Funct Integr Genomics (2000) 1:56–69
Digital Object Identifier (DOI) 10.1007/s101420000005

ORIGINAL PAPER

Stuart C. G. Rison · T. Charles Hodgman
Janet M. Thornton

Comparison of functional annotation schemes for genomes

Received: 25 October 1999 / Accepted: 15 December 1999 / Published online: 12 April 2000

Abstract In this paper we survey a number of functional classification schemes applicable to genomes. We present the concepts of depth, breadth and resolution as descriptors of the schemes' scope and architecture and compare selected classifications according to these criteria. We also generate a 'Combined Scheme' against which we map six classifications which we believe are representative of the range currently available. The mapping allows the generation of 'FuncWheels', which are graphical representations of hierarchical classification schemes. They are used to illustrate similarities and differences in functional space coverage. This survey highlights many issues related to the design and implementation of gene product functional classifications, which are discussed in the light of emerging 'second-generation' schemes.

Key words Functional genomics · Protein function · Gene product classification · Ontology

Introduction

The analysis of genes and gene products is usually performed in order to discover, confirm or clarify their function. The function of a gene product is its *raison d'être*; understanding this function is key to understanding how a limited number of interacting gene products can generate life, from simple unicellular organisms to the incredibly complex multi-cellular *Homo sapiens*.

The association of functional data with gene products (functional annotation) first appeared in databases of gene products such as SWISS-PROT or PIR, in which protein entries are accompanied by careful human-generated annotations of their empirically determined or predicted role (Bairoch and Apweiler 1999; Barker et al. 1999). However, although these annotations include keywords chosen from a controlled vocabulary, they are currently not formally organised in a functional annotation scheme, although there have been many efforts to classify such databases on the basis of their annotation (Tamames et al. 1998; Eisenhaber and Bork 1999; Licciulli et al. 1999).

The first extensive gene product functional classification scheme was devised in 1993 to catalogue the 1171 *Escherichia coli* genes known at the time (Riley 1993). This was some 4 years before the complete genome for *E. coli*, currently estimated to have approximately 4,300 genes, was sequenced (Blattner et al. 1997). An updated version of the classification scheme was published in 1996 (Riley and Labedan 1996) and regular updates can be found in GenProtEC (Riley 1998a) and EcoCyc (Karp et al. 1999). More recently, genome sequencing projects have been the driving force behind the development of alternative functional annotation schemes.

Once a genome is sequenced, the first step is to identify genes and attempt to annotate the functions of their products. However, in order to understand the overall mechanisms operating, the genes need to be organised according to the biological processes they perform. Such an organisation needs a standardised functional annotation scheme. Functional classification schemes are usually simple hierarchies which begin by defining function in very general terms and become increasingly specific as one progresses down the hierarchy. When dealing with genomes, such schemes allow the gene complement of an organism to be sub-divided into sets of functionally relat-

S.C.G. Rison
Ludwig Institute for Cancer Research, 91 Riding House Street, London, W1P 8BT, UK

J.M. Thornton · S.C.G. Rison
Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK
e-mail: rison@biochem.ucl.ac.uk

T.C. Hodgman
Global Research Information Systems, GlaxoWellcome Medicines Research Centre, Gunnels Wood Road, Stevenage, SG1 2NY, UK

J.M. Thornton
Department of Crystallography, Birkbeck College, Malet Street, London, WC1E 7HX, UK

Table 1 List of gene product classification schemes: references and URLs

Functional classification	URL	Reference(s)
GenProtEC	http://genprotec.mbl.edu/start/	Riley 1998a
EcoCyc	http://ecocyc.pangeasystems.com/	Karp et al. 1999
Sanger Centre (*M. tuberculosis*)	http://www.sanger.ac.uk/Projects/M_tuberculosis/	Cole et al. 1998
Institut Pasteur: SubtiList	http://bioweb.pasteur.fr/GenoList/SubtiList/	Moszer et al. 1996
Institut Pasteur: TubercuList	http://bioweb.pasteur.fr/GenoList/TubercuList/	Cole et al. 1998
MIPS: Yeast Genome Database (MYGD)	http://www.mips.biochem.mpg.de/proj/yeast/	Mewes et al. 1997
MIPS: *Arabidopsis thaliana* Database (MATDB)	http://www.mips.biochem.mpg.de/proj/thal/	Mewes et al. 1999
MIPS: PEDANT	http://pedant.mips.biochem.mpg.de/	Frishman and Mewes 1997
Proteome.com: YDP and WormPD	http://www.proteome.com/databases/	Hodges et al. 1999
MGI: Mouse Genome Database (MGD)	http://www.informatics.jax.org/	Blake et al. 1999
TIGR: Microbial databases	http://www.tigr.org/tdb/mdb/mdb.html	See TIGR genome papers, e.g. Fleischmann et al. 1995 (*H. influenzae*) Fraser et al. 1995 (*M. genitalium*)
TIGR: Expressed Gene Anatomy Database (EGAD)	http://www.tigr.org/tdb/egad/egad.html	n/a
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg/	Ogata et al. 1999
What Is There (WIT)	http://wit.mcs.anl.gov/WIT2/	Selkov et al. 1998
COG	http://www.ncbi.nlm.nih.gov/COG/	Tatusov et al. 1997
Gene Ontology	http:/www.geneontology.org/	Ashburner et al. 1999, other documents available on-line

ed gene products and also help to provide an overview of the biology of an organism. There are currently many different schemes used to annotate genomes. Even the interpretation of the *Mycoplasma genitalium* genome, which with 470 genes is the smallest completed (Fraser et al. 1995), greatly profits from organisation into a scheme.

We surveyed a number of WWW sites with functional classification schemes, and literature references and URLs for these are given in Table 1 . We described a number of these schemes in terms of their resolution, depth and breadth; these terms help determine the scope and architecture of the schemes. We then focused on six functional classification schemes which we considered representative of the range currently available: EcoCyc (essentially identical to GenProtEC), TIGR, SubtiList, MIPS/PEDANT, KEGG and WIT. EcoCyc and GenProtEC are updated versions of Riley's original scheme (Riley 1998a; Karp et al. 1999), while TIGR (Fleischmann et al. 1995) and SubtiList (Moszer et al. 1996) are adaptations of it. The MIPS/PEDANT scheme was developed by the researchers at the Munich Information Centre for Protein Sequences (MIPS) (Frishman and Mewes 1997; Mewes et al. 1997). Finally, KEGG (Ogata et al. 1999) and WIT (Selkov et al. 1998) mainly address regulation and metabolic pathways. Mapping of these schemes onto a 'Combination Scheme' allowed us to compare them. This analysis included the generation of FuncWheels, a novel way of graphically depicting gene product functions. Certain schemes, although independently implemented, were not included on the basis of their similarity with schemes already present in the selection. For example, the *Mycobacterium tuberculosis* genome classification scheme employed at the Sanger Centre is essentially the same as the Riley scheme (Cole et al. 1998) and the COGs scheme is a 'generalisation' of the Riley scheme into broader functional categories (Tatusov et al. 1997). Therefore in both of these cases we would expect to see a similar coverage of functional space. Furthermore, we did not include the 'Gene Ontology' scheme (Gene Ontology Consortium 1999) for mapping, even though it represents a separate type of functional classification scheme. There were two reasons for this: first, its scope is much larger, and its structure more complex, than the chosen schemes; secondly, the 'Gene Ontology' represents a radical rethink of gene product function classification. Therefore, its direct comparison with the chosen schemes would have been difficult and ineffective.

