Published online October 14, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 18 ***5539–5545***
doi:10.1093/nar/gkh894

The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes

Andreas Ruepp[1,*], Alfred Zollner[2], Dieter Maier[2], Kaj Albermann[2], Jean Hani[2], Martin Mokrejs[1,3], Igor Tetko[1], Ulrich Güldener[1], Gertrud Mannhaupt[4], Martin Münsterkötter[1] and H. Werner Mewes[1,4]

[1]Institute for Bioinformatics (MIPS), GSF National Research Center for Environment and Health, Ingolstaedter Landstraße 1, D-85764 Neuherberg, Germany, [2]Biomax Informatics AG, Lochhamerstr. 11, D-82152 Martinsried, Germany, [3]Faculty of Science, Charles University, Vinicna 5, 128 42 Prague, Czech Republic and [4]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received July 30, 2004; Revised September 9, 2004; Accepted September 28, 2004

ABSTRACT

In this paper, we present the Functional Catalogue (FunCat), a hierarchically structured, organism-independent, flexible and scalable controlled classification system enabling the functional description of proteins from any organism. FunCat has been applied for the manual annotation of prokaryotes, fungi, plants and animals. We describe how FunCat is implemented as a highly efficient and robust tool for the manual and automatic annotation of genomic sequences. Owing to its hierarchical architecture, FunCat has also proved to be useful for many subsequent downstream bioinformatic applications. This is illustrated by the analysis of large-scale experiments from various investigations in transcriptomics and proteomics, where FunCat was used to project experimental data into functional units, as 'gold standard' for functional classification methods, and also served to compare the significance of different experimental methods. Over the last decade, the FunCat has been established as a robust and stable annotation scheme that offers both, meaningful and manageable functional classification as well as ease of perception.

INTRODUCTION

In recent years, molecular biology has become an information-rich science relying on computational information management. At the same time, progress in computational power allows the integration of all available information into qualitative and quantitative models, thus transforming biology from descriptive to predictive science. To this end, all available biological information has to be integrated to generate Biological Information Systems (BIS) (1). A BIS requires that biological information to be represented must be structured in such form that it becomes accessible for systematic computational analysis. While primary sequence and expression data are already available in this way, the information about the functional attributes of genes and proteins is traditionally hidden in the prose of biological literature. As free text is inherently difficult to mine with automatic methods, manual efforts to extract and structure information in specialized databases have been made. To harvest the full power of automatic information management, the databases must consistently present the information content using standardized vocabularies based on biological concepts. In addition to consistent name spaces, such vocabularies fulfil the additional task to reflect biological interdependences and therefore can be used as classification systems with respect to the functional interpretation of biological systems. Such a vocabulary needs to fulfil a number of criteria, such as human usability, computer readability, independence on organism, breadth and depth of coverage, stability and extendibility.

Since 1956, enzymes are classified by the EC nomenclature system, a hierarchical scheme based on the chemistry of the reaction they catalyse (2). In addition, a nomenclature scheme for membrane transport proteins [TC system, (3)] has been included by the EC commission.

The first attempts to generate standardized functional vocabularies that are not restricted to certain types of proteins such as enzymes or transporters have been made by the databases of PIR and SWISS-PROT (4,5). The objective of these databases is to associate gene products and functional data beyond the often inconsistent naming conventions used in the title lines for historical reasons. Here, the cellular function of proteins is described by keywords, which are neither implemented in a formal structure nor provide a framework to describe the relations between individual terms of the keyword lists. The first formally organized functional annotation scheme was proposed to annotate proteins of *Escherichia coli* (6). This scheme was also used as the basis for annotation when in

*To whom correspondence should be addressed. Tel: +49 89 3187 3189; Fax: +49 89 3187 3585; Email: andreas.ruepp@gsf.de

1995 the first completely sequenced genomes were published (7,8).

While coordinating the *Saccharomyces cerevisiae* genome project at MIPS, a hierarchically structured controlled vocabulary, the Functional Catalogue (FunCat), was developed. At that time, FunCat contained only those categories required to describe yeast biology (e.g. no multicellular functionality) (9,10). While the design principle of the catalogue has remained stable since then, its content has been extended for plants to annotate genes from the *Arabidopsis thalania* genome project and furthermore to cover prokaryotic organisms and finally animals too (11–14).

