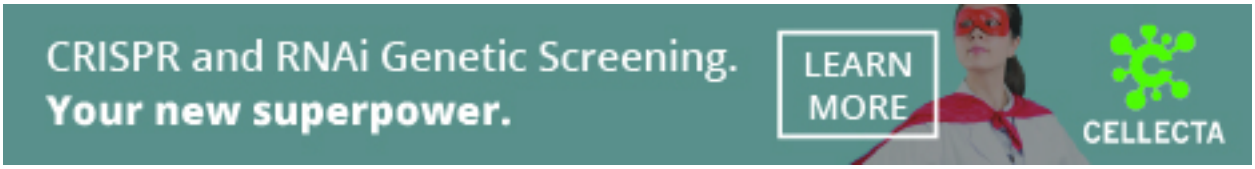
Assessment of Genome-Wide Protein Function Classification for *Drosophila melanogaster*

Huaiyu Mi, Jody Vandergriff, Michael Campbell, et al.

Genome Res. 2003 13: 2118-2128
Access the most recent version at doi:10.1101/gr.771603



Letter

Assessment of Genome-Wide Protein Function Classification for *Drosophila melanogaster*

Huaiyu Mi,[1] Jody Vandergriff,[1] Michael Campbell,[1] Apurva Narechania,[1] William Majoros,[1] Suzanna Lewis,[2] Paul D. Thomas,[1,5] and Michael Ashburner[3,4,5]

[1]*Protein Informatics, Celera Genomics, Foster City, California 94404, USA;* [2]*Molecular and Cell Biology, University of California Berkeley, California 94720-3200, USA;* [3]*Department of Genetics, University of Cambridge, Cambridge CB2 3EH, England;* [4]*European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, England*

The functional classification of genes on a genome-wide scale is now in its infancy, and we make a first attempt to assess existing methods and identify sources of error. To this end, we compared two independent efforts for associating proteins with functions, one implemented by FlyBase and the other by PANTHER at Celera Genomics. Both methods make inferences based on sequence similarity and the available experimental evidence. However, they differ considerably in methodology and process. Overall, assuming that the systematic error across the two methods is relatively small, we find the protein-to-function association error rate of both the FlyBase and PANTHER methods to be <2%. The primary source of error for both methods appears to be simple human error. Although homology-based inference can certainly cause errors in annotation, our analysis indicates that the frequency of such errors is relatively small compared with the number of correct inferences. Moreover, these homology errors can be minimized by careful tree-based inference, such as that implemented in PANTHER. Often, functional associations are made by one method and not the other, indicating that one of the greatest challenges lies in improving the completeness of available ontology associations.

[The full set of associations analyzed for this paper is available as Supplemental Material on the GO (http://www.geneontology.org/) and Celera (http://panther.celera.com/publications/gr7716_03_suppl) Web sites.]

Recent advances in whole-genome sequencing and assembly have begun to revolutionize our understanding of biology. Complete or draft eukaryotic genome sequences are known for *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Schizosaccharomyces pombe* (Wood et al. 2002), *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), *Drosophila melanogaster* (Adams et al. 2000), *Arabidopsis thaliana* (*Arabidopsis* Genome Initiative 2000), human (Lander et al. 2001; Venter et al. 2001), *Caenorhabditis briggsae* (http://www.ensembl.org/Caenorhabditis_briggsae/), *Fugu rubripes* (Aparicio et al. 2002), mouse (Mouse Genome Sequencing Consortium 2002; Mural et al. 2002; http://www.ncbi.nih.gov/genome/guide/mouse/), and rat (http://www.hgsc.bcm.tmc.edu/projects/rat/), containing a total of >200,000 known or predicted genes. Accessing and interpreting these large, complex data sets require computational methods, which in turn require the data to be well-structured. Ontologies have been used for some time in computer science as structured representations of objects and their relationships.

Because of their small genome size and relatively simple biology, microbes were the first free-living organisms to be sequenced (Fleischmann et al. 1995; Goffeau et al. 1996), and the first to have their functions represented in ontologies (Karp and Riley 1996; Mewes et al. 1997). These ontologies have proved to be exceptionally useful in several fields, including comparative biology and understanding gene expression changes during cellular processes or adaptation to external conditions. In order to apply these techniques to more complex metazoa such as *Drosophila* or humans, more extensive ontologies are required. The Gene Ontology (GO; Ashburner et al. 2000) is emerging as a standard across eukaryotic biology, with all of the major eukaryotic model organism databases as consortium members. GO defines a controlled vocabulary for classifying the molecular functions, biological processes, and cellular components of gene products (mostly proteins). Although the vocabularies are controlled, through the use of unique and persistent identifiers, the process of defining and refining the ontology is dynamic. The improvements to the vocabulary are driven by the use of the terms in practical applications. The primary application is associating terms with gene products as comprehensively as possible, given the present state of our knowledge.

