294–298 Nucleic Acids Research, 2002, Vol. 30, No. 1

GTOP: a database of protein structures predicted from genome sequences

Takeshi Kawabata[1], Satoshi Fukuchi[1,2], Keiichi Homma[1,2], Motonori Ota[1], Jiro Araki[3], Takehiko Ito[3], Nobuyuki Ichiyoshi[3] and Ken Nishikawa[1,*]

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1-111 Yata, Mishima, Shizuoka 411-8540, Japan, [2]Japan Science and Technology Corporation, 1–8 Honcho, 4-chome, Kawaguchi City, Saitama, 332-0012, Japan and [3]Mitsubishi Research Institute, Inc., 3–6 Otemachi, 2-chome, Chiyoda-ku, Tokyo 100-8141, Japan

Received August 14, 2001; Revised and Accepted October 2, 2001

ABSTRACT

Large-scale genome projects generate an unprecedented number of protein sequences, most of them are experimentally uncharacterized. Predicting the 3D structures of sequences provides important clues as to their functions. We constructed the Genomes TO Protein structures and functions (GTOP) database, containing protein fold predictions of a huge number of sequences. Predictions are mainly carried out with the homology search program PSI-BLAST, currently the most popular among high-sensitivity profile search methods. GTOP also includes the results of other analyses, e.g. homology and motif search, detection of transmembrane helices and repetitive sequences. We have completed analyzing the sequences of 41 organisms, with the number of proteins exceeding 120 000 in total. GTOP uses a graphical viewer to present the analytical results of each ORF in one page in a 'color-bar' format. The assigned 3D structures are presented by Chime plug-in or RasMol. The binding sites of ligands are also included, providing functional information. The GTOP server is available at http://spock.genes.nig.ac.jp/~genome/gtop.html.

INTRODUCTION

As numerous genome projects give us an enormous volume of hypothetical amino acid sequences, their functional analyses become crucial in the post-sequence era. An effective approach to characterizing protein functions is to determine their 3D structures first. Even if the 3D structure cannot be definitively assigned to a protein, it can be predicted when its homologs are found in 3D structure databases. A rapid accumulation of 3D structure data (1), combined with development of sensitive homology search methods (2), enables us to predict a large portion of protein sequences (3,4). Databases of predicted protein structures against a large number of protein sequences will be useful in annotating their functions. Recently, a large-scale structure determination project, called 'structural genomics' was initiated by several organizations worldwide (5). The enterprise aims to understand molecular functions of uncharacterized proteins through solution of their 3D structures. As the total number of such proteins is almost unmanageably large, a practical goal of the undertaking is to determine the structures of all the representative proteins in each family. If necessary, the structures of the rest can be predicted computationally using the representative structures. Therefore, in order to gain the most from structure genomics, we need to develop a database containing predictions of all the available protein sequences.

We report here development of the Genomes TO Protein structures and functions database (GTOP), containing an extensive repository of protein fold predictions (sequence versus structure alignments) obtained chiefly by the program PSI-BLAST (6). GTOP also provides the outcome of other analyses such as homology and motif search, detection of transmembrane helices, coiled-coil region and repetitive sequences, among others. These data, combined with predicted 3D structures, constitute effective tools in characterizing protein functions.

Recently, several other databases of automatic sequence analysis were developed. For example, the SUPERFAMILY database (http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/) provides a large collection of fold assignments using hidden Markov models, MODBASE (7) and 3DCrunch (http://www.expasy.ch/swissmod/SM_3DCrunch.html) give comparative protein structure models with full atoms. However, the GTOP database differs from them in that they focus exclusively on 3D structures, without providing any other analyses. Despite the conceptual similarity to other general databases of automatic analyses, such as PEDANT (8), GeneQuiz (9) and Genecensus (http://bioinfo.mbb.yale.edu/genome/tree/tree.cgi), GTOP has the advantage of supplying a well-designed interface dealing with predicted 3D structures.

