

Inter-database annotation linkages in model organism databases

W. John MacMullen

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360 Email: macmw@ils.unc.edu

Inter-database linkage via annotations is one approach to the integration of knowledge in the biomedical domain, which is often fragmented by specialization. This pilot study examined annotations in ten model organism databases and the Gene Ontology (GO) to assess explicit and implicit linkages between organisms. The databases had similar annotation processes, content, and knowledge, with some variation due to organizational objectives and technology infrastructure. While all databases had the potential to be linked to all others via GO, only some databases had non-GO links to others. This may be due in part to a lack of biologically significant relationships among some of the organisms, or that putative relationships between them that exist in GO have not yet been explored.

Research goals

This project is a component of the larger Annotation of Structured Data project described in MacMullen (2005). It explores the types of, and relationships among, annotations in specific model organism databases (MODs) and the Gene Ontology resource to gain an understanding of these resources and how their utility and usability might be improved through enhanced annotation functionality.

Information and knowledge in biomedical research become fragmented due to specialization. For instance, accumulated biological knowledge is often captured in databases constrained to particular organisms used as biological models. This knowledge can become isolated from related information in other organisms unless linkages are created. Inter-database linkage of annotations is explored here as one method of ameliorating knowledge fragmentation across such boundaries as organism or species. Annotation activity associated with these databases can be viewed as being of three distinct types: processes, entities, and knowledge (MacMullen, 2005).

Model organism databases

Model organisms are biological organisms which have high research utility due to certain features, such as relative simplicity, small genome size, or functional similarity to aspects of human biology. Within biomedical research they are valued for their use as surrogates for human gene expression analysis. Following the recent development of full-genome sequencing activities, the available quantities of raw sequence data and experimental information have grown significantly. Highly curated public repositories of gene sequence data and related annotations have been developed to assemble this vast body of knowledge into integrated resources for use by researchers. The databases investigated in this project are among the most well known and developed: for the fruitfly *Drosophila melanogaster* (Drysdale, et al., 2005), the mouse *Mus musculus* (Eppig, et al., 2005), the rat *Rattus norvegicus* (de la Cruz, et al., 2005), the roundworm *Caenorhabditis elegans* (Chen, et al., 2005), the zebra-fish *Danio rerio* (Sprague, et al., 2003), the yeasts *Saccharomyces cerevisiae* (Christie, et al., 2004) and *Schizosaccharomyces pombe* (Hertz-Fowler, et al., 2004), the mold *Dictyostelium discoideum* (Kreppel, et al., 2004), the *Oryza* species of rice (Ware, et al., 2002), and the mustard weed *Arabidopsis thaliana* (Rhee, et al., 2003).

Methods

This project used structure- and content analyses as the main approaches in assessing the types of annotations present in selected model organism databases, and the extent to which annotations of similar types were linked across individual databases. Information on MOD annotation processes was obtained from the websites of the individual databases, as well as from discussions with individual curators. The ten databases listed above were analyzed, and the role of the Gene Ontology (GO) database (Harris, et al., 2004) in cross-database linkage was examined. Many model organism databases exist, but only those that are members of the GO Consortium were analyzed for this project, given the focus on the role of the GO project on inter-database annotation linkages.

Results

Annotation processes

Annotation is part of a larger process of curation of biological information. In the context of model organism databases, 'annotation' typically refers to the process by which value is added to raw sequence data by associating information and evidence from a variety of sources. The value added by annotations in the MODs varies from characterizations of the function of genomic sequences to the provision of links to organism strain information and how to obtain physical samples from labs. A keyword search in PubMed on the root 'annotat*' yields 5,163 documents (as of 2005-06-30) whose topics range from in-depth annotation of specific gene sequences to documentation of software tools used to perform both manual and automatic annotation.

Most of the model organism databases have Ph.D.-level biologists on staff whose sole focus is on database curation. They extract information and evidence from published literature and other sources (such as other primary databases) in order to annotate individual gene records. Other researchers within each model organism community supply the curators with additional relevant information. Prior to human curation, software algorithms are sometimes used to make a first pass at identifying putative entities and relations in raw data sets. Some automated and semi-automated text-mining tools also provide functionality to make links to underlying evidence explicit (see, e.g., Chen & Sharp, 2004). While some of the newer MODs use a standard toolkit (Stein, et al., 2002) and have similar intellectual goals, those with longer histories have significant variation in their underlying database infrastructures, physical workflows, and user interfaces.

