On the representation of "gene function" in databases

A discussion paper for ISMB, Montreal, 1998.
Version 1.2 -- June 19 1998.
Michael Ashburner, EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

---

0. This is not a final document.

It is a discussion paper. I welcome input, be it criticism or suggestions for change. Known bugs are that a few of the URLs do not work. This paper lives at http://www.ebi.ac.uk/~ashburn/ontology_discussion.html.

1. Introduction.

One of the major challenges facing genomic research today is that of integrating sequence data with the vast, and growing, body of data from functional analyses of genes. One step towards meeting this challenge is to design an ontology to represent gene function, and to implement this within a genomic database. The advantages that would flow from this are several. One major advantage is that if different 'single-organism' databases adopted the same ontology, then the community would have a powerful method for exploring the functional aspects of the genomes of several different organisms. Another major advantage is that this will provide an aide to the discovery of the function of new sequences. Genes are expressed in temporally and spatially characteristic patterns. Their products are (often) located in specific cellular compartments and may be part of one or more multi-component complexes. Gene products possess one or more biochemical, physiological or structural functions. Genes may have more than one product and these may be functionally distinct. These are all attributes of genes which are of great interest to all biologists. Genes have orthologs, paralogs and functional 'homologs' (which may or may not share sequence similarity) in other species, and may be assigned by a variety of methods (algorithmic, experimental). We need a way to describe these attributes in a rigorous way that will enable biologists to annotate genomes and to explore the universe of genomes. In the ideal world sequence (nucleic acid, protein), structure, and genomic databases would all agree on how this should be done. That would promote interconnections between databases based upon attributes richer than sequence or structure. There are several attributes of genes that need to be represented systematically. These can be illustrated by some of the types of question a user might ask of a database:

- where is a gene expressed? the spatial problem, described in terms of an organism's anatomy.

- what is the (sub)-cellular localisation of a gene product?, described in terms of subcellular anatomy.
- when is a gene expressed? the temporal problem, described in terms of an organism's ontogeny.
- what is the function of a gene product?, described in terms of a functional classification of gene products.
- of what larger process is the function of a gene's product a part?, described in terms of a process hierarchy.
- by what processes is a gene's activities controlled ? what genes does a gene's product regulate ?, described by a regulatory hierarchy.
- of what larger complex is this function a component?, described by a parts-list of multi-component (RNA, protein, etc) complexes.
- what genes in species a have the function of gene x in species b? represented by species a and b sharing a functional classification of gene products, with the necessary links between the databases.

I will describe a possible solution to the problem in terms of Drosophila. The reason for this choice is that FlyBase is probably the most comprehensive of the 'single-organism' genetic/genomic databases. Moreover, FlyBase has made considerable progress in this field already. In particular, FlyBase already describes genes in terms of the spatial and temporal properties of their expression. However, we stress that there is nothing in this solution that is Drosophila specific. The solutions are designed to be general.

2. Progress by FlyBase in the design and implementation of a controlled vocabulary for anatomy and ontogeny.

FlyBase has built and uses a graph of anatomy and subcellular anatomy that incorporates ontogeny (1). This is, for flies, essentially complete (although it is continually updated as knowledge changes). (A similar, but strictly hierarchical, anatomical controlled vocabulary has also been built for the mouse by the Edinburgh group and will be used by the Mouse Gene Expression Database (2)). When originally written, this controlled vocabulary was organized as a simple hierarchy. This has proved to be insufficiently expressive and its data structure is now rather more complex. The limitations of a simple (tree-shaped) hierarchy are (a) that it can express only one type of relationship between pairs of terms, and (b) that the relationship cannot be many-to-many, only one-to-many. The controlled vocabulary originally satisfied these constraints, because we represented only the relationship x is a component of y and we defined only non-intersecting terms, so that x could never be a component of both y and z unless y was itself a component of z (or z of y). The present controlled vocabulary, in contrast, represents three distinct types of relationship, all of which can be many-to-many. These relationships are: instance (isa), part-whole and ontogenetic (the last of these may demand some explanation, it is the relationship: structure x develops from structure y).

