Gene Ontology: tool for the unification of biology

The Gene Ontology Consortium*

Genomic sequencing has made it clear that a large fraction of the genes specifying the core biological functions are shared by all eukaryotes. Knowledge of the biological role of such shared proteins in one organism can often be transferred to other organisms. The goal of the Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing. To this end, three independent ontologies accessible on the World-Wide Web (http://www.geneontology.org) are being constructed: biological process, molecular function and cellular component.

The accelerating availability of molecular sequences, particularly the sequences of entire genomes, has transformed both the theory and practice of experimental biology. Where once biochemists characterized proteins by their diverse activities and abundances, and geneticists characterized genes by the phenotypes of their mutations, all biologists now acknowledge that there is likely to be a single limited universe of genes and proteins, many of which are conserved in most or all living cells. This recognition has fuelled a grand unification of biology; the information about the shared genes and proteins contributes to our understanding of all the diverse organisms that share them. Knowledge of the biological role of such a shared protein in one organism can certainly illuminate, and often provide strong inference of, its role in other organisms.

Progress in the way that biologists describe and conceptualize the shared biological elements has not kept pace with sequencing. For the most part, the current systems of nomenclature for genes and their products remain divergent even when the experts appreciate the underlying similarities. Interoperability of genomic databases is limited by this lack of progress, and it is this major obstacle that the Gene Ontology (GO) Consortium was formed to address.

Functional conservation requires a common language for annotation

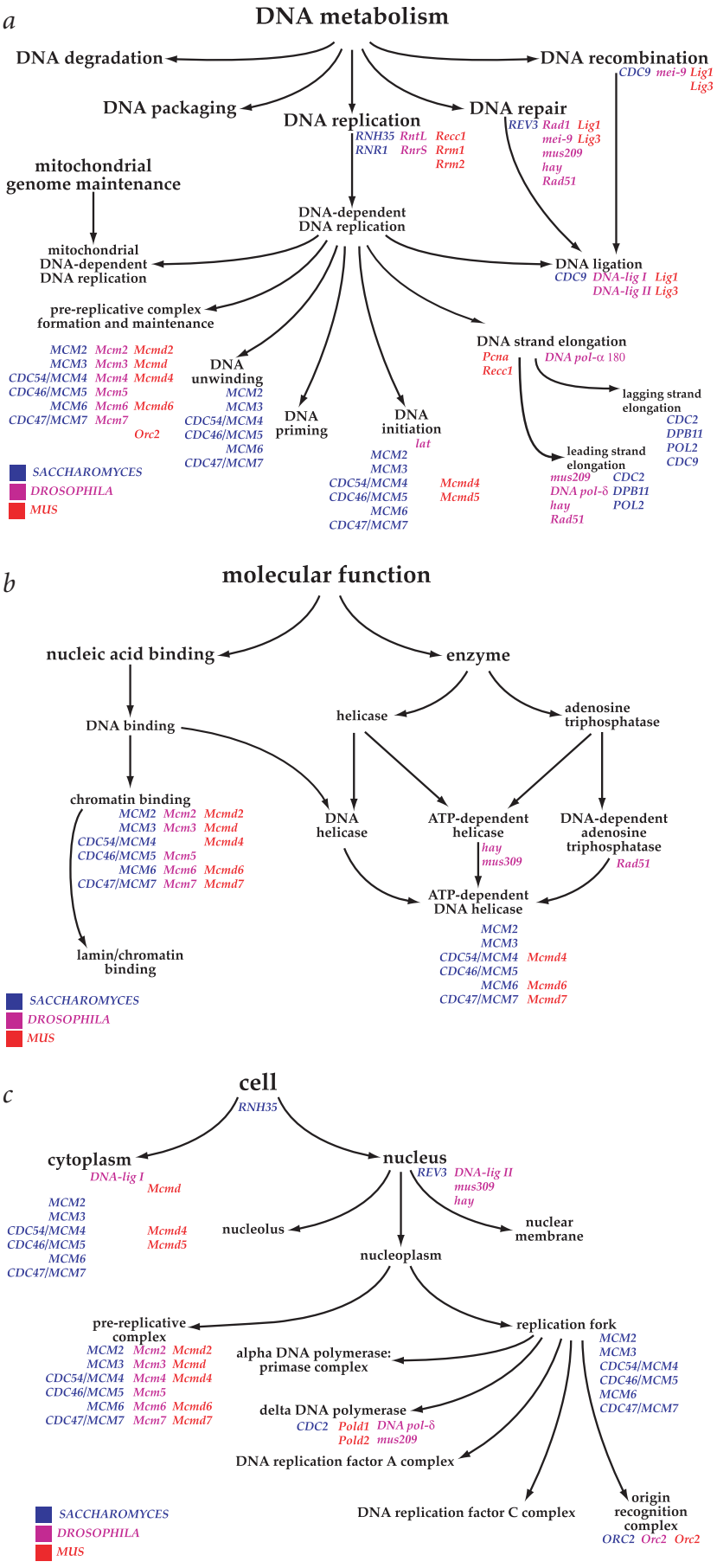
Nowhere is the impact of the grand biological unification more evident than in the eukaryotes, where the genomic sequences of three model systems are already available (budding yeast, *Saccharomyces cerevisiae*, completed in 1996 (ref. 1); the nematode worm *Caenorhabditis elegans*, completed in 1998 (ref. 2); and the fruitfly *Drosophila melanogaster*, completed earlier this year[3]) and two more (the flowering plant *Arabidopsis thaliana*[4] and fission yeast *Schizosaccharomyces pombe*) are imminent. The complete genomic sequence of the human genome is expected in a year or two, and the sequence of the mouse (*Mus musculus*) will likely follow shortly thereafter.

