D196–D199 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh043

The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms

Liam J. McGuffin, Stefano A. Street, Kevin Bryson, Søren-Aksel Sørensen and David T. Jones*

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Received August 11, 2003; Revised and Accepted September 22, 2003

ABSTRACT

Currently, the Genomic Threading Database (GTD) contains structural assignments for the proteins encoded within the genomes of nine eukaryotes and 101 prokaryotes. Structural annotations are carried out using a modified version of GenTHREADER, a reliable fold recognition method. The GenTHREADER annotation jobs are distributed across multiple clusters of processors using grid technology and the predictions are deposited in a relational database accessible via a web interface at http://bioinf.cs.ucl.ac.uk/GTD. Using this system, up to 84% of proteins encoded within a genome can be confidently assigned to known folds with 72% of the residues aligned. On average in the GTD, 64% of proteins encoded within a genome are confidently assigned to known folds and 58% of the residues are aligned to structures.

INTRODUCTION

Comprehensive and reliable annotation databases play an essential role in the interpretation and exploitation of the deluge of information resulting from genome sequencing projects.

A number of annotation resources, which include structural assignments for genes within complete genomes, have been developed over the past few years, such as GeneQuiz (1), PEDANT (2), MAGPIE (3) and GeneWeaver (4). These resources use an ensemble of methods to annotate biological sequences as well as utilizing BLAST (5) to assign structures to obvious sequence homologues.

More recently, dedicated structural annotation databases including 3D-GENOMICS (6) and Gene3D (7) have been developed which primarily use PSI-BLAST (8) in order to assign protein folds to more distantly related protein sequences.

The Genomic Threading Database (GTD) is a new dedicated structural annotation database available on the web at http://bioinf.cs.ucl.ac.uk/GTD. It differs from 3D-GENOMICS and Gene3D in that GenTHREADER (9,10) is the key part of the annotation system, which is a more reliable, sensitive and selective method for detecting remote homology between protein sequences and known folds. In addition, grid technology is harnessed to speed up the process of accurately assigning structures to proteins from complete proteomes.

METHODS

GenTHREADER

GenTHREADER is a widely used protein fold recognition method which is available on the PSIPRED server (11), intended to predict the folds of individual protein sequences with distant homology to known structures. A distributed and improved version (10) of the method was implemented for the GTD in order to ensure relatively fast and reliable annotation of whole proteomic sequences.

The GenTHREADER method consists of a feed-forward neural network which is trained to combine sequence alignment scores, length information, pairwise and solvation potentials derived from threading into a single score representing the likelihood of an evolutionary relationship between two proteins. The recently improved version also makes use of profiles seeded by structural alignments, bidirectional scoring and PSIPRED predicted secondary structure in order to maximize reliability and coverage (10).

Distribution system

The trade-off for more reliability and coverage is a slow down in the speed at which predictions are made. To counter this, annotation jobs are distributed across clusters of processors at University College London and Imperial College, London using grid technology.

The Globus toolkit—GT2 (http://www.globus.org)—is used to communicate between the remote sites. GT2 provides security between sites and secure job submission. In conjunction with the GT2, the jobs are scheduled using Sun Grid Engine (SGE, http://www.sun.com/gridware). The combination and use of these technologies results in a markedly improved throughput performance.

*To whom correspondence should be addressed. Tel: +44 20 7679 7982; Fax: +44 20 7387 1397; Email: dtj@cs.ucl.ac.uk

Measuring the reliability of the annotation using *p*-values

It is essential to provide a quantitative measure of the confidence we have in any particular fold assignment. For this we used a similar approach to that of BLAST, based on hypothesis testing. We determined the statistical significance of obtaining a fold match with a given score or better when compared with a null model. Our null model is that a match of this score has occurred by chance and does not actually signify that the sequence has the specified fold. Clearly the alternative model is that the match score is due to the sequence actually having the given fold.

In more detail, we generated random pairings of sequences which are known not to have the same fold. Applying GenTHREADER to these provided a score distribution for the null model, which was modelled by a generalized extreme value distribution using R statistical package (12) with the 'evd' library. This allowed us to determine the statistical significance of any score using a one-sided test based on this distribution. The statistical significance, or *p*-value, obtained tells us the likelihood of the fold being incorrectly assigned. Finally, we form the following confidence ranges based on the *p*-value: certain ($0 \leqslant p < 0.01\%$), high ($0.01\% \leqslant p < 0.1\%$), medium ($0.1\% \leqslant p < 1\%$), low ($1\% \leqslant p < 10\%$) and guess ($p \geqslant 10\%$).

