Opinion

Genome cartography through domain annotation

Chris P Ponting and Nicholas J Dickens

Address: MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK.

Correspondence: Chris P Ponting. E-mail: Chris.Ponting@anat.ox.ac.uk

Published: 3 July 2001

Genome ***Biology*** 2001, **2(7)**:comment2006.1–2006.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/7/comment/2006

Abstract

The evolutionary history of eukaryotic proteins involves rapid sequence divergence, addition and deletion of domains, and fusion and fission of genes. Although the protein repertoires of distantly related species differ greatly, their domain repertoires do not. To account for the great diversity of domain contexts and an unexpected paucity of ortholog conservation, we must categorize the coding regions of completely sequenced genomes into domain families, as well as protein families.

Delivery of the human genome draft sequence by publicly funded [1] and corporate [2] projects promises to precipitate significant biomedical advances this century. To rise to this challenge, biologists must become adept at navigating the vast expanses of genomic DNA data that may seem, at first glance, to be devoid of features. Yet lying beneath this facade of uniformity are rich veins of knowledge awaiting exploitation. Surveying and signposting this apparently bland genomic landscape should guide investigators towards experiments that address specific hypotheses about gene function.

But in what language are the signposts to be written? Different communities of biologists speak in dialects that are not always mutually comprehensible, particularly with respect to the umbrella term ‘function’ [3]. Where one investigator might be interested in designing active-site inhibitors using high-resolution protein structural data, another might require information on gene pathways, and another might be focused on relating genotype to phenotype. If the genome is to offer up its secrets to all scientific communities, its surveyors need to adopt universal vocabularies.

Prediction or experimental finding?

Broadly speaking, there are two common ways of annotating (assigning names and functions to) genes. The first is the association of a gene with relevant experimental findings. Websites such as LocusLink [4], GeneCards [5], euGenes [6] and Ensembl [7] (see Table 1) integrate information from diverse sources relevant to individual human genes. Thus one may browse sequence information, for example concerning single-nucleotide polymorphisms (SNPs), alongside both descriptions of molecular and cellular function and information relevant to human disease. Such annotation is essential, yet it is currently restricted to the minority of human genes - those that have been characterized experimentally [1].

The second type of annotation relies not on empirical observations but rather on predicted evolutionary relationships. All genes that are thought to have arisen from a common ancestor are defined as homologs: where additional copies have arisen by gene duplication within a single genome they are defined as paralogs, whereas corresponding genes in different species are orthologs. Sometimes homologous gene products have strong sequence similarities, such that an inference of homology is straightforward; one such example is the *Drosophila melanogaster* gene *branchless*, which encodes a homolog of human fibroblast growth factor (FGF) [8]. On other occasions, protein homologs have subtle or indiscernible sequence similarities that try the patience and expertise of genome wayfarers. For example, human FGF and interleukin-1α have highly similar tertiary structures

