Nucleic Acids Research, 2005, Vol. 33, Database issue ***D279–D283***
doi:10.1093/nar/gki016

PFD: a database for the investigation of protein folding kinetics and stability

Kate F. Fulton[1], Glyn L. Devlin[1], Rachel A. Jodun[1], Linda Silvestri[3], Stephen P. Bottomley[1], Alan R. Fersht[3] and Ashley M. Buckle[1,2,*]

[1]The Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Faculty of Medicine, [2]Victorian Bioinformatics Consortium, PO Box 53, Monash University, Clayton, Victoria 3800, Australia and [3]Cambridge Centre for Protein Engineering and Cambridge University Chemical Laboratory, MRC Centre, Hills Road, Cambridge, CB2 2QH, UK

Received July 17, 2004; Revised and Accepted September 17, 2004

ABSTRACT

We have developed a new database that collects all protein folding data into a single, easily accessible public resource. The Protein Folding Database (PFD) contains annotated structural, methodological, kinetic and thermodynamic data for more than 50 proteins, from 39 families. A user-friendly web interface has been developed that allows powerful searching, browsing and information retrieval, whilst providing links to other protein databases. The database structure allows visualization of folding data in a useful and novel way, with a long-term aim of facilitating data mining and bioinformatics approaches. PFD can be accessed freely at http://pfd.med.monash.edu.au.

INTRODUCTION

Understanding the rules that govern protein folding is one of the great challenges of molecular biology. Studies of protein folding, combining experiment and simulation, have led to a solid understanding of the physical process of folding and the forces that stabilize proteins. The last 10 years have witnessed a revolution in our understanding of the pathway and stability of protein folding (1). The sustained growth of folding studies is fuelled by the availability of new sequences, rapid structure determination and radical developments in experimental methods. Furthermore, recent successes in folding simulations have improved our understanding of the protein folding process at atomic resolution (2), providing further avenues for experimental investigation. Analysis of the folding mechanisms and pathways of proteins within homologous families has propelled protein folding into the post-genomic era (3).

Traditionally, kinetic and thermodynamic data are collected and analysed on an individual protein basis, and is published in an unstructured fashion, despite the best efforts to tabulate it. Clearly, this presents an enormous challenge for data analysis, even simple searching for trends requires exhaustive manual inspection of the literature. With the exception of ProTherm [Thermodynamic Database for Proteins and Mutants (4)], the vast majority of web-accessible databases focus on sequences and structures. There are currently no tools that bring together both kinetic and thermodynamic folding data for proteins and mutants.

A comparison of the folding properties for more than 50 proteins represents the most comprehensive compilation of folding data to date (5). This painstaking analysis uncovered some general trends but also highlighted the great diversity in folding behaviour. The speed at which a protein folds and the pathway it takes are dictated by its structural and energetic characteristics. Recent work suggests that the fundamental physics underlying folding may be relatively simple: the mechanism of folding appears to be dictated by the low-resolution features (or *topology*) of the folded protein structure (6). Topology can be described by the parameter contact order, which is defined as the average sequence separation between contacting residues in the 3D structure. Proteins having a low contact order, e.g. α-helical bundles, fold faster than those with a high contact order, e.g. β-sandwiches (6). Topology has been found to be the overriding determinant of folding rate for a wide range of proteins (6–9). However, studies on the topologically similar members of the immunoglobulin family have shown that they fold with rate constants which correlate better with stability (10). Studies on horse and yeast cytochrome *c* also suggest that stability is an important factor (11). Furthermore, protein engineering studies show that mutations which do not affect the contact order can change the folding rate by many orders of magnitude (5). Thus, in many cases, factors other than topology must also be significant.