This work highlighted many of the issues involved in functional scheme design and implementation and we discuss these with particular focus on recent developments in this area.

Methods

Classification scheme uploading

All analysed functional classification schemes were available on-line during the course of August 1999 when

the data for this paper were collected. Where relevant and possible, they were uploaded locally and converted to a format suitable for storage in the publicly available postgreSQL relational database management system (PostgreSQL 1999).

We conceptualised the schemes as trees – sets of connected nodes organised hierarchically. The nodes are functions or functional categories (e.g. 'DNA synthesis' or 'Transport'). Progression from the top (level-1) nodes down to the terminal nodes represents increasingly specific functions. The functions can be identified by means of a hierarchical key (for example, function 5.3.1) in which the first number (5) refers to level-1, the second (3) to level-2 etc.

All the uploaded classification schemes were easily stored in such a format with the exception of the 'Gene Ontology' (Gene Ontology Consortium 1999). The latter is implemented as a directed acyclic graph (DAG), which has a more complex data structure than a tree. A DAG allows a node to have more than one parent and for the edges to distinguish between different types of relationships between nodes. It is not possible to covert a DAG data structure onto a tree structure without some concessions; in particular, we lost the capacity to distinguish between relationship types, and nodes with more than one parent had to be duplicated and inserted separately within the tree structure. Nevertheless, we used this conversion to estimate the depth, breadth and resolution of the 'Gene Ontology'.

Design of the 'Combination Scheme' and scheme mapping

In order to compare the six chosen functional classification schemes, we designed a 'Combination Scheme' (CS) of gene product functions. The CS is not intended as a replacement scheme but was designed solely to facilitate a comparison of the current schemes to appreciate their similarities and differences.

The generation of the CS was iterative. It involved the collation of all functional nodes described in the selected schemes and their organisation into a tentative scheme. Because we wanted the CS to be as simple as possible, the first attempted scheme had two levels only. However, it was soon evident that such a scheme was not viable. We therefore designed a three level scheme that was modified during two rounds of mapping. This generated a CS with a broad coverage of all functions described in the selected schemes without excessive bias towards any one of them. The details of the design of the CS are given below.

All the nodes in all the schemes investigated were collected and obviously duplicated functions or functional categories were eliminated. The complete list was reorganised into a three-level tree with six nodes at the top-level and 73 level-3 nodes. The scheme was manually generated with care but remained arbitrary in many respects. For each scheme investigated, all level-1, 2, 3 and 4 nodes (a total of 1,315 nodes) were compared with nodes in the CS and mapped to the lowest (most specific) CS node possible. To simplify the mapping process, we only allowed a one to one relationship between a node in the mapped schemes and the CS. In certain cases, such a rule made mapping impossible. For example, the node 'Cell growth, Cell division and DNA synthesis' in the MIPS/PEDANT scheme could be mapped onto three different CS nodes. In some instances, where such multi-functional categories overwhelmingly pointed towards one CS node, we mapped onto that node but we usually skipped these functions rather than assign them incorrectly.

In order to keep the CS as universal as possible, we tried to avoid including functions as a separate node which tended to be species specific. For example, 'Sporulation', a property specific to certain organisms including *Bacillus subtilis*, was present as a function in the SubtiList scheme and could have justifiably been included as an additional 'Organism process' in the CS. However, because the function is specific to a very limited number of organisms, it was subsumed into the more generalised 'Adaptation' category on the basis that 'Sporulation' is usually initiated in response to nutrient starvation.

To identify and eliminate scheme-specific nodes from the CS, we analysed the results of the first mapping and identified all CS nodes associated with only one or two distinct schemes. Each of these nodes was reviewed and either subsumed into another node, combined with other scheme-specific nodes, deleted, reclassified or, in rare instances where the function was considered critical, left unchanged. We then repeated the mapping process to determine coverage of the CS by each of the six selected schemes. A similar iterative process has previously been used to classify SWISS-PROT function annotations (Tamames et al. 1996) in which SWISS-PROT entry keywords were mapped onto a one level scheme, based on the segregation of the Riley scheme into three nodes: 'Energy', 'Information' and 'Communication'. All our mapping procedures were performed using database backed Perl scripts (Wall et al. 1996) and further details on the mapping process, the mapping results and the mapped schemes can be found at http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/.

Generation of FuncWheels

FuncWheels are a graphical representation of all the nodes in a three level classification scheme. The wheel is separated into differently coloured segments each representing a top-level node and proportional in size to the number of level-3 nodes in them. The wheel is also divided into an inner disc and an outer ring. The inner disc of the wheel is divided into segments representing level-2 nodes, again of a size proportional to the number of level-3 nodes in them, whilst the outer ring is divided into equally sized segments each representing a level-3 node (see Fig. 2 for an example of a FuncWheel).

To illustrate the coverage of the CS by mapped schemes, we used FuncWheels in which non-matched CS level-3 nodes were blanked out. In addition, level-3 nodes were considered unoccupied and blanked out if more than two-thirds of their child (level-3) nodes were blanked, unless they spanned only two level-3 nodes when they were considered unoccupied only if neither of the two level-3 nodes was occupied (see Fig. 3 for examples of such 'coverage' FuncWheels).

All FuncWheels were generated using a modified version of the software used to generate 'CATH wheels' (Martin et al. 1998). Data for the generation of these wheels were extracted from the scheme database using SQL queries (Bowman et al. 1996) and Perl scripts (Wall et al. 1996).

Results

Scheme survey

The functional classification scheme data gathered during this survey are summarised in Table 2 , which also includes some data on related classification schemes (e.g. classification of gene products by subcellular localisation or by Enzyme Commission code). Schemes discussed in this paper are in bold in the table, but the other schemes are included for completeness. Table 2 also lists means of accessing the assigned function classification(s) of gene products (e.g. by gene name, by EMBL code etc.).

The surveyed classification schemes were for the most part related to genome sequencing initiatives or analysis of genomes. It is worth pointing out that the 'Gene Ontology' scheme – designed by a consortium of researchers affiliated with the '*Saccharomyces cerevisiae* Genome Database' (SGD), the *Drosophila melanogaster* database 'FlyBase' and the 'Mouse Genome Informatics' (MGI/MGD) group – is actually composed of three parts (Gene Ontology Consortium 1999). These are schemes concerning cellular localisation, biological processes and biological function. This distinction of biological process and function is extremely pertinent to the design and implementation of functional classification schemes and will be discussed later.