The version of the following FlyBase controlled vocabularies now made available (1) covers the following domains:

- $publication_cv
- $genes_cv
- $allele_cv
- $aberration_cv
- $wild-type_stock_descriptor
- $transcript_cv
- $kilobase_cv
- $protein_cv ! this is empty see (ref: 4)
- $gene_function_cv ! this is empty see (ref: 5)
- $molecular_cv
- $subcellular_location_cv
- $anatomical_descriptive_qualifier_cv
- $anatomical_descriptive_term_cv
- $developmental_stage_cv
- $anatomy_cv

The following conventions are used:

! <text> text to the right of ! is a comment; ! alone on a line is simply for ease of viewing.

/= <text> text to the right of /= is a synonym of the term; always follows !; may be more than one per line.

[ ] <integer> the integers are references for the term; always follows !; may be more than one per line (the references are not included in this version).

$<text> the name of the domain of terms.

%<term> an instance of the term that is the immediate parent term.

% <term> an instance of a term; may be more than one per term.

<<term> a part of the term that is the immediate parent term.

< <term> a part of a term; may be more than one per term.

~ <term> the ontogenetic (developmental) parent of the term; may be more than one per term. Any combination of % <term> < <term> ~ <term> may occur on a line; but these are always in order % < ~.

Despite its structure this ontology is used very simply in FlyBase to provide controlled vocabularies for various domains. FlyBase is in the process of writing tools to allow users to exploit the ontologies (a prototype has been written by Aubrey de Grey). None of the terms have definitions. These will be written. In addition, for the anatomy_cv, FlyBase will link the ontology to pictures of the objects (see (3)).

3. Description of 'gene function' in FlyBase - the current model.

FlyBase now uses a small keyword list - that is a simply a flat list of controlled terms - to describe functional attributes of genes (4). Examples of these are enzyme names (using those of the Enzyme Commission with a database cross-reference to the ENZYME database via the EC number) and, a controlled vocabulary of terms, e.g. "transcription-factor", "GABA-receptor" or "chromatin-binding-protein". In FlyBase there are now about 950 different terms and over 3200 genes are indexed by one or more term. FlyBase WWW interfaces allow searches to be made on these terms. This is, very clearly, a poor representation of rich data, since a keyword list cannot represent any relationship between objects. [As an aside, FlyBase uses PROSITE terms for the description of structural attributes of gene products, see (5)).]

4. What do we mean by function ?

Conceptually, this has been a major hurdle during discussions of this subject. Distinctions have been made between 'function', 'role' and 'process'. Examples of 'function' would be: "transcription factor", "transporter", "enzyme"; examples of 'role' would be: "transcription factor regulating HOXA1", "transporting sucrose"; examples of 'process' would be: "specification of thoracic segmental identity", "catabolism of carbohydrates". At the Les Treilles meeting the discussion concluded that 'role' was probably not needed, a conclusion reflected by Letovsky in (6). Gelbart, in a post Les Treilles personal contribution, suggested the term "integrative processes". These would be a subset of Letovsky's "biological processes". We conclude, following the discussions at Les Treilles, that the major dimensions of the classification are: molecular function biological process These can be best illustrated by a few examples:

"molecular function" would (at high levels) include:

transporter enzyme
transcription_factor
motor_protein
signaling_molecule
receptor

"biological process" would (at high levels) include:

intermediary_metabolism
intracellular_protein_traffic
organogenesis
stress_response
sex_determination
gametogenesis
behavior

5. Design.

Letovsky set out in (6) some of the design criteria; I paraphrase and add to these:

- machine readable.
- multidimensional, with more than one axis of classification.
- applicable across species.
- allow multiple classes of relationship, for example part-whole relationships, ISA relationships, temporal relationships, synonymies, taxonomical extent.
- allow expression of regulatory and reaction relationships.
- allow representation of incomplete knowledge.

