Using ontology visualization to understand annotations and reason about them

Mary E. Dolan, Ph.D., and Judith A. Blake, Ph.D.,
Mouse Genome Informatics [MGI], The Jackson Laboratory,
Bar Harbor, ME 04609 USA
mdolan@informatics.jax.org

Biomedical ontologies not only capture a wealth of biological knowledge but also provide a representational system to support the integration and retrieval of biological information. Various biomedical ontologies are used by model organism databases to annotate biological entities to the literature and have become an essential part of high throughput experiments and bioinformatics research. We are exploring the power of ontology visualization to enhance the understanding of annotations by placing annotations in the graph context of the broader biological knowledge the ontology provides. Presenting annotations in this context provides a better understanding of the annotations because humans are adept at extracting patterns and information from graphical representations of complex data.

INTRODUCTION

Biological systems can be very complex but many aspects of biological system characterization have a wealth of biomedical knowledge accumulated over years of clinical and laboratory experience. Ontologies provide a shared understanding of a domain that is human intelligible and computer readable and, consequently, a representational system to support the integration and retrieval of this knowledge.

As techniques of large-scale genomic analysis and functional gene annotation have progressed and are becoming more common, it is essential to find approaches to provide a comprehensive view of annotation sets. We are exploring the power of several widely used ontologies to provide a comprehensive graphical view of annotations by presenting the annotations visualized within an ontology relationship structure. By presenting annotations in the graph context we hope to provide a better understanding of the annotations because humans are adept at extracting patterns and information from graphical representations of complex data.

BACKGROUND

Ontologies can be used to abstract knowledge of a domain in a way that can be used by both by humans and computers by providing an explicit representation of the entities of interest and the relationships among them. In particular, biomedical ontologies representing various aspects of biology are being used for annotating entities to the literature and for integrating the diverse information resulting from the analysis of high-throughput experiments.

Open Biomedical Ontologies (OBO) is an umbrella repository for well-structured controlled vocabularies for shared use across different biological and medical domains [1]. The OBO website contains a range of ontologies that are designed for biomedical domains. Some of the OBO ontologies, such as the Gene Ontology (GO), apply across all organisms. Others are more restricted in scope; for example, the Mammalian Phenotype Ontology (MP) is a phenotype ontology designed for specific taxonomic groups.

The GO Project was established to provide structured, controlled, organism-independent vocabularies to describe gene functions [2] and, as a consequence, provides semantic standards for annotation of molecular attributes in different databases. Members of the GO Consortium supply annotations of gene products using this vocabulary. The GO and annotations made to GO provide consistent descriptions of gene products and a valuable resource for comparative functional analysis research.

Currently, the three ontologies of GO contain nearly 20,000 terms [3]. The terms are organized in structures called directed acyclic graphs (DAGs) which differ from strict hierarchies in that a more specialized (granular) child term can have more than one less specialized parent term. In the GO a child can be related to a parent by either a ‘part of’ or ‘is a’ relationship. Mouse Genome Informatics (MGI) curators use the GO to annotate mouse genes from the literature. Currently, MGI has more than 100,000 annotations to more than 17,000 genes; approximately half of the annotations are manual

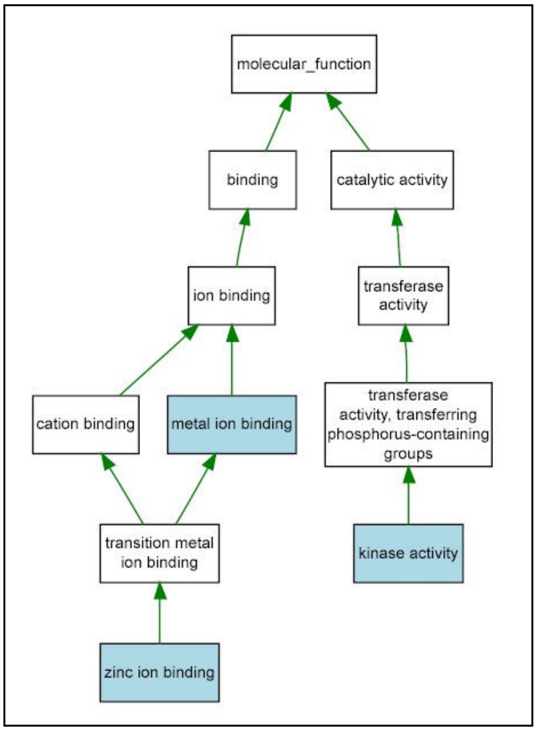


Figure 1. GO annotation graph for mouse Hgs (HGF-regulated tyrosine kinase substrate) provides an alternative to tabular or text views. Blue/shaded nodes in the GO graph indicate mouse annotations. Full graph and annotation set available at: http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=GOMarkerGraph&id=MGI:104681

annotations from the literature, the balance from automated data loads. An MGI user has the option of viewing the full set of GO annotations for a particular gene in three formats: as a table, as automatically generated text, and as a graph. The graph presents relevant parts of the GO with direct annotations indicated as colored nodes, as shown in figure 1. The graphical format allows a user to easily see, for example, whether a gene product appears to participate in a broad range of molecular functions or in only a narrow, specialized function.

