GENIUS II: A Database System Including Complete Genome ORFs that are Assigned to Protein Three-Dimensional Domains

Yukimitsu Yabuki[1,2] **Yuri Mukai**[1] **Makiko Suwa**[1]

yukimitsu-yabuki@aist.go.jp yuri-mukai@aist.go.jp m-suwa@aist.go.jp

[1] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan

[2] Information and Mathematical Science Laboratory (IMS), Inc., 2-43-1 Ikebukuro, Toshima-ku, Tokyo 171-0041, Japan

Keywords: multiple intermediate sequence search, genome analysis, CATH

1 Introduction

About 90 kinds of complete genomes are now available due to the rapid increase of genome sequences. These genomes have enormous hypothetical or uncharacterized proteins, whose functional annotation is the most important issue in the post-genome era. The essential information to elucidate function is the three-dimensional structure of proteins, but comprehensive structure determination is not realistic even with the structural genomic projects. Therefore the computational structure assignment to protein sequences will greatly enhance our understanding and give essential insight to the function of novel or uncharacterized genes.

Our GENIUS II is completely automated and high quality system for assigning protein coding regions in genomes to three dimensional structures by using a multiple intermediate sequence search (MISS) [6]. It is the method in which the query and the target PDB sequences are linked by multiple intermediate sequences, which are gathered by an iterative PSI-BLAST search [1]. It had been previously reported that the MISS methods are intended to obtain results with minimum false positives other than the assignment by only using PSI-BLAST search [3, 6]. The specificity of links between the query, the intermediate and the target sequences has been optimized for more than 99% with high sensitivity by employing safe thresholds, being evaluated by using CATH single domains. Although with tight requirement for being high accurate system, the MISS method still can detect distant homologues even with the sequence similarity below twilight zone.

2 Method and Results

To assign tertiary domains to protein coding regions in genome sequences, we performed the following automated MISS procedures:

1) First, we selected amino acid sequences with more than 40 residues long, from PDB database. Transmembrane helix, coiled-coil and low-complexity regions were masked, using the SOSUI [2], MULTICOIL [7] and SEG [8] programs, respectively, and subsequently compared in a pair wise manner using FASTA [4]. Sequences, each having pair wise scores above an optimized threshold (normalized score = 88) [4] and 80% overlaps were clustered into same groups and PDB representatives were selected.

2) Using PSI-BLAST, each representative PDB sequence was searched against the nr-aa database. All aligned regions with below the safe threshold (E = 0.001) [5], were stored in the psi_PDB database linking information about which PDB sequence was associated with each aligned region.

3) Protein coding regions for all genomes were taken from the GenBank annotation and were then searched against this psi_PDB database using the FASTA with thresholds as optimized in this program (normalized score = 88) [5]. In all cases the aligned region consisted of either more than 100 residues or more than 50% of the query sequence (genome ORF) are selected. Because of the way how we created psi_PDB, whatever sequence that hit to this database, show significant match to protein three-dimensional domain.

4) For further functional evaluation, PROSITE patterns are assigned to protein coding regions of each genome. Since PROSITE patterns are written by regular expression, we determined the P value, which is calculated as the multiplication of each residue frequency in SWISSPROT. P value is employed to access the motif quality for functional assignment.

In the 87 genomes applied by MISS method, an average of 37.3% protein coding regions of each genome (39.1% for case of only prokaryote) show significant matches to proteins of known structure. This performance represents a significant increase over our previous assignments (23 – 32%) [6], and is reasonable because the results strongly depend on the growth of the PDB and nr-aa database size.

The detailed results for the application of this system to 87 complete genomes are summarized in GENIUS II (`http://genius.cbrc.jp/`) which navigates the following menus.

(A) View results: For each genome the percentage of reliable hits of protein coding regions to known 3D structures is displayed as pie charts, and all protein coding regions are listed with (i) the name of intermediate sequence, (ii) PDB code of domain structure, (iii) alignment between intermediate sequence and PDB sequence, (iv) the location of functional sites characterized by PROSITE motifs, and (v) detailed structural and functional information by links to the PDBsum and CATH database. We have furthermore prepared for a graphical visualization of the 3D domain structure which matched with the intermediate and genomic (query) sequences using Rasmol or Chime viewers.

(B) Key word search: Protein coding regions in different genomes are linked to each other and their relationship can be easily retrieved using keywords or PDB codes.

(C) Sequence Search: We have made the MISS procedure available on this web page for user-supplied sequences.

(D) Set up for using Rasmol: This supports to download and set up the Rasmol or Chime viewers.

References

[1] Altschul, S.F., Adden, T.L., Schaffer, A.A., Zhang, Z., Miller, W., and Lipman, D.J., Gapped blast and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, 25:3389–3402, 1997.

[2] Hirokawa, T., Boon-Chieng, S., and Mitaku, S., SOSUI: Classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, 14:378–379, 1998.

[3] Park, J., Karplus, K., Barret, C., Hughey, R., Haussler, D., Hubbard, T., and Chotia, C., Sequence comparisons using multiple sequences detect three times as many remote homologues as pair wise methods, *J. Mol. Biol.*, 284:1201–1210, 1998.

[4] Pearson, W.R. and Lipman, D.J., Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

[5] Salamov, A.A., Suwa, M., Orengo, C.A., and Swindells, M.B., Combining sensitive database searches with multiple intermediates to detect distant homologues, *Protein Eng.*, 12:95–100, 1999.

[6] Salamov, A.A., Suwa, M., Orengo, C.A., and Swindells, M.B., Genome analysis: Assigning protein coding regions to three-dimensional structures, *Protein Sci.*, 8:771–777, 1999.

[7] Wolf, E., Kim, P.S., and Berger, B., MultiCoil: A program for predicting two- and three-stranded coiled coils, *Protein Sci.*, 6:1179–1189, 1997.

[8] Wooton, J.C., Non-globular domains in protein sequences: Automated segmentation using complexity measures, *Comput. Chem.*, 18:269–285, 1994.