During the last three years, another catalogue has become widely used for the annotation of eukaryotic genomes, the Gene Ontology (GO) (15). The organization of the GO catalogue annotation scheme differs substantially in its general structure from the previously described schemes, as it is not strictly hierarchical but organized as acyclic graphs.

In this paper, we describe FunCat as an efficient and comprehensive tool for the annotation of functional genome information. We explain the structure of FunCat and demonstrate its use for various applications, including manual genome annotation, automatic functional annotation of predicted genes and analysis of data from large-scale transcriptome and proteome studies. In addition, FunCat is compared with other annotation schemes.

RESULTS AND DISCUSSION

Structure of the FunCat

Any classification scheme employs attributes to assign membership to the individual categories. Selection of the attributes and description of the relation of the categories in the classification scheme are the key issues for its design. Some attributes are well defined, computable and highly selective, others may be associated and being descriptive only. A decision must be taken for the purpose of the classification scheme. If all attributes associated to an object are employed to describe classes, the classification scheme will rather serve as a descriptive tool instead of being useful to identify objects with similar features. Incorporating a large number of attributes into the classification has the advantage of building a powerful retrieval and navigation tool, as long as all attributes can be assigned to all objects with high confidence. Nevertheless, such a scheme has two major drawbacks: first, it is very hard to serve for the automatic assignment for large sets of non-identical objects (as described below) and second, the categories become rather sparsely populated. However, in recent work these attributes have been used for the automatic assignment of proteins to functional classifications which use kernel methods (16).

Sequencing of complete genomes concomitantly raised a demand for comprehensive description of the functional aspects of the associated protein complements. Taking into account the broad and highly diverse spectrum of known protein functions, the FunCat annotation scheme consists of 28 main categories (or branches) that cover general features like cellular transport, metabolism and protein activity regulation (see Table 1; the main categories of the FunCat). Each of the main functional branches is organized as a hierarchical, tree like structure (see Figure S1). This basic concept has been retained since the annotation of the yeast genome and proved to be well suited for the annotation of other genomes (see Table 2). The FunCat provides a general, stable annotation scheme and serves as a database retrieval environment with only four major extensions since 1996. In analogy to some textbooks, high-level branches of biochemical and molecular functions and their subcategories are structured within FunCat into sections which contain chapters and paragraphs.

A general consideration at the design of an annotation scheme is the balance between human usability, specificity of the categories and requirement for subsequent bioinformatic applications. In order to keep the FunCat descriptive, but compact, it has been decided to classify protein functions not down to the most specific level. When a more detailed description of proteins is required, additional catalogues with a specific focus can be included in the annotation process. An example for such a resource is the well-established Enzyme

Table 1. Main functional categories of the FunCat

Functional classification catalogue (FunCat) version 2.0	
Metabolism	
01	Metabolism
02	Energy
04	Storage protein
Information pathways	
10	Cell cycle and DNA processing
11	Transcription
12	Protein synthesis
14	Protein fate (folding, modification and destination)
16	Protein with binding function or cofactor requirement (structural or catalytic)
18	Protein activity regulation
Transport	
20	Cellular transport, transport facilitation and transport routes
Perception and response to stimuli	
30	Cellular communication/signal transduction mechanism
32	Cell rescue, defense and virulence
34	Interaction with the cellular environment
36	Interaction with the environment (systemic)
38	Transposable elements, viral and plasmid proteins
Developmental processes	
40	Cell fate
41	Development (systemic)
42	Biogenesis of cellular components
43	Cell type differentiation
45	Tissue differentiation
47	Organ differentiation
Localization	
70	Subcellular localization
73	Cell type localization
75	Tissue localization
77	Organ localization
78	Ubiquitous expression
Experimentally uncharacterized proteins	
98	Classification not yet clear-cut
99	Unclassified proteins

With the exception of categories 78, 98 and 99, all main categories are the origin of hierarchical, tree-like structures. To make the introduction of new main categories possible, the numbering of the categories is not strictly sequential.