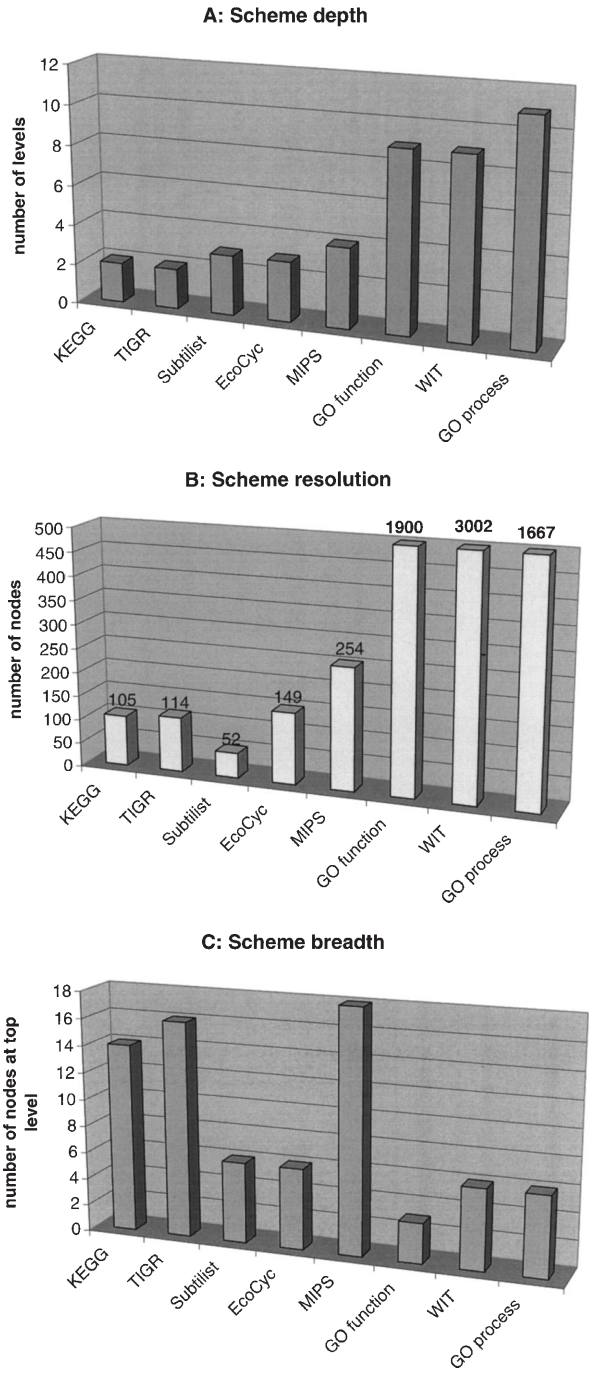

Fig. 1 **A** Depth, **B** resolution, **C** breadth of the six mapped schemes and of the 'Gene Ontology' process and function classifications

Single vs multi-organism schemes

Some of the genome-related WWW sites were single organism databases, and the others dealt with multiple organisms. Whilst the MIPS databases included two single organism databases (MYGD and MATD for *S. cerevisiae* and *Arabidopsis thalliana* respectively) and a multiple organism database (PEDANT), they all shared one functional classification scheme 'FunCat' originally based on yeast gene products but adapted to be applicable to a number of other organisms (Mewes et al. 1997, 1999). Whilst the 'FunCat' is used in PEDANT to classify many gene complements, including the partially completed human one, it nevertheless remains yeast orientated, although efforts are being made to extend the scope of the classification (D. Frishman, personal communication). The 'Gene Ontology' is being developed with the aim of being applicable to many organisms (Gene Ontology Consortium 1999; Riley 1998b). We note that it is considerably more complex than previous schemes by an order of magnitude.

Table 2 An overview of gene product classification schemes identified during a survey of 16 genome related WWW sites. All the gene product classification schemes encountered are listed; those discussed in the paper are in bold, and where applicable three examples of nodes in the top level of the classification are given. The table also indicates the breadth, depth and resolution of the schemes and lists alternative access routes to gene product data. Schemes empirically determined to support multiple functional annotations of single gene products are flagged in the '*1:M annotation'* column

Database	Classifications	Examples of top nodes	Depth	Breadth	Resolution	1:M annot.	Alternative access routes to gene product information
Single organism							
GenProtEC	**Physiological role**	Cell process, Metabolism of small mols., Structural elements	3	10	118	Y	Gene/protein name, Blattner number, SwissProt ID
	Gene type	Enzyme, RNA, Membrane	1	21	21		
	EC scheme	Oxidoreductases, Transferases, Ligases	4	6			
EcoCyc	**Taxonomy of genes**	Cell process, Lipid biosynthesis, Broad regulatory functions	3	6	150	Y	Gene/protein name
	Taxonomy of pathways	Signal-transduction pathways, Intermediary metabolism, Biosynthesis	3	6	30		
	Taxonomy of reactions	EC-Reactions, Binding reactions, Transport reactions	4	6	305		
	Taxonomy of s compound	Hormones, Lipids, All carbo-hydrates	5	16	85		
Institut Pasteur: SubtiList	**Functional categories**	Intermediary metabolism, Information pathways, Cell envelope and cell processes	3	6	52		Gene name, Gene chromosomal location, Text search
Institut Pasteur: TubercuList	Functional categories	Intermediary metabolism and respiration, Information pathways, Cell wall and cell processes	1	11	11		Gene name, Gene chromosomal location, Text search
MIPS: MYGD	**Functional catalogue (FunCat)**	Transport facilitation, Energy, Cellular organisation	4	14	75	Y	Gene name, Yeast gene code, Sequence similarity, PIR code, EMBL code, CDYS code
	Protein classes (ClassCat)	ATPases, Transcription factors, Molecular chaperones	4	22	187	Y	
	Subcellular localisations (SubcellCat)	Plasma membrane, Nucleus, Mitochondria	2	15	42		
	EC scheme	As above	4	6			
	Protein complexes (CompCat)	Replication complexes, Transcription complexes, Cell-cycle checkpoint	4	70	317		
	Phenotypes (PhenCat)	Stress response defect, Sensitivity to antibiotics, Auxotrophy defects	4	11	179		
	Pathways	Amino acid metabolism, Energy, Signal transduction	2	10			
MIPS: MATD (Arabidopsis thaliana)	**Functional catalogue (FunCat)**	Transport facilitation, Energy, Cellular organisation	4	17	255	Y	Keyword, BAC/cosmid clone code, MIPS code, Protein entry code
	Genetic element type	tRNA, 5'UTR, Gene/protein	1	14	14		

Table 2 (Continued)

Database	Classifications	Examples of top nodes	Depth	Breadth	Resolution	1:M annot.	Alternative access routes to gene product information
YPD (*S. cerevisiae*) / WormPD (*C. elegans*)	**Functional categories**	Binding protein, Ligase, Protein kinase	1	1	55/55	Y	Gene name, Keyword, Sequence similarity
	Cellular role	Amino-acid metabolism, Pol l transcription, Signal transduction	1	1	42/45	Y	
	Subcellular localisations	Plasma membrane, Golgi, Nuclear nucleolus	1	1	24/32	Y	
	Molecular environment	Actin-cytoskeleton associated, Protein synthesis factor	1	1	10/10		
	Genetic properties	XX animals are male, 1 intron, Null lethal	1	1	23/119		
	Post-translational modifications	Phosphorylation, Ubiquitination, N-linked glycosylation	1	1	30/25		
Sanger Centre (*M. tuberculosis*)	**Gene list**	Cell process, Lipid biosynthesis, Broad regulatory function	4	6	119		Gene name, Sequence similarity
MGI: MGD	Phenotypic classification	Biochemical, Anatomical, Physiological	2	8	34		Gene/Protein name, Gene/Protein ID, Sequence
Multiple organisms dbs							
TIGR	**Gene identification list**	Amino acid biosynthesis, Regulatory functions, Cell envelope	2	16	114		Gene name, Function text search, Locus search
	EGAD Cellular roles	Metabolism, Cell signalling/ cell communication, Cell structure/motility	3	6	49	Y	Gene name, Sequence ID
PEDANT databases	**"Yeast" FunCat**	Transport facilitation, Energy, Cellular organisation	4	16	240	Y	Sequence ID, Text search, PIR keywords/ superfamilies, PROSITE patterns, PFAM domains
	EC scheme	As above	4	6			
	Structural classes	All alpha, All beta, Alpha beta	1	4	4		
	SCOP scheme	All alpha proteins, Alpha plus beta protein, Membrane and cell surface proteins	5	10			
Pathway related							
KEGG: GENES	**Gene catalogue (functional and metabolic)**	Energy metabolism, Membrane transport, Signal transduction	2	14	105	Y	Gene names, EC numbers, Via DGET/LinkDB, Sequence similarity
KEGG: PATHWAYS	Pathway classification (metabolic)	Energy metabolism, Metabolism of complex lipids, Metabolism of macromols	2	10	90		
KEGG: PATHWAYS	Pathway classification (regulatory)	Signal transduction, Ligand-receptor interaction, Molecular assembly	2	4	10		
KEGG: LIGAND/ COMPOUND	Ligand (compound)	Carbohydrate, Lipid, Nucleic acid	3	5			
KEGG: LIGAND/ ENZYME	**Ligand (enzyme EC)**	Oxidoreductases, Transferases, Ligases	4	6			
WIT	**General overview**	Intermediate metabolism and bioenergetics, Information pathway, Structure and function of the cells	9	6	3,002		Text search in ORFs, pathways, enzymes, overviews, Ortholog clusters, Operon clusters, Sequence

