Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W393–W397***
doi:10.1093/nar/gkm236

DBAli tools: mining the protein structure space

Marc A. Marti-Renom[1,*], Ursula Pieper[2], M. S. Madhusudhan[2], Andrea Rossi[2], Narayanan Eswar[2], Fred P. Davis[2], Fátima Al-Shahrour[3], Joaquín Dopazo[3] and Andrej Sali[2]

[1]Structural Genomics Unit, [2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94158-2330, USA and [3]Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain

Received January 30, 2007; Accepted March 31, 2007

ABSTRACT

The DBAli tools use a comprehensive set of structural alignments in the DBAli database to leverage the structural information deposited in the Protein Data Bank (PDB). These tools include (i) the DBAlit program that allows users to input the 3D coordinates of a protein structure for comparison by MAMMOTH against all chains in the PDB; (ii) the AnnoLite and AnnoLyze programs that annotate a target structure based on its stored relationships to other structures; (iii) the ModClus program that clusters structures by sequence and structure similarities; (iv) the ModDom program that identifies domains as recurrent structural fragments and (v) an implementation of the COMPARER method in the SALIGN command in MODELLER that creates a multiple structure alignment for a set of related protein structures. Thus, the DBAli tools, which are freely accessible *via* the World Wide Web at http://salilab.org/DBAli/, allow users to mine the protein structure space by establishing relationships between protein structures and their functions.

INTRODUCTION

The number of known protein structures deposited in the Protein Data Bank (PDB) has grown exponentially over the years (1). This trend is expected to continue, partly due to the structural genomics efforts (2,3). Currently, there are ~41 000 protein structures deposited in the PDB, containing ~88 000 protein chains. These protein structures constitute a structural space that can be mined to facilitate the understanding, assignment and modification of protein function. Previously developed databases for the classification of protein structure domains, such as SCOP [http://scop.mrc-lmb.cam.ac.uk/scop/ (4)] or CATH [http://www.cathdb.info (5)], and servers for functional annotation of protein structures, such as ProFunc [http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/ (6,7)], ProKnow [http://www.doe-mbi.ucla.edu/Services/ProKnow (8)] and Phunctioner [http://www.sbg.bio.ic.ac.uk (9)], provide an effective way of describing and annotating the protein structure space. However, none of these servers combine a comprehensive database of protein structural alignments with tools for automatically annotating protein structures.

Here, we describe five tools that aid in the analysis of the data stored in DBAli, our comprehensive relational database of pairwise and multiple structural alignments (10). These tools include (i) the DBAlit program that allows users to input their structure for comparison by MAMMOTH (11) against all chains in the PDB; (ii) the AnnoLite and AnnoLyze programs that annotate a target structure based on its stored relationships to other structures; (iii) the ModClus program that clusters structures by sequence and structure similarities; (iv) the ModDom program that identifies recurrent fragments, including domains, from structure; and (v) an implementation of the COMPARER method (12) in the SALIGN command in MODELLER that creates a multiple structure alignment for a set of related protein structures. The DBAli tools allow users to establish relationships between protein structures and their fragments in a flexible and dynamic manner.

The DBAli database is briefly introduced first. Next, we describe each of the five tools that make use of the structural alignments deposited in DBAli. Finally, we discuss the use of the DBAli tools to analyze a structure determined by the New York Structural Genomics Research Consortium (NYSGXRC).

*To whom correspondence should be addressed. Tel: +34 96 3289680; Fax: +34 96 3289701; Email: mmarti@cipf.es

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

THE DBAli SERVER

The DBAli server (http://salilab.org/DBAli/) is divided into four main sections: the DBAli database, search pages, tools pages and special pages, each one of which is dedicated to a specific tasks (Figure 1).

The DBAli database

The DBAli database contains pairwise and multiple structure alignments of protein structures in the PDB. Pairwise alignments are updated weekly and multiple alignments are updated monthly. Currently (January 2007), DBAli contains a total of 86,277 PDB chains in ~1.38 billion pairwise alignments with a MAMMOTH *P*-value higher than 2.0 (Table 1). DBAli also stores multiple structure alignments for 11,615 families with 30,900 non-redundant PDB chains representing 86,277 chains in PDB (Table 1). Structural and functional annotations are obtained from our databases LigBase (13) and PIBASE (14) and external databases such as CATH (5), SCOP (4), InterPro (15), PFam (16), EC (17) and GO (18). Cross-links between the external databases is adopted from the MSD database (19) (Table 1).