Table 2. Manual FunCat annotation of whole genomes

Organism	Number of manually annotated proteins	Proteins with annotated function	Reference
Archaea			
Thermoplasma acidophilum	1507	668	(13)
Bacteria			
Bacillus subtilis 168	4106	3002	http://www.biomax.de
Listeria monocytogenes EGD	2846	1956	http://pedant.gsf.de
Listeria innocua Clip 11262	2968	1959	http://pedant.gsf.de
Helicobacter pylori KE26695 (ATCC 700392)	1567	957	http://pedant.gsf.de
Eukaryotes			
Saccharomyces cerevisiae	6157	4033	(9)
Neurospora crassa	8348	3713	(11)
Arabidopsis thaliana	26444	17717	(14)
Homo sapiens	24910	14633	http://www.biomax.de

This list contains complete sequenced genomes that were manually annotated at MIPS or Biomax. Proteins with annotated function contain meaningful functional categories, i.e. entries with categories like '98 classification not yet clear-cut' or '99 unknown protein' were not counted.

Nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme/). The approach of EC catalogue is different from the one of FunCat because it classifies the information based on the underlying chemical mechanism, whereas FunCat classification is based on the pathway where the enzyme acts (with respect to the biochemical pathway where it is involved). The annotation of proteins carrying enzymatic activity in a cellular perspective is not always unambiguous. Although we might be able to predict the mechanism of the function of a certain enzyme based on structural data (or presence of a known domain in a sequence), we are not always able to figure out how the enzyme acts in the metabolic context. On its most specific level, FunCat follows a more intuitive approach and attempts to store information if an enzyme is used for biosynthesis or degradation of a certain metabolite. This assignment is often difficult, as enzymes catalyse the same reaction in both directions. Biological systems, however, often favour most reactions only in one direction (typically while removing the products they shift reaction equilibria constantly to the right). It happens that an enzyme catalyses the very same reaction in anabolic direction but in another tissue (or even in the same tissue but in different metabolic state) in catabolic direction as it was found for glutamate dehydrogenase in liver mitochondria (17). In such cases, the hierarchical approach of FunCat allows to assign the enzyme to more unspecific categories, which include both, the anabolic and the catabolic direction of the enzymatic reaction. However, the drawback of the current version lies in the fact that we cannot distinguish between the case that a protein has both features of the subsequent classes or the alternative that there is no evidence for one or the other way, and due to missing additional information, the assignment must be left open.

In total, the superset of FunCat currently contains 1307 categories (see Tables S2 and S3). These are not species-specific because the aim of FunCat is not to be dedicated to a single organism but to allow annotation of a large spectrum of organisms. However, where required it contains subcategories concerned with functional peculiarities that are specific to organism groups. In addition, one main category is assigned to cover viral, transposon and plasmid gene products. Each of the functional categories is assigned to a unique two-digit number. The upward context of the hierarchical tree consists of the prefix of the preceding nodes, located in the upper levels in the hierarchy. The levels of categories are separated by dots, e.g. *01 metabolism* is a representative of the highest level, and *01.01.03.02.01 biosynthesis of glutamate* belongs to the most specific level of FunCat. For the majority of the proteins, the cellular function cannot be entirely described with a single functional category but has to be considered as the sum of different properties. Since a context-oriented protein annotation using only one functional category exceeds the capacity of a compact and manageable annotation scheme, FunCat enables the assignment of multiple categories for a single protein.

The architecture of the FunCat allows smooth and flexible extensions to be incorporated as shown by the recent development performed during annotation of the human genome (http://www.biomax.de/), when several categories have been newly introduced (e.g. *41 development*, *43 cell type differentiation* and *77 organ localization*).

Applications of the FunCat for the analysis of genome data

Manual annotation of genomes illustrated by the annotation of the yeast genome. In the course of the *S.cerevisiae* genome project, the FunCat was developed which enables in depth annotation of the still growing experimental data. The setup of FunCat was a prerequisite for annotation of the genome using a controlled vocabulary in a systematic way. In addition, further catalogues were used or developed to annotate this genome encompassing most detailed experimental evidence of any organism in the eukaryotic kingdom: Protein Classes, Protein Complexes, Localization, EC, Transport (18) and Phenotype Catalogue.