Genes that share close evolutionary relationships are likely to function in similar ways. As a complement to our previous work [4] on the assessment of annotation consistency of independently developed annotation sets for curated mammalian orthologs [5], we provided comparative graphical visualizations of annotations, one graph for each mouse-human-rat ortholog triple with nodes colored according to organism annotated. Coloring nodes to distinguish among annotations extends the usefulness of the visualization for pattern recognition by users. The graphical format, as shown in figure 2, allows a user to assess the consistency, inconsistency and level of detail of annotations made to different model organisms.

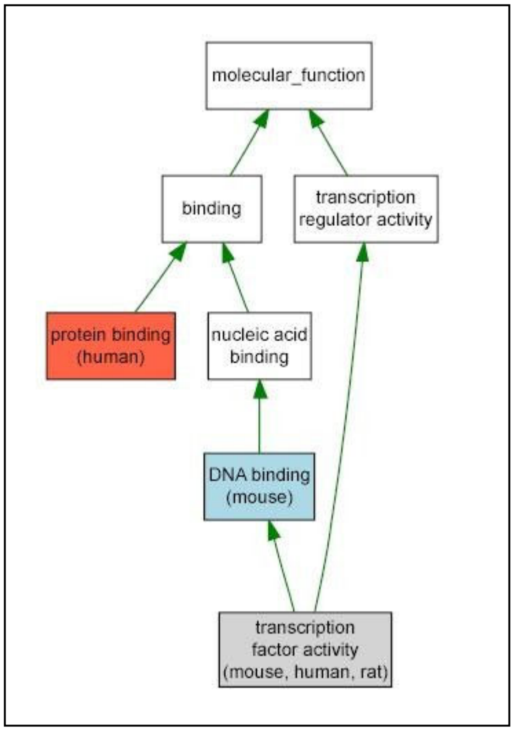


Figure 2. GO comparative graph for MGI curated orthologs to mouse Pax6 (paired box gene 6). The nodes are color-coded according to organism: mouse annotations shown in blue/lighter shading, human annotations in red/darker shading, multiple organisms in gray. Full graph available at: http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=GOOrthologyGraph&id=MGI:97490

Our examination of the comparative graphs led to the observation that annotations are often complementary, reflecting the fact that the different model organisms are used to study different aspects of biology. Since biologists are often species-blind and assemble their initial picture of a gene and its function without regard to the taxonomic origin of the gene that was studied in a particular experiment, this suggested the broader application of such graphs as ‘summary’ rather than ‘comparative’ graphs that might be used to answer the request: “Show me everything that is known about this gene.” The power of this representation is that it provides a view of the summary of information derived from species-specific experimental results.

In addition to the ability to visualize comparative annotation sets, graphs can be used to coordinate information for animal models of human

