Genius: An Automated System for Assigning Genome ORFs to 3D Structures, Using a Novel Multiple Intermediate Sequence Search

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, Welcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[3] Biomolecular Structure and Modeling Unit, Department of Biochemistry, University College London, Gower Street, London, UK
[4] Inpharmatica Ltd., 60 Charlotte Street, London, W1P2AX, UK

Keywords: structural assignment, genome analysis, multiple intermediate sequence search

1 Introduction

The procedure to assign protein sequences to known 3D structure enable us to obtain essential information concerning the function of the novel gene product.s Based only on standard sequence search methods such as FASTA and BLAST, complete genomes have currently been analyzed and 10-20% of the ORF sequences have been linked to a domain of known structure [1, 2]. Other approaches using sequence searches, fold recognition components, or combinations of these have significantly increased the performance to almost 40 % [3, 4, 5].

On the contrary, our strategy is to maximize the number of relationships identified by a method based only upon sequence searches, because these are more sensitive and faster than the fold-recognition procedures, and they can be immediately applied to all finished genomes.

We described here the "Genius" system for assigning genome ORFs to three dimensional protein structures. It is based on the the high performance of the MISS (Multiple Intermediate Sequence Search) method [6] in which the query and the target sequences are linked by multiple intermediate sequences which have been gathered by an iterative PSI-blast search. The reliability of links between the query, the intermediate and the target sequences has been optimized for high sensitivity and specificity, using safe thresholds, which were determined in previous study [6]. Using the MISS method we can link the query and it's distance homologues even if they only share weak sequence similarity below the twilight zone.

2 Method

To assign 3D protein domains to genome ORFs, we performed the following automated procedures of the MISS system. 1) All sequences from PDB were first masked for transmembrane regions and coiled-coils, and then compared in a pair wise manner using FASTA. Sequences having scores above the specified threshold [6] were clustered into families and a representative selected. 2) Using PSI-blast, each representative PDB sequence was run against the Owl database (ver. 31.4). All aligned regions from sequences below the safe threshold (E=0.001) [6], were stored together with information about which PDB sequence was associated with each aligned region (psi_PDB database). 3) Every open reading frame in each genome was then searched against this 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 Genius System

The detailed solution on the application of this system to 23 genomes are summarized in "Genius" (http://www.hri.co.jp/ genius1.2/), which is the second updated release of the first version [7]. For each genome the reliable hit information of ORFs and domain structures are listed with alignment results and the location of functional sites characterized by PROSITE motifs, and detailed structural information by linking to the CATH database. A visualization of the 3D structure and interface region of the PDB – intermediate – ORF is available using Rasmol or Chime viewers. ORFs of different genomes are linked to each other and the relation can be easily retrieved using keywords or PDB codes. This workbench also support the MISS procedure for the Web user's original sequence to detect quite distant homologues.

In the 23 genomes summarized in the "Genius" web page, approximately 40% on average of the ORFs of each genome show significant matches to proteins of known structure. It is reasonable that the performance is significantly increased following the recalculation since our first assignment version [9], as the results strongly depends on the growth of the data in PDB and Owl. This compares favorably with the results reported previously [1, 2, 5]. Although it is difficult to compare our results directly, because of the influence of size of the database as described above, our results are significantly higher than most previously reported methods. More than 25 structures were repeatedly hit (> 100) with open reading frames. It was interesting that most of these were phosphate containing coenzymes relating to gene transcription, and their structural features are unusual, because not only do most of the single domain proteins belong to the alpha/beta class of structures, but so do nearly all of the domains hit within multi-domain proteins. 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.

4 Future Enhancements

Genomes can be analyzed quickly and are updated at regular intervals. In our future versions we hope to increase the search sensitivity by taking full advantage of all the sequences available from sequencing projects rather than the current approach which only uses OWL during the PSI-Blast step. We hope that this database system will be a popular contribution to the 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*, PEDANTic genome analysis, 13:415–416, 1997.

[3] Ficher, D. and Eisenberg, D., Predictiong structures for genome proteins, *Current Opinion in Structural Bology*, 9:208-211, 1999.

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

[5] Rychlewski, L., Zhang, B. and Godzik, B., Fold and functions for *Mycoplasma genitalium*, *Folding and Design*, 3:229–238, 1998.

[6] 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.

[7] 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.