectively.

Evaluate the first raked model

We calculated the cumulative GDT_TS score of the first ranked models by FAMSD_QA score and compared with that of other automatic servers (Table 1). FAMSD_QA ranked at second following Zhang-Server.

Rank	team name	Sum of GDT_TS	average
1	Zhang-Server	[illegible]	82.77
2	[illegible]	[illegible]	81.[illegible]
3	[illegible]	[illegible]	[illegible]
4	RAPTOR	[illegible]	[illegible]
5	METATASSER	[illegible]	59.8

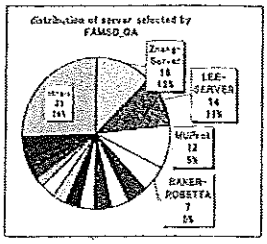

Table 1. Server ranking by the cumulative GDT_TS score

Furthermore, we evaluated the accuracy of side chain torsion angles for easy targets. We calculated the cumulative number of residues that have sufficiently accurate chi1 angle (within 30 degrees from native). As a result, FAMSD_QA ranked first of all server teams and 3D-Jury ranked fourth (Figure 2). This shows that FAMSD_QA scoring can select good models in terms of not only backbone geometry but also side chain torsion angles.

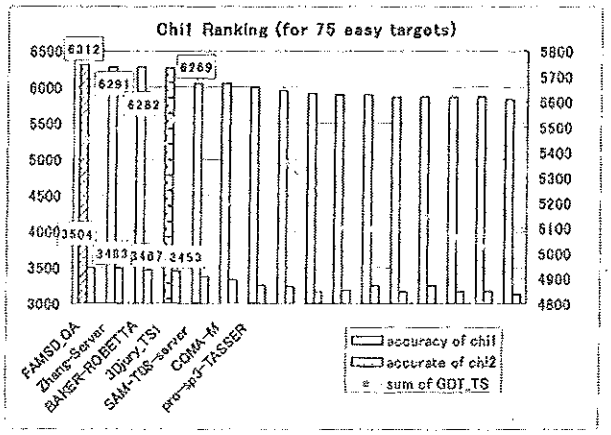

Figure 2. Server ranking by the accuracy side chain.

CONCLUSION

We developed an alternative consensus score for the purpose of selecting good models that have accurate side chain atoms. The new consensus score considers the side chain environment. We participated in recent CASP8 experiment and evaluated our method. As a result, side chain accuracy of the first ranked models by our new method was the best of all servers including 3D-Jury. It was proved that our consensus method using the side chain environment can select better side chain models.

REFERENCE

[1] Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 2003; 19:1015–8

[2] http://www.predictioncenter.org/casp8/index.cgi

[3] Ogata, K. and Umeyama, H. (2000). An automatic homology modeling method consisting of database searches and simulated annealing J Mol Graph Model/J Mol Graph Model 18, 258-272, 305-256.

[4] Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.

[5] http://robetta.bakerlab.org/CASP8_eval/index.html