The first comparison between two complete eukaryotic genomes (budding yeast and worm[5]) revealed that a surprisingly large fraction of the genes in these two organisms displayed evidence of orthology. About 12% of the worm genes (~18,000) encode proteins whose biological roles could be inferred from their similarity to their putative orthologues in yeast, comprising about 27% of the yeast genes (~5,700). Most of these proteins have been found to have a role in the 'core biological processes' common to all eukaryotic cells, such as DNA replication, transcription and metabolism. A three-way comparison among budding yeast, worm and fruitfly shows that this relationship can be extended; the same subset of yeast genes generally have recognizable homologues in the fly genome[6]. Estimates of sequence and functional conservation between the genes of these model systems and those of mammals are less reliable, as no mammalian genome sequence is yet known in its entirety. Nevertheless, it is clear that a high level of sequence and functional conservation will extend to all eukaryotes, with the likelihood that genes and proteins that carry out the core biological processes will again be probable orthologues. Furthermore, since the late 1980s, many experimental confirmations of functional conservation between mammals and model organisms (commonly yeast) have been published[7–12].

This astonishingly high degree of sequence and functional conservation presents both opportunities and challenges. The main opportunity lies in the possibility of automated transfer of biological annotations from the experimentally tractable model organisms to the less tractable organisms based on gene and protein sequence similarity. Such information can be used to improve human health or agriculture. The challenge lies in meeting the requirements for a largely or entirely computational system for comparing or transferring annotation among different species. Although robust methods for sequence comparison are at hand[13–15], many of the other elements for such a system remain to be developed.

*Michael Ashburner[1], Catherine A. Ball[3], Judith A. Blake[4], David Botstein[3], Heather Butler[1], J. Michael Cherry[3], Allan P. Davis[4], Kara Dolinski[3], Selina S. Dwight[3], Janan T. Eppig[4], Midori A. Harris[3], David P. Hill[4], Laurie Issel-Tarver[3], Andrew Kasarskis[3], Suzanna Lewis[2], John C. Matese[3], Joel E. Richardson[4], Martin Ringwald[4], Gerald M. Rubin[2] & Gavin Sherlock[3]

[1]FlyBase (http://www.flybase.bio.indiana.edu). [2]Berkeley Drosophila Genome Project (http://fruitfly.bdgp.berkeley.edu). [3]Saccharomyces Genome Database (http://genome-www.stanford.edu). [4]Mouse Genome Database and Gene Expression Database (http://www.informatics.jax.org). Correspondence should be addressed to J.M.C. (e-mail: cherry@stanford.edu) and D.B. (e-mail: botstein@genome.stanford.edu), Department of Genetics, Stanford University School of Medicine, Stanford, California, USA.