In addition to these differences, curators often make a distinction between general MOD annotation activities and those specifically related to Gene Ontology annotation. The GO resource was explicitly developed to address the problems of disciplinary knowledge fragmentation (Harris, et al., 2004), but GO annotation is only one task of the overall MOD curation activities. When annotating to GO, the decision criteria involve whether the evidence is appropriate to GO, and if so, which ontologies, terms, and evidence codes should be used. Thus, each piece of evidence of interest to a MOD may or may not be appropriate for GO annotation.

Annotation entities and attributes

Since MODs typically develop as a result of large-scale gene sequencing, the basic unit of analysis in each database is frequently an individual gene (or gene product) record. The curation process distills the accumulated knowledge about the organism into highly concentrated records with significant intra-database links. Each field in a record may be considered an instance of an annotation, but the record itself is also viewed as an annotation to the underlying gene. Field-level annotations may be characterized into types relating to nomenclature, location, structure, function, phenotype, interactions, relationships, and evidence, for example. Most MODs contain tens of thousands of annotations in order to capture this knowledge. Because they differ in their definitions of what an annotation is, it is difficult to compare quantities of annotations across MODs. To provide a baseline context, Table 1 presents a summary of the numbers of gene products that each MOD has annotated to the Gene Ontology, and the numbers of literature references used in making GO annotations. Note that each gene product may be annotated to zero or more terms in each of the three GO ontologies, and not every annotation is derived from a published document. GO contains a total of 18,852 terms (as of 2005-06-30).

Table 1. Total gene products in MODs annotated to the GO ontologies, and references used as evidence (ordered by number of gene products)

| MOD | Gramene | Arabidopsis | Mouse | Rat | C. elegans |
|---------------|----------------|--------------------|--------------|------------|-------------------|
| Gene products | 38,273 | 33,690 | 16,135 | 12,384 | 11,812 |
| References | 2,381 | 3,066 | 5,225 | 3,624 | 755 |

| MOD | Drosophila | Zebrafish | S. cerevisiae | Dictyostelium | S. pombe |
|---------------|-------------------|------------------|----------------------|----------------------|-----------------|
| Gene products | 11,215 | 9,507 | 6,456 | 5,660 | 4,019 |
| References | 7,442 | 530 | 5,296 | 387 | 1,132 |

Source: Gene Ontology Consortium

Knowledge represented in annotations

Examples of knowledge captured by annotations in the types described above include: narrative descriptions of the gene and its associated products and processes; gene name synonyms; GO annotations; individual gene products and variations (e.g., alleles, polymorphisms); genetic markers; genetic and physical maps showing chromosomal location; molecular clones and probes; putative orthologous loci (i.e., similar genes in different species which evolved from a common ancestor); links to underlying evidence, such as literature citations; intra-database links to related products and processes; links to relevant information in external molecular biology resources, such as protein structures; and links to other MODs.

The GO annotation process classifies evidence for annotations into thirteen types, called evidence codes, both to differentiate their sources as well as to provide a certain confidence level. Evidence from a peer-reviewed publication whose underlying experiment used a direct assay to show functional similarity between two genes, for example, has an implicitly higher level of confidence than a similarity score assigned automatically by a sequence alignment program that evaluated the two gene sequences and inferred a relationship.

Links between MODs

Apart from the internal relationships noted in the preceding 'entities' section, MODs explicitly reference related external objects of various types via hypertext links, such as literature citations, homologous genes in other organisms, and protein sequences and structures in other online resources, such as those from NCBI. As described above, the databases studied here have significant numbers of GO annotations and inter-database linkages for each of the three GO ontologies: process, function, and cellular component. While all of the databases have the potential to be linked to all others via GO, only some databases have non-GO links to others. This may be due in part to a lack of biologically significant relationships among some of the organisms, or that putative relationships between them that exist in GO have not yet been explored.