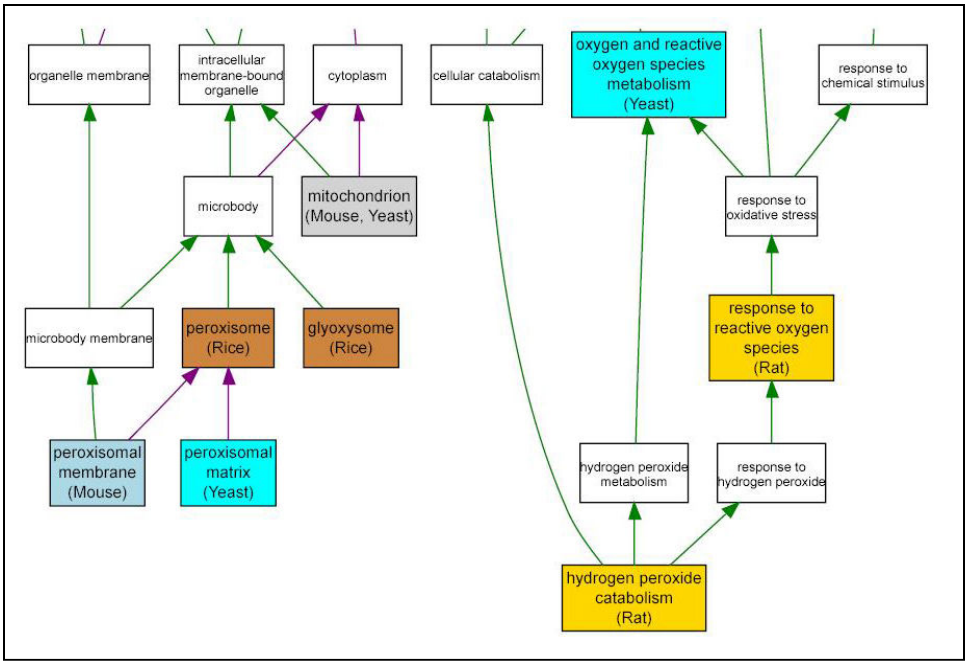

organelle membrane
intracellular membrane-bound organelle
cytoplasm
cellular catabolism
oxygen and reactive oxygen species metabolism (Yeast)
response to chemical stimulus
microbody
mitochondrion (Mouse, Yeast)
response to oxidative stress
microbody membrane
peroxisome (Rice)
glyoxysome (Rice)
response to reactive oxygen species (Rat)
peroxisomal membrane (Mouse)
peroxisomal matrix (Yeast)
hydrogen peroxide metabolism
response to hydrogen peroxide
hydrogen peroxide catabolism (Rat)

Figure 3. GO annotation graph for OMIM gene CATALASE; CAT. The graph coordinates GO annotations for thirteen model organisms with nodes colored by organism. Full graph and annotation set available at: http://www.spatial.maine.edu/~mdolan/OrthoDisease_Graphs/OMIM_GeneGraphs/CAT.html

diseases. The primary purpose of performing experiments that study the consequences of mutations in a particular organism is that these experiments provide valuable models for the understanding of human disease. We have extended our ontology visualization approach [6] to the orthology sets developed in the resource OrthoDisease [7], a comprehensive database of model organism genes that are orthologous to human disease genes derived from the OMIM database [8], a continuously updated catalog of human genes and inherited, or heritable, genetic diseases. We have abstracted orthology information on thirteen organisms for which curated GO annotation sets are publicly available. By combining all GO annotations for the orthologs associated with each disease gene or with each disease, we obtain a comprehensive annotation set for each disease gene and for each disease. Each annotation set is presented on the GO graph with nodes having annotation colored according to the organism that is the source of the annotation. Figure 3 shows part of the graph for OMIM gene CAT that demonstrates the degree of similarity annotations to diverse organisms can show. Of course, in some sense, it is the differences that are of more interest in this case since we are interested in collecting together as much information as possible.

DESCRIPTION OF CURRENT WORK

While each annotation group develops curation standards to meet the needs of their community, one of the important results of various ontology projects has been an attempt to develop a common vocabulary and shared annotation standards that enhance the utility of these annotations for analysis. We have found that regardless of the ontology, presenting terms in a graphical context makes the relationships of ontology terms clear, provides context for annotations, and makes the examination of large annotation sets feasible. The long-term objective, now, is to build consensus for curation standards that will strengthen the utility of data integration capabilities of this approach.

We have generalized our GO visualization approach to other ontologies and annotation data sets. First, we construct a complete graph to represent the ontology. Second, we color nodes that have annotations and limit the graph to the sections necessary to show all annotations. By limiting the graphs to annotated sections we do not have to deal with scalability issues that might arise if we were to attempt to represent an entire ontology that includes thousands of terms. Finally, we build a web page for each gene that includes an image of the graph and a table of annotations. In addition, to facilitate the examination