The last six years have witnessed a huge increase in the number of proteins being studied, and is set to grow further as

*To whom correspondence should be addressed. Tel: +61 3 9905 3781; Fax: +61 3 9905 3781; Email: Ashley.Buckle@med.monash.edu.au
Present address:
Glyn L. Devlin, Biological Physics Group, Cavendish Laboratory, Cambridge CB3 0HE, UK

structural genomics projects gain momentum. In order to exploit this wealth of data so that data mining efforts may uncover further relationships between folding behaviour and structural character, a central repository is critical. Indeed, recent benchmarking of predicted folding rates (12), together with comparisons of the folding behaviour of two- and three-state folding proteins (13), emphasizes the need for a centralized database. In order to address this issue, here, we describe the design and implementation of a relational database for protein folding, Protein Folding Database (PFD). Entries are heavily annotated, particularly with experimental, structural and functional details. A user-friendly web interface to the database allows querying using many parameters, as well as retrieval and presentation of data. The database will have three distinct roles: (i) data repository: new data can be rapidly deposited, validated and made available to the folding community and wider scientific arena; (ii) experimental resource, the database will be of use to the biophysicist seeking to compare new folding data with the current dataset for similar proteins, bypassing the relatively slow and inefficient examination of the literature. The database will play a useful role in the design of folding experiments, e.g. both as a guide in the design of experimental methodology and in the selection of proteins belonging to homologous families; and (iii) theoretical resource, all experimental folding data will be at the disposal of theoreticians, strengthening the emerging conspiracy between experiment and simulation.

PFD DESCRIPTION

Our approach is to create a database that captures as much as possible of the relevant information important for a folding experiment: kinetic rates of folding and unfolding; equilibrium free energies; experimental methods such as spectroscopic technique (probe) and method of perturbation (e.g. denaturant), and instrument details; publication information; protein details, such as fold, structural class, biological function and mutation information. Relationships are made between entities using standard relational database techniques. PFD was created using open-source MySQL relational database server software, version 4.0.16 (www.mysql.com), running on an Apple Dual 2.0 GHz G5/OS X Server (version 10.3.4). A web-based query interface to the database was created using the Java programming language and Apple WebObjects software (version 5.2.2) and the Xcode development environment (Figure 1).

USE OF PFD IN FOLDING RESEARCH

The essence of our approach is to allow a diverse collection of folding data to be searched via multiple parameters, and the results presented in a structured fashion. Typical queries that can be formulated are 'compare the folding behaviour of monomeric α-helical proteins?'; and 'which beta proteins larger than 60 residues have folding rates greater than 10^3 s^{-1}?'. The web interface allows a detailed, spreadsheet-like list of results allowing quick visualization of general trends in data (Figure 2). The results of a search can be sorted on any heading, which is useful, e.g. when inspecting the variability of folding rates among proteins within a family. Each entry also contains information of the publication and a URL to the entry in NCBI PubMed literature database.

Annotation of proteins exploits the hierarchy used by the Structural Classification of Proteins database [SCOP (14)]: proteins belong to families, which in turn belong to a structural class (e.g. all alpha proteins). This was performed to minimize