Database interface and comparison

GenTHREADER was used to predict the folds of proteins within the genomes of *Homo sapiens*, *Mus musculus*, *Anopheles gambiae*, *Drosophila melanogaster*, *Oryza sativa*, *Caenorhabditis elegans*, *Fugu rubripes*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and over 100 prokayotes. Proteomic sequences were downloaded from a variety of sources; version numbers and the locations of the sequences are listed at http://www.e-protein.org.

The resulting predictions and corresponding sequence alignments were uploaded into tables within a MySQL relational database (http://www.mysql.com). A web interface was developed to allow users to search the database.

The database was queried to produce summary tables displaying coverage of sequences and residues (see website for a summary of predictions). Summary statistics from the GTD were then compared with PSI-BLAST-based annotations of 14 organisms from the 3D-GENOMICS website (http://www.sbg.bio.ic.ac.uk/3dgenomics). Where it was possible, the versions of sequences and strains of organism were kept the same.

RESULTS

Comparison of coverage by databases

On 8 August 2003, the total number of organisms annotated in the GTD exceeded that of Gene3D and 3D-GENOMICS (Table 1).

Table 1. The numbers of organisms annotated by each database that are available to search via the web

	Gene3D	3D-GENOMICS	GTD
Prokaryotes	64	84	101
Eukaryotes	2	9	9

Both the percentage of sequences assigned to structures and the percentage of residues aligned in the GTD was seen to be greater than that of PSI-BLAST-based annotation databases such as 3D-GENOMICS (Fig. 1a and b).

In the GTD, a maximum of 84% of sequences are assigned at a *p*-value of <1% and over 72% of the residues are aligned, as shown in the case of *Escherichia coli* K12 (Fig. 1a and b). For all organisms assigned in the GTD, on average, 64% of sequences are confidently assigned to known folds and 58% of the residues can be confidently aligned.

The frequency of folds assigned

Figure 2 shows the frequency of the top 10 fold types assigned to proteins encoded by four representative genomes in the GTD: a multicelled eukaryote, *H.sapiens*; a single-celled eukaryote, *S.cerevisiae*; a eubacterium, *E.coli* K12 and an archeabacterium, *Methanococcus jannaschii.*

For the single-celled organisms, the P-loop-containing nucleotide triphosphate hydrolases is found to be the most commonly occurring SCOP (13) fold group. Conversely, in the human proteome, immunoglobulin-like β sandwiches and C2H2 and C2HC zinc fingers are the two most commonly occurring fold groups. In general, multicellular organisms are shown to have similar rankings of frequencies of the top fold types to human. This observation is in agreement with the findings of Müller *et al.* (6).

DISCUSSION

The GTD provides a comprehensive, dedicated resource of reliable structural annotations of the proteins encoded by the genomes of over 100 recently sequenced organisms. By using the GenTHREADER method for genomic scale fold recognition we are able to assign more proteins with distant sequence homology to known folds than could be assigned using simple sequence-based methods such as PSI-BLAST. In addition, the use of grid technology greatly increases the rate at which newly released genomes can be annotated and at which old annotations can be updated.

The structural annotations in the 3D-GENOMICS and Gene3D databases are both currently based primarily on PSI-BLAST searches. However, it is anticipated that both of these databases will employ fold recognition analysis of whole proteomes in the near future.

Since more accurate protein structure predictions can be gained from a consensus of methods (14) it is pertinent to combine the results from several annotation databases together. The e-Protein project (http://www.e-protein.org) is a pilot initiative which proposes to combine structural and functional annotation databases from University College London, Imperial College London and the European Bioinformatics Institute, through a single interface using the Distributed Annotation System (15). In addition, it is proposed that the workload of all annotation jobs will eventually be distributed across processing clusters at each site using a similar grid system to the GTD.

