Genome Analysis: Assigning Protein Coding Regions to 3D Structures

Asaf A. Salamov [12]
salamov@sanger.ac.uk
Makiko Suwa [1]
suwa@hri.co.jp
Christine A. Orengo [3]
orengo@biochem.ucl.ac.jp
Mark B. Swindells [14]
swintech@biochem.ucl.ac.jp

[1] Helix Research Institute, 1532-3 Yana, Kisarazu-shi, Chiba, 292, Japan
[2] The Sanger Center, Wellcome TrustGenome Campus, Hinxton, Cambridge, CB101SA, 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

The advent of complete genome sequencing has emphasised the need for comparable developments in automated analysis, for finding approximate 3D structure of a protein combined with the functional data. The first attempt to provide an automated procedure for genome analysis was GeneQuiz [3]and Pedant [4], which reported that several ten % of the ORFs have at least one region linked to a known structure, by using only standard sequence search. Recent approach using fold recognition [5] or PSI-blast [8] assigned more high % of ORFs to a known structure. We are interested in maximising the number of relationships that can be identified on the basis of sequence analysis alone because these always precede the use of slower fold-recognition. Furthermore we want to provide a method that can be immediately applied to all finished genomes. Our work takes advantage of several developments: (1) CATH database [6]: which classifies homologues on the basis of structural and functional similarities. (2) The benefits of using intermediate sequences [2, 7] in sequence searching. (3) PSI-blast: the automated iterative search techniques [1] which allows more sensitive searches. This work describes an automated system that uses these developments to maximise the number of sequences that are linked to a known structure. We first describe how safe cutoffs were determined for the selected sequence alignment methods. Then we show how these methods were applied to 11 microbial genomes.

2 Making a flexible system for genome analysis

We determined optimal thresholds for the fasta, gapped-blast and PSI_blast. Using several matrices, fasta and gapped-blast were assessed by linking 841 proteins in in CATH with intermediates from the OWL sequence database. First, unrelated pairs were used to determine thresholds, and then related pairs employed to assess the sensitivity. We found that the most sensitive combinations for database searching were fasta/blosum62 and gapped-blast/blosum50, and, normalized thresholds for safe database searching were found to be 88 and 18, respectively. For PSI-blast, running each test sequence against a database consisting of both OWL and the 841 test sequences, we found that E-value thresholds of 0.01 and 0.001 both gave a specificity of around 99%. With these results in hand, we set about generating a procedure that could analyse large quantities of sequence data.combine as following steps: 1) All sequences from PDB were clustered into families and a representative selected. 2) Using PSI-blast, each PDB representative was run against OWL database. All aligned regions from sequences below the safer E, were stored together with information about PDB sequence associated with each aligned region (psi_PDB). 3) For each genome we took the sequences and using the gapped blast, searched the resulting sequence against psi_PDB. 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 11 genomes

Applying our system, we show how many ORFs could be linked to a PDB file by searching psi_PDB databases. All detailed data are available as an internet solution: Genius (http://www.hri.co.jp/genius). In this manner, information is available for each hit, when both the PDB/intermediate and intermediate/gene scores are beyond the required thresholds and emphasis that aligned region consisted of either >100 residues or >50% of the smaller sequence. In most cases, the results are more than double those of GeneQuiz and Pedant. For example, we can assign nearly 32% of the *M. genitalium* ORFs to a region of known structure. This compares favorably with the results of 12% from GeneQuiz, 16% from Pedant and 22% from frsvr [5]. Although Rychlewski et al. [8], report a higher sensitivity (38%), one should note that they ran PSI-Blast with a threshold ($E < 0.1$) without additional checks. It is valuable to look at the structures which are hit most frequently. We found that vast majority of hits is ab class structure, which are mostly corresponding to coenzyme. In structural terms, the results are unusual, because this class is not so major in PDB database. Another major hits correspond to P-loop motifs of ABC transporter, which do not pass the requirement of covering either >50% of the smaller sequence or >100 residues. It is almost certainly true that many of the P-loops in known structures may have arisen from convergent evolution, as their overall topologies can be quite different. Therefore in our opinion, there is currently no clear evidence for a homologous relationship between ABC transporters and a protein of known structure.

4 Future enhancements

Our automatically generated results are extremely competitive, even without requiring fold recognition techniques. Genomes can be calculated quickly and updated at regular intervals. In our future versions we hope to increase the sensitivity by taking full advantage of all the sequences available from sequencing projects.

References

[1] 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.

[2] Bork, P., Gellerich, J., Groth, H., Hooft, R., and Martin, F., Divergent evolution of a beta/alpha-barrel subclass: detection of numerous phospate-binding sites by motif search, *Protein Science*, 4:268–274, 1995.

[3] 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.

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

[5] 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.

[6] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M., CATH– a hierarchic classification of protein domain structures, *Structure*, 5:1093-1108, 1997.

[7] Park, J., Teichmann, S.A., Hubbard, T., and Chothia, C., Intermediate sequences increase the detection of homology between sequences, *J. Mol. Biol.*, 273:349-354, 1997.

[8] Rychlewski, L., Zhang, B. and Godzik, A., Fold and functions for *Mycoplasma genitalium* proteins, *Folding and Design*, 3:229-238, 1998.