These distinct catalogues ease the annotation process by providing a self-organizing system compliant but restricted to functional modules such as pathways. As the main goal is to classify the function of a gene, and not an exhaustive description of attributes, one does not need to handle large numbers of categories and their complex dependences. The increase in the number of classes results in an increase in the absolute number of missed or missing assignments. By focusing on a compact scheme of features such as functional categories, the time and cost-consuming process of annotation is simplified and the rapidly growing number of available genomes can be covered using automatic or supervised methods for the assignment. The majority of entries in the Comprehensive Yeast Genome Database (CYGD, http://mips.gsf.de/genre/proj/yeast/index.jsp) are assigned to multiple FunCat categories resulting in a multidimensional annotation. Protein kinase SNF1, as an example for a manually annotated protein, is involved in signalling transduction (FunCat 30.01.05; enzyme mediated signalling transduction). In addition, its different cellular roles are assigned by using FunCats 01.05.04 (regulation of C-compound and carbohydrate

utilization), 32.01.11 (nutrient starvation response; stress response), 40.01.05 (growth regulators/regulation of cell size), 34.07.02 (cell-matrix adhesion) and 40.20 (cell aging). Interaction of SNF1 with other proteins and mode of action are assigned by using FunCats 16.01 (protein binding; SNF1 is part of a protein complex), 14.07.03 (modification by phosphorylation, dephosphorylation and autophosphorylation; the kind of protein modification), 18.01.01 (modification; the mechanism of protein activity regulation) and 18.02.09 (regulator of transcription factor; the target of regulation). All statements to a particular entry are now referenced to a PubMed entry if possible. Additionally, an evidence tag is assigned, for example clearly indicating the type and character of experiments, such as individual, high-throughput experiments etc.

Since the first annotation of the yeast genome, the FunCat has been used for manual annotation of nine genomes (see Table 2). To allow efficient bioinformatic analysis of large datasets, the FunCat was also implemented in a semi-automatic annotation suite, the Pedant system (19). Although the Pedant assignments, which are based on the assignment of functional categories by sequence similarity, are not always reliable, they improve the quality and the efficiency of tedious manual inspection.

Analysis of data from large-scale transcriptome/proteome experiments. The capabilities of the FunCat as a controlled (structured) vocabulary and classification scheme are not only limited to the functional annotation of genomes but also provide a powerful tool to analyse genome- and proteome-wide data which have been generated by large-scale transcriptome/proteome experiments (15,20–22) as well as computational analysis of the functional networks (23,24). Transcriptome analysis allows monitoring the transcriptional level of each gene of the genome at a certain point of time. While the proteins encoded by a genome represent the whole physiological capability of an organism, the concerted up- or down-regulation of transcripts under certain physiological conditions such as growth on a certain substrate or reaction to stress like heat-shock describes the physiological response of the organism to the specific conditions under investigation. Monitoring transcript levels over time adds the dynamic dimensions of time and experimental condition to the static genomic information. As microarray experiments generate vast amounts of data to be analysed and interpreted, the gene products have to be classified into functional groups to be able to uncover the functional dependences of the genes that are synchronized in order to obtain an appropriate cellular response (25,26). This type of functional projection is facilitated by the scalable architecture of FunCat as the relationship of proteins can be detected by the general functional context given in the first digits in FunCat number describing the respective main categories. Numerous articles are published in the literature, where FunCat annotation was used for this kind of experiments (27,28).

The dataset of yeast protein–protein interactions (29) and the associated FunCat annotation is also used for the analysis of proteome data. The results from large-scale experiments like mass-spectrometry analysis or two-hybrid experiments are used to identify novel protein–protein interactions and protein complexes (30–33). FunCat annotation has also been used to evaluate the reliability of the methods and the significance of the individual data (34) since a functional relation of two gene products is presumably correlated with coinciding functional assignments.