Table 2 (Continued)

Database	Classifications	Examples of top nodes	Depth	Breadth	Resolution	1:M annot.	Alternative access routes to gene product information
Ontologies							
Gene Ontology	**Functional primitive**	Protein, Ribozyme, Nucleic acid	9	3	1,740		N/A
	Cellular component	Extracellular, Intracellular, Unlocalised	9	3	385		
	Process	Cell growth and maintenance, Cell communication	11	6	1,667		
Misc.							
COGs	**Functional annotation**	Information storage and processing, Cellular processes, Metabolism	2	3	21		Sequence similarity, Functional class

Scheme depth, resolution and breadth

To gain an understanding of the scope and structure of the surveyed schemes, we collected data on the number of levels, the number of nodes at the top level, and the total number of nodes (see Table 2). These three elements can be used to represent the depth, breadth and resolution of the classifications respectively and are, for a selection of schemes, plotted in Fig. 1.

Depth can be thought of as the potential of the scheme for division into subsets: the greater the depth, the further the scheme allows subdivision into functional groups. For example, the MIPS/PEDANT scheme has a depth of four, and when applied to the *S. cerevisiae* gene complement yields sets of 742 ORFs involved with transcription (level-1), 539 ORFs involved with mRNA transcription (level-2), 411 ORFs involved with mRNA synthesis (level-3) and 30 ORFs involved with chromatid modification (level-4). The depth of a scheme represents the amount of magnification that can be applied to functions; much like a microscope, the higher the magnification, the more specifically one can resolve a particular subset of functions. The depths of the mapped schemes together with that of the 'Gene Ontology' function and process classifications are plotted in Fig. 1A. The depth indicated in the bar chart is the maximum depth encountered and not all branches of the functional tree necessarily extend that far. Depths ranged from two (TIGR and EcoCyc) to 11 ('Gene Ontology' process scheme). Only the 'Gene Ontology' schemes and the WIT scheme have depths greater than four levels. The WIT database, constructed to aid the reconstruction of metabolic pathways, contains a painstakingly detailed classification of metabolism and information pathway functions (404 terms related to these functions are found at a depth greater than six). The 'Gene Ontology' function and process schemes had a maximum depth of nine and 11 respectively which reflect the intricacy of the scheme.

The next parameter is the resolution of a scheme (Fig. 1B). We used the intuitive hypothesis that schemes with a large number of function nodes were likely to have more specific functional descriptions. To use an analogy from the computer world, if all gene functions are represented as a screen – where the fundamental unit is function rather than pixels – the greater the number of function nodes, the higher the resolution. Resolutions ranged from 52 for SubtiList to 3,002 for WIT. Again the size of the WIT scheme is apparent, as is that of the 'Gene Ontology' schemes, which have a combined resolution of over 3,500 nodes, illustrating the minutiae that has gone into designing these schemes. Depth and resolution are closely linked: the greater the depth and resolution of a scheme, the finer its granularity (Gene Ontology Consortium 1999).

The breadth of the schemes, plotted in Fig. 1C, and represented by the number of nodes at the top level, helps to illustrate the coverage of the scheme. The broadest schemes, TIGR and MIPS/PEDANT, had 16 nodes at the top level and the narrowest is the section of the 'Gene Ontology' dedicated to gene product function with a breadth of three. TIGR and MIPS/PEDANT do offer good coverage of function but judging a scheme by its breadth can be misleading. Whilst the 'Gene Ontology' function ontology has a depth of three, this is because, at the top level, the ontology distinguishes between proteins, ribozymes and nucleic acids. The protein node itself has 16 level-2 nodes (e.g. 'signal transduction' and 'structural protein') many of which tend to be top-level nodes in the other schemes. Therefore, a scheme with a limited breadth does not necessarily have a narrow coverage of function.

The Combination Scheme

The CS, designed to allow comparison between schemes, was generated by compiling all level-1, level-2, level-3 and level-4 nodes in the six selected schemes and joining, splitting, deleting or renaming them during two

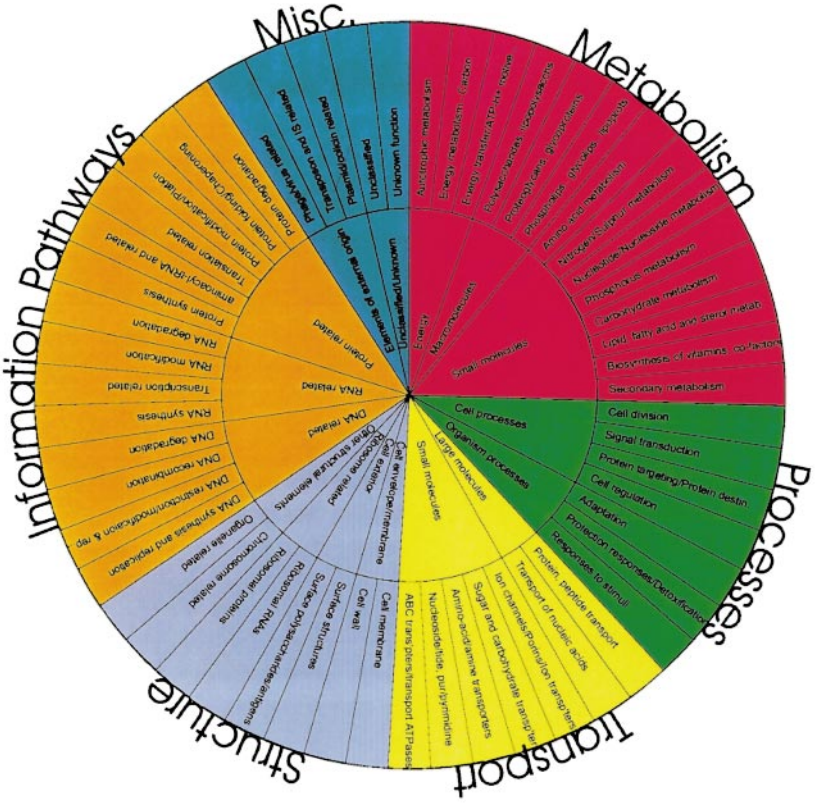

Fig. 2 The 'Combination Scheme' FuncWheel. Level-3 nodes are labelled in the outer ring, level-2 nodes in the inner disc. All identically coloured segments belong to the same level-1 node; these nodes are labelled on the edge of the FuncWheel

rounds of mapping. The first pass mapping was performed to identify CS nodes biased towards one particular node. Of the first-pass level-3 CS nodes, 17 were found to be associated with only one scheme and 12 nodes to be associated with two. As our aim was to avoid such bias, we modified the CS to reduce their incidence. Of the 29 level-3 nodes identified as potentially scheme-specific, 15 nodes were variously grouped into combined nodes, four nodes were deleted, three nodes were subsumed into other nodes and seven nodes were kept unchanged. In the first pass mapping, we skipped 149 of the 864 nodes.