The functional classification of genes on a genome-wide scale is now in its infancy, and experimental characterization of molecular function and cellular or physiological role has not yet been scaled to match the progress of genomic and transcript sequencing. There is a need to understand which methods of classification work well and which do not. Furthermore, it would be instructive to understand the shortcomings of various methods to classify genes by the function of their products. At present, most classification is based on curating information from scientific papers, or by inference based on sequence similarity to known gene products.

In this paper, we compare two methods for associating proteins with functions, one implemented by FlyBase (The FlyBase Consortium 2002; http://www.flybase.org) and the other in PANTHER (Thomas et al. 2003a,b; http://panther.celera.com). The two methods are alike in very general terms: They both make inferences based on sequence similarity and the available experimental evidence. However, they were performed completely independently and differ considerably in methodology and process. We believe that these differences are large enough to allow

[5]Corresponding authors.
E-MAIL paul.thomas@fc.celera.com; FAX (650) 554-2344.
E-MAIL ma11@gen.cam.ac.uk; FAX +44 1223 333992.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.771603.

13:2118–2128 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

us to analyze the types of errors that appear in large-scale efforts to classify the functions of gene products. To this end, we exhaustively compared, for all the presently predicted protein products of *Drosophila melanogaster*, >11,000 separate associations with molecular functions and biological process terms of the Gene Ontology (http://www.geneontology.org).

RESULTS

Two different methods (which we will refer to as "PANTHER" and "FlyBase") were applied independently for associating *Drosophila* proteins with GO terms describing molecular functions and biological processes. The process for comparing these two methods is outlined in Figure 1. Details are provided in Methods, but the major components are: (1) mapping the PANTHER/X ontology to GO; (2) for all *Drosophila* proteins, translating PANTHER/X associations to GO associations using this mapping; (3) using the GO ontology structure to automatically classify PANTHER–GO and FlyBase–GO associations as either "matched" or "unmatched"; and (4) manually reviewing the unmatched associations to determine which associations were correct and which were not.

Coverage of Classifications

The coverage of each method, in terms of the number of proteins that were associated with GO terms, is represented in Figure 2. As of December 2001, FlyBase associated 6301 (44%) of the *Drosophila* proteins with at least one GO molecular function term (Fig. 2A), and 2794 (19%) of the proteins with GO biological process terms (Fig. 2D). For PANTHER, we report two sets of numbers. PANTHER contains two separate ontologies. The one (PANTHER/LIB), describing protein sequence family/subfamily relationships, is structured as a hierarchy of HMM models. The other (PANTHER/X) is a curated set of terms describing molecular function and biological process that was built by manually reviewing the sequences used in constructing the PANTHER/LIB, often using GO as a reference. PANTHER/LIB families and subfamilies are mapped to PANTHER/X terms, which have been mapped into GO (see Methods).

Although 8127 (57%) *Drosophila* proteins have a significant hit to a PANTHER/LIB HMM, not all of these models have been associated with a PANTHER/X term, and not all PANTHER/X terms have been associated with GO terms. This means that although many proteins could be associated with a PANTHER/LIB term, most but not all proteins could subsequently be mapped to GO. The reasons for this incomplete mapping are: (1) the manual curation process associating PANTHER/LIB families and subfamilies to functional terms in the ontologies is still in progress; (2) many proteins can be assigned to a family of uncharacterized or ambiguous function; and (3) some of the proteins are associated with PANTHER/X terms that have no clear mapping into GO. A total of 4862 proteins (34%) have at least one meaningful GO molecular function association (Fig. 2B), whereas 3658 (26%) have at least one GO biological process association (Fig. 2E).

Figure 2, C and F, illustrates the overlap between the sets of *Drosophila* proteins classified by PANTHER and FlyBase. For molecular function, there are a total of 3283 proteins (23%) that are associated with GO terms by both methods, whereas an additional 4597 proteins (32%) are associated with GO terms by only one of the methods and not the other. For biological process, the overlap is much smaller: 1156 proteins (8%) have GO terms from both methods, whereas 5293 (37%) have GO terms from at least one method.