Using the FunCat for comparative genomics. The wealth of genomic data that has been produced during the last years allows predicting the physiological capabilities of organisms *in silico*. The systematic comparative genome analysis (i) allows us to get insight into the evolutionary principles of gene duplication and gene loss (35) and (ii) permits to formulate hypotheses and subsequently to perform instructive experiments to unravel novel insights into the lifestyle of organisms. The bases of the organism comparison are the protein sequence similarity searches which have to be supplemented by automated functional assignments. The relation between sequence similarity and functional conservation has been assessed for protein domains and complete proteins in several investigations (36–38). It was shown that above certain thresholds of sequence similarity, the transfer of functional annotation is highly reliable. However, the transfer of functional information from experimentally characterized proteins to predicted proteins is yet to be performed with care since minor modifications in the sequence can result in highly specific consequences for the interaction with substrates on the atomic level. Sequence similarity found for pairs of membrane-bound transporters, for example well justifies the assignment of both proteins as involved in directed transport but not necessarily which substrate is being transported. Taking these constraints into account, it is feasible to transfer information from manually annotated genomes to genomes of phylogenetically related organisms. The accuracy of the assignment can be improved if synteny between genomes is taken into account, and protein function is predicted between orthologous proteins. Even under favourable conditions of conserved gene order, manual annotation is needed to avoid overinterpretation of the sequence relationship. However, the exhaustive and consistent assignment of proteins into functional classes allows for a functional interpretation of genomes compared.

Comparison of functional annotation schemes

Comparison of FunCat and the Riley scheme. Among the most widely used annotation schemes is the one developed by Monica Riley (6). The primary intention for the development of this catalogue was the description of the known set of proteins from *E.coli*. Later, this annotation scheme was adapted by other databases like TIGR (6,8) or SubtiList (39). In the beginning, the Riley scheme allowed only the assignment of one functional category per gene product, but this turned out to be insufficient for comprehensive protein annotation. Therefore, in a new classification scheme from this group, the MultiFun (40), assignment of multiple categories for one gene product is possible. In addition, MultiFun incorporated the EC and a modified TC nomenclature (3) for the description of enzymes and transport proteins, respectively.

Comparison of annotation schemes is difficult since the quality of its content can hardly be assessed and the general design to some extent reflects the preferences of the respective scientists. Depending on the focus on research, different aspects of protein function will be treated especially with

care and more in-depth than others. However, Thornton and co-workers (41) made an attempt to compare at least scope and architecture of different annotation schemes. Hence, a 'Combined Scheme' (CS) was generated, in which the most general level (level-1) consists of six protein function branches such as 'metabolism', 'process', 'transport', 'structure', 'information pathways' and 'miscellaneous'. In addition, two more layers with increasing specificity contain 16 level-2 and 55 level-3 functional categories, respectively. Mapping of different annotation schemes to the CS generated the so-called functional wheels (FuncWheels), which are graphical representations of hierarchical annotation schemes. Comparison of six annotation schemes showed that the difference in the scope between genome annotation schemes like FunCat and the ones based on the Riley scheme on the one hand and databases like KEGG or WIT on the other hand. The latter are especially strong in fields like e.g. metabolism but lack complete sections of the protein function spectrum which make them unsuitable at least for manual annotation of whole genomes. In the survey of annotation schemes by Thornton and co-workers, FunCat turned out to cover the categories of the functional wheel most comprehensively of the catalogues analysed. Comparison of FunCat and functional catalogues based on the Riley scheme revealed for all three categories, depth, resolution and breadth of the annotation schemes the FunCat as the catalogue with the highest scores. Another major difference between the databases that use the Riley scheme and the FunCat is indicated by their scope. The former annotation scheme is focussed on prokaryotes, whereas the FunCat covers the complete spectrum of organisms from prokaryotes to mammals. Due to its different architecture, the GO annotation scheme was not included in this analysis.

Comparison of FunCat and GO. A recently published annotation scheme, GO (15), was developed with a focus on functional annotation of eukaryotic genomes. In contrast to other annotation schemes, the GO is constructed as a set of acyclic graphs, allowing more than one parent class per child. The direct acyclic graph theory uses the terms 'parent' and 'child', where each 'child' may belong to one or more 'parents'. Another feature of the GO architecture is the description of proteins by three different ontologies, namely biological process, molecular function and cellular localization. For comparison of GO and FunCat, the genome annotation of *S.cerevisiae* is ideally suited since the protein complement of this organism is well characterized by two groups, MIPS and the Saccharomyces Genome Database (SGD) (42) that use those different schemes for annotation.

As the annotation of yeast at MIPS relies on distinct catalogues, FunCat can be compared only to parts of the 'Molecular Function' and 'Biological Process' terms of the GO system. The 'Cellular Compound' section of GO is represented by the MIPS Localization and Complex Catalogues. For instance, the Localization Catalogue is structured into 58 categories which contain more than 13 000 localization assignments for the protein complement of *S.cerevisiae*.