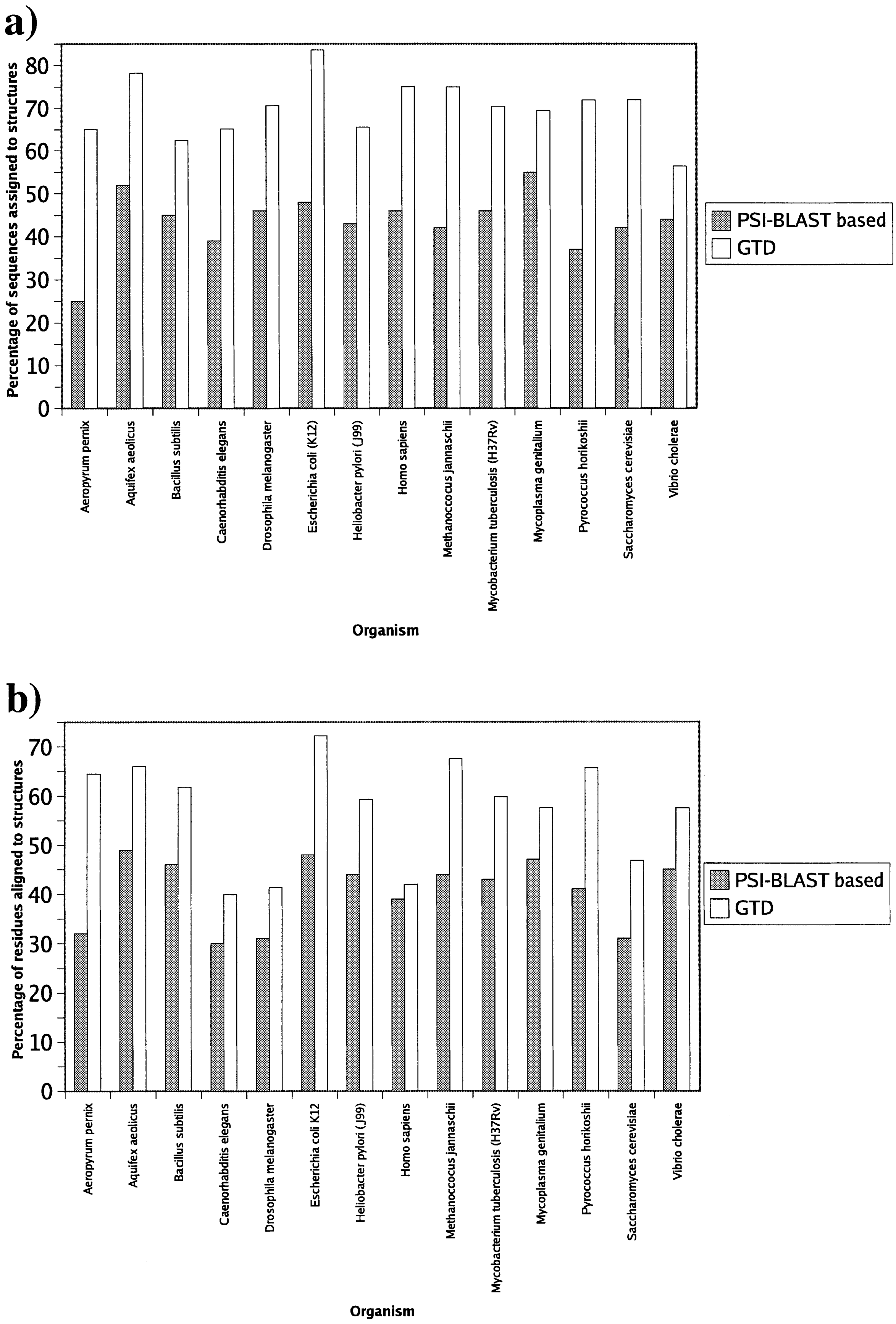

Figure 1. The difference in coverage of structural assignments between the PSI-BLAST based annotations from 3D-GENOMICS, and those in the GTD. (**a**) The coverage of sequences assigned to structures. In the case of the GTD, only GenTHREADER assignments with p < 1% have been counted (see Methods). (**b**) The coverage of residues aligned to sequences.