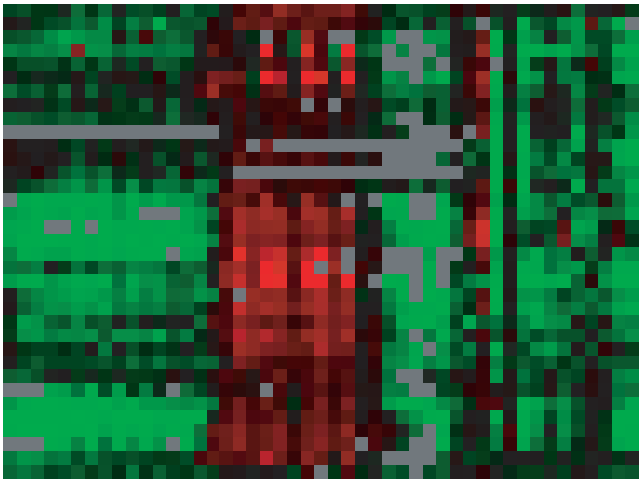


A dynamic gene ontology

The GO Consortium is a joint project of three model organism databases: FlyBase[16], Mouse Genome Informatics[17,18] (MGI) and the *Saccharomyces* Genome Database[19] (SGD). It is expected that other organism databases will join in the near future. The goal of the Consortium is to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism. Early considerations of the problems posed by the diversity of activities that characterize the cells of yeast, flies and mice made it clear that extensions of standard indexing methods (for example, keywords) are likely to be both unwieldy and, in the end, unworkable. Although these resources remain essential, and our proposed system will continue to link to and depend on them, they are not sufficient in themselves to allow automatic transfers of annotation.

Each node in the GO ontologies will be linked to other kinds of information, including the many gene and protein keyword databases such as SwissPROT (ref. 20), GenBank (ref. 21), EMBL (ref. 22), DDBJ (ref. 23), PIR (ref. 24), MIPS (ref. 25), YPD & WormPD (ref. 26), Pfam (ref. 27), SCOP (ref. 28) and ENZYME (ref. 29). One reason for this is that the state of biological knowledge of what genes and proteins do is very incomplete and changing rapidly. Discoveries that change our understanding of the roles of gene products in cells are published on a daily basis. To illustrate this, consider annotating two different proteins. One is known to be a transmembrane receptor serine/threonine kinase involved in p53-induced apoptosis; the other is known only to be a membrane-bound protein. In one case, the knowledge about the protein is substantial, whereas in the other it is minimal.

Fig. 1 Examples of Gene Ontology. Three examples illustrate the structure and style used by GO to represent the gene ontologies and to associate genes with nodes within an ontology. The ontologies are built from a structured, controlled vocabulary. The illustrations are the products of work in progress and are subject to change when new evidence becomes available. For simplicity, not all known gene annotations have been included in the figures. ***a***, Biological process ontology. This section illustrates a portion of the biological process ontology describing DNA metabolism. Note that a node may have more than one parent; for example, 'DNA ligation' has three parents, 'DNA-dependent DNA replication', 'DNA repair' and 'DNA recombination'. ***b***, Molecular function ontology. The ontology is not intended to represent a reaction pathway, but instead reflects conceptual categories of gene-product function. A gene product can be associated with more than one node within an ontology, as illustrated by the MCM proteins. These proteins have been shown to bind chromatin and to possess ATP-dependent DNA helicase activity, and are annotated to both nodes. ***c***, Cellular component ontology. The ontologies are designed for a generic eukaryotic cell, and are flexible enough to represent the known differences between diverse organisms.



YOR124C	UBP2	protein degradation	ubiquitin-specific protease	cytoplasm
YBR082C	UBC4	protein degradation	ubiquitin conjugating enzyme	unknown
YPR158W	—	unknown	unknown	unknown
YGR142W	BTN2	unknown	unknown	unknown
YOL032W	—	unknown	unknown	unknown
YDR171W	HSP42	stress response	chaperone	cytoskeleton
YKR011C	—	unknown	unknown	unknown
YGL087C	MMS2	protein degradation	ubiquitin conjugating enzyme	unknown
YFR003C	—	unknown	unknown	unknown
YFR010W	UBP6	protein degradation	ubiquitin-specific protease	unknown
YKL213C	DOA1	protein degradation	unknown	unknown
YGR136W	—	unknown	unknown	unknown
YIL076W	SEC28	vesicle coat assembly	vesicle transporter	coatomer
YOR007C	SGT2	unknown	unknown	unknown
YNL007C	SIS1	protein folding	unknown	unknown
YOR027W	STI1	protein folding	unknown	unknown
YPL240C	HSP82	protein folding	chaperonin	unknown
YMR186W	HSC82	protein folding	chaperone	unknown
YER103W	SSA4	protein folding	chaperone	cytoplasm
YDR258C	HSP78	protein folding	chaperone	mitochondrial matrix
YLL026W	HSP104	protein folding	cochaperone	cytoplasm
YLR217W	—	unknown	unknown	unknown
YLR216C	CPR6	protein folding	peptidylprolyl isomerase	cytoplasm
YLR259C	HSP60	protein folding	chaperonin	mitochondrion
YDR214W	—	protein folding	unknown	unknown
YOR020C	HSP10	protein folding	chaperone	mitochondrial matrix
YNL281W	HCH1	protein folding	unknown	unknown
YJL021C	—	unknown	unknown	unknown
YAL028W	—	unknown	unknown	unknown
YJL034W	KAR2	protein folding	chaperone	endoplasmic reticulum lumen
YLL024C	SSA2	protein folding	chaperone	cytoplasm
YAL005C	SSA1	protein folding	chaperone	cytoplasm
YAL004W	—	unknown	unknown	unknown
YFL016C	MDJ1	protein folding	chaperone	mitochondrial inner membrane
YNL006W	LST8	transport	unknown	unknown