6. Implementation.

There has been considerable work within the AI community in the design of tools for ontologies. A good example is the 'Ontolingua' project at Stanford (7). In addition, there are major projects in the medical field for the rigorous representation of concepts and facts, for example the National Library of Medicine's 'Unified Medical Language System' (UMLS) (8). There would be obvious advantages were the ontology for the representation of gene function be designed in a format that can be translated, for example, by Ontolingua. Schulze-Kremer (9, 10) is developing a general ontology for molecular biology (OMB); he has also developed an ontology editor. In general, this ontology is 'top-down' and has not, to my knowledge, been implemented within a biological context.

7. What should we do now ?

For FlyBase the priorities are:

(i) to design and evaluate abstract schema for the representation of ontologies relevant to the domain of FlyBase. One of the challenges here is to define the classes of relationship between objects. A very interesting set of classes of biological relationship has been developed by the Cytoskeletal Protein Interactions Database at Yale (11).

(ii) to translate existing FlyBase ontologies into this schema.
(iii) to build an ontology for the functional domain.
(iv) to represent FlyBase data within the context of this ontology.

I have made available preliminary draft function and process ontologies for Drosophila at (5a) and (5b).

The existing functional information in FlyBase will be the first guide for representing data within the ontology. The names (symbols) of genes and their protein products will be attributes of objects in the functional schema. [FlyBase curation attaches unique symbols and identifier numbers to each different protein product of a gene; this means that 'function' objects can be linked to both.] In addition to structured FlyBase data, the text information attached to FlyBase genes, the hierarchical index made by FlyBase for the InterActive Fly (12) and the KEYWORD list of SWISSPROT, can all be used as sources of data for population. Although the primary links will be made from gene products as attributes of functional terms, it will be a simple parse to replace these attributes with others, for example, protein sequences, protein domains or profiles and 3D-structures. This will be easy because links between these attributes and gene names

(symbols) exist both in FlyBase and in the sequence data bases (Genbank/EMBL and SWISSPROT). [Every Genbank/EMBL nucleic acid sequence and every SWISSPROT/TREMBL record from Drosophila has a database_xref to and from FlyBase; the only exceptions to this statement are very recent data (< circa 1 month) and some of the Berkeley Drosophila Genome Project ESTs]. It will, as a consequence, be straightforward to go from a new sequence to the functional roles in Drosophila played by Drosophila genes of related sequence (as determined by a similarity algorithm to a sequence or pattern database) and, hence, to make inferences concerning the function of related (e.g. by sequence or structure) proteins in distant species. In practice, we would imagine prototyping the function domain with a carefully limited set of sub-domains.

8. Other work in this field.

There are several related projects, particularly in the bacterial and yeast fields.

The field of metabolic reconstruction and metabolic databases clearly overlaps this project, but differs in emphasis. With the exception of the MIPS and YPD efforts for yeast, and EcoCyc, there are few attempts to integrate such functional classifications with genetic databases.

There have been some proposals for a simple hierarchical classification of gene products, based on the Enzyme Commission system. These include a proposal for the classification of plant gene products (13) and a draft classification of the proteins involved in the process of protein synthesis by Amos Bairoch (14). The pioneer in this field was, of course, Monica Riley who, in 1993, published a very influential classification of gene function for E. coli (15). This has served as the basis for, for example, the classifications of gene function used by the TIGR groups in their analyses of the complete sequences of several bacterial genomes (e.g. (16)) and by others (e.g. M. pneumoniae group in Heidelberg (17); the Japanese group working on cyanobacterial sequences (18)). The NCBI is working on a related project for sequenced bacterial genomes, 'Clusters of Orthologous Groups' (19) based on Riley's work.