The DBAli tools

DBAli incorporates five tools that use the structural relationships in DBAli.

DBAlit. The DBAlit program takes a user input structure in the PDB format and compares it against all chains in the PDB using the MAMMOTH algorithm. On average, a user-input structure is compared against all known protein structure chains in ~200 min using 10 Xeon CPUs. To preserve the privacy of the coordinates, DBAli generates a random and unique chain identifier that is returned to the user by e-mail. The new chain identifier

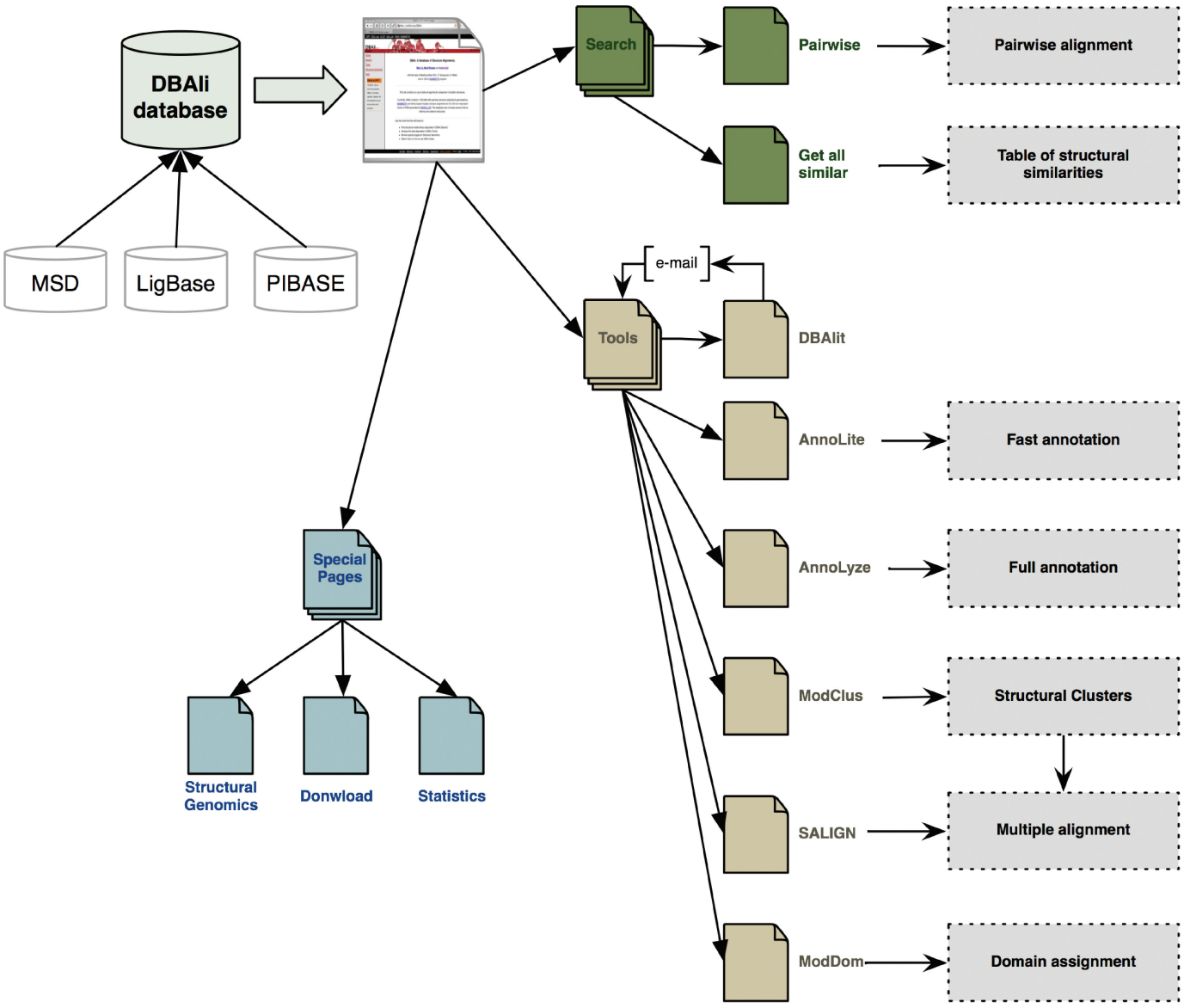


Figure 1. DBAli server organization.

can then be used to retrieve data from the DBAli database as well as to use the other tools outlined below.