The fact that the methods do not overlap to a great degree may be initially surprising. By and large this reflects the different strategies used to make the associations, as well as the nascent

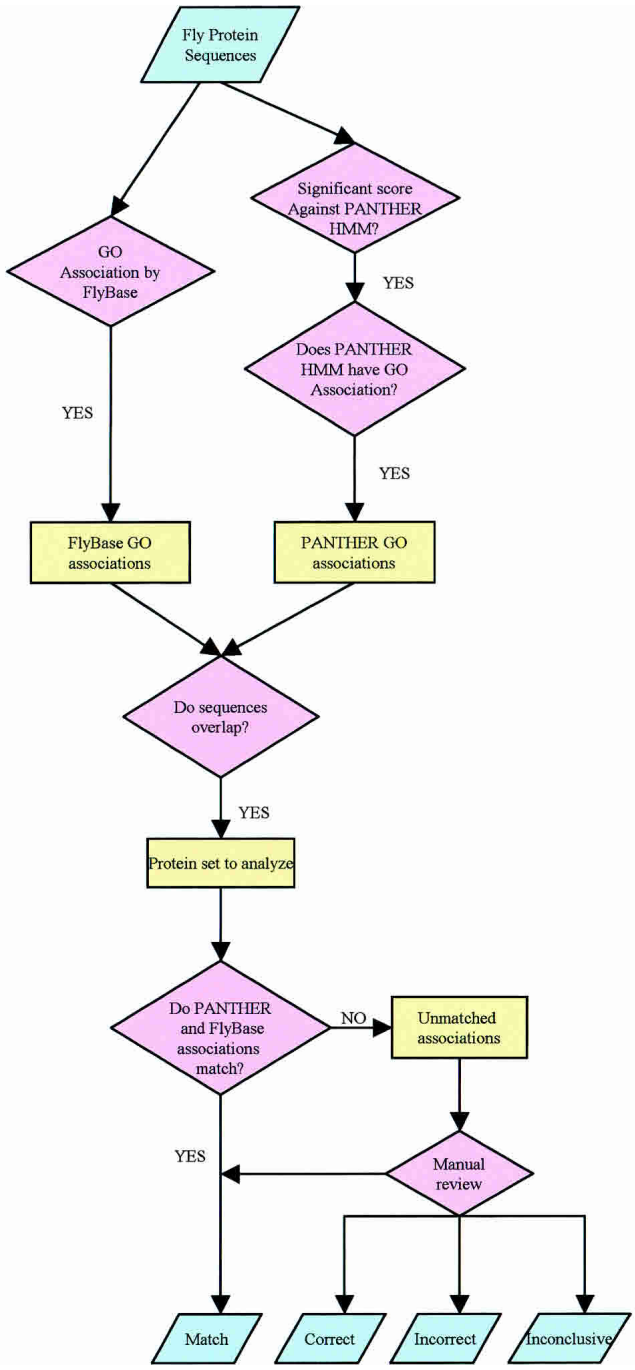


Figure 1 Process for comparing FlyBase–GO and PANTHER–GO associations for *Drosophila* proteins.

state of genome-scale functional classification. FlyBase associations are largely based on curation from the literature and curator-reviewed sequence similarities; associations based solely on the former would not be expected to be represented to the same degree in the PANTHER set. The relatively low overlap means that the coverage of the combined classifications is much greater than that of either method alone. This indicates that, as long as different methods do not have widely different error rates, independent efforts can be usefully combined using a common standard such as GO. It is also significant that the overlap between