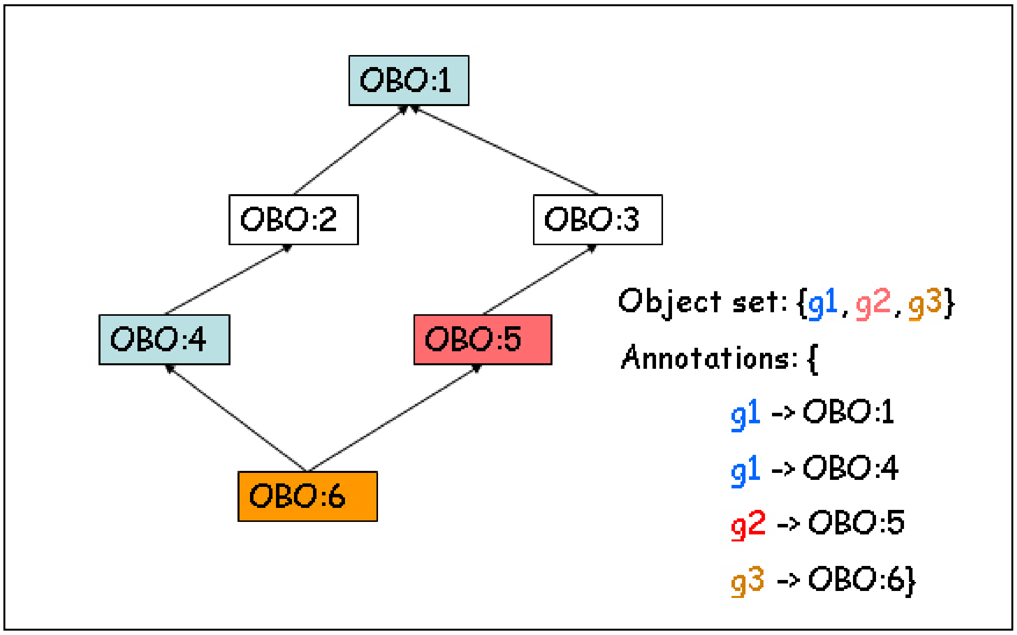


Figure 4. The comparative graph paradigm: an ontology provides the relationship structure among terms; a grouping idea defines the object set; and discriminating idea distinguishes objects whose annotations will be color-coded in the graph.

of larger graphs, we provide scalable vector graphics (SVG) images, which include pan-zoom-search functionality that allow a user to examine specific sections of the graphs. The graph images are generated using GraphViz, a freely available, open source graph layout program [9].

Gene expression data sets describe when and where particular genes are active. Providing a comprehensive picture of the level of gene expression across developmental stages and anatomical structures will facilitate investigation of regulation of gene expression.

We have applied our simple graphical display approach to gene expression data with annotations to both the Adult Mouse Anatomical Dictionary (MA) [10] and the Edinburgh atlas of mouse embryonic development (EMAP) [11]. For each gene with annotation data, the resulting graph shows the mouse anatomy ontology with anatomical structure nodes colored to indicate where that gene is expressed. In addition, in the case of the EMAP graphs, we have attempted to tease apart time dependence of gene expression patterns by separating annotations to different developmental stages by producing graphs for each Theiler stage.

The laboratory mouse is an important model organism for a broad range of human diseases and disorders, including diabetes, heart disease, and cancer. Genomic and genetic investigations of particular mouse models (phenotypes) reveal the contribution of particular genomic variants (alleles) to the presentation of disease phenotypes. The annotation of genotype-phenotype associations is an essential part of assessing mouse models for human disease.

We have adapted our comparative GO annotation approach to phenotype annotations made to different mouse gene alleles to create Mammalian Phenotype (MP) Ontology [12] graphs. As in the case of GO comparative graphs (figure 2), the generalized approach to comparative graphs requires three things: an ontology to provide the relationship structure, a grouping idea to connect the annotated objects, and a distinguishing idea (see figure 4). First, we construct a complete graph to represent the ontology. Second, we color nodes that have annotations according to the distinguishing characteristic and limit the graph to the sections necessary to show all annotations. Finally, we build a web page for each gene that includes an image of the graph and a table of annotations.

In the case of the GO comparative graphs the grouping idea is orthology and the distinguishing idea is organism: mouse annotations in blue, human annotations in red and so forth. In the case of MP graphs the grouping idea is the gene and the distinguishing idea is the allele: each allele's annotated nodes are colored differently. In a similar way to color coding of GO nodes by organism, color-coding of MP nodes by allele allows a user to easily

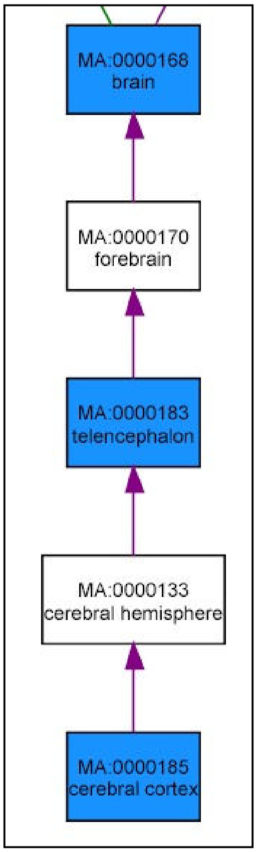


Figure 5. Part of the Adult Mouse Anatomical Dictionary (MA) annotation graph for postnatal expression data for mouse gene Abcg2 (ATP-binding cassette, sub-family G (WHITE), member 2).
Full graph and annotation set available at: http://www.spatial.maine.edu/~mdolan/GXD_Graphs/Abcg2.html

see similarities and differences in alleles annotated to different phenotypes. Our purpose in creating such graphs is to move beyond simply providing another representation of a phenotype data set to add potential value to this data set as a method of assessing mouse models for human disease.