AnnoLite and AnnoLyze. The AnnoLite and AnnoLyze programs (20) annotate the structures processed by DBAli. A given structure is characterized by annotations of related structures in the MODBASE database of comparative models (21), the PIBASE database of structurally defined protein interfaces (14), the LIGBASE database of small molecule binding sites (13) and the MSD database of macromolecular structures (19). The inputs for the annotation programs are a PDB chain code and thresholds for filtering sequence and structure similarities. The AnnoLite program searches the DBAli database for structurally similar proteins and collects their known annotations. Next, a *P*-value score is calculated for each transferred annotation using a Fisher's exact test for 2×2 contingency tables comparing two groups of annotated chains (i.e. the group of similar chains to the query and the group of all annotated chains in the PDB)(22). Currently, AnnoLite annotates the input protein structure with CATH (5) and SCOP (4) fold assignments, EC numbers (17), InterPro entries (15), PFAM families (16) and Gene Ontology codes (23). The accuracy and coverage of AnnoLite were benchmarked with a set of fully annotated 1,879 non-redundant PDB chains. AnnoLite can reliably annotate a structure for all of the functional properties, with the exception of the GO cellular component term. For example, the CATH fold can be recovered for 96% of the dataset with 89% reliability and direct functional annotation with EC numbers and Gene Ontology molecular function codes can be recovered with reliabilities of 81 and 74% for 83 and 88% of the dataset, respectively. Additionally, AnnoLyze inherits ligands from LIGBASE and interacting partners from PIBASE. The output from the two programs provides an automatic annotation of the protein structures.

ModClus. The ModClus program clusters protein structures based on their sequence and structure similarities. The input to ModClus is a list of PDB chain codes. The output is a list of clusters of the input chains. The clustering depends on user-defined thresholds for structure and sequence similarities. ModClus implements a greedy algorithm as follows: (i) the first chain in the list seeds the first cluster; (ii) the next chain is compared by sequence and/or structure to all chains in each of the existing clusters; it either joins the first sufficiently similar cluster or seeds a new cluster if it is not sufficiently similar to all of the chains in any of the other clusters; (iii) the clustering continues with step (ii) until all chains are processed.

ModDom. The ModDom program assigns domain boundaries in a given structure using the superpositions stored in DBAli. The input is a PDB chain code, which is used as a query to identify structurally similar chains in DBAli. ModDom relies on the relationship between recurrent structures and structural units to predict domain boundaries. The program first builds a residue co-occurrence matrix based on structural alignments selected from the DBAli database and then clusters residue co-occurrences to find common fragments in the query protein structure. The ModDom program has been benchmarked for domain assignment with a non-redundant set of protein structures. ModDom assigns 80% of residues identically to the domain assignments in the SCOP database. The user is provided with all domain assignments and their scores, a structural conservation profile based on the retrieved alignments with similar structures, and a JMol window for inspection of the domain assignments.

SALIGN. The SALIGN command, an implementation of the COMPARER program (12) in MODELLER, generates a multiple structure alignment given a list of PDB chain codes. The alignment is displayed through the JMol applet. In addition, an HTML frame presents an easy-to-read sequence alignment corresponding to the structural superposition. The user is provided with options to download the alignment in the HTML or PIR formats and the superposed coordinates in the PDB format.