The advantage of the simple and hierarchical FunCat structure is the intuitive category structure. With a question in mind, it is easily possible to browse through the main categories down to the specific level of question and access the annotated entries. The annotation of yeast uses only 18 of the main categories and 258 distinct categories. Excluding the GO terms corresponding to EC numbers, SGD uses 1551 GO terms in their yeast annotation which reflects the different aims of the systems. GO is applied nearly exclusively for the annotation of the genome, describing the function, process and component of a gene. FunCat clearly focuses on the functional process not describing the molecular function on the atomic level which is achieved by applying further catalogues and/or free text. As a result, the database user can browse intuitively through FunCat categories, finding quickly the relevant parts of the 'functome'.

Apart from the yeast genome, there are other considerations while comparing both annotation schemes. GO aims at representing a fine granular description of proteins that provides annotation with a wealth of detailed information. The statement 'GO describes how gene products behave in a cellular context' (www.gene-ontology.org) indicates that such a description should be as detailed as possible. This results in two major difficulties to achieve this goal. On the one hand, a detailed description leads to a large number of terms, the ontology for biological processes alone contains more than 8000 terms and such a plethora of terms is very difficult to be handled for annotators. On the other hand, the large number of possible assignments is prone to inconsistent or even erroneous assignments that tend to propagate subsequently. For different genomes, coverage of the GO assignments differs. A recent investigation quantified the extent of non-uniform annotation using GO. Annotation of *Drosophila melanogaster* was independently performed by two groups, both using GO (43). The result for the ontology 'biological process' was that only 1156 proteins were annotated consistently by both groups, but GO assignments for 4137 proteins were assigned differently.

Future directions

Although most of the FunCat functional categories are self-explanatory, it is required to define all terms semantically unambiguous. This is addressed by establishing a publicly accessible repository, the FunCatDB, which will store various kinds of information that are linked to functional categories such as a formal description of the functional category, frequently associated functional categories or GO terms that can be mapped to functional categories. In addition, interfaces will supply scientists with tables of proteins and their homologues that are assigned to respective functional categories. We intend to map data from external resources like Biochemical Pathways, KEGG and TC database into functional categories whenever possible. In addition, genomes that were annotated at MIPS will serve as a resource for manually curated proteins. The association of functional categories is a critical issue in both manual and automated transfer of protein annotations since it is frequently species-specific or depends on the phylogeny of organisms. For example, this is found in proteins of the tricarboxylic acid cycle (TCA cycle) that are required for biosynthesis of several amino acids. Furthermore, many aerobically growing organisms are able to additionally use the TCA cycle in the energy generation process. Thus, the FunCatDB will not only be a repository of FunCat associated data but also defines cross-correlation of functional categories in a species-specific

manner. The current version of the FunCatDB is available at http://mips.gsf.de/proj/funcatDB.

CONCLUSION

We present the FunCat as an appropriate comprehensive and functional classification scheme for the description of proteins using a structured controlled vocabulary. FunCat proved its value and usability during the annotation of genomes during the last eight years. Comparison with other annotation schemes proved FunCat to be a well-balanced compromise between extensive depth, breadth and resolution but without being too granular. Further developments of FunCat include a description as well as example proteins for individual functional categories. As scientific progress is constantly revealing new insights from the biology of organisms, the development of functional classification schemes is an ongoing process and we will release updates of FunCat in appropriate intervals. The current version of FunCat version 2.0 is available via the World Wide Web (http://mips.gsf.de) and is also presented in the Supplementary Material (Table S3).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education, Science, Research and Technology (BMBF) and European Commission (BFAM: 031U112C, HNB: 01SF9985, EU-Framework 5: QLRI-CT1999-01333).