RESULTS

Graphical representations of expression data sets using anatomy ontologies

The Mouse Anatomical Dictionary provides ontologies that provide a standardized nomenclature for anatomical parts to describe the complex patterns of gene expression in the developing and adult mouse and how they relate to the emerging tissue structure. Terms that describe embryonic developmental stages (Theiler Stages 1 through 26) have been developed by the Edinburgh Mouse Atlas Project (EMAP) [11]. Terms that describe mice at postnatal stages, including adult (Theiler stage 28) have been developed as the Adult Mouse Anatomical Dictionary (MA) [10].

Adult Mouse Anatomical Dictionary graphs display relationships of annotations

The Adult Mouse Anatomical Dictionary (MA) is an anatomy ontology that can be used to provide standardized nomenclature for anatomical terms in the postnatal mouse. It was developed as part of the Gene Expression Database (GXD) resource of information from the mouse [12]. The Adult Mouse Anatomical Dictionary organizes anatomical structures for the postnatal mouse spatially and functionally. Each MGI gene detail page includes links to gene expression data; the user can select data for the postnatal mouse and obtain a tabular view of available expression data.

Our graphical representations present another view of the data, as shown in figure 5. This partial view of the graph for Abcg2 (ATP-binding cassette, sub-family G (WHITE), member 2) clearly shows the relationship of three annotations as variations in granularity. Note that the colored nodes indicate only direct annotations made by curators from the literature, although indirect annotation can be inferred from the ontology structure.

EMAP graphs provide information on developmental stage specific expression

The Edinburgh Mouse Atlas Project (EMAP) annotation of gene expression data can be used to capture the complex and ever-changing patterns throughout the development of the mammalian embryo and how they relate to the emerging tissue structure at each developmental stage.

We have adapted the EMAP ontology to separate annotations associated with different Theiler stages and created EMAP annotation graphs for each stage, effectively treating each stage as a separate ontology structure. With this approach we can, within the limits of incomplete annotation, see stage separated annotations as a time series of expression patterns. For example, figure 6 shows expression annotations for mouse gene Shh (Sonic hedgehog) for Theiler stages 11 (figure 6, upper panel) and 12 (figure 6, lower panel). A user might consult such graphs to explore changes in expression pattern between stages or determine the earliest stage at which the gene is known to be expressed in a particular anatomical structure. The way these graphs are presented at our web site, a user can move forward or back to adjacent Theiler stage.