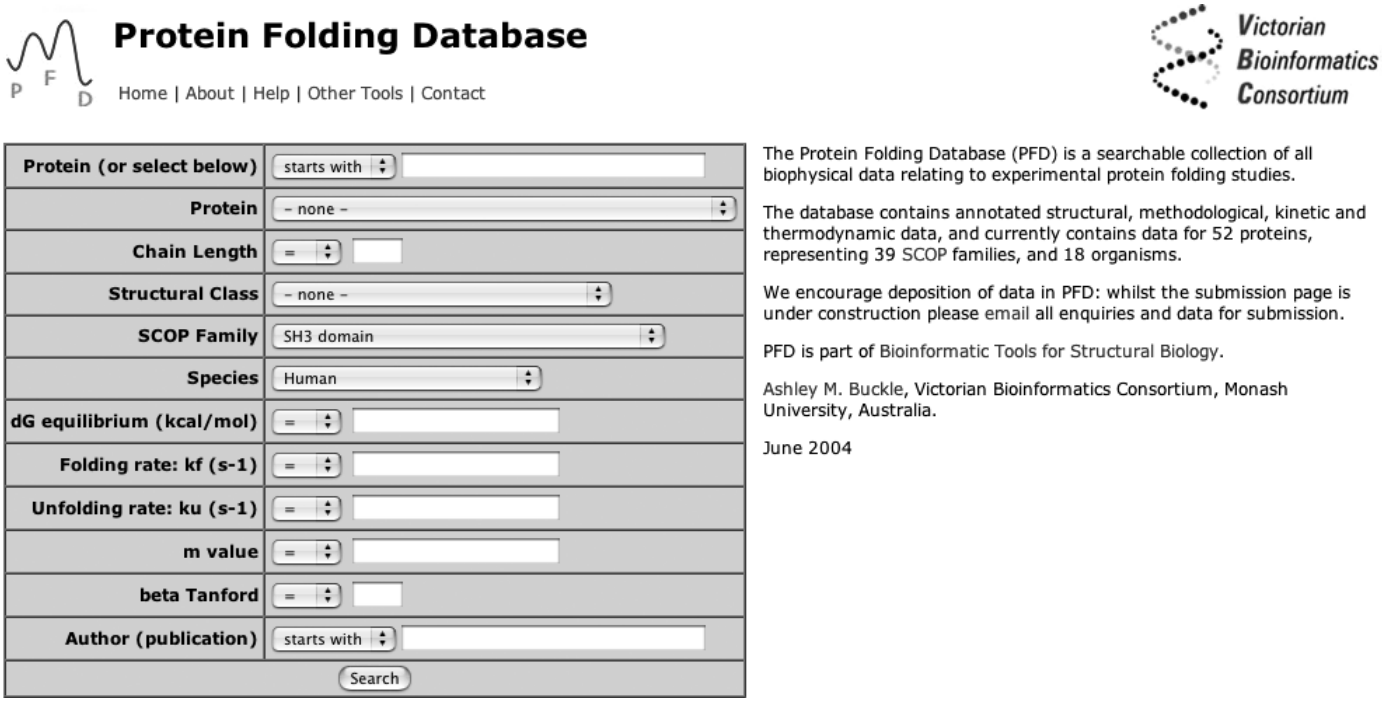


Figure 1. The web-query interface to PFD. The database can be searched using multiple parameters relating to structural, thermodynamic and kinetic attributes.

51 Results Display 25 items Page 2 of 3

The table can be sorted on any heading by clicking on the ≣ icon. Click Inspect to show details of the refolding methods for each entry.

Class and Family are as defined by the SCOP database.

	Protein	Class	Family	Species	Length	CO	Methods	dG-eq	m	[D]50%	kf	ku
Inspect	ACBP	Alpha	Acyl-Co A Binding Protein	Rat	86	12.0	Equilibrium Kinetics	6.05	3.40	1.8	395.44	0.00667
Inspect	CspA	Beta	Cold shock DNA-binding domain-like	E. coli	69	18.0	Equilibrium	3.10	1.70	1.8		
Inspect	HPr	Alpha+Beta	HPr-like	E. coli	85	18.4	Equilibrium	5.00	2.00	2.0		
Inspect	HPr	Alpha+Beta	HPr-like	E. coli	85	18.4	Kinetics Equilibrium	5.56	2.77	2.0	14.90	0.00209
Inspect	HPr	Alpha+Beta	HPr-like	E. coli	85	18.4	Equilibrium	4.56	2.17	2.1		
Inspect	FKBP12	Alpha+Beta	FKBP immunophilin/proline isomerase	Human	107	18.0	Equilibrium Kinetics	8.20	3.90	2.1	1.40	1.50000
Inspect	HPr	Alpha+Beta	HPr-like	E. coli	85	18.4	Equilibrium Kinetics	4.78	2.22	2.1		
Inspect	ACBP	Alpha	Acyl-Co A Binding Protein	Bovine	86	12.0	Equilibrium Kinetics	7.06	3.03	2.3	278.66	0.00010
Inspect	Bc-Csp	Beta	Cold shock DNA-binding domain-like	Bacillus caldolyticus	66	11.0	Equilibrium Kinetics	4.80	1.79	2.7	1370.00	0.64000
Inspect	Tm-Csp	Beta	Cold shock DNA-binding domain-like	Thermotoga maritima	66	17.8	Equilibrium Kinetics	6.26	1.88	3.3	565.00	0.01800