The resulting version of the CS still had three levels and six level–nodes but now only had 55 level-3 nodes and, with minor modifications, became the working version shown in Table 3 and illustrated as a FuncWheel in Fig. 2. Second-pass mapping of the selected schemes to this CS confirmed that the incidence of over-specific nodes had been minimised. Only 139 nodes were skipped during this second round mapping. The mapping also generated the data used in the generation of FuncWheels for the six selected schemes. Nevertheless, mapping of selected schemes onto the CS was difficult to complete. All schemes use umbrella terms (especially at the higher levels) and some of these did not resolve well onto our CS; by extension, it was not always trivial to unambiguously reclassify the children nodes of such umbrella terms within the CS. The CS is amongst the 'smallest' of the schemes, with only 77 nodes, and yet it could accommodate all the other schemes combined, even those with markedly more nodes such as the 254 belonging to the MIPS/PEDANT scheme. This is a good indicator of the level of subsuming involved in generating and mapping to the CS, and explains why we do not recommend its use as a substitute scheme.

The mapping was also subjected to a number of arbitrary assignments, e.g. when distinguishing functions relating to energy metabolism from those concerned with small molecule metabolism. As far as possible, we tried to be consistent; for example, the tricarboxylic acid (TCA) cycle, a functional node found in many of the schemes, was always mapped onto 'energy metabolism – carbon' (CS 1.1.2) regardless of whether it was under a different parent node (e.g. 'Carbohydrate metabolism') in the mapped scheme.

It is interesting to note that the final CS is similar to the eight top-node scheme employed by Tamames et al. (1997) in their analysis of functionally related genes in *Haemophilus influenzae* and *E. coli* although it was designed entirely independently. The Tamames scheme was adapted from the TIGR scheme (Fleischmann et al. 1995) and found to be a good compromise between functional specificity and ease of use for the analysis of genomes.

Functional scheme comparison

A full list of mapping assignments, along with further details regarding the mapping process, can be found in our WWW site (http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/). The mapping allowed us to compare

Table 3 The 'Combination Scheme' (CS). The hierarchical CS was used as a common reference to compare various classification schemes. The CS has six level-1 nodes, 16 level-2 nodes and 55 level-3 nodes. Numbers represent the key of the functions (e.g. CS 4.1.1 is 'cell membrane')

1 Metabolism

1.1 Energy
1.1.1 autotrophic (energy) metabolism
1.1.2 energy metabolism (carbon)
1.1.3 energy transfer/ATP-proton motive force
1.2 Macromolecules
1.2.1 polysaccharides, lipopolysaccharides
1.2.2 proteoglycans, glycoproteins
1.2.3 phospholipids, glycolipids, lipoproteins
1.3 Small molecules
1.3.1 amino acid metabolism
1.3.2 nitrogen/sulphur metabolism
1.3.3 nucleotide/nucleoside metabolism
1.3.4 phosphorus metabolism
1.3.5 carbohydrate metabolism
1.3.6 lipid, fatty acid and sterol metabolism
1.3.7 biosynthesis of vitamins, co-factors and prosthetic groups
1.3.8 secondary metabolism

2 Processes

2.1 Cell processes
2.1.1 cell division
2.1.2 signal transduction
2.1.3 protein targeting/protein destination
2.1.4 cell regulation
2.2 Organism processes
2.2.1 adaptation
2.2.2 protection responses/detoxification
2.2.3 responses to stimuli

3 Transport

3.1 Large molecules
3.1.1 protein, peptide transport
3.1.2 transport of nucleic acids
3.2 Small molecules
3.2.1 ion channels/porins/ion transporters
3.2.2 sugar and carbohydrate transporters
3.2.3 amino-acid/amine transporters
3.2.4 nucleoside, nucleotide, purine and pyrimidine transporters
3.2.5 ABC transporters/transport ATPases

4 Structure and organisation of structure

4.1 Cell envelope/membrane
4.1.1 cell membrane
4.1.2 cell wall
4.2 Cell exterior
4.2.1 surface structures
4.2.2 surface polysaccharides/antigens
4.3 Ribosome related
4.3.1 ribosomal RNAs
4.3.2 ribosomal proteins
4.4 Other structural elements
4.4.1 chromosome related
4.4.2 organelle related

5 Information Pathways

5.1 DNA related
5.1.1 DNA synthesis and replication
5.1.2 DNA restriction/modification and repair
5.1.3 DNA recombination
5.1.4 DNA degradation
5.2 RNA related
5.2.1 RNA synthesis
5.2.2 transcription related
5.2.3 RNA modification
5.2.4 RNA degradation
5.3 Protein related
5.3.1 protein synthesis
5.3.2 aminoacyl-tRNA synthetases/transferases and aminoacyl-tRNA
5.3.3 translation related
5.3.4 protein modification/phosphorylation
5.3.5 protein folding/chaperoning
5.3.6 protein degradation

6 Miscellaneous

6.1 Elements of external origin
6.1.1 phage/virus related
6.1.2 transposon and is related
6.1.3 plasmid/colicin related
6.2 unclassified/unknown
6.2.1 unclassified
6.2.2 unknown function

schemes by generating a set of coverage FuncWheels as shown in Fig. 3. In these, CS nodes not represented in each of the six selected schemes are blanked out (see the Methods section for full details on the generation of FuncWheel). Blanked nodes can be determined by comparison of the coverage FuncWheels with the CS FuncWheel in Fig. 2.

The most extensive coverage of the CS is provided by the MIPS/PEDANT scheme with only five level-3 nodes unoccupied (Fig. 3D). The MIPS/PEDANT scheme is also the only scheme to have all its level-2 nodes occupied. Conversely, the KEGG scheme, with 11 out of 16 level-2 nodes blanked, has the lowest overall coverage: level-1 segments for 'Processes', 'Transport', and 'Information pathways' are almost entirely blanked out although the scheme has good coverage of metabolism (Fig. 3E).

The WIT scheme has good overall coverage except for the 'Processes' level-1 segment (and the relatively trivial 'Miscellaneous' level-1 segment). WIT is also unsurpassed in its coverage of metabolism and is the only scheme with no blanks in that segment (Fig. 3F).

It comes as no surprise that EcoCyc (Fig. 3A) and TIGR (Fig. 3B) exhibit very similar coverage as they are both based on the Riley scheme. The SubtiList scheme is partially based on the Riley scheme but a number of functions have been combined and adapted for *B. subtilis* with consequent partial loss of CS coverage (Fig. 3C).

In the future, we will use FuncWheels to graphically depict the functional coverage of fully sequenced genomes and we hope to gain insight into the functional distinctions that characterise genomes. A similar comparison of genomes on the basis of their gene product function distribution performed on 44 organisms using a very simplified three-node scheme, highlighted differences between viruses, bacteria, eukaryotic unicellular organisms, plants and animals (Tamames et al. 1996).