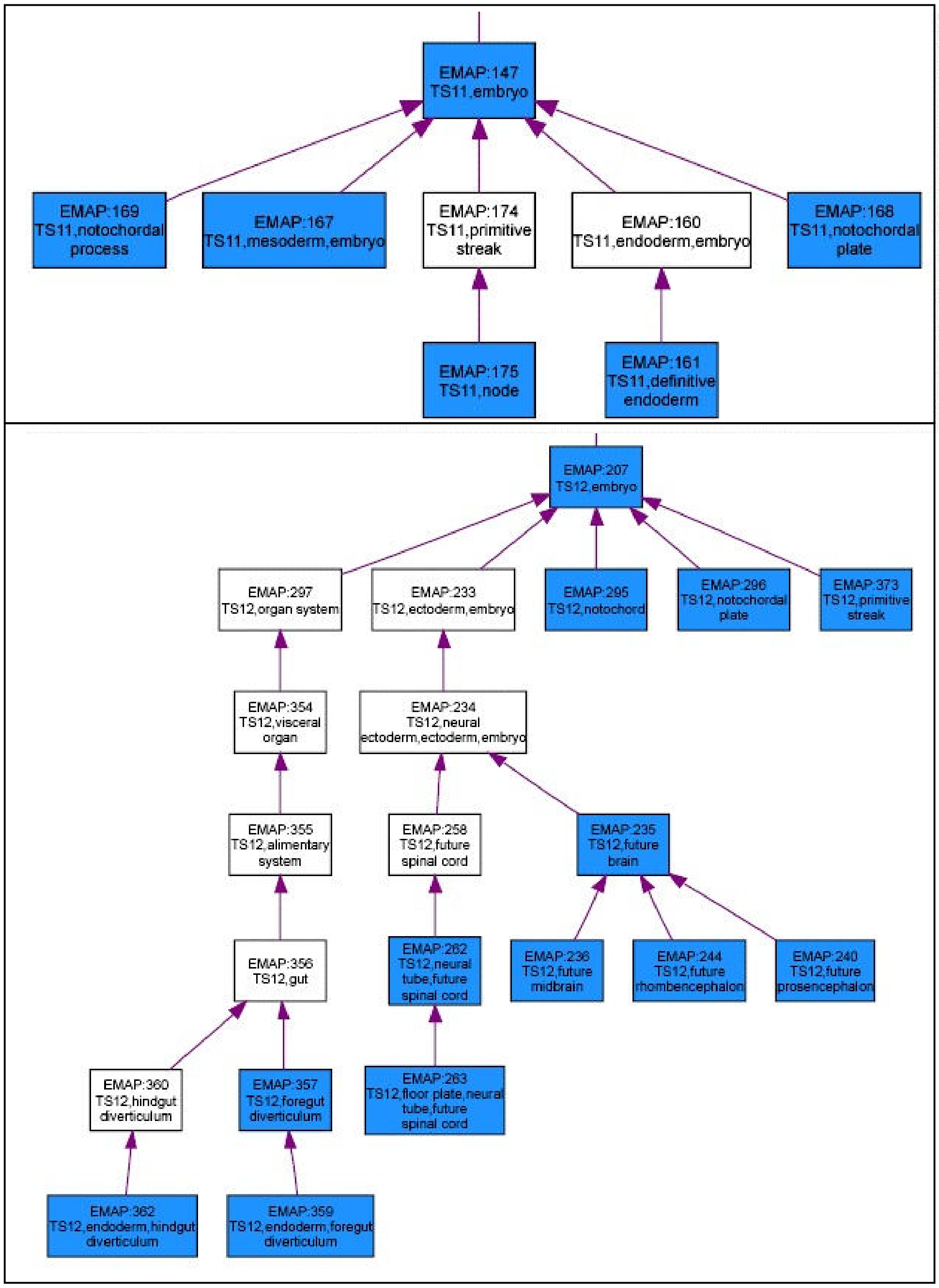


Figure 6. EMAP ontology graphs for Theiler stages 11 (upper) and Theiler stage 12 (lower) displaying expression patterns for mouse Shh (Sonic hedgehog). (Annotations available from GXD.)Full graph and annotation set available at: http://www.spatial.maine.edu/~mdolan/GXD_Graphs/TimeSlices/TS11.html