Figure 2. A typical results page listing is shown. This is a summarized table, containing most of the important folding data. However, by clicking 'Inspect' you can show all the available folding data. This will also provide the link to the publication details.

redundancy in the database so that all structural information for an entry can be retrieved via the SCOP link. SCOP and PDB provide an array of links to other databases (such as Entrez, Pfam and ASTRAL), as well as an array of tools that operate on the data (e.g. 3D visualization). The hierarchical classification of *structural class*/*family*/*protein* allows convenient browsing (akin to browsing proteins belonging to a particular fold in SCOP): folding data for proteins are grouped under their fold or structural class, which may prove convenient when examining the folding behaviour of proteins within a family. Effective use of hyperlinks in search results pages allows useful browsing. For example, simple searching may retrieve results for several proteins. Examining any entry in more detail yields information on the protein structure, folding thermodynamics and kinetics, experimental methods, mutations (if any), publication(s) and annotations (Figure 3). The power of the relational database approach allows us to visualize folding data in a novel way.

Availability and submissions: PFD is freely available at http://pfd.med.monash.edu.au. Submissions and enquiries should be emailed to Ashley.Buckle@med.monash.edu.au.

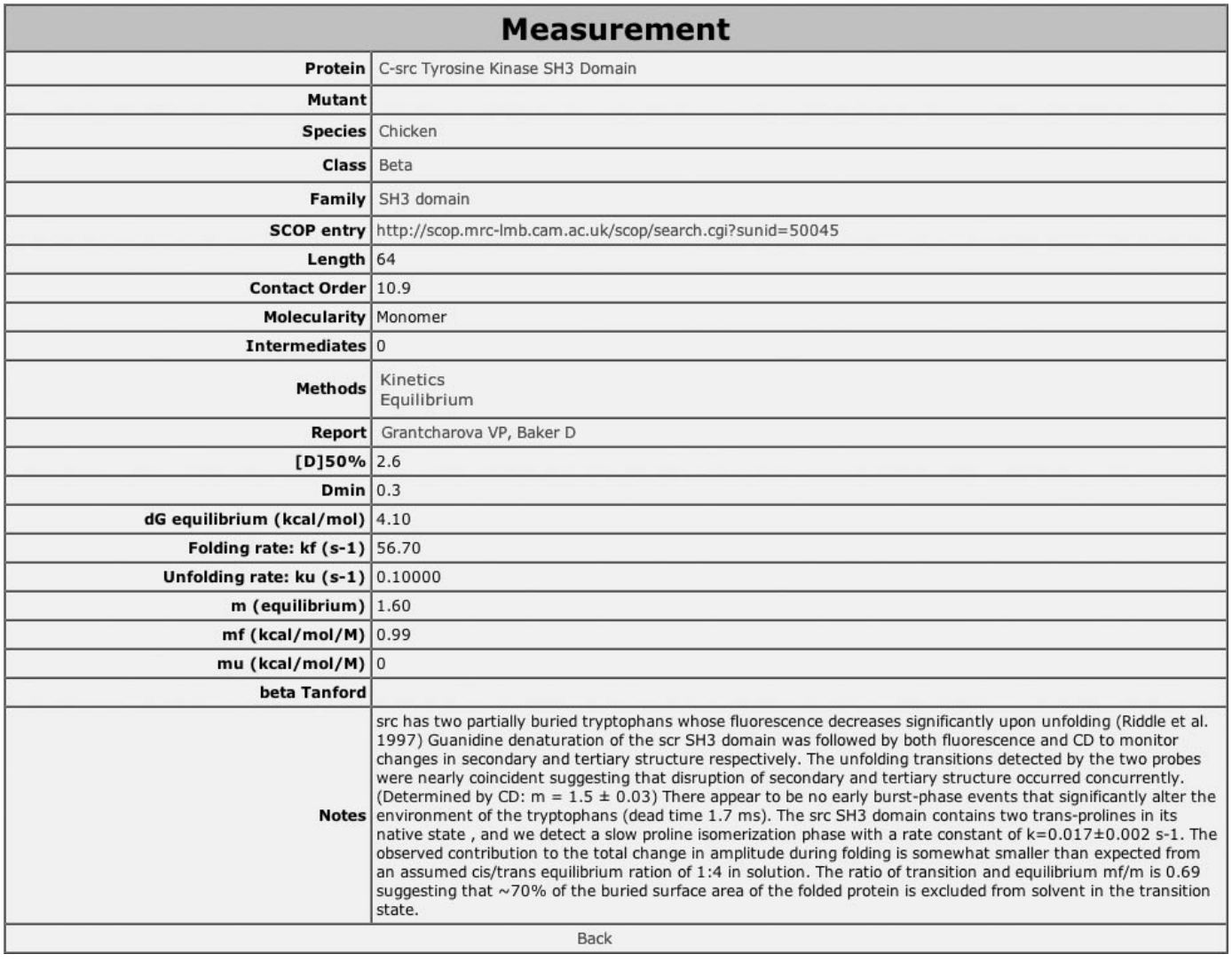
CONCLUSIONS AND FUTURE DIRECTIONS

The constructed database and web-based query interfaces have demonstrated the applicability and usefulness of the database design. The ability to query the database with important folding 'questions' indicates that its design accurately reflects the organization of data in a real folding experiment. Future work will focus on the following areas:

(i) *Functional annotation*: An analysis of folding data must take into account the biological function of the protein. Any trends uncovered must also be considered in the context of function. To enable these entries will be linked to the Gene Ontology database (15), which annotates database entries on molecular function, biological process and cellular location.