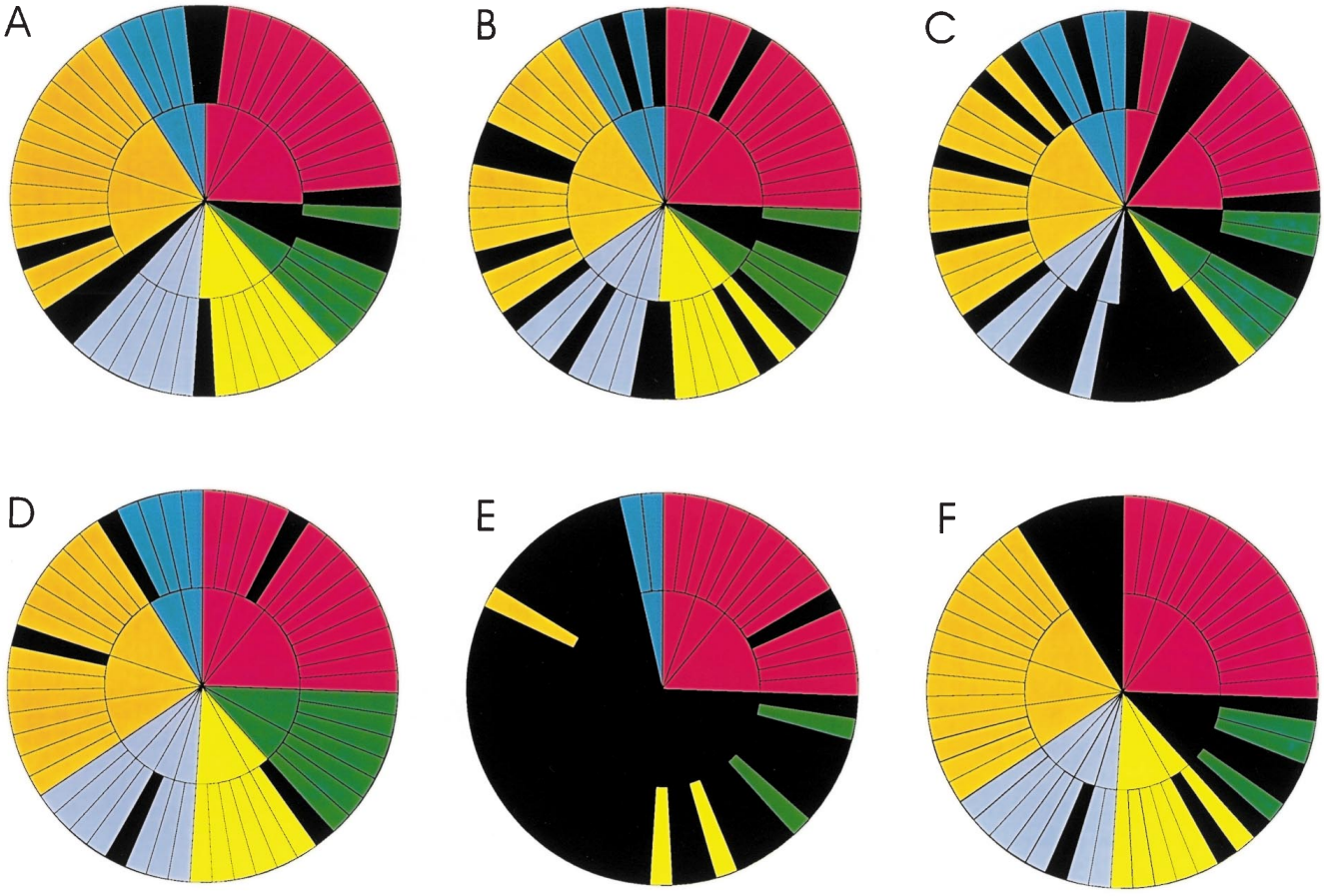

Fig. 3A–F Coverage of the 'Combination Scheme' (CS) illustrated using FuncWheels. **A** EcoCyc; **B** TIGR; **C** SubtiList; **D** MIPS; **E** KEGG; **F** WIT. In these FuncWheels, nodes in the CS not represented by the illustrated scheme are blanked out. CS functions present and absent can be identified by reference to Fig. 2

Discussion

Mapping and scheme comparison limitations

Clearly, the comparison of schemes depends on their mapping to the CS. This mapping is not straightforward and is constrained by the requirement for one-to-one correspondence between a node in the mapped schemes and a node in the CS. Therefore, the absence of mapping to a CS node can mean one of three things:

1. The CS node is not represented in the mapped scheme. For example, the KEGG scheme, at the time of data gathering, did not explicitly describe functions pertaining to 'Structure', and therefore this segment in the KEGG FuncWheels (Fig. 3E) is completely blanked out.
2. The mapping process has assigned nodes that could have mapped to the 'missing' CS node elsewhere. For example, Fig. 3B shows that the CS nodes 'Cell membrane' (CS 4.1.1) has not been 'mapped to' by the TIGR scheme, yet this scheme has a 'Cell Envelope' node. Two of the nodes under the TIGR 'Cell envelope' node could have been mapped to CS 4.1.1: 'biosynthesis of surface polysaccharides and lipopolysaccharides' and 'lipoproteins'. The former was more accurately mapped to 'Surface polysaccharides/antigens' (CS 4.2.2), the latter was mapped to 'metabolism of phospholipids, glycolipids and lipoproteins' (CS 1.2.3) and therefore, the CS 4.1.1 node appears unoccupied.
3. The mapping process could not resolve ambiguity of broad-coverage nodes. In the SubtiList scheme FuncWheel, the majority of transport related functions (third segment) are blanked out. However, the SubtiList scheme does include the function node 'Transport/binding proteins and lipoproteins'; most of the CS level-3 transport related nodes could be subsumed by this broad function, but because we cannot map this node specifically to any of them, they appear unoccupied.

In view of these limitations, we reiterate that the comparison of the mapped schemes to the CS is a means of getting an approximate overview of the schemes' coverage. A different person repeating the mapping would doubtless have emerged with somewhat different coverage FuncWheels but not, we believe, to the extent of changing the overall conclusions that could be drawn from them.

In addition, although the breadth, depth and resolution descriptors used herein offer a good handle to compare

functional classification schemes, they do not reflect the quality of schemes. It would be unwise to assume that a wide, deep scheme with high resolution is necessarily better than one with small breadth, depth and low resolution. Broad schemes tend to be used to offer users rapid access to large functional categories, but this means that super-sets of these categories must be constructed manually. For example, to generate the equivalent of the 'Information Pathways' node in the CS, the 'Transcription', 'Translation' and 'DNA metabolism' nodes of the TIGR scheme must be combined. Deeper schemes allow users to identify gene products associated with quite specific functions without having to resort to alternative functional information databases (e.g. SWISS-PROT), but complicate access to gene product data. High-resolution schemes may indicate focus on a particular area of functional classification (e.g. the WIT scheme and metabolism) or simply reflect the extent of the scheme. A high-resolution scheme may be crucial for the expert user but may prove dauntingly complex to others. Different depths, breadths and resolutions reflect different functional classification strategies and goals on the part of their implementers and cater to different needs on the part of their users. This is well illustrated by the COGs scheme (Tatusov et al. 1997) where the combination of some nodes in the 'Riley scheme' generates nodes with broader functional coverage (i.e. coverage remains the same but depth, breadth and resolution are reduced). This is needed to classify the COGs, which group together related proteins, with similar but sometimes non-identical functions. Conversely, the WIT scheme requires very detailed description of function (i.e. a deep, broad scheme with high resolution) to allow the development of metabolic models (Selkov et al. 1998).