Table 1

A key to the databases mentioned in this article

Database	URL	Description
COGs	http://www.ncbi.nlm.nih.gov/COG/	Clusters of orthologous groups of proteins, generated from the comparison of protein sequences encoded in 34 complete genomes, representing 26 major phylogenetic lineages.
DAS	http://stein.cshl.org/das/	A distributed sequence annotation system software client and database server for the annotation of protein sequences.
Ensembl	http://www.ensembl.org/	Software for the automatic annotation of eukaryotic genomes. Annotation and searching with gene, SNP, and cross-genome comparative information.
EuGenes	http://iubio.bio.indiana.edu:8089/	Automatic annotations of sequence databases with gene and genomic information, including chromosome, genetic and molecular maps.
Gene Ontology	http://www.geneontology.org/	A dynamic, controlled vocabulary applicable to the annotation of eukaryotic genomes. Includes knowledge of the role of genes and proteins within cells.
GeneCards	http://bioinformatics.weizmann.ac.il/cards	A database of human genes that maps genes, proteins and diseases. Provides information on gene function.
InterPro	http://www.ebi.ac.uk/interpro/	Proteome analysis database based on Pfam, SMART, Prosite, PRINTS and ProDom protein and domain family databases and the SWISS-PROT and TrEMBL sequence databases. Also contains software for the annotation of protein sequences using these databases.
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink	Interface to a database of sequence and descriptive information correlated with genetic loci.
Mammalian Homology	http://www.informatics.jax.org/menus/homology_menu.shtml	Mammalian homology and comparative maps. Tools and databases from the Jackson Laboratory for the comparison of mammalian genomes.
Pfam	http://www.sanger.ac.uk/Pfam	Protein families database containing multiple sequence alignments and hidden Markov models.
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	Database of protein fingerprints based on protein motifs.
ProDom	http://www.toulouse.inra.fr/prodom.html	Protein domain database based on an automatic compilation of homologous domains.
Prosite	http://www.isrec.isb-sib.ch/profile/	Prosite profiles are protein domain profiles constructed from multiple sequence alignments of proteins from families of related sequences.
SMART	http://smart.embl-heidelberg.de/	Protein domain families database containing multiple sequence alignments and hidden Markov models, based on a smaller set of domains than Pfam but designed to find domains that are more difficult to detect.
SMD	http://genome-www4.stanford.edu/MicroArray/SMD	The Stanford microarray database of raw and normalized data from microarray experiments, including interfaces for data retrieval and analysis.
SWISS-PROT and Trembl	http://www.expasy.ch/sprot	Protein sequence databases. SWISS-PROT represents a 'gold standard' of annotation.
TIGRFAMs	http://www.tigr.org/TIGRFAMs/	Database of protein families based on hidden Markov models of multiple protein sequence alignments.

despite insignificantly similar sequences but, as a result of their similar growth-factor-type functions, they are in fact very likely to be homologs [9].

The importance of annotating the genomic landscape on the basis of homology is that protein homologs invariably have similar tertiary structures and frequently also have similar functions. The surveying and way-marking of each new gene, therefore, needn't always be an arduous process of discovering structure and function from scratch, since clues can be inferred from what has already been experimentally gleaned from its homologs. By applying the concept of homology, the problem of broadly predicting the functions of all genes is brought within the realms of possibility.

There is a pitfall to be avoided when annotating genes by homology: homology is defined on the basis of evolution, rather than function. On one hand, homologs may be related only by evolution and not by similarities in molecular mechanism; relatives of enzymes that now lack catalytic sites are just such examples. On the other hand, examples abound of divergent homologs, or even non-homologs, whose functions overlap. Consequently, homology assignment indicates only an approximate direction for future empirical determination of function, perhaps analogous to laying down a compass bearing rather than an exact map reference.

Protein domains

The first completely sequenced genomes, such as that of *Haemophilus influenzae* [10], were annotated following searches of all available sequence databases with each predicted gene product. This approach assumes that matters are straightforward: although homologous genes may not always be similar in function, genes are either homologous or they are not. A major problem in gene annotation arises, however, when one encounters sequence-similar, homologous portions of genes embedded in otherwise sequence-dissimilar, non-homologous contexts. The presence of these domains demonstrates that evolution has constantly fused and divided genes, using a repertoire of pre-existing components (Figure 1).

Domains are homologous portions of sequences that are encoded in different gene contexts and have survived the evolutionary tests of time without fragmentation. In three dimensions, domains are observed to be compact units of structure, often with a hydrophobic interior and a hydrophilic exterior, and they are not divisible into smaller units. Consequently, domains represent the finite vocabulary of protein evolution: if domains are words, then multidomain proteins are complete sentences.

Just as there is a dictionary or lexicon for every language, there is one - or in this case several - for the vocabulary of domains. Pfam [11,12] is the widest-ranging lexicon of domain families, predicting at least one domain for more than two-thirds of all entries in the SWISS-PROT [13] protein database. SMART [14,15] is a more concise collection, focusing on those domains that are widespread and difficult to detect. Prosite [16,17] also has a dictionary of domain profiles, as does Celera [18] (called Panther) and The Institute for Genomic Research, TIGR [19,20].