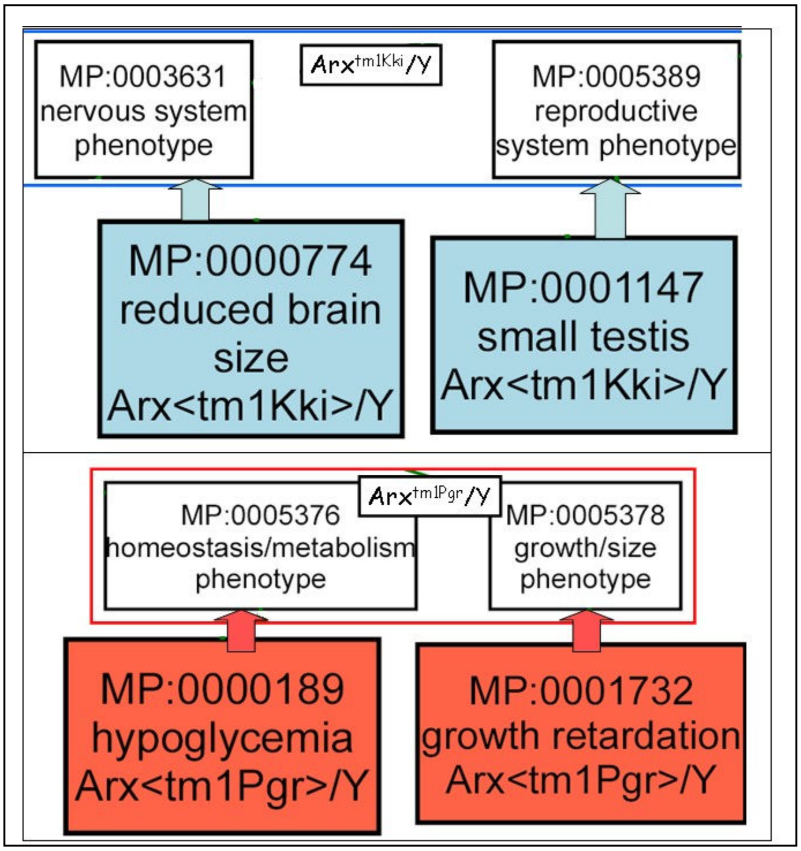


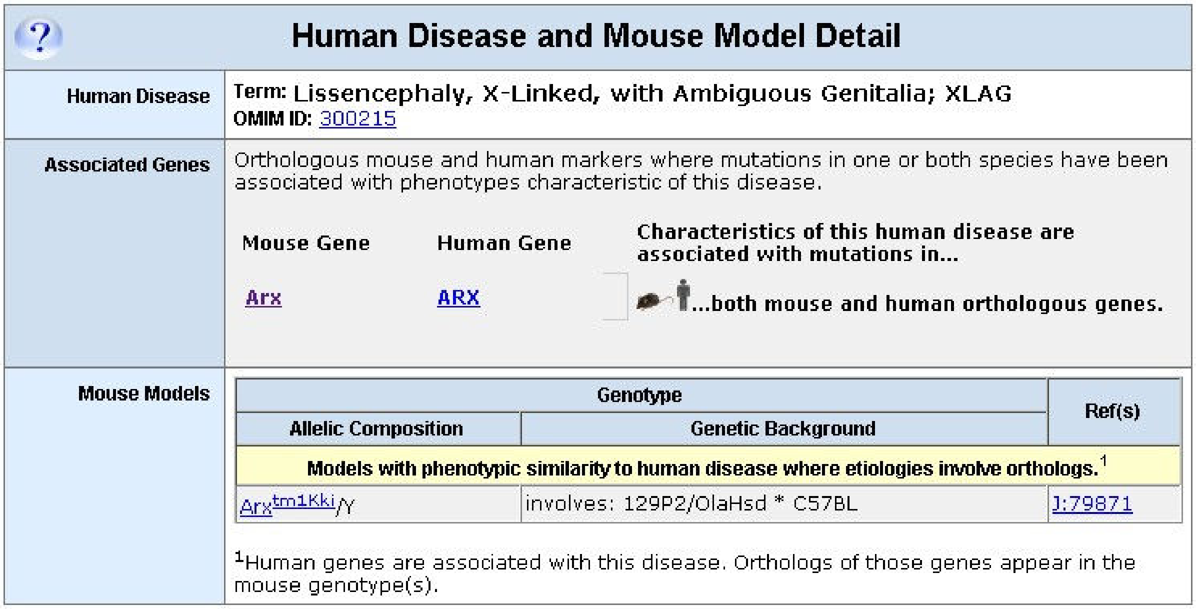
Figure 7. Detail of the Mammalian Phenotype (MP) Ontology annotation graph for two alleles of mouse gene Arx representing allelic compositions Arx^{tm1Kki}/Y (blue/lighter shading) and Arx^{tm1Pgr}/Y (red/darker shading). We observe that the allele annotations segregate in separate ontology branches. Only the allelic composition Arx^{tm1Kki}/Y high-level phenotypes correspond to nervous system and reproductive system phenotypes, while only the allelic composition Arx^{tm1Pgr}/Y corresponds to homeostasis/metabolism and growth/size phenotype. Full graph and annotation set available at: http://www.spatial.maine.edu/~mdolan/GenoPheno_Graphs/Arx.html

Using graphical representations to reason about annotations: assess mouse models for human disease

The Mammalian Phenotype (MP) Ontology [13] is used by MGI to represent phenotypic data. The MP Ontology enables annotation of mammalian phenotypes in the context of mutations and strains that are used as models of human disease and supports different levels of phenotypic knowledge. For example, among the highest levels of the MP Ontology are terms for: growth/size phenotype, homeostasis/metabolism phenotype, nervous system phenotype, and reproductive system phenotype.

So for example, the mouse gene Arx (aristaless related homeobox gene (Drosophila)) has 2 alleles, Arx^{tm1Kki} and Arx^{tm1Pgr}, both of which have been annotated to MP by curators at MGI. We might ask: how do the annotations to the different alleles compare? Applying the comparative graph methodology and indicating MP annotations to terms by color-coding according to allelic composition Arx^{tm1Kki}/Y and Arx^{tm1Pgr}/Y results in the graph detail shown in figure 7. (Information on mouse strain background is not indicated in the graph but is given in a complete annotation table that accompanies the graph.) We observe that in the graph the allele annotations segregate in separate branches reflecting the fact that the phenotype annotations associated with the two alleles fall into distinct high-level phenotypes. Only the allelic composition Arx^{tm1Kki}/Y corresponds to high-level nervous system and reproductive system phenotypes, while only the allelic composition Arx^{tm1Pgr}/Y corresponds to homeostasis/metabolism and growth/size phenotypes.

.



Human Disease and Mouse Model Detail

Human Disease

Term: Lissencephaly, X-Linked, with Ambiguous Genitalia; XLAG
OMIM ID: 300215

Associated Genes

Orthologous mouse and human markers where mutations in one or both species have been associated with phenotypes characteristic of this disease.

Mouse Gene	Human Gene	Characteristics of this human disease are associated with mutations in...
Arx	ARX	...both mouse and human orthologous genes.

Mouse Models

Genotype		Ref(s)
Allelic Composition	Genetic Background	
Models with phenotypic similarity to human disease where etiologies involve orthologs.[1]		
Arx^{tm1Kki}/Y	involves: 129P2/OlaHsd * C57BL	J:79871

[1] Human genes are associated with this disease. Orthologs of those genes appear in the mouse genotype(s).