Meaning of scheme level

In some classification schemes, levels have a semantic value. For example, in the Enzyme Commission (EC) scheme, a four-level hierarchical scheme of enzyme-catalysed reactions, the first level represents the major class of enzyme activity (e.g. 'transferases' or 'hydrolases'), and the second, the group or bond acted upon (e.g. 'transferring phosphorus containing groups' or 'acting on peptide bonds') (IUBMB 1992). Such semantic 'level-meaning' is absent in the surveyed schemes. Levels are often used to divide functions into subsets, but the rationale for this subdividing is dependent on the parent node (e.g. if the parent node is 'amino-acid metabolism', the children nodes usually relate to the metabolism of a specific amino-acid) rather than an intrinsic property of the level. Resolution and depth in the schemes is therefore not consistent for all branches of the functional trees. It seems very unlikely that an overall functional classification scheme could be designed with semantically meaningful levels. Perhaps though, such meaning could be implemented within specific subsets of the scheme, for example by classifying all transport related functions using a system such as the 'Transport Commission' system (Saier 1998).

Function, apples, and oranges

One of the main issues bearing on functional classification schemes derives from a more philosophical question: "What is function?" Function is an umbrella term, e.g. a gene product can be described in terms of its biochemistry, molecular activity, cellular function and physiological role (Rastan and Beeley 1997). These functions are distinct and different. Consider the human serine protease trypsin: biochemically it catalyses the hydrolysis of peptide bonds following lysine or arginine residues in peptides, its molecular activity is as a proteolytic enzyme, its cellular function is protein degradation, and its physiological role is to aid digestion. Such distinctions are rare in functional classification schemes. In her review of systems for cataloguing the functions of gene products, Riley (1998b) points out that many schemes juxtapose the 'apples and oranges' of function and combine different aspects of gene product function, such as biochemical and physiological function, into a one-dimensional list. This problem is inherent in the surveyed schemes, which all mix 'apples and oranges'. Similarly the CS includes, for example, the nodes 'cell regulation' (CS 2.1.4), a physiological function, and 'ion channels' (CS 3.2.1) a molecular function. The current schemes cannot be merely re arranged to tackle this; separating the apples from the oranges requires a fundamental reclassification. This remains one of the most pressing and complex issues to be resolved for effective functional classification of gene products.

The 'Gene Ontology' illustrates a possible solution to this problem by distinguishing function in terms of three organising principles: gene product function, process and cellular localisation (Gene Ontology Consortium 1999). The function of a gene product is defined as 'a capability that a physical gene product (or gene product group) carries as a potential'. To avoid confusion with the more general use of the term function, this organising principle is also known as a 'functional primitive'. Examples of functional primitives include broad terms (e.g. 'enzyme' and 'transporter') and narrower ones (e.g. 'adenylate cyclase'). Process is defined as 'a biological objective accomplished via one or more ordered assemblies of functions'; e.g. 'cell growth and maintenance', or more specifically 'pyrimidine metabolism'. The division between organising principles is, however, not always definitive. The term 'signal transduction', for example, exists within both the function and process categories.

Multi-dimensionality and multi-functionality

The obvious solution to dealing with the umbrella term 'function' would be to distinguish carefully all these different aspects of function and to describe a gene prod-

uct's function in terms of each of them. This solution is encapsulated in the concept of multi-dimensionality of classification schemes as proposed by Riley (1998b). The three organising principles of the 'Gene Ontology' represent three functional dimensions (biochemical for 'functional primitive', cellular and physiological for 'process' and spatial for the 'cellular localisation'). Such a classification is invaluable in understanding the role of a gene product. This is illustrated by the comprehensively annotated Yeast Protein Database (YPD) (Hodges et al. 1999). Each gene product in the YPD is annotated in up to six different dimensions: genetic properties, functional category, post-translational modification, cellular role and subcellular localisation (see Table 2 for examples of nodes in these categories). Although each of these dimensions is only a list (i.e. a scheme with only one level, and resolution equal to breadth), the combined information described by these six parameters permits the gene product to be positioned very accurately within the functional space.

Another aspect of multi-dimensionality concerns the hierarchical classification of functions within schemes; certain functions can be involved in a number of more generalised functional classes. In the 'Gene Ontology', the functional node 'ATP-binding and phosphorylation-dependent chloride channel' is an example of an 'intracellular ligand-gated ion channel', a 'chloride channel' and a 'transmembrane conductance regulator'. This is handled in the 'Gene Ontology' by conceptualisation of the scheme as a DAG; a simple tree-like hierarchy could not contend with such complexity.

Finally, many proteins are multi-functional: capable of performing a variety of biological roles, sometimes simultaneously (particularly with multidomain proteins). The biological role of a protein may also be dependent on its environment or localisation (Todd et al. 1999). In Table 2 , we indicate the schemes that we have found empirically to include multiple functional assignments for gene products.

Current schemes

In this paper we have focused on six schemes mapped to the CS (EcoCyc, TIGR, SubtiList, MIPS, WIT and KEGG) and the 'Gene Ontology'. Two broad families of schemes emerged from our survey: (1) genome related schemes and (2) schemes related to the interaction networks of gene products.

The genome related schemes are EcoCyc, TIGR, SubtiList, MIPS and the 'Gene Ontology'. Two of them (EcoCyc and TIGR) are current implementations or derivations of Riley's original classification (Riley 1993). As a consequence, they can essentially be thought to represent the same scheme (implemented with trees of different breadth, depth and resolution). The SubtiList scheme was derived from an adapted combination of parts of the WIT related 'Metabolic Pathways Database (MPW) and of the Riley scheme (Moszer et al. 1996; Selkov et al. 1998). In addition, the scheme includes a number of functions specific to *B. subtilis*. In terms of their coverage of the CS, no doubt because of their relation to the Riley scheme, the three schemes are quite similar even though SubtiList appears to have a noticeably smaller coverage of the CS than the other two schemes. This is partly due to mapping limitations and partly because the SubtiList scheme was designed with the specific needs of the *B. subtilis* research community in mind, and therefore focuses on functional aspects of major relevance to them. The original Riley scheme was designed for the unicellular prokaryotic eubacteria *E. coli*, and this bias will exist in all derivative schemes. With schemes such as TIGR that are applied to diverse gene complements, such a bias could be problematic.

The MIPS scheme shares a lot of the Riley scheme but extends it to encompass a number of further functions. Some of these functions (e.g. signal transduction) exist in all organisms but are not explicitly listed in the Riley based schemes, whilst others are present to allow better coverage of eukaryotic functions by the scheme (e.g. organelle related functions). The MIPS scheme can, in essence, be thought of as a superset of Riley schemes which begins to address the issue of generating functional schemes applicable to multiple and diverse organisms.

As we have previously mentioned, genome sequencing initiatives are the main driving force in the development of functional classification schemes. Nevertheless, the vast majority of genome-sequencing initiatives have been focused on unicellular micro-organisms. Of the 24 complete genomes listed in NCBI's 'Entrez Genomes', only one, *Caenorhabditis elegans*, is multicellular (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/org.html). Both Riley's scheme and the MIPS scheme were designed for classification of the genomes of unicellular organisms. Therefore, there is a great paucity of functional nodes concerning the interaction between cells in many schemes.

WIT and KEGG are databases of gene product interactions. They deal with functions performed by the concerted actions of gene products in pathways and complexes. Both the WIT and the KEGG functional classification schemes generally classify the function of a gene product by association with a pathway or complex. This helps explain why both these classification schemes have good coverage of metabolism. At the time of data gathering, the KEGG scheme had only minimal coverage of non-metabolism related functions, but a recent visit to the KEGG homepage (http://www.genome.ad.jp/kegg) confirmed that the KEGG scheme is being extended to include a number of non-metabolism-related functions. The WIT scheme had good coverage of the CS dealing with transport, structure and information-related pathways in addition to metabolism. This association of gene products to pathways and complexes is very relevant to their function: all but the simplest of biological roles in cells are performed by interactions of gene products.