Each of these resources detects domains using numerical representations of multiple sequence alignments, either hidden Markov models (HMMs) or generalized profiles (GPs) [21]. Although the constructions of HMMs and GPs are very different, formally they are equivalent. Homology assignments are guided by comparisons of HMMs or GPs with protein sequence databases, and by implementation of an upper threshold value for *E*, the number of unrelated sequences expected purely by chance that are aligned with a particular score, or higher, in the search. This procedure has been shown on many occasions to identify subtle, yet informative, sequence similarities among distant homologs.

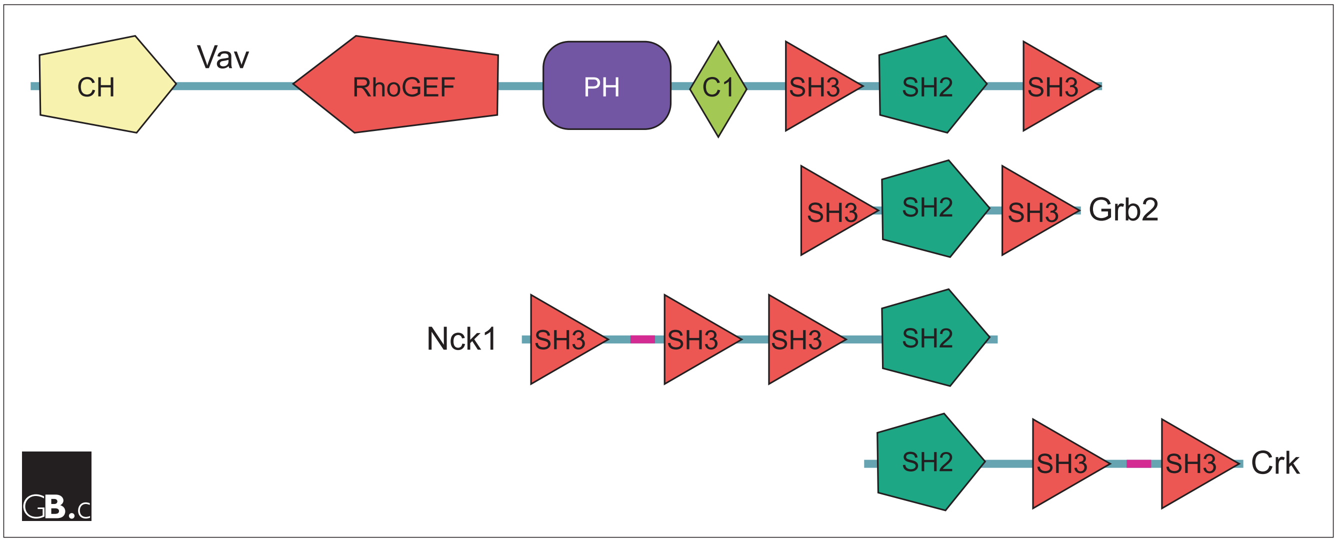

Figure 1
Domain architectures of the human signaling proteins Vav, Grb2, Nck1 and Crk, as predicted using SMART [14,15]. Although these proteins share two common domains (namely Src-homology 2 and 3, or SH2 and SH3, domains) they are not all homologous to one another. Rather, their SH2 domains are homologous and their SH3 domains are homologous. The complex architectures of these proteins imply that the co-occurrence of these two domain types has arisen on more than one occasion, and thus that the domain combinations seen in the four proteins are not likely to have arisen from a common ancestor with whom they share a common architecture.

Taken together, Pfam [11,12], SMART [14,15] and the other domain resources carry considerable redundancy, although each has its own merits. An additional resource, InterPro [22,23], has been derived in part to avoid the onerous task of querying each of these domain lexicons separately for each protein sequence of interest. InterPro (release 3.0) combines the domain and motif sets of Prosite [16,17], Pfam [11,12], ProDom [24], PRINTS [25] and SMART [14,15] in a hierarchical manner and is thereby able to provide annotation for 74% of all SWISS-PROT (protein database) and TrEMBL (translated DNA sequence database) entries [13].