Fig. 2 Correspondence between hierarchical clustering of expression microarray experiments with GO terms. The coloured matrix represents the results of clustering many microarray expression experiments[32]. In the matrix, each row represents the yeast gene described to the right, and each column represents the expression of that gene in a particular microarray hybridization. For each gene in the matrix, the table at right lists the systematic ORF name, the standard gene name (if known), and the GO biological process, molecular function and cellular component annotations for that gene. The GO annotations suggest that this experimental expression cluster groups gene products involved in the biological process of protein folding. In contrast, the molecular function and cellular component annotations of these gene products correlate less well with the clustered expression patterns of these gene products.

We need to be able to organize, describe, query and visualize biological knowledge at vastly different stages of completeness. Any system must be flexible and tolerant of this constantly changing level of knowledge and allow updates on a continuing basis.

Similar considerations suggested that a static hierarchical system, such as the Enzyme Commission[30] (EC) hierarchy, although computationally tractable, was also likely to be inadequate to describe the role of a gene or a protein in biology in a manner that would be either intuitive or helpful for biologists. The hierarchical EC numbering system for enzymes is the standard resource for classifying enzymatic chemical reactions. The EC system does not address the classification of non-enzymatic proteins or the ability to describe the role of a gene product within a cell; also, the system has little facility for describing diverse protein interactions. The vagueness of the term 'function' when applied to genes or proteins emerged as a particular problem, as this term is colloquially used to describe biochemical activities, biological goals and cellular structure. It is commonplace today to refer to the function of a protein such as tubulin as 'GTPase' or 'constituent of the mitotic spindle'. For all these reasons, we are constructing three independent ontologies.

Three categories of GO

Biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation, in the sense that something goes into a process and something different comes out of it. Examples of broad (high level) biological process terms are 'cell growth and maintenance' or 'signal transduction'. Examples of more specific (lower level) process terms are 'translation', 'pyrimidine metabolism' or 'cAMP biosynthesis'.

Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition also applies to the capability that a gene product (or gene product complex) carries as a potential. It describes only what is done without specifying where or when the event actually occurs. Examples of broad functional terms are 'enzyme', 'transporter' or 'ligand'. Examples of narrower functional terms are 'adenylate cyclase' or 'Toll receptor ligand'.

Cellular component refers to the place in the cell where a gene product is active. These terms reflect our understanding of eukaryotic cell structure. As is true for the other ontologies, not all terms are applicable to all organisms; the set of terms is meant to be inclusive. Cellular component includes such terms as 'ribosome' or 'proteasome', specifying where multiple gene products would be found. It also includes terms such as 'nuclear membrane' or 'Golgi apparatus'.

Ontologies have long been used in an attempt to describe all entities within an area of reality and all relationships between those entities. An ontology comprises a set of well-defined terms with well-defined relationships. The structure itself reflects the current representation of biological knowledge as well as serving as a guide for organizing new data. Data can be annotated to varying levels depending on the amount and completeness of available information. This flexibility also allows users to narrow or widen the focus of queries. Ultimately, an ontology can be a vital tool enabling researchers to turn data into knowledge. Computer scientists have made significant contributions to linguistic formalisms and computational tools for developing complex vocabulary systems using reason-based structures, and we hope that our ontologies will be useful in providing a well-developed data set for this community to test their systems. The Molecular Biology Ontology Working Group (http://wwwsmi.stanford.edu/projects/bio-ontology/) is actively attempting to develop standards in this general field.