TARGET PROTEIN SEQUENCES

In principle, GTOP aims to analyze the entire protein sequences in all the completely sequenced genomes. The current version of GTOP provides analyses of 41 organisms: 28 eubacteria, 8 archaebacteria, 3 eukaryotes, bacteriophage

*To whom correspondence should be addresssed. Tel: +81 559 81 6859; Fax: +81 559 81 6889; Email: knishika@genes.nig.ac.jp

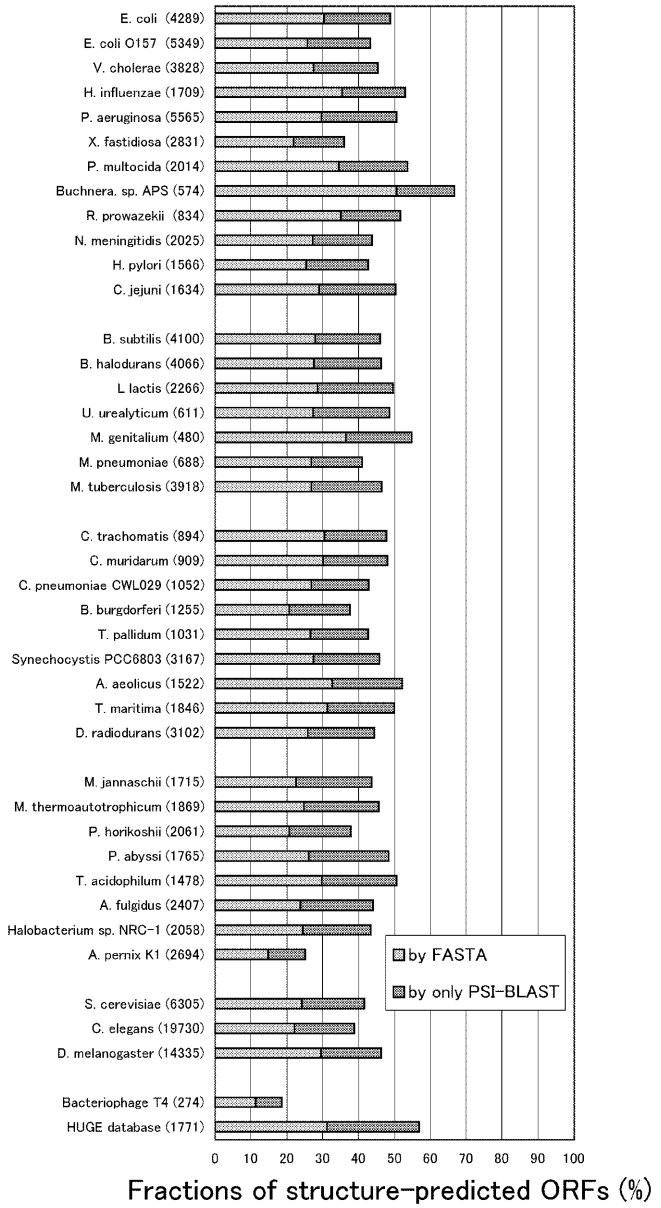

Figure 1. Fractions of structure-predicted ORFs in GTOP database. The digits in parentheses are the numbers of ORFs in the corresponding organism.

T4 and human cDNA data compiled from the HUGE database (10). The organisms are listed on the left of Figure 1. The sequences of *Synechocystis* and *Caenorhabditis elegans* were obtained from Cyanobase (11) and Wormpep (12), respectively, while the rest of the sequences were downloaded from the NCBI FTP site (ftp://ncbi.nlm.nih.gov/genbank/genomes/).

STRUCTURE PREDICTIONS CONTAINED IN GTOP

Table 1 catalogs the kinds of analyses whose results are included in the GTOP database. The 3D structure was mainly predicted by PSI-BLAST (6) detecting weak sequence homologies against proteins of known structure. Profiles used in PSI-BLAST were constructed for each open reading frame (ORF) sequence from a non-redundant sequence database, as described previously (13). Using these profiles, we scanned the sequences of the Protein Data Bank (PDB; 1) to predict their structures. We also scanned the domain sequences of the SCOP database (14) to assign SCOP taxonomy to each ORF. The domain sequences were collected using the information stored in the 'dir.dom' file in the SCOP database. We also applied PSI-BLAST in the reverse direction: we constructed profiles for each of the representative domains in the SCOP database and scanned these profiles using the Reverse PSI-BLAST program in the BLAST program package. The combination of the two directions of PSI-BLAST has been reported to result in prediction with higher sensitivity than either direction alone (3,4). Moreover, we predicted regions with special structures: transmembrane helices by SOSUI (15), coiled-coil regions by MULTICOIL (16) and low-complexity regions by SEG (17). Prior to PSI-BLAST application, we masked these regions to avoid making erroneous predictions. We also utilized the standard pairwise sequence search FASTA (18) or BLAST (6), to confirm the results of PSI-BLAST for pairs with relatively high similarity.