Conclusions and future work

Annotations linked across multiple databases provide one approach to address the problem of knowledge fragmentation by specialization, and provide a way to integrate related knowledge found in specialized resources. Measurement of the quantity of annotations in a MOD, and comparisons of types of annotations across MODs, is difficult due to differing work processes and disagreements about what constitutes an instance of an annotation. In addition, the definitions of what genes and gene products are is also contentious across organisms, exacerbated by the fact that in many cases work on characterizing genes is still in its early stages. One of the ten MODs studied that indicates a future direction for the evolution of these resources is Gramene, which is based on the rice genome, but whose intent is to leverage the knowledge of that genome to integrate and annotate information from other agriculturally important grains and grasses, such as maize, barley, and wheat. This model of horizontal integration into a single resource seems like a logical next step after MODs have vertically integrated a broad variety of information within a specialty.

The work in this project can be extended to include other model organism databases, and to annotation within other generalized resources, such as NCBI's GenBank (and other components of the Entrez suite of resources), to further understand annotation processes, entities, and knowledge. More work is required on user needs and potential interface innovation. Further research is also needed on measuring and evaluating consistency facets of annotations across multiple MODs, methods for quantifying and visually representing and navigating interlinked GO annotations, and mappings of other annotation and classification systems to GO.

Acknowledgments

This work was partially funded by an unrestricted research gift from Microsoft Research to the Annotation of Structured Data research group in the School of Information and Library Science at the University of North Carolina at Chapel Hill. The project's website is available at: <http://ils.unc.edu/annotation>

References

- Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 5, 147.
- Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., et al. (2005). WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Research* 33(1), D383-D389. Available: <http://www.wormbase.org/>
- Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S. et al. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Research* 32(1), D311-D314. Available: <http://www.yeastgenome.org/>

- de la Cruz, N., Bromberg, S. Pasko, D., Shimoyama, M., Twigger, S., Chen, J., et al. (2005). The Rat Genome Database (RGD): Developments towards a phenome database. *Nucleic Acids Research* 33(1), D485-D491. Available: <http://rgd.mcw.edu/>
- Drysdale, R.A., Crosby, M.A., & the FlyBase Consortium (2005). FlyBase: Genes and gene models. *Nucleic Acids Research* 33:D390-D395. Available: <http://flybase.org/>
- Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A., & the Mouse Genome Database Group. (2005). The Mouse Genome Database (MGD): From genes to mice—A community resource for mouse biology. *Nucleic Acids Research* 33(1), D471-D477. Available: <http://www.informatics.jax.org/>
- Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32(1), D258-261. Available: <http://geneontology.org/>
- Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., et al. (2004). GeneDB: A resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Research* 32(1), D339-D343. Available: <http://www.genedb.org/genedb/pombe/>
- Kreppel, L., Fey, P., Gaudet, P., Just, E., Kibbe, W.A., Chisholm, R.L., et al. (2004). dictyBase: A new Dictyostelium discoideum genome database. *Nucleic Acids Research* 32: D332-D333. Available: <http://dictybase.org/>
- MacMullen, W.J. (2005). Annotation as Process, Thing, and Knowledge: Multi-domain studies of structured data annotation. SILS Technical Report TR-2005-02. Chapel Hill: University of North Carolina, School of Information and Library Science, Technical Report Series.
- Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., et al. (2003). The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research* 31(1), 224-228. Available: <http://www.arabidopsis.org/>
- Sprague, J., Clements, D., Conlin, T., Edwards, P., Frazer, K., Schaper, K., et al. (2003). The Zebrafish Information Network (ZFIN): The zebrafish model organism database. *Nucleic Acids Research* 31(1), 241-243. Available: <http://zfin.org/>
- Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., et al. (2002). The generic genome browser: A building block for a model organism system database. *Genome Research* 12(10):1599-1610.
- Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, et al. (2002). Gramene: A resource for comparative grass genomics. *Nucleic Acids Research* 30(1):103-105. Available: <http://www.gramene.org>

Inter-database annotation linkages in model organism databases

W. John MacMullen

School of Information and Library Science, University of North Carolina at Chapel Hill
100 Manning Hall, CB# 3360, Chapel Hill, NC 27599-3360 USA



Abstract

Inter-database linkage via annotations is one approach to the integration of knowledge in the biomedical domain, which is often fragmented by specialization. This pilot study examined annotations in ten model organism databases and the Gene Ontology (GO) to assess explicit and implicit linkages between organisms. The databases had similar annotation processes, content, and knowledge, with some variation due to organizational objectives and technology infrastructure. Types and quantities of inter-database links varied considerably, as did the target resources to which each MOD linked. While all databases had the potential to be linked to all others via GO, only some databases had explicit direct links to others, or via GO. This may be due in part to a lack of biologically significant relationships among some of the organisms, or that putative relationships between them that exist in GO have not yet been explored.