The pros and cons of the domain-centric view of a genome

Pfam, SMART and InterPro were recently chosen by the public and private consortia to annotate their human genome draft sequences [1,2]. To be more precise, annotations were applied to the proteome, the current predicted set of all expressed proteins encoded by the draft. The proportion of the proteome that could be annotated, even minimally, using these resources is low, at approximately 40-60%. The resources were used to annotate the proteome according to lists of component domains rather than protein or gene names. Thus, protein sequences were identified not, for example, as the product of the Lbc oncogene, but rather as containing RhoGEF (Dbl-homology) and pleckstrin homology (PH) domains (Figure 2). This is protein annotation viewed through the evolutionary lens, rather than faithful extracts from the current body of experimental knowledge. By implication, proteome annotation by domain content implies that a gene product's function is a synthesis of the generalized functions of its component parts (see Figure 2). Although such descriptions of molecular function are approximate and pale in comparison to accounts of experimentally derived characteristics, they provide the best predictions that can be mustered for the uncharacterized majority of human genes.

It is important to emphasize that annotation of individual proteins or complete proteomes using domains is achieved automatically rather than by manual curation. This is relevant, because a newly sequenced genome's proteome is generally in a high state of flux, with additions and deletions resulting from sequence updates, enhanced understanding of gene structure and identification of previously overlooked genes. The animal genome sequencing projects already underway will proceed in the same manner as the publicly funded human project, through numerous draft stages and on towards completion. Providing up-to-date, and necessarily automatic, annotation of incomplete genomic data will be essential.

One might suggest that gene annotation by use of domains is a relatively short-term measure that will be made redundant by results from high-throughput studies of non-vertebrate model organisms. It can be argued that detailed predictions of the functions of most human genes can be inferred from studies of their orthologs, as these are the most likely members of a family to have similarities in molecular and cellular roles. In contrast to previous expectations, the human genome draft publications found that the great majority of human genes have no orthologs in each of three important

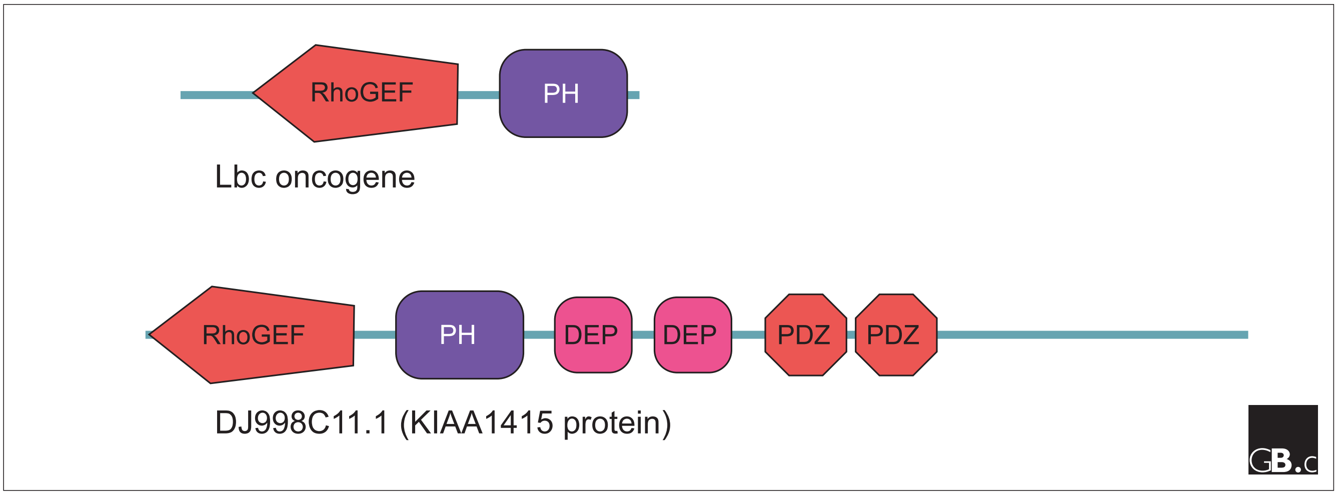

Figure 2
Domain architectures of the Lbc oncogene product, and the experimentally uncharacterized protein KIAA1415, as predicted using SMART [14,15]. The predicted domains in KIAA1415 suggest that it facilitates guanine nucleotide exchange on Rho-type small GTPases (it has Rho guanine-nucleotide exchange factor, or RhoGEF, and pleckstrin homology or PH domains), it is localized to the plasma membrane (PH domain) and it may interact with the carboxyl termini of transmembrane receptors or channels (it has PDZ domains). The generalized functions of DEP domains are not currently known. It should be noted that prediction of this protein's domain architecture does not allow prediction of either its cellular function or the phenotypic effect of its deficiency.