The fractions of structure-predicted ORFs are shown in Figure 1. We used the PDB updated on March 2001 and set the E-value threshold to 0.001. On average, we could predict structures for 42.3% of ORFs in a total of 41 organisms: FASTA predicted ~26.5% of the ORFs, while PSI-BLAST added another 15.8%. These figures are a little larger than those reported previously (2,3). The increase is mostly attributable to the fact that we used a later version of the PDB. In the case of *Escherichia coli*, the fraction of structure-predicted ORFs in April 2000 was 45.8%, whereas that in March 2001 was 48.9%. That is, accumulation of PDB data in ~1 year increased the predicted fraction by 3.1%. The observation demonstrates that we need to frequently update this kind of database to maintain its value.

OTHER ANALYSES CONTAINED IN GTOP

Besides structure prediction, GTOP contains results of numerous analyses (Table 1). PSI-BLAST search against the well-annotated sequence database SWISS-PROT (19) was also performed using the profiles constructed for fold prediction. As detection of repetitive sequences in proteins is important in determining their domain structures, RepAlign, a program specifically developed in our group for this purpose, was put to use. Furthermore, we performed a motif search against PROSITE (20) and a exploration of known domains in Pfam (21) using HMMER (http://hmmer.wustl.edu). A BLAST search within all the sequences in GTOP generates an organism pattern, i.e. the numbers of homologs in the 41 organisms (shown as 'OrgPattern' in Fig. 2A). This information is valuable in examination of the phylogeny of a protein and prediction of its function (22,23).

OVERVIEW OF THE DATABASE

GTOP data are stored in two kinds of files, namely raw result files and flat summary files. All the raw outputs of each analysis method, such as BLAST, FASTA, PSI-BLAST and HMMER, are stored in the GTOP database. Due to a very large size, the raw files cannot be quickly searched and viewed. To