UTILITY AND DISCUSSION

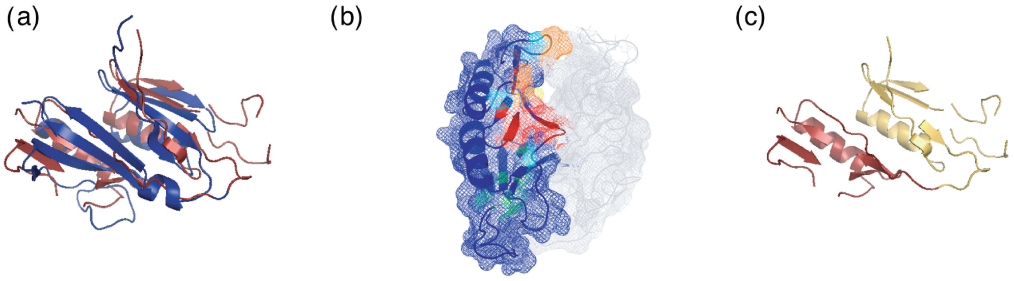
Target selection strategies for structural genomics have led to the experimental determination of many protein structures whose functions are not yet known. The tools in DBAli can be employed to annotate the functions of such structures, as illustrated by the following example.

The New York Structural Genomics Research Consortium (NYSGXRC) selected a hypothetical protein from *Pseudomonas aeruginosa* as a target for structure determination (target T1794). The structure was

Table 1. Internal and external data sources used by the DBAli tools

Database	Last update	Type of information	Number of entries	Reference
DBAli pairwise	January 2007	Structural alignments	1,379,352,642	(10)
DBAli multiple	January 2007	Structural alignments	11,615	(10)
MSD	November 2006	PDB chains	79,170	(19)
CATH	July 2006	Superfamilies	2,359	(5)
SCOP	April 2006	Domains	94,779	(4)
InterPro	July 2006	Domains and motifs	13,057	(15)
PFam	July 2006	Protein families	8,376	(16)
EC	November 2006	Enzymes	32,137	(17)
GO	July 2006	Functional terms	21,017	(18)
LigBase	February 2004	Protein ligands	101,359	(13)
PIBASE	September 2004	Protein interactions	158,915	(14)

successfully determined and deposited in the PDB database (code 1u6l, release date 14 December 2004). Searches by PSI-BLAST (24) and threading by GenThreader (25) indicated similarity to the glyoxalase/bleomycin domain (PfamA family PF00903). This domain is found in several proteins including the bleomycin resistance protein and dioxygenases. The DBAli tools confirm and add to these findings. A search for structures similar to chain A of 1u6l (1u6lA) results in 306 related structures. The first annotated hit in the list of similar structures corresponds to a bleomycin resistance protein (PDB code 1xrkB). As a result, the AnnoLyze program (all parameters set to their default values except for minimal sequence identity set to 15%) predicts a binding site on 1u6l that may bind a bleomycin-like ligand. The AnnoLite program predicts that the 1u6lA chain adopts the glyoxalase/bleomycin resistance fold (SCOP code d.32.1.1). ModDom detects that the protein in fact contains two glyoxalase/bleomycin resistance fold domains. In summary, ModDom, AnnoLyze and AnnoLite, adding to the results from sequence-based searches, annotate 1u6lA as a two-domain antibiotic resistance protein and localize a putative binding site for the antibiotic bleomycin (Figure 2).



(d)

	Conf.	P-value	Link	Description
CATH:	●	2.6e-66	3.10.180.10	2,3-Dihydroxybiphenyl 1,2-Dioxygenase, domain 1
SCOP:	●	3.2e-66	d.32.1.2	Antibiotic resistance proteins
PFAM:	●	1.1e-63	PF00903	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily
	●	4.6e-20	PF06983	3-demethylubiquinone-9 3-methyltransferase
InterPro:	●	8.3e-57	IPR004360	Glyoxalase/bleomycin resistance protein/dioxygenase
	●	5.6e-27	IPR000335	Bleomycin resistance protein
	●	5.4e-19	IPR009725	3-demethylubiquinone-9 3-methyltransferase
	●	9.8e-12	IPR011588	Glyoxalase/extradiol ring-cleavage dioxygenase
EC Number:	●	3.5e-30	2.5.1.18	Glutathione transferase.
GO Molecular Function:	●	1.8e-23	0030955	potassium ion binding
	●	1.4e-19	0008144	drug binding
	●	6.8e-19	0004364	glutathione transferase activity
	●	2.8e-11	0030145	manganese ion binding
	●	1.7e-2	0046872	metal ion binding
	●	6.5e-1	0016740	transferase activity
GO Biological Process:	●	5.9e-39	0046677	response to antibiotic
	●	7.5e-10	0006313	DNA transposition