model organisms whose genome sequences are known, namely *Drosophila* (fruitfly), *Caenorhabditis elegans* (nematode worm) and *Saccharomyces cerevisiae* (baker's yeast) [1,2]. Thus, the contribution of domain identification to gene-function prediction will remain until such a time as reliable results from high-throughput studies on a more closely related organism, such as the mouse, are available [26].

Comparing proteomes using domain families

Deconvolution of human proteins into their constituent domains has also played a major role in understanding the evolution of chordates [1,2]. Three significant differences were detected between the repertoires of human domain families and those of the nematode worm, fruitfly, the mustard cress *Arabidopsis thaliana* and yeast. First, only a small proportion (7%) of human domain families are absent from the other proteomes. Second, numerous domain families were greatly expanded in terms of the number of members in humans, whereas others were considerably reduced. Finally, the human proteome contained significantly more combinations of domains. This finding demonstrated that domain 'invention' in the chordate lineage has made only a minor contribution to proteome diversity, whereas expansions and contractions of domain families, and domain additions and deletions, have all added greatly to proteome innovation.

Such studies demonstrate the power of comparative proteomics in relating gene content to the evolution of eukaryotic organisms. But as with the argument that the complexity of organisms is only loosely coupled to total gene number, the question of whether the number of representatives of a domain family in a proteome is directly related to that family's contribution to cellular and organismal function remains open. It is hoped that future proteome comparisons will progress beyond the simple enumeration of homologous genes and domains towards an understanding of the biology that underlies the variability of domain family sizes.

Domain-centric annotation is but one of many methods. Although it provides information about molecular structure, function and evolution, by and large it is unable to predict functional aspects, such as cellular or organismal role, protein-binding partners or post-translational modifications. Fortunately, other views that address these aspects have been incorporated with domain predictions into web-based resources such as LocusLink [4], GeneCards [5], euGenes [6] and Ensembl [7]. Each of these sites represents a confluence of diverse information sources that are mapped to specific regions of genomic sequence. Whilst navigating these sites it should be borne in mind that they often fail explicitly to distinguish between annotations that are experimentally derived and those that are predicted by homology-based methods. Nevertheless, these sites are a significant boon to biologists since they provide views of genomes from multiple vantage points, from protein tertiary structure through to SNPs and on to human disease.

An improved navigation

As viewed now, the human genome appears to be a relatively featureless landscape, punctuated by islands of annotations for well-characterized and biomedically important genes. As biological sciences progress into a more knowledge-rich, as well as data-rich, era, the cartography of this genome will become more complex with numerous different functional characteristics being assigned to a growing fraction of human genes. It will be increasingly important to restrict descriptions of function to a common and broad vocabulary that is compatible with computational approaches. Fortunately, a cross-community approach to dealing with this issue is already underway. The Gene Ontology project (GO) [27,28] has created an initial hierarchy of defined terms that encompasses many of the flavours of 'function' commonly described in biology. GO has begun to permeate throughout genomics and it will do so more rapidly as its scope and attention to detail improves.

Two further domain-centric approaches look set to guide the efficient navigation of genomes: orthology prediction and the partitioning of a domain family into subfamilies with distinct functions. Orthology is a valuable concept from which to infer functional information between species, and orthologs from the genomes of 30 bacteria, archaea and the yeast *Saccharomyces cerevisiae* can be predicted directly using the COGs database [29,30]. Pairs of orthologs from animal genomes are also available on the web (for example from the Jackson Laboratory [31]). No resource is yet available that accurately predict sets of orthologs for several multicellular eukaryotes, however, such as the fruitfly, worm, *Arabidopsis*, mouse and human. This situation will inevitably change on completion of the human and mouse genome sequences.