REFERENCES

1. Endy,D. and Brent,R. (2001) Modelling cellular behaviour. *Nature*, **409**, 391–395.
2. Barrett,A.J. (1997) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions. *Eur. J. Biochem.*, **250**, 1–6.
3. Saier,M.H. (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.
4. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
5. Barker,W.C., Garavelli,J.S., Huang,H.Z., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,C.L., Yeh,L.S.L., Ledley,R.S., Janda,J.F. *et al.* (2000) The protein information resource (PIR). *Nucleic Acids Res.*, **28**, 41–44.
6. Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *FEMS Microbiol. Rev.*, **57**, 862–952.
7. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
8. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
9. Mewes,H.W., Albermann,K., Bahr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G. *et al.* (1997) Overview of the yeast genome. *Nature*, **387**, 7–8.
10. Mewes,H.W., Frishman,D., Gruber,C., Geier,B., Haase,D., Kaps,A., Lemcke,K., Mannhaupt,G., Pfeiffer,F., Schuller,C. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 37–40.
11. Galagan,J.E., Calvo,S.E., Borkovich,K.A., Selker,E.U., Read,N.D., Jaffe,D., FitzHugh,W., Ma,L.J., Smirnov,S., Purcell,S. *et al.* (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.
12. Mewes,H.W., Albermann,K., Bahr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G. *et al.* (1997) Overview of the yeast genome. [Erratum (1997) Nature, **387**, 9.] *Nature*, **387**, 737.
13. Ruepp,A., Graml,W., Santos-Martinez,M.L., Koretle,K.K., Volker,C., Mewes,H.W., Frishman,D., Stocker,S., Lupas,A.N. and Baumeister,W. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*, **407**, 508–513.
14. Salanoubat,M., Lemcke,K., Rieger,M., Ansorge,W., Unseld,M., Fartmann,B., Valle,G., Blocker,H., Perez-Alonso,M., Obermaier,B. *et al.* (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 820–822.
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. *Nature Genet.*, **25**, 25–29.
16. Pavlidis,P., Weston,J., Cai,J. and Noble,W.S. (2002) Learning gene functional classifications from multiple data types. *J. Comput. Biol.*, **9**, 401–411.
17. Michal,G. (1998) *Biochemical Pathways: An Altlas of Biochemistry and Molecular Biology*. Wiley & Sons, Inc. Hoboken, NJ.
18. De Hertogh,B., Carvajal,E., Talla,E., Dujon,B., Baret,P. and Goffeau,A. (2002) Phylogenetic classification of transporters and other membrane proteins from *Saccharomyces cerevisiae*. *Funct. Integr. Genomics*, **2**, 154–170.
19. 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.
20. Balazsi,G., Kay,K.A., Barabasi,A.L. and Oltvai,Z.N. (2003) Spurious spatial periodicity of co-expression in microarray data due to printing design. *Nucleic Acids Res.*, **31**, 4425–4433.
21. Clare,A. and King,R.D. (2002) How well do we understand the clusters found in microarray data? *In Silico Biol.*, **2**, 511–522.
22. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
23. Dobrin,R., Beg,Q.K., Barabasi,A.L. and Oltvai,Z.N. (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics*, **5**, 10.
24. Yeger-Lotem,E., Sattath,S., Kashtan,N., Itzkovitz,S., Milo,R., Pinter,R.Y., Alon,U. and Margalit,H. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. *Proc. Natl Acad. Sci. USA*, **101**, 5934–5939.
25. Tornow,S. and Mewes,H.W. (2003) Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.*, **31**, 6283–6289.
26. Jansen,R. and Gerstein,M. (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res.*, **28**, 1481–1488.
27. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al.* (1999) Functional characterization of the *S.cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
28. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Veronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., Andre,B. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
29. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Güldener,U., Mannhaupt,G., Münsterkotter,M., Pagel,P., Strack,N., Stümpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
30. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

31. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
32. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
33. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
34. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
35. Koonin,E.V., Wolf,Y.I. and Karev,G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223.
36. Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.
37. Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.
38. Wilson,C.A., Kreychman,J. and Gerstein,M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, **297**, 233–249.
39. Moszer,I., Jones,L.M., Moreira,S., Fabry,C. and Danchin,A. (2002) SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res.*, **30**, 62–65.
40. Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
41. Rison,S.C., Hodgman,T.C. and Thornton,J.M. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, **1**, 56–69.
42. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al.* (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.
43. Mi,H.Y., Vandergriff,J., Campbell,M., Narechania,A., Majoros,W., Lewis,S., Thomas,P.D. and Ashburner,M. (2003) Assessment of genome-wide protein function classification for *Drosophila melanogaster*. *Genome Res.*, **13**, 2118–2128.