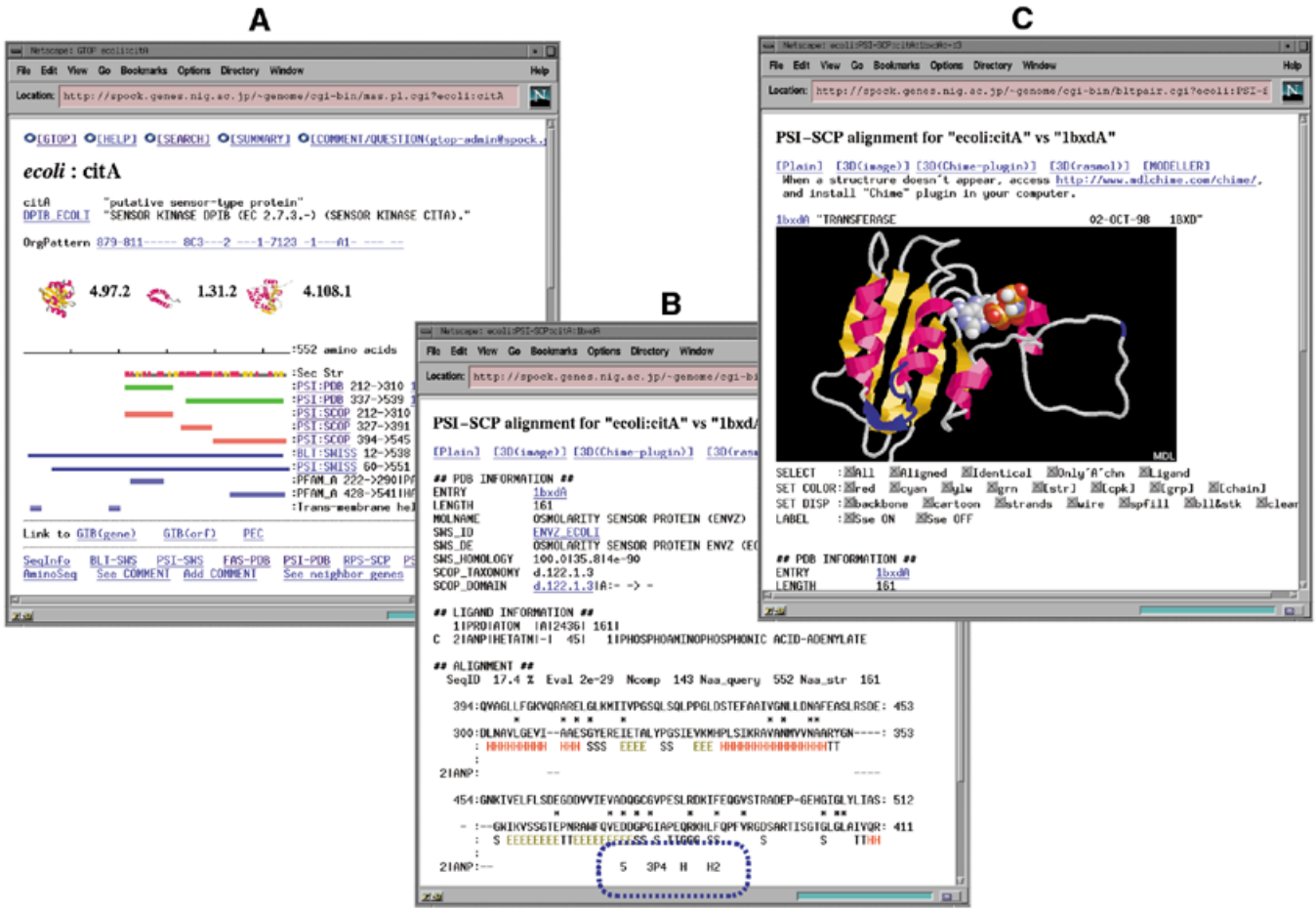

Figure 2. Snapshots of GTOP web pages. (**A**) The summary page of citA, an ORF in *E.coli*. Results of various kinds of analysis are shown in colored bar formats. If SCOP structure domains are assigned, small structure icons are shown above the colored bars. The structure icons were prepared for each superfamily of SCOP. The string indicated by 'OrgPattern' is the number of homologs of the given protein in the organisms stored in GTOP database. (**B**) An alignment of citA and the 3D structure of its homolog 1bxdA found by PSI-BLAST. The characters under the alignment (shown in a blue box) indicate the numbers of atomic contacts between a given residue and the ligand, ANP. (**C**) A Chime plug-in view of the 3D structure of 1bxdA, a homolog of citA. The regions of insertion and deletion are colored blue.

Table 1. Analyses contained in GTOP database

Analysis	Program	Target database
Fold prediction	FASTA (18), PSI-BLAST (6)	PDB (1)
Assignment of SCOP domain	PSI-BLAST, Reverse PSI-BLAST	SCOP (14)
Homology search versus well-annotated proteins	BLAST (6), PSI-BLAST	SWISS-PROT (19)
Transmembrane helices	SOSUI (15)	–
Coiled-coil region	MULTICOIL (16)	–
Low-complexity region	SEG (17)	–
Motif	(In-house program)	PROSITE (20)
Known domains	HMMER	Pfam (21)
Repetitive sequence	(In-house program)	–
Organism pattern	BLAST (6)	Sequences in GTOP

facilitate search, we introduced flat summary files (called the 'master files'), which are a brief extract of all the analytical results of ORFs. In the case of homology searches, the master files store only family names and representative entry names from a large list of homologous proteins. A keyword search against the master files is available on the web server.

VIEWERS OF THE RESULTS

A graphical user interface is a necessary tool in browsing a database containing an enormous amount of data. We developed several such tools to browse the GTOP database via the World Wide Web. A viewer of master files shows information of master files in a color-bar format. A screen view of the *E.coli* citA gene is presented in Figure 2A. From this view, we can easily grasp the overall structure of each ORF: the location of transmembrane helices, predicted 3D structures, Pfam domains, to name a few. Links to other related databases, such as HUGE (http://www.kazusa.or.jp/huge/), PEC (http://www.shigen.nig.ac.jp/ecoli/pec/) and GIB (http://gib.genes.nig.ac.jp/) are placed in this page. Upon clicking a link in the master file view, another view appears, providing more detailed information. Figure 2B shows an alignment between a query protein and a homologous structure found by PSI-BLAST. We developed a device to identify ligand binding sites (encircled by a blue dotted line in Fig. 2B). Besides the usefulness of this information in predicting binding sites of proteins, we can check functional similarity through examinations of residue conservation in binding sites. A predicted structure can be displayed in the browser with the use of a Chime plug-in (http://www.mdlchime.com/chime/) (Fig. 2C). We prepared two additional ways to show structures: by image and using RasMol (http://www.OpenRasMol.org/) as an external application. Aligned regions are colored differently. The GTOP database provides alignments between sequences and 3D structures, but not the coordinates of all the atoms. For users who want to build full atom models, we developed a tool with which to make MODELLER (24) input files.