*Figure 8. MGI integrates data on mouse models of human disease from OMIM with existing data for mouse genes and strains. For example, as shown on this "Associated Human Diseases" information page for Arx, Arx^{tm1Kki} /Y on the strain background 129P2/OlaHsd * C57BL is a known mouse model for OMIM human disease, "Lissencephaly, X-Linked, with Ambiguous Genitalia; XLAG" characterized by nervous system and reproductive system phenotypes. The visualization methodology as shown in figure 7 is consistent with the known association of this particular human disease and the Arx^{tm1Kki} mouse model. (This page is available at: http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=humanDisease&key=850912)*

This distinction is confirmed by seeing that, indeed, Arx^{tm1Kki} is a known mouse model for OMIM human disease, "Lissencephaly, X-Linked, with Ambiguous Genitalia; XLAG" (see figure 8), which is characterized by nervous system and reproductive system phenotypes. The visualization methodology outlined here is consistent with the known association of this particular human disease and the Arx^{tm1Kki} mouse model. Our hope is that examination of the MP graphs for specific disease associated phenotypes would help point to good mouse models. To facilitate this, we have created an index to all genes and alleles indicating high-level phenotypes. For example, a user can search the index for all genes and alleles annotated for "nervous system phenotype" and examine the linked MP graphs for segregation of allele phenotypes and a potential novel mouse model for a human disease characterized by nervous system abnormality. In this way we have extended the usefulness of the graphical representations beyond just another way of presenting the data to a method that allows a user to reason about annotations.

Availability of graphs

All graphs presented in this work are publicly available.

- The GO graphs are available for each gene from the gene detail pages at MGI.
- The OrthoDisease graphs are available at: http://www.spatial.maine.edu/~mdolan/OrthoDisease_Graphs/
- The Adult Mouse Anatomical Dictionary (MA) graphs for GXD data for selected genes are available http://www.spatial.maine.edu/~mdolan/GXD_Graphs/
- The Theiler stage separated Edinburgh Mouse Atlas Project (EMAP) graphs displaying GXD data for Shh are available at: http://www.spatial.maine.edu/~mdolan/GXD_Graphs/TimeSlices
- The Mammalian Phenotype (MP) graphs for all MGI genes with phenotype annotations are available at: http://www.spatial.maine.edu/~mdolan/GenoPheno_Graphs/

CONCLUSIONS

Biological systems can be very complex but many aspects of biological system characterization have a wealth of biomedical knowledge accumulated over years of clinical and laboratory experience. Ontologies provide a shared understanding of a domain that is human intelligible and computer readable that can help support the integration and retrieval of this knowledge.

Here we provide a methodology to visualize sets of annotations as provided by a model organism database curation system to aid researchers in better comprehending and navigating the data. The result is a comprehensive view of available knowledge. As more annotations are made and become available, such tools will be both more necessary, to handle larger data sets, and more useful, as annotation approaches completeness. We believe that this approach to coordinating biological knowledge available in model organism resources will provide a valuable resource in medical research and contribute to understanding these systems.

Acknowledgements

This work is funded by NIH/NHGRI (HG-002273).

References

1. Open Biomedical Ontologies (OBO) [http://obo.sourceforge.net/]
2. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 2000, 5: 25-29.
3. The Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res 2006, 34: D322-D326
4. Dolan ME, Ni L, Camon E, and Blake JA. A procedure for assessing GO annotation consistency. Bioinformatics 2005, 21(Suppl 1):i136-i143.
5. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, et al. The Mouse Genome Database (MGD): from genes to mice -- a community resource for mouse biology. Nucleic Acids Research 2005, 33: D471-5.
6. Dolan ME and Blake JA. Using Ontology Visualization to Coordinate Cross-species Functional Annotation for Human Disease Genes. Proceedings Nineteenth IEEE International Symposium on Computer-based Medical Systems: Ontologies for Biomedical Systems 2006, 583-587.
7. O'Brien KP, Westerlund I, Sonnhammer EL. OrthoDisease: a database of human disease orthologs. Human mutation 2004, 24(2):112-9.
8. Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research. 2005, 33: D514-D517.
9. GraphViz [http://www.graphviz.org/]
10. Hayamizu TF, Mangan M, Corradi JP, Kadin JA and Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biology 2005, 6:R29 1-8.
11. Baldock RA, Bard JB, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, et al. EMAP and EMAGE: a framework for understanding spatially organized data. Neuroinformatics 2003, 1:309-325.
12. Hill DP, Begley DA, Finger JH, Hayamizu TF, McCright IJ, Smith CM, Beal JS, Corbani LE, Blake JA, Eppig JT, et al. The mouse Gene Expression Database (GXD): updates and enhancements. Nucleic Acids Res 2004, 32: D568-D571.
13. Smith CL, Goldsmith CW and Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biology 2004, 6:R7 1-9.