The 'Gene Ontology' is representative of the 'next generation' of functional classification schemes. Rather

than updating existing schemes, the 'Gene Ontology' has been designed from scratch and addresses many of the problems and issues we have discussed in this paper. The 'Gene Ontology' is multi-dimensional, separating the concepts of 'functional primitive', 'process' and 'localisation'. Its more complex architecture allows it to accommodate functional descriptions that are examples of more than one parent node. The scheme is being developed for classification of the gene complements of both unicellular and multi-cellular organisms. We did not attempt to map the 'Gene Ontology' to the CS, but we are confident that it would have occupied all of the CS nodes.

The most extensive scheme currently in use and applied to a significant number of genomes is the MIPS scheme. However, perhaps one of the most notable conclusions it that all genome related schemes (other than the 'Gene Ontology') cover broadly the same set of functions and there is little to make one scheme overwhelmingly superior to another. The schemes are tantalisingly similar but unfortunately different enough to make direct comparisons between them difficult. With respect to the pathway and metabolism schemes, WIT has the most extensive functional classification but KEGG has built more generalised pathways that may be more accessible to many researchers. Certainly, consideration of gene-product interactions in pathways and complexes will play an important part in any future functional classification scheme.

The recent application of the 'Gene Ontology' for annotation of gene products identified in the *Adh* region of the *D. melanogaster* genome (Ashburner et al. 1999) illustrates the potential of such a scheme.

Future schemes

Functional classification schemes will become an increasingly critical element of genome databases. We believe that future schemes should have a controlled vocabulary and be integrated within an ontology which will not only classify functional nodes but control their grammar and semantics (Baker et al. 1999; Schulze-Kremer 1998). Ideally, they should be applicable to all species but still be capable of accommodating very specific functions and allow cross-species functional correspondence where possible. They will also have to be able to contend with environment and location dependent changes in the function of gene products. This will only be possible if multiple functional assignments for gene products are permitted. Furthermore, the most effective functional classification schemes will be multi-dimensional which will allow for accurate positioning of gene products in the function space. To deal with these multiple parameters, schemes will undoubtedly need to explore more complex structures than simple trees.

The increasing availability of multi-cellular genomes demands the development of more complete schemes that will have to classify not only the functions related to intercellular communication but also those related to the more complex organisation of multi-cellular organisms (e.g. tissues and organs).

As the quantity of information on gene products increases at an unparalleled pace, it is imperative that the quality of functional annotation improves. The 'Gene Ontology' represents a promising development in this area. On the other hand, its very complexity and scope may be an obstacle to its widespread implementation. There is perhaps a need for a less extensive scheme, spanning the gap between simple, tree-like classification schemes and the 'Gene Ontology', or perhaps we should consider having both highly detailed and simplified functional schemes catering to different sets of users (Gelbart 1998).

Whichever schemes are developed, there is certainly a need to try and use a single, standardised format or to provide tools to map between the different schemes. The logical progression is towards functional classifications that are applicable to all organisms and cover all functional dimensions. The power of such schemes will only be realised when applied consistently over multiple genomes to allow comparison of organisms. Functional genomics will then be ready for its golden age.

Acknowledgements We thank Dr. Neil Stoker for helpful discussions. S.R. is supported by a GlaxoWellcome studentship. This is a publication from the Bloomsbury BBSRC structural biology centre.

References

Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N, Hartzell G, Harvey D, Hong L, Houston K, Hoskins R, Johnson G, Martin C, Moshrefi A, Palazzolo M, Reese MG, Spradling A, Tsang G, Wan K, Whitelaw K, Celniker S, Rubin GM (1999) An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the *Adh* region. Genetics 153:179–219

Bairoch A, Apweiler R (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res 27:49–54

Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A (1999) An ontology for bioinformatics applications. Bioinformatics 15:510–20

Barker WC, Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LS, Ledley RS, Mewes HW, Pfeiffer F, Tsugita A, Wu C (1999) The PIR-International Protein Sequence Database. Nucleic Acids Res 27:39–43

Blake JA, Richardson JE, Davisson MT, Eppig JT, and the Mouse Genome Database Group (1999) The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. Nucleic Acids Res 27:95–98

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden Ma, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453–1474

Bowman JS, Emerson SL, Darnovsky M (1996) The practical SQL handbook. Addison-Wesley Developers Press, Reading, Mass.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R,

Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream M-A, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Suqares S, Sulston JE, Taylor K, Whitehead S, Barell BG (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393:537–544

Eisenhaber F, Bork P (1999) Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. Bioinformatics 15:528–535

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L-I, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu P-C, Lucier TS (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

Frishman D, Mewes HW (1997) PEDANTic genome analysis. Trends Genet 13:415–416

Gelbart WM (1998) Databases in genomic research. Science 282:659–661

Gene Ontology Consortium (1999) http://www.geneontology.org/

Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. Nucleic Acids Res 27:69–73

IUBMB (1992) Enzyme nomenclature: recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Academic Press, New York

Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M (1999) EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res 27:55–58

Licciulli F, Catalano D, D'Elisa D, Lorusso V, Attimonelli M (1999) KEYnet: a keywords database for biosequences functional organization. Nucleic Acids Res 27:365–367

Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM (1998) Protein folds and functions. Structure 6:875–884

Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A (1997) Overview of the yeast genome. Nature 387[Suppl]:7–8

Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D (1999) MIPS: a database for genomes and protein sequences. Nucleic Acids Res 27:44–48

Moszer I, Kunst F, Danchin A (1996) The European *Bacillus subtilis* genome sequencing project: current status and accessibility of the data from a new World Wide Web site. Microbiology 142:2987–2991

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34

PostgreSQL homepage (1999) http://www.postgresql.org/

Rastan S, Beeley LJ (1997) Functional genomics: going forwards from the databases. Curr Opin Genet Dev 7:777–783

Riley M (1993) Functions of the gene products of *Escherichia coli*. Microbiol Rev 57:862–952

Riley M (1998a) Genes and proteins of *Escherichia coli* K-12 (GenProtEC). Nucleic Acids Res 26:54

Riley M (1998b) Systems for categorizing functions of gene products. Curr Opin Struct Biol 8:388–392

Riley M, Labedan B (1996) *E. coli* gene products: physiological functions and common ancestries. In: Neidhardt FN, Curtiss R III, Lin ECC, Ingraham JL, Low KB, Magasanik B, Reznikoff W, Riley M, Schaechter M, Umbarger E (eds) *Escherichia coli* and Salmonella: cellular and molecular biology, 2nd edn. ASM Press, Washington, DC

Saier MH Jr (1998) Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. Adv Microb Physiol 40:81–136

Schulze-Kremer S (1998) Ontologies for molecular biology. Pac Symp Biocomput 3:695–706

Selkov E Jr, Grechkin Y, Mikhailova N, Selkov E (1998) MPW: the Metabolic Pathways Database. Nucleic Acids Res 26:43–45

Tamames J, Casari G, Ouzounis C, Valencia A (1996) Genomes with distinct function composition. FEBS Lett 389:96–101

Tamames J, Casari G, Ouzounis C, Valencia A (1997) Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol 44:66–73

Tamames J, Ouzounis C, Casari G, Sander C, Valencia A (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. Bioinformatics 14:542–543

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Todd AE, Orengo CA, Thornton JM (1999) Evolution of protein function, from a structural perspective. Curr Opin Chem Biol 3:548–556

Wall L, Christiansen T, Schwartz RL (1996) Programming Perl, 2nd edn. O'Reilly, Sebastopol, Calif.