Genius: System for Assigning Protein Coding Regions to 3D Structures

Makiko Suwa [1]
suwa@hri.co.jp
Henrik T. Yudate [1]
yudate@hri.co.jp
Yasuhiko Masuho [1]
masuho@hri.co.jp
Asaf A. Salamov [2]
salamov@sanger.ac.uk
Christine A. Orengo [3]
orengo@biochemistry.ucl.ac.uk
Mark B. Swindells [4]
swintech@biochemistry.ucl.ac.uk

[1] Helix Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba 292, Japan
[2] The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[3] Biomolecular Structure and Modelling Unit, Department of Biochemistry, University College, London, Gower Street, London, UK
[4] Inpharmatica Ltd., 60 Charlotte Street, London, W1P2AX, UK

1 Introduction

Several attempts to provide a completely automated procedure for genome analysis categorize the hits to globular domains into those matching a sequence of known three dimensional structure. Based only on standard sequence searching procedures such as FASTA, BLAST, complete genomes have currently been analyzed and 10-20% of the ORF sequences linked to a domain of known structure [1,2]. More Recently, many projects [3] joining sequence search and fold recognition components have been largely increased the fraction ratio to almost 40% [4,5,6]. On the contrary, our interest is to maximize the number of the relationships identified by the method based on sequence search alone, because sequence search is more sensitive and quick than the fold-recognition procedures, and then it can be immediately applied to all finished genomes. Our method is based on the high performance of MISS (Multiple Intermediate Sequence Search) method [7] in which the query and the target sequences are linked by multiple intermediate sequences which were gathered by iterative sequence search (PSI-blast) [8]. The reliability of links between the query – the intermediate – the target sequence were already promised by safer threshold of sequence search algolisums with high sensitivity and specificity, evaluated in the previous work [8]. Using MISS method we can detect distance homologues although they share quite weak sequence similarity below twilight zone.

2 Method

To assign 3D protein domains to genome ORFs, we performed an following automated system of MISS procedure. 1) All sequences from PDB were first masked for transmembrane regions and coiled-coils, and then compared with one another in a pairwise manner using FASTA. Sequences having scores above the specified threshold [8] were clustered into families and a representative selected. 2) Using PSI-blast, each representative PDB sequence was run against Owl database (ver. 31.4). All aligned regions from sequences below the safer threshold (E=0.001) [8], were stored together with information about which PDB sequence was associated with each aligned region (psi_PDB database). 3) Every open reading frame for each genome was searched against psi_PDB database, using the gapped blast algorithm. Because of the way that psi_PDB were made, whatever sequence is hit, a link can be immediately made to a region of a known structure.

3 Application to 23 genomes

Applying this system to 23 genomes, about 40% on average of the ORFs of each genomes had significant matches to proteins of known structure. It is reasonable that the ratio is significantly larger than our first assignment of previous version [9], since the results strongly depends on the growth of data size of PDB and Owl. If we see the case of *M. genitalium*, nearly 53% of the ORFs were asigned to a region of known structure. This compares favorably with the results reported previously (12% [1], 22% [2], 38% [6]). Although it is difficult to compare directly our results and them, because of the influence of database grouth as descrived above, our results were significantly higher than most previously reported methods.

More than 60 proteins showed frequent hit number (> 50) to open reading frames. Most of them relating to gene transcription, with phosphate containing coenzymes. Reasons for this bias can be gleaned from an analysis of known structures, where despite the vast array of enzymes, coenzyme functionality is most frequently provided by one of only a small set of distinct domains. It would seem likely therefore that these domains that have been continually reused during evolution, through repeated genetic rearrangement.

Their structural feature is unusual, because not only do all the single proteins belong to the ab class of structures, but so do nearly all of the domains hit within multi-domain proteins. This was not anticipated, as typically a non-homologous database contains only 50% in the ab class, with the remainder divided equally between mainly a and mainly b structures.

The reliable hit information between complete genome ORFs and protein domain structures. was summarized in "Genius" (http://www.hri.co.jp/genius1.2/). This is a new release from the first version [9]. If some ORF have significant match to tertiary structure, you can visualize the 3Dstructure and interface region of (PDB – intermediate - ORF) sequence. Furthermore, the advanced usage such as the orf retreeval using keywords or the MISS search for your own sequence are also available. We hope this database system is applicable to comparative structural genomics field.

References

[1] Casari, G., Andrade, M.A., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, J., Valencia, A., and Sander, C., Challenging times for bioinformatics, *Nature*, 376:647–648, 1995.

[2] Frishman, D. and Mewes, H.W., *Trends in Genetics*, 13:415–416, 1997.

[3] Ficher, D. and Eisenberg, D., Predicting structures for genome proteins, *Current Opinion in Structural Biology*, 9:208–211, 1999.

[4] Fischer, D. and Eisenberg, D., Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*, *Proc. Natl. Acad. Sci. USA*, 94:11929–11934, 1997.

[5] Gribskov, M., Translational initiation factors IF-1 and eIF-2 alpha share an RNA-binding motif with prokaryotic ribosomal protein S1 and polynucleotide phosphorylase, *Gene*, 119:107–111, 1992.

[6] Rychlewski, L., Zhang, B., and Godzik, B., Fold and function predictions for *Mycoplasma genitalium* proteins, *Folding and Design*, 3:229–238, 1998.

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

[8] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., 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.

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