(ii) *Data exchange*: How will other databases be able to use data from the folding database? This is a serious challenge because of the vast heterogeneity in database standards and data structure. This can be addressed by making folding data available using extensible markup language (XML). XML provides the capability of representing protein data in a single, standardized data structure that is easily transmitted over a network. This will require the construction of a specification language for protein folding data that will allow for portable, system-independent, machine-parsable and human-readable representation of essential features of protein folding. All folding data can then be made available in XML format.

(iii) *Data visualization*: As the dataset grows, visualization of text becomes cumbersome. This will require the development of graphical representations of the data, such as Chevron plots. In particular, graphical methods allowing



Measurement	
Protein	C-src Tyrosine Kinase SH3 Domain
Mutant	
Species	Chicken
Class	Beta
Family	SH3 domain
SCOP entry	http://scop.mrc-lmb.cam.ac.uk/scop/search.cgi?sunid=50045
Length	64
Contact Order	10.9
Molecularity	Monomer
Intermediates	0
Methods	Kinetics Equilibrium
Report	Grantcharova VP, Baker D
[D]50%	2.6
Dmin	0.3
dG equilibrium (kcal/mol)	4.10
Folding rate: kf (s-1)	56.70
Unfolding rate: ku (s-1)	0.10000
m (equilibrium)	1.60
mf (kcal/mol/M)	0.99
mu (kcal/mol/M)	0
beta Tanford	
Notes	src has two partially buried tryptophans whose fluorescence decreases significantly upon unfolding (Riddle et al. 1997) Guanidine denaturation of the scr SH3 domain was followed by both fluorescence and CD to monitor changes in secondary and tertiary structure respectively. The unfolding transitions detected by the two probes were nearly coincident suggesting that disruption of secondary and tertiary structure occurred concurrently. (Determined by CD: m = 1.5 ± 0.03) There appear to be no early burst-phase events that significantly alter the environment of the tryptophans (dead time 1.7 ms). The src SH3 domain contains two trans-prolines in its native state , and we detect a slow proline isomerization phase with a rate constant of k=0.017±0.002 s-1. The observed contribution to the total change in amplitude during folding is somewhat smaller than expected from an assumed cis/trans equilibrium ration of 1:4 in solution. The ratio of transition and equilibrium mf/m is 0.69 suggesting that ~70% of the buried surface area of the folded protein is excluded from solvent in the transition state.
Back	