RESEARCH QUESTIONS

- How integrated (via inter-links) is the knowledge in MODs?
- How many inter-DB links exist in a MOD entry?
- How many inter-MOD links exist in a MOD entry?
- What types of knowledge do inter-DB links link?

Definitions:

- Annotations studied here are the links in MOD entries
- Links can be intra-database (within) or inter-database (between)
- Inter-database links can be explicit (physical link) or implicit (intellectual link)

METHODS

Since preliminary investigations showed that the amount and type of annotation in MODs could vary by gene function and degree to which the underlying gene had been characterized, a strategy was desired by which MOD entries selected for comparative analysis could be more representative than by selecting records at random.

Entry selection

The Gene Ontology database was searched via AmiGO to find a relatively granular GO term to which a large number of MODs had created annotations. GO term GO:000166, 'nucleotide binding' annotated one or more times by direct association by all MODs except ZFIN. In each of the MODs with links, one entry that contained the annotation to GO:000166 was examined. In cases where more than one annotation was present, the first entry in the result set was selected. As an exploratory project, no attempt was made to find 'representative' entries, and the results below may vary significantly if different entries are substituted.

Link quantification

The raw numbers of inter-database and inter-MOD links were counted in each MOD entry. Each MOD's most detailed entry type was used. The actual number of links within an entry to a resource were counted, even if the entry linked to the same resource more than once. The directionality of inter-MOD links was collected and appears in Table 1. Other annotation types and features were also collected, such as link facets of any experimental articles associated with an entry, which are shown in Table 3.

Link typing

Nine categories of inter-MOD links were observed in the entries: Gene Ontology, Sequence, Structure, Orthology, Homology, Expression/Interaction, Literature, Visualization, and Definition. These non-MOD inter-database links are summarized in Table 2.

RESULTS

Inter-MOD link quantities

A total of fourteen inter-MOD links were present across four of the nine MODs for the selected entries with the annotation to GO term GO:000166. Only one of the four MODs, Wormbase, had links to multiple MODs – one each to Flybase, GeneDB, S. pombe, and SGD. RGD had nine links, all to MGD, but conversely, MGD had only one link to RGD. (However, the nine physical RGD links were to the same underlying MGD entry.) The final inter-MOD link was from GeneDB, S. pombe to SGD. Table 1 shows the distribution and directionality of inter-MOD links.

Table 1. Inter-MOD links for GO:000166

| DB | DictyBase | Flybase | GeneDB | Gramene | MGD | RGD | SGD | TAIR | Wormbase | GO | All |
|-----------------|-----------|---------|--------|----------|---------|------|------|--------|----------|----|-----|
| Gene product | dictyBase | CG17904 | po4 | 49D11.23 | 1914137 | Diat | HNT1 | ARPC1A | UNC-53 | | |
| DictyBase | - | - | - | - | - | - | - | - | - | 1 | 1 |
| Flybase | - | - | - | - | - | - | - | - | - | 1 | 1 |
| GeneDB S. pombe | - | - | - | - | - | - | - | - | - | 1 | 1 |
| Gramene | - | - | - | - | - | - | - | - | - | 1 | 1 |
| MGD | - | - | - | - | - | 9 | - | - | - | 1 | 1 |
| RGD | - | - | - | - | - | 1 | - | - | - | 1 | 1 |
| SGD | - | - | - | - | - | - | - | - | 1 | 1 | 1 |
| TAIR | - | - | - | - | - | - | - | - | - | 1 | 1 |
| Wormbase | - | - | - | - | - | - | - | - | - | 1 | 1 |
| Zfin | - | - | - | - | - | - | - | - | - | 1 | 1 |
| GO | 11 | 1 | 13 | 8 | 9 | 21 | 5 | 5 | 3 | 30 | 148 |
| All MODs | 0 | 0 | 1 | 0 | 1 | 9 | 0 | 0 | 3 | 3 | 14 |
| All MODs + GO | 11 | 1 | 14 | 8 | 10 | 30 | 5 | 5 | 33 | 9 | 126 |