The EcoCyc database of metabolic pathways (20) developed by Peter Karp and Monica Riley and the PUMA project at the Argonne National Laboratory (21) both integrate functional classifications (what PUMA calls a functional 'overview') and metabolic pathways databases (see also the KEGG project (22)). The PUMA project has now been succeeded by WIT ('What Is There') at ANL (23). TIGR have also developed a simple classification of cellular roles (distinguishing as here between function and role). This (EGAD) was used for their classification of human EST sequences (24).

For Saccharomyces cerevisiae the MIPS group (A. Zollner) has developed a hierarchical classification (with three levels) of the products of yeast genes (25, see also 26). The Yeast Proteome Database similarly classifies gene products by function and subcellular location (27).

In the plant world there has been considerable discussion about classifications of gene products, largely under the auspices of the MENDEL group (28). SoyBase, a genetic database for Soy bean, has a classification of metabolism (29). There are also efforts towards classifications within functional domains, for example the Cell Signaling Networks Database (30).

9. Acknowledgements.

I thank first Aubrey de Grey for his help; without it nothing would have been done. I thank and acknowledge Monica Riley for her lead; I also thank Terri Gaasterland and Natalia Maltsev for discussions, Stan Letovsky for his wisdom, Suzi Lewis for her encouragement and the participants at the Banbury Meeting in April 1997 and the Les Treilles meeting in October 1997 for making this fun, as well as for their knowledge and advice. FlyBase is supported by grants from the NIH (W.M. Gelbart, PI) and the Medical Research Council, London.

10. References.

(1) http://www.ebi.ac.uk/~ashburn/drosophila.ontology.txt.v263
(2) http://genex.hgu.mrc.ac.uk/
(3) http://flybase.bio.indiana.edu/images/
(4) http://flybase.bio.indiana.edu/genes/lk/function/function.html
(5) http://flybase.bio.indiana.edu/genes/lk/function/structure.html
(5a) http:/www.ebi.ac.uk/~ashburn/function.ontology.txt.v1.1
(5b) http:/www.ebi.ac.uk/~ashburn/process.ontology.txt.v1.1
(6) http://info.gdb.org/letovsky/provence.html
(7) http://ksl-web.stanford.edu/knowledge-sharing/ontolingua/
(8) http://www.nlm.nih.gov/research/umls/
(9) Schulze-Kremer, S. (1998). Ontologies for molecular biology. Pacific Symp. Biocomputing '98. pp. 695-706.
(10) http://igd.rz-berlin.mpg.de/~www/oe/mbo.html
(11) http://paella.med.yale.edu:80/cypsid/
(12) http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly/hierarchy/test0.html
(13) Dure, L. (1991). Plant Molec. Biol. Reporter. 9: 220-228.
(14) Bairoch, A., Protein biosynthesis functional tree; unpublished.
(15) Riley, M. (1993). Microbiol. Revs 57: 862-952.
(16) http://www.tigr.org/docs/tigr/tigr-scripts/hi_scripts/hi_gene_table.pl
(17) http://www.zmbh.uni-heidelberg.de/M_pneumoniae/Herrmann/Results.html
(18) http://www.kazusa.org.jp/cgi-1/get_htext?S.PCC6803+B
(19) http://ncbi.nlm.nih.gov/COG/
(20) http://ecocyc.PangeaSystems.com/ecocyc/server.html
(21) http://www-c.mcs.anl.gov/home/compbio/PUMA/Production/puma.html>/A>
(22) http://www.genome.ad.jp/htbin/show_man?pathway
(23) http://www-c.mcs.anl.gov/home/compbio/WIT/wit.html
(24) http://www.tigr.org/docs/tigr-scripts/egad_scripts/role_reports.spl
(25) http://elk.mips.biochem.mpg.de/ycd/function/
(26) http://pedant.mips.biochem.mpg.de/frishman/pedant.html
(27) http://www.proteome.com/YPD_contents_by_category.html
(28) http://probe.nalusda.gov:8000/plant/aboutmendel.html
(29) http://129.186.26.94/
(30) http://geo.nihs.go.jp/csndb/class.html