Biological process, molecular function and cellular component are all attributes of genes, gene products or gene-product groups. Each of these may be assigned independently and, indeed, we believe that simply recognizing that biological process, molecular function and cellular location represent independent attributes is by itself clarifying in many situations, as in the annotation of gene-expression data. The relationships between a gene product (or gene-product group) to biological process, molecular function and cellular component are one-to-many, reflecting the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular

functions, and participate in multiple alternative interactions with other proteins, organelles or locations in the cell.

The ontologies are developed for a generic eukaryotic cell; accordingly, specialized organs or body parts are not represented. Full integration of the ontologies with anatomical structures will occur as the ontologies are incorporated into each species' database and are related to anatomical data within each database. GO terms are connected into nodes of a network, thus the connections between its parents and children are known and form what are technically described as directed acyclic graphs. The ontologies are dynamic, in the sense that they exist as a network that is changed as more information accumulates, but have sufficient uniqueness and precision so that databases based on the ontologies can automatically be updated as the ontologies mature. The ontologies are flexible in another way, so that they can reflect the many differences in the biology of the diverse organisms, such as the breakdown of the nucleus during mitosis. In this way the GO Consortium has built up a system that supports a common language with specific, agreed-on terms with definitions and supporting documentation (the GO ontologies) that can be understood and used by a wide biological community.

Examples of GO annotation

As one example, consider DNA metabolism, a biological process carried out by largely (but not entirely) shared elements in eukaryotes. The part of the process ontology (with selected gene names from *S. cerevisiae*, *Drosophila* and *M. musculus*) shown is largely one parent to many children (Fig. 1*a*). One notable exception is the process of DNA ligation, which is a child of three processes, DNA replication, DNA repair and DNA recombination. The yeast gene product Cdc9p is able to carry out the ligation step for all three processes, whereas it is uncertain whether the same enzyme is used in the other species. From the point of view of the ontology, it matters not, and a computer (or a human searcher) will find the appropriate nodes in either case using as the query either the enzyme, the gene name(s) or the GO term (or, if available, the unique GO identifier, in this case, GO:0003910).

Also shown are the molecular function ontology for the MCM protein complex members that are known to regulate initiation of DNA replication in the three organisms (Fig. 1*b*), and a portion of the cellular component ontology for these proteins (Fig. 1*c*). These ontologies reflect the finding that Mcm2–7 proteins are components of the pre-replicative complex in several model organisms, as well as sometimes localizing to the cytoplasm[30]. The ontology supports both biological realities, and yet the molecular functions and the biological processes of the MCM homologues are conserved nevertheless.

The usefulness of the GO ontologies for annotation received its first major test in the annotation of the recently completed sequence of the *Drosophila* genome. Little human intervention was required to annotate 50% of the genes to the molecular function and biological process ontologies using the GO method. Another use for GO ontologies that is gaining rapid adherence is the annotation of gene-expression data, especially after these have been clustered by similarities in pattern of gene expression[32,33]. The results of clustering about 100 yeast experiments (of which about half are shown; Fig. 2) grouped together a subset of genes which, by name alone, convey little to most biologists. When the full short GO annotations for process, molecular function and location are added, however, the biological reason and import of the co-expression of these genes becomes evident.

The GO project is currently using a flat file format to store the ontologies, definitions of terms and gene associations. The ontologies, gene associations, definitions and documentation are available from the GO web site (http://www.geneontology.org), which also describes the principles and objectives used by the project. The ontologies are by no means complete. They are being expanded during the association of gene products from the collaborating databases and we expect them to continue to evolve for many years. GO requires that all gene associations to the ontologies must be attributed to the literature; for each citation the type of evidence will be encoded. As of early April 2000 there were 1,923, 2,094 and 490 nodes in the process, function and component ontologies, respectively. The three organism databases have made substantial progress to link gene products. Thus far the process, function and component ontologies have associations with 1,624, 1,602 and 1,577 yeast genes; 741, 2,334 and 1,061 fly genes; and 1,933, 2,896 and 1,696 mouse genes, respectively. A running table of these statistics can be found at the web site.