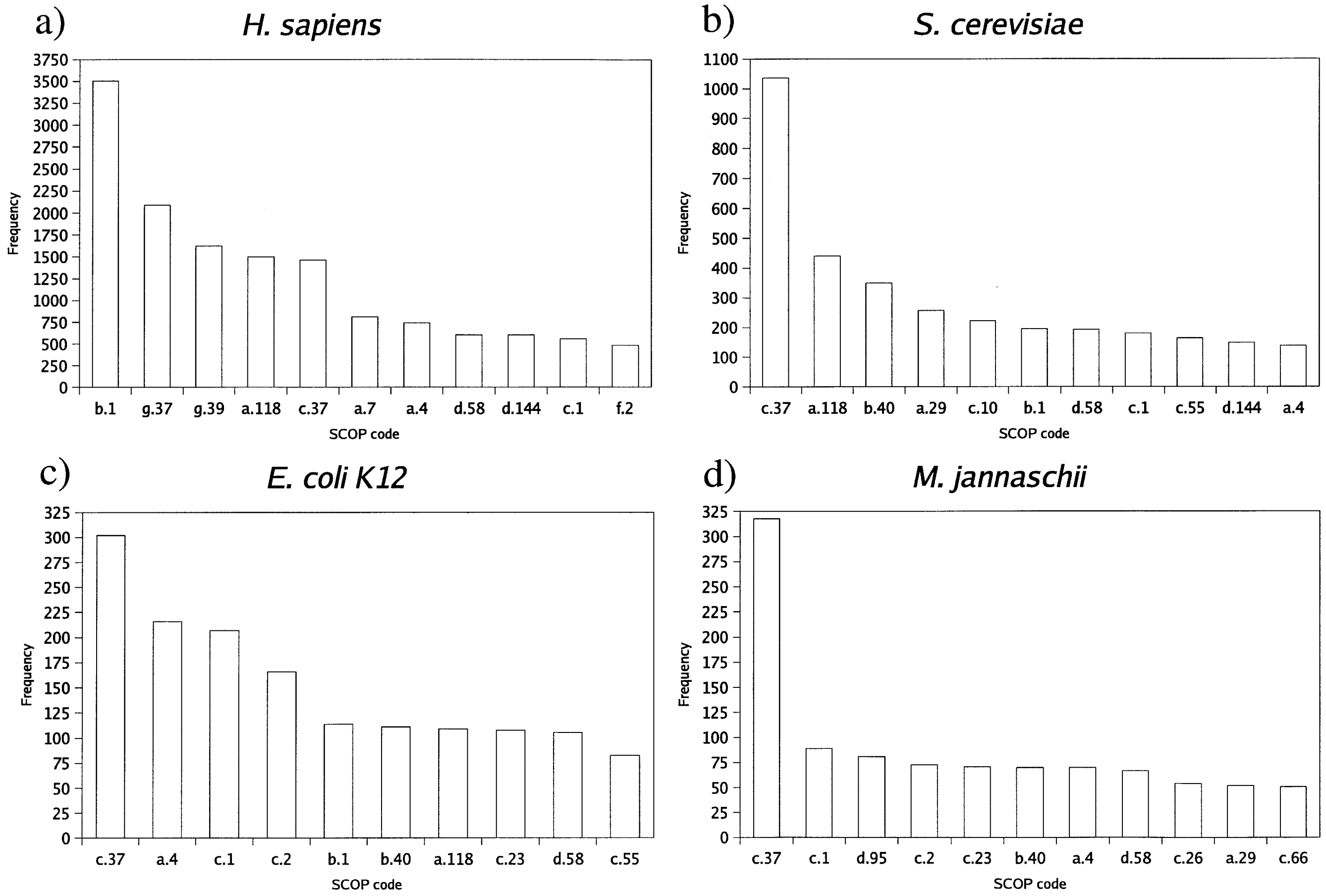

Figure 2. The frequency of the different fold categories of proteins encoded within four representative genomes. Each SCOP (13) code relates to the following folding types: a.4, DNA/RNA-binding 3-helical bundle; a.7, spectrin repeat-like; a.29, bromodomain-like; a.118, α-α superhelix; b.1, immunoglobulin-like β-sandwich; b.40, OB-fold; c.1, TIM β/α-barrel; c.2, NAD(p)-binding Rossmann-fold domains; c.10, leucine-rich repeat; c.23, flavodoxin-like; c.26, adenine nucleotide α hydrolase-like; c.37, P-loop-containing nucleotide triphosphate hydrolases; c.55, ribonuclease H-like motif; c.66, *S*-adenosul-L-methionine-dependent methyltransferases; d.58, ferredoxin-like; d.95, homing endonuclease-like; d.144, protein kinase-like (PK-like); f.2, membrane all-α; g.37, C2H2 and C2HC zinc fingers; g.39, glucocorticoid receptor-like (DNA-binding domain).

ACKNOWLEDGEMENTS

This work was supported by the UK Biotechnology and Biological Sciences Research Council and the UK Department of Trade and Industry.

REFERENCES

1. Hoersch,S., Leroy,C., Brown,N.P. andrade,M.A. and Sander,C. (2000) The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem. Sci.*, **25**, 33–35.
2. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al.* (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.
3. Gaasterland,T., Sczyrba,A., Thomas,E., Aytekin-Kurban,G., Gordon,P. and Sensen,C.W. (2000) MAGPIE/EGRET annotation of the 2.9-Mb *Drosophila melanogaster* Adh region. *Genome Res.*, **10**, 502–510.
4. Bryson,K., Luck,M., Joy,M. and Jones,D.T. (2000) Applying agents to bioinformatics in GeneWeaver. *Lect. Notes Artif. Int.*, **1860**, 60–71.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
6. Muller,A., MacCallum,R.M. and Sternberg,M.J.E. (2002) Structural characterization of the human proteome. *Genome Res.*, **12**, 1625–1641.
7. Buchan,D.W., Rison,S.C., Bray,J.E., Lee,D., Pearl,F., Thornton,J.M. and Orengo,C.A. (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.*, **31**, 469–473.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
9. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.
10. McGuffin,L.J. and Jones,D.T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–881.
11. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
12. Ihaka,R. and Gentleman,R. (1996) R: A language for data analysis and graphics. *J. Comp. Graph. Stat.*, **5**, 299–314.
13. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
14. Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19** 1015–1018.
15. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.