SEARCH AND SUMMARY PAGE

A search tool is very important in making large databases like GTOP accessible. We prepared two search devices. The first one is a keyword search, which scans the master file. A gene name, ORF name or family name can be used as a query. The second one is the sequence homology search using BLAST. Despite their usefulness, they are unsuitable for comparisons of organisms from multiple view points. We therefore prepared a summary page, which shows in one line the numbers of various domains in organisms, such as SCOP folds, Pfam families, PROSITE motifs and transmembrane helices.

PREDICTION SERVICE

In addition to the results of completed analyses, we provide an automatic structure prediction service using BLAST and Reverse PSI-BLAST. If a user pastes a sequence of interest into the form on the web page, the protein structures of its homologs will promptly be returned. This service is important for two reasons. First, GTOP cannot cover all kinds of proteins, however hard we try to expand the number of target proteins. Secondly, GTOP cannot reanalyze the PDB data every week, as too much computational costs would be incurred, while the PDB itself is updated weekly. On both of these counts, the prediction service can remedy the intrinsic defect in the ready-made portion of GTOP.

FUTURE DIRECTION

Because of the ever increasing nature of both genome sequences and target databases (such as the PDB and SWISS-PROT), maintaining the GTOP database is no easy task. We are now modifying the structure of the GTOP database to decrease the computational costs for updating and to store results of a larger number of analyses. One way to decrease computational costs is to omit one of the two directions of PSI-BLAST, as doing so can dramatically decrease the computational time with a relatively small loss in sensitivity. As mentioned above, the prediction service can compensate for the delay in updating. We also intend to include more organisms, especially eukaryotes, which will further enhance the value of the database.

ACKNOWLEDGEMENTS

We are grateful to Drs Osamu Ohara and Reiko Kikuno, who provided us with the latest sequences of the HUGE database and gave us useful advice. We also thank Drs Akiko Nishimura and Hiroshi Nakashima for helpful biological insights, and Dr Shigeki Mitaku for the SOSUI program. This work was supported by the ACT-JST Program, a grant-in-aid from the Japan Science and Technology Corporation (JST), and grant-in-aid for scientific research on priority areas of 'Genome Information Science' of the Ministry of Education, Science, Sports and Culture. T.K. is a research fellow of the Japan Society for the Promotion of Science.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.
2. Park,J., Karplus,K., Barret,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparison using multiple sequences detect three times as many homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
3. Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res.*, **9**, 17–26.
4. Teichmann,S.A., Chothia,C. and Gerstein,M. (1999) Advances in structural genomics. *Curr. Opin. Struct. Biol.*, **9**, 390–399.
5. Burley,S.K. (2001) An overview of structural genomics. *Nature Struct. Biol.*, **7** (suppl.), 932–934.
6. 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.
7. Sanchez,R., Peiper,U., Mirkovic,N., de Bakker,P.I.W., Wittenstein,E. and Sali,A. (2000) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **28**, 250–253. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 255–259.
8. Frishman,D., Albermann,K., Jani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.-W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.
9. Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
10. Kikuno,R., Nagase,T., Suyama,M., Waki,M., Hirosawa,M. and Ohara,O. (2000) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **28**, 331–332. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 166–168.
11. Nakamura,Y., Kaneko,T., Hirosawa,M., Miyajima,N. and Tabata,S. (1998) Extension of CyanoBase. CyanoMutants: repository of mutant information on *Synechocystis* sp. strain PCC6803. *Nucleic Acids Res.*, **26**, 63–67.

12. Sonnhammer,E.L. and Durbin,R. (1997) Analysis of protein domain families in *Caenorhabditis elegans*. *Genomics*, **46**, 200–216.
13. Kawabata,T., Arisaka,F. and Nishikawa,K. (2000) Structural/functional assignment of unknown bacteriophage T4 proteins by iterative database searches. *Gene*, **259**, 223–233.
14. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
15. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
16. Wolf,E., Kim,P.S. and Berger,B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.*, **6**, 1178–1189.
17. Wooton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
18. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
19. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
20. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 235–238.
21. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
22. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
23. Pellegrini,M., Marcotte,E.M., Thompson.,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
24. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.