Figure 3. The full entry can be retrieved for individual records, and hyperlinks allow efficient browsing (e.g. by SCOP family, molecularity, etc.).

the visualization of relationships between structural parameters, such as contact order and folding kinetics, will prove very useful.

(iv) *Data deposition and validation*: It is vital that new folding data is deposited in the same timeframe as publication (as is the case of the PDB). This means that the data becomes readily available to the community and amenable to analysis. This can be achieved using a forms-based system that will allow data to be deposited, via a web-browser, directly by the originator of the data, again in an analogous manner to the PDB. Validation logic can also be built into the deposition process, providing both a useful service to the depositor as well as an indication on data quality to users. The latter two aims are particularly important for database functionality and growth, respectively, and will be given priority. This approach will allow us to achieve a high degree of uniformity in the structure of folding data, which will benefit experimentalists in data acquisition and handling.

ACKNOWLEDGEMENTS

S.P.B. is a Monash University Senior Logan Fellow and R.D. Wright Fellow of the NH&MRC. We thank Christina Mitchell for financial support, and James Whisstock and Ross Coppel for continuing support.

REFERENCES

1. Fersht,A. (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman and Company, New York, NY.
2. Fersht,A.R. and Daggett,V. (2002) Protein folding and unfolding at atomic resolution. *Cell*, **108**, 573–582.
3. Gunasekaran,K., Eyles,S.J., Hagler,A.T. and Gierasch,L.M. (2001) Keeping it in the family: folding studies of related proteins. *Curr. Opin. Struct. Biol.*, **11**, 83–93.
4. Bava,K.A., Gromiha,M.M., Uedaira,H., Kitajima,K. and Sarai,A. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **32**, D120–D121.
5. Jackson,S.E. (1998) How do small single-domain proteins fold? *Fold. Des.*, **3**, R81–R90.

6. Plaxco,K.W., Simons,K.T. and Baker,D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.
7. Chiti,F., Taddei,N., White,P.M., Bucciantini,M., Magherini,F., Stefani,M. and Dobson,C.M. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.*, **6**, 1005–1009.
8. Martinez,J.C. and Serrano,L. (1999) The Folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.*, **6**, 1010–1016.
9. Riddle,D.S., Grantcharova,V.P., Santiago,J.V., Alm,E., Ruczinski,I. and Baker,D. (1999) Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.*, **6**, 1016–1024.
10. Clarke,J., Cota,E., Fowler,S.B. and Hamill,S.J. (1999) Folding studies of the immunoglobulin-like beta-sandwich proteins suggest their share a common folding pathway. *Structure Fold. Des.*, **7**, 1145–1153.
11. Mines,G.A., Pascher,T., Lee,S.C., Winkler,J.R. and Gray,H.B. (1996) Cytochrome *c* folding triggered by electron transfer. *Chem. Biol.*, **3**, 491–497.
12. Ivankov,D.N. and Finkelstein,A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 8942–8944.
13. Kamagata,K., Arai,M. and Kuwajima,K. (2004) Unification of the folding mechanisms of non-two-state and two-state proteins. *J. Mol. Biol.*, **339**, 951–965.
14. 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**, D226–D229.
15. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.