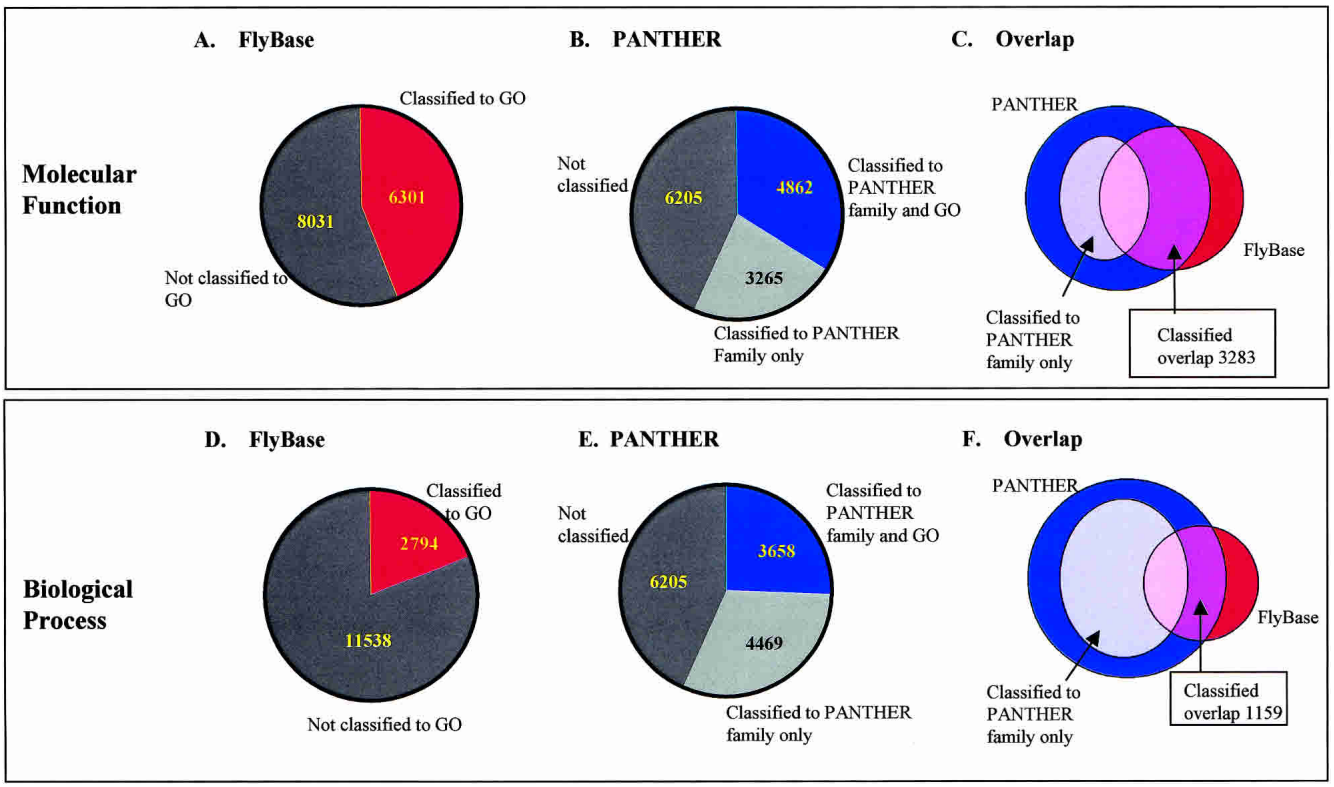


Figure 2 Coverage of *Drosophila* proteins classified by FlyBase and PANTHER. (*A–C*) Classification coverage for molecular function categories. (*A*) FlyBase associated 6301 proteins (red area) with at least one GO molecular function. (*B*) PANTHER associated 4862 proteins (blue area) with a GO molecular function. The light gray area indicates proteins that hit a PANTHER HMM, but were not associated with a GO term (see text for details), whereas the dark gray area indicates proteins that did not hit any PANTHER HMMs. (*C*) Venn diagram illustrating the overlap between proteins classified by FlyBase and PANTHER. (*D–F*) Classification coverage for biological process categories. (*D*) number of proteins classified by FlyBase, (*E*) number classified by PANTHER, and (*F*) the overlap between sets of proteins classified by the two methods.

the two methods is much smaller for biological process classifications than for molecular function. This reflects the relative difficulty in making these associations. Biological process has a broad definition, from cellular pathways to physiological processes; many proteins are involved in multiple processes. Biological process is also often less straightforward to infer from sequence similarity than is molecular function. Additionally, the roles of many proteins in particular biological processes are yet to be elucidated, despite characterization of their molecular functions.

Association Concordance and Reasons for Disagreement

To assess the accuracy of functional associations, we analyzed in detail the set of *Drosophila* proteins that were associated with GO terms by both PANTHER and FlyBase (Fig. 2C,F). This is the most informative set for analysis, as we have two independently proposed hypotheses to compare. We also verified that, because this set is a representative sampling of all of the functional associations made by the two methods, we do not expect that analysis of this set will lead to any bias in our conclusions. PANTHER makes a total of 3867 unique associations between GO molecular function terms and the 3283 proteins (an average of 1.18 associations per protein), whereas FlyBase makes a total of 3700 unique associations for the same 3283 proteins (an average of 1.13). We therefore assessed all of the 3867 + 3700 = 7567 molecular function associations, as summarized in Figure 1 and described in Methods. First, we used the ontology structure to establish automatically whether a PANTHER association is equivalent to a FlyBase association. If both PANTHER and FlyBase associate a given protein with the same GO term, or they associate the protein with terms that have a direct relationship in the