Inter-DB link types

The most common link types observed were those related to the underlying sequence (142) and structure (140) of the entry, although in each case one database accounted for the majority of links. The SGD entry had the most overall annotations with 176, and entries in more types (seven of nine) than any other MOD. Three other MODs had links in six categories. Table 2 shows the distribution of link types by and among the nine MODs. A total of 602 inter-DB links were observed across the nine MODs, with 233 unique resources were linked to by the nine entries. Resources with the most links from multiple MODs were the NCBI Nucleotide (51), and InterPro (22) resources. Some resources had larger numbers of links, but they came almost entirely from single MODs; for example, PDB with 44 from SGD; SCOP with 23 from SGD; and NCBI Blink with 22 from Gramene.

Table 2. Inter-DB link types for GO:000166

| DB | DictyBase | Flybase | GeneDB | Gramene | MGD | RGD | SGD | TAIR | Wormbase | GO | All |
|------------------|-----------|---------|--------|----------|---------|------|------|--------|----------|-----|-----|
| Gene product | dictyBase | CG17904 | po4 | 49D11.23 | 1914137 | Diat | HNT1 | ARPC1A | UNC-53 | | |
| Gene Ontology | 11 | 1 | 13 | 8 | 9 | 22 | 5 | 5 | 3 | 30 | 104 |
| Sequence | 4 | 29 | 2 | 90 | 6 | 8 | 2 | 1 | 1 | 1 | 142 |
| Structure | 1 | 11 | 3 | 8 | 99 | 1 | 1 | 1 | 1 | 16 | 140 |
| Orthology | 1 | 1 | 1 | 3 | 1 | 1 | 25 | 1 | 5 | 4 | 41 |
| Homology | 1 | 1 | 4 | 22 | 1 | 1 | 4 | 1 | 1 | 1 | 31 |
| Express/Interact | 3 | 5 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 12 |
| Literature | 36 | 0 | 9 | 0 | 6 | 2 | 34 | 3 | 21 | 111 | 111 |
| Visualization | 1 | 1 | 1 | 0 | 3 | 7 | 1 | 1 | 1 | 16 | 12 |
| Definition | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 10 |
| Total | 53 | 34 | 42 | 35 | 119 | 39 | 176 | 11 | 72 | 21 | 602 |

Literature-related annotation features

Experimental papers in the biomedical literature provide evidence of relationships that is used when creating annotations in MODs. Table 3 provides details on literature-related links observed in the nine MOD entries. The Wormbase entry had the greatest number of unique citations (63) and internal links (126). The DictyBase entry had the greatest number of external links (36). External links were typically to the paper's abstract in PubMed, or (in two MODs) to a source of the paper's full text. Internal links were typically to a separate MOD entry where the details of that paper are described. Four of the nine MODs assigned keywords to papers to aid in searching, and the DictyBase entry had a two-level category/keyword indexing system.

Table 3. Literature-related annotation features for GO:000166

| DB | DictyBase | Flybase | GeneDB | Gramene | MGD | RGD | SGD | TAIR | Wormbase | GO | All |
|---------------------|-----------|---------|--------|----------|---------|------|------|--------|----------|-----|-----|
| Gene product | dictyBase | CG17904 | po4 | 49D11.23 | 1914137 | Diat | HNT1 | ARPC1A | UNC-53 | | |
| Total unique cites | 23 | 1 | 1 | 11 | 5 | 1 | 7 | 3 | 3 | 63 | 90 |
| Internal links | 23 | 1 | 1 | 3 | 9 | 1 | 17 | 3 | 3 | 126 | 183 |
| External links | 36 | 1 | 8 | 3 | 6 | 1 | 34 | 3 | 21 | 109 | 109 |
| Full text links | 13 | 1 | 1 | 1 | 1 | 1 | 11 | 1 | 1 | 24 | 24 |
| Categories/keywords | 3/9 | | | | | | 1 | 20 | 17 | | |