Figure 2. Functional annotation of the target T1794 from the NYSGXRC consortium (PDB code 1u6l chain A). (**a**) The structure closest to the 1v6lA chain with known annotation is an antibiotic inhibitor from *Streptoallateichus hindustanus* (PDB code 1xrk chain B). It is shown in the ribbon representation in red, superposed on 1u6lA in blue (RMSD of 3.5 Å over 106 residues, sequence identity of 14.3% and MAMMOTH *P*-value = 13.8). (**b**) Predicted binding site for bleomycin A2 ligand (residues colored in red and orange: Lys80, Gly81, Cys82, Ser83, Ser85, Ans87, Gln108, Phe115, Trp116, Ser119, Gly121, Thr124, Gly128, Val129, Ala130 and Val133). (**c**) Domain boundaries of 1u6lA as assigned by ModDom. The protein structure is predicted to have two domains: residues 1–76 (yellow) and residues 77–123 (red). (**d**) Screen capture of the AnnoLite results for the 1u6lA chain.

CONCLUSIONS

The DBAli database stores a comprehensive and up-to-date comparison of protein structures in the PDB. The data are stored in a MySQL relational database and can be accessed and downloaded *via* a web server. Several tools have been developed that interact with the data deposited in DBAli, including the DBAlit, AnnoLite and AnnoLyze, ModClus, ModDom, and SALIGN programs. The design of DBAli allows easy cross-referencing to other databases. DBAli already includes links to the MODBASE, LigBase, PIBASE, PDB, CATH, SCOP, PFamA, InterPro, Enzyme and GO databases. Through further integration of new tools and other databases, DBAli and its tools are becoming a valuable resource to the structural biology community. As of January 2007, the DBAli tools have been used by more than 1,700 unique users from 79 different countries worldwide who performed on average ~600 tasks per month.

AVAILABILITY AND REQUIREMENTS

DBAli is freely available on the Internet at http://salilab.org/DBAli and requires a web browser that is capable of running the JMol applet. The web interface is programmed in PHP and MySQL supports the underlying database.

ACKNOWLEDGEMENTS

We are grateful to Dr Angel Ortiz for the MAMMOTH program. We would also like to thank the JMol team, Erik Bosrup and Stijn van Dongen for JMol, overLib and MCL, respectively. We appreciate valuable comments from Prof Helen M. Berman and Prof. Wayne F. Anderson. We acknowledge funding by The Sandler Family Supporting Foundation, NIH grants GM 62529, GM074929, GM71790 and GM54762, as well as hardware gifts from IBM, Intel, HP and Network Appliance. Funding to pay the Open Access publication charges for this article was provided by Generalitat Valenciana Starting Grant.

Conflict of interest statement. None declared.

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.
2. Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. *et al.* (1999) Structural genomics: beyond the human genome project. *Nat.Genet.*, **23**, 151–157.
3. Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.*, **8**, 559–566.
4. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32** (Database issue), D226–D229.
5. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
6. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.
7. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2003) From protein structure to biochemical function? *J. Struct. Funct. Genomics*, **4**, 167–177.
8. Pal,D. and Eisenberg,D. (2005) Inference of protein function from protein structure. *Structure*, **13**, 121–130.
9. Pazos,F. and Sternberg,M.J. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.
10. Marti-Renom,M.A., Ilyin,V.A. and Sali,A. (2001) DBAli: a database of protein structure alignments. *Bioinformatics*, **17**, 746–747.
11. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
12. Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.
13. Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, **18**, 200–201.
14. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33** (Database Issue), D201–D205.
16. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32** (Database issue), D138–D141.
17. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
18. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
19. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33** (Database Issue), D262–D265.
20. Marti-Renom,M.A., Rossi,A., Al-Shahrour,F., Davis,F.P., Pieper,U., Dopazo,J. and Sali,A. (2007) The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. *BMC Bioinformatics* (Suppl 4):S4.
21. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
22. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
23. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662–672.
24. 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.
25. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.