The GO concept is intended to make possible, in a flexible and dynamic way, the annotation of homologous gene and protein sequences in multiple organisms using a common vocabulary that results in the ability to query and retrieve genes and proteins based on their shared biology. The GO ontologies produce a controlled vocabulary that can be used for dynamic maintenance and interoperability between genome databases. The ontologies are a work in progress. They can be consulted at any time on the World-Wide Web; indeed, their availability to human and machine alike is essential to maintain their flexibility and allow their evolution along with increased understanding of the underlying biology. It is hoped that the GO concepts, especially the distinctions between biological process, molecular function and cellular component, will find favour among biologists so that we can all facilitate, in our writing as well as our thinking, the grand unification of biology that the genome sequences portend.

Acknowledgements

We thank K. Fasman and M. Rebhan for useful discussions, and Astra Zeneca for financial support. SGD is supported by a P41, National Resources, grant from National Human Genome Research Institute (NHGRI) grant HG01315; MGD by a P41 from NHGRI grant HG00330; GXD by National Institute of Child Health and Human Development grant HD33745; and FlyBase by a P41 from NHGRI grant HG00739 and the Medical Research Council, London.

Received 20 March; accepted 5 April 2000.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546 (1996).
2. Worm Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *The C. elegans* Sequencing Consortium. *Science* **282**, 2012–2018 (1998).
3. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
4. Meinke, D.W. *et al*. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282**, 662–682 (1998).
5. Chervitz, S.A. *et al.* Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res*. **27**, 74–78 (1999).
6. Rubin, G.M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
7. Tang, Z., Kuo, T., Shen, J. & Lin, R.J. Biochemical and genetic conservation of fission yeast Dsk1 and human SR protein-specific kinase 1. *Mol. Cell. Biol*. **20**, 816–824 (2000).
8. Vajo, Z. *et al*. Conservation of the *Caenorhabditis elegans* timing gene clk-1 from yeast to human: a gene required for ubiquinone biosynthesis with potential implications for aging. *Mamm. Genome* **10**, 1000–1004 (1999).
9. Ohi, R. *et al*. Myb-related Schizosaccharomyces pombe cdc5p is structurally and functionally conserved in eukaryotes. *Mol. Cell. Biol.* **18**, 4097–4108 (1998).
10. Bassett, D.E. Jr *et al*. Genome cross-referencing and XREFdb: implications for the identification and analysis of genes mutated in human disease. *Nature Genet.* **15**, 339–344 (1997).
11. Kataoka T. *et al*. Functional homology of mammalian and yeast RAS genes. *Cell* **40**,19–26 (1985).
12. Botstein, D. & Fink, G.R. Yeast: an experimental organism for modern biology. *Science* **240**, 1439–1443 (1988).
13. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. **28**, 33–36 (2000).
14. Andrade, M.A. *et al*. Automated genome sequence analysis and annotation. *Bioinformatics* **15**, 391–412 (1999).
15. Fleischmann, W., Moller, S., Gateau, A. & Apweiler, R. A novel method for automatic functional annotation of proteins. *Bioinformatics* **15**, 228–233 (1999).
16. The FlyBase Consortium. The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res*. **27**, 85–88 (1999).
17. Blake, J.A. *et al*. The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Res.* **28**, 108–111 (2000).
18. Ringwald, M. *et al*. GXD: a gene expression database for the laboratory mouse—current status and recent enhancements. *Nucleic Acids Res*. **28**, 115–119 (2000).
19. Ball, C.A. *et al*. Integrating functional genomic information into the *Saccharomyces* Genome Database. *Nucleic Acids Res*. **28**, 77–80 (2000).
20. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
21. Benson, D.A. *et al*. GenBank. *Nucleic Acids Res.* **28**, 15–18 (2000).
22. Baker, W. *et al*. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **28**, 19–23 (2000).
23. Tateno, Y. *et al*. DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.* **28**, 24–26 (2000).
24. Barker, W.C. *et al*. The Protein Information Resource (PIR). *Nucleic Acids Res.* **28**, 41–44 (2000).
25. Mewes, H.W. *et al*. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
26. Costanzo, M.C. *et al*. The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28**, 73–76 (2000).
27. Bateman, A. *et al*. The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).
28. Lo Conte, L. *et al*. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**, 257–259 (2000).
29. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
30. *Enzyme Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. NC-IUBMB.* (Academic, New York, 1992).
31. Tye, B.K. MCM proteins in DNA replication. *Annu. Rev. Biochem.* **68**, 649–686 (1999).
32. Eisen, M., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
33. Spellman, P.T. *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).