Model Organisms

Model organisms are biological organisms which have high research utility due to certain features, such as relative simplicity, small genome size, or functional similarity to aspects of human biology. Within biomedical research they are valued for their use as surrogates for human gene expression analysis. Model organism databases (MODs) provide rich collections of professionally curated information about specific model organisms. The ten MODs studied in this project are:

- DictyBase, for the mold *Dictyostelium discoideum*
- Flybase, for the fruitfly *Drosophila melanogaster*
- GeneDB from Sanger Institute for the fungus *Schizosaccharomyces pombe*
- Gramene, for the rice *Oryza sativa*
- MGD, the Mouse Genome Database, for *Mus musculus*
- RGD, the Rat Genome Database, for *Rattus norvegicus*
- SGD, the Saccharomyces Genome Database, for *Saccharomyces cerevisiae* (yeast)
- TAIR, the Arabidopsis Information Resource, for *Arabidopsis thaliana* (mustard plant)
- Wormbase, for the roundworm *Caenorhabditis elegans*
- Zfin, the Zebrafish Information Network, for *Danio rerio*

Visualizing inter-DB links and relationships

Even in a single-node analysis of inter-DB links, the graphical display of linkages can be useful in visualizing the network of relationships. Figure 1 provides a partial view of the number of links between the MODs, the Gene Ontology, and selected third-party resources to illustrate the complexity of links in only nine MOD entries. Links to external resources (both shown and not shown) from some of the MODs were omitted for readability.

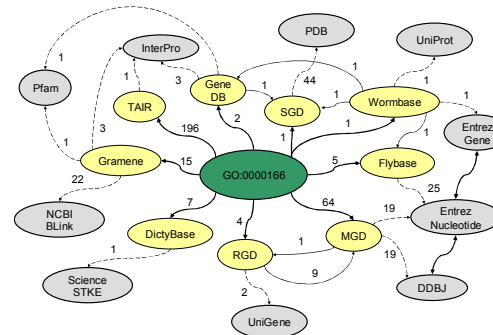


Figure 1. Partial inter-DB annotation network for GO:000166

CONCLUSIONS / FUTURE RESEARCH

As a small pilot study, these results are informative but not complete or generalizable. Further research, described below, is needed to better characterize the amounts, types, and relationships of annotations in MODs. This work provides some potential hypotheses that could be evaluated in future projects:

- Linking, full-text access, and literature classification are inconsistent in quantities and types across MODs.
- Quantities of literature citations vary with the level of characterization of the underlying gene product.
- Homologous and orthologous genes in multiple MODs are not necessarily reciprocally linked.

An interesting example of the latter statement is the linkage between the entries of RGD and MGD. RGD's Diat entry links to MGD's Diat entry as a putative ortholog. MGD in turn links its Diat back to RGD's. But MGD's Diat is not annotated to GO as having a 'nucleic binding' function as RGD's is. Instead, MGD associates that GO term with a different gene, MGI:1914137, which is orthologous to RGD's entry 1305466.

Likewise, the entry for S. pombe, po4, is linked to a homologous version in SGD, which is not annotated to 'nucleic binding' and does not have a reciprocal link to S. pombe. Similarly, the Wormbase entry links to different entries in Flybase, SGD, and S. pombe that those that are annotated to 'nucleic binding'. This project does not try to assess the reasons for these types of relationships, but the phenomenon in general would benefit from further research. We can imagine that technical and/or biological reasons could explain these types of linkage relationships.

Future work could investigate methods for large-scale, automated characterization and representation of inter-DB relationships for more than one GO term, including ways to discover and visualize clusters of annotations shared by multiple MODs. Other facets could include exploring the relative position of annotations in the GO ontologies, and quantifying distance between annotations by the number of nodes and branches in which two or more annotations are separated.

Acknowledgments

- This work benefited from discussions with curators from multiple model organism databases at the 2005 Gene Ontology Annotation Camp held at Stanford University in June. The Stanford Department of Genetics and the Saccharomyces Genome Database (SGD) provided travel support.
- W.J.M. was supported in part by an unrestricted research gift from Microsoft Research to the Annotation of Structured Data research group in the School of Information and Library Science at the University of North Carolina at Chapel Hill. The project's website is available at: <http://ils.unc.edu/annotat.ion>

Poster and abstract available at: <http://ils.unc.edu/~macm/wasist>