A more difficult problem is the partitioning of a homologous domain family into multiple subfamilies representing multiple functions. Homologous proteins with divergent sequences frequently have distinct functions [32] that are characterized by contrasting patterns of conserved amino acids. One productive approach to the analysis and prediction of functional subtypes identified key sites in multiple protein sequence alignments that specify the different functional subtypes. This method has performed well in defining functional subtypes with prediction accuracies of up to 96% [33].

A key element in ensuring the general utility of genomic data lies in collating predicted with experimentally derived observations. A central function will be provided by sophisticated web forums that accumulate and automatically present integrated functional data. The best example so far is the Distributed Annotation System, or DAS [34], which seeks to amalgamate annotations donated by experimentalists worldwide. Even these schemes, however, will face a major challenge in integrating functional information from the huge datasets that arise, for example, from microarray [35,36] and proteomics [37] experiments. The human genomic landscape, relatively featureless now, will soon be teeming with evidence, pointers, and clues. Ultimately, the success of the genome projects will be measured not in the completion of sequences, but in how access to integrated data opens up new avenues of research and therapy.

References

1. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
3. Jacq B: **Protein function from the perspective of molecular interactions and genetic networks**. *Briefings Bioinformatics* 2001, **2**:38-50.
4. **LocusLink** [http://www.ncbi.nlm.nih.gov/LocusLink]
5. **GeneCards** [http://bioinformatics.weizmann.ac.il/cards]
6. **euGenes** [http://iubio.bio.indiana.edu:8089/]
7. **Ensembl** [http://www.ensembl.org/]
8. Sutherland D, Samakovlis C, Krasnow MA: ***branchless* encodes a *Drosophila* FGF homolog that controls tracheal cell migration and the pattern of branching.** *Cell* 1996, **87**:1091-1101.
9. Zhang J, Cousens LS, Barr PJ, Sprang SR: **Three-dimensional structure of human basic fibroblast growth factor, a structural homolog of interleukin 1α.** *Proc Natl Acad Sci USA* 1991, **88:**3446-3450.

10, Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1996, **269**:496-512.

11. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2000, **28**:263-266.
12. **Pfam** [http://www.sanger.ac.uk/Pfam]
13. **SWISS-PROT** [http://www.expasy.ch/sprot/]
14. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART: a simple modular architecture research tool: identification of signaling domains.** *Proc Natl Acad Sci USA* 1998, **95**:5857-5864.
15. **SMART** [http://smart.embl-heidelberg.de/]
16. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999*,* **27**:1215-1219.
17. **Prosite** [http://www.isrec.isb-sib.ch/profile/]
18. **Celera, Inc.** [http://www.celera.com]
19. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins.** *Nucleic Acids Res* 2001, **29**:41-43.
20. **TIGRFAMs** [http://www.tigr.org/TIGRFAMS/]
21. Hofmann, K: **Sensitive protein comparisons with profiles and hidden Markov models.** *Briefings Bioinformatics* 2000, **1**:167-178.
22. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29**:37-40.
23. **InterPro** [http://www.ebi.ac.uk/interpro]
24. **ProDom** [http://www.toulouse.inra.fr/prodom.html]
25. **PRINTS** [http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/]
26. Jackson IJ: **Mouse genomics: making sense of the sequence**. *Curr Biol* 2001, **11**:R311-R314.
27. The Gene Ontology Consortium. **Gene ontology: tool for the unification of biology.** *Nature Genet* 2000, **25**:25-29.
28. **Gene Ontology** [http://www.geneontology.org]
29. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22-28.
30. **COGs** [http://www.ncbi.nim.nih.gov/COG/]
31. **Mammalian Homology and Comparative Maps** [http://www.informatics.jax.org/menus/homology_menu.shtml]
32. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**:1113-1143.
33. Hannenhalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments.** *J Mol Biol* 2000, **303**:61-76.
34. **Distributed Sequence Annotation System (DAS)** [http://stein.cshl.org/das/]
35. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA: **The Stanford microarray database.** *Nucleic Acids Res* 2001, **29**:152-155.
36. **The Stanford Microarray Database (SMD)** [http://genome-www4.stanford.edu/MicroArray/SMD]
37. Hoogland C, Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, Appel RD: **The 1999 SWISS-2DPAGE Database.** *Nucleic Acids Res* 2000, **28**:286-288.