FAMSBASE: Modeling Database of 161 Genomes

Mitsuo Iwadate **Kazuhiko Kanou** **Daisuke Takaya**
iwadatem@pharm.kitasato-u.ac.jp
Katsuichiro Komatsu **Mayuko Takeda-Shitaka** **Hideaki Umeyama**

Department of Biomolecular Design, Kitasato University, Minato-ku, Tokyo 108-8641, Japan

Keywords: FAMS, homology modeling, CAFASP, FAMSBASE

1 Introduction

We had participated in the CASP (Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) that is the competition of the protein structure prediction in 2000 and 2002. Our server were estimated almost the best as the individual homology modeling server in the CAFASP3 (Critical Assessment of Fully Automated Structure Prediction) that is the full automatic homology modeling section of CASP5. In this competition we used homology modeling software FAMS [1] (Full automatic protein modeling system) that we had been developed for several years [3]. Thus, usefulness of the software FAMS was clearly shown in these competitions.

We also have developed FAMSBASE that is web interface of FAMS model structure database. In this article we describe that proteins coded in the 161 species genomes were modeled.

The commercial version of FAMS modeling software service is available in In-Silico Sciences, Inc. (`http://www.pd-fams.com`). FAMSBASE is available in `http://famsbase.bio.nagoya-u.ac.jp/` [4] or `http://www.pdfams.jbic.or.jp/`.

2 Method and Results

We constructed FAMSBASE, database of the model structures calculated by FAMS program. Input data of FAMS are alignments between amino acid sequences of query protein and amino acid sequences of PDB. A website GTOP [2] `http://spock.genes.nig.ac.jp/~genome/` provides alignments calculated by many kind of alignment software, FASTA, BLAST and so on. On website GTOP top 5 hits alignments of RPS-BLAST data about 161 species were used for FAMSBASE.

For such a large size calculation large amount of computational resources were required. Therefore FAMS1K, PC cluster system constructed with 1000 personal computers (NEC: mate Pentium III 869 MHz), were developed. The total number of models is approx. 800 thousands (shown in Table 1).

3 Discussion

Software FAMS produces model structure based on alignments with each reference structure solved with experimental procedure, X-ray diffraction, NMR or other techniques. In enough high homology targets (More than 20 % sequence identity) with known structures on CASP5 competition, FAMS produced good models. Lower e-values threshold means more accurate homology statistically, and models in FAMSBASE have enough accuracy to analyze protein structures and functions relationships. In this database the RPS-BLAST method on 161 species were used with e-value threshold 0.001.

Table 1 shows the summary of RPS-BLAST hits (e-value < 0.001). In 88 species of bacteria, 14 species of archaea and 14 species of eukaryotes, the rates of modeled ORFs are roughly half. However, corresponding value in 43 species of viruses are below the 20%. This is due to the less number of the PDB structural homologs.

In Table 2, the rate of modeled residues of archaea and bacteria are 42.6% and 46.4%, respectively. These are roughly same to corresponding rates of ORFs (see Table 1). And the values of viruses also keep the 22.7% (18.4% in Table 1). But the rate of modeled residues of eukaryotes (32.5%) is significantly less than corresponding hit rate of ORFs (48.5%).. This means that hit ORFs contains many non-modeled region in eukaryotes genomes. Moreover, the average ORF length of eukaryotes genomes, 385.8 residues, is significantly longer than corresponding lengths in archaea, bacteria or viruses

Currently, roughly half of genomes ORFs are available in structural models. This is true. But modeled residue rates are 30% and 20% in eukaryotes and viruses genomes, respectively. The eukaryotes genomes, including human, are 15% backward in structural genomics against bacteria genomes. And most of archaea genomes have intermediate characters of bacteria and eukaryotes.

Table 1: Summary of RPS-BLAST hits (e-value < 0.001) and modeled ORFs of 161 species.

Group	Number of ORFs	Number of Hit ORFs	Number of Models	Hit rate (%)	Models per ORF	Number of Species
Archaea	38249	17932	58404	46.9	3.26	16
Bacteria	261341	135034	455690	51.7	3.37	88
Eukaryotes	174091	84387	281360	48.5	3.33	14
Viruses	2602	479	1001	18.4	2.09	43
Total	476283	237832	796455	49.9	3.35	161

Table 2: Summary of modeled residues of 161 species.

	Number of total residues	Number of modeled residues	Rate of modeled residues (%)	Average ORF length
Archaea	10845376	4624458	42.6	283.5
Bacteria	81022067	37552007	46.4	310.0
Eukaryotes	67162057	21820488	32.5	385.8
Viruses	565754	128656	22.7	217.4
Total	159595254	64125609	40.2	335.1

References

[1] Iwadate, M., Ebisawa, K., and Umeyama, H., Comparative modeling of CAFASP2 competition, *Chem-Bio Informatics Journal*, 1:136–148, 2001.

[2] Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N., and Nishikawa, K., GTOP: a database of protein structures predicted from genome sequences, *Nucleic Acids Res.*, 30:294–298, 2002.

[3] Ogata, K. and Umeyama, H., Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms, *Protein Eng.*, 10:353–359, 1997.

[4] Yamaguchi, A., Iwadate, M., Suzuki, E., Yura, K., Kawakita, S., Umeyama, H., and Go, M., Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species, *Nucleic Acids Res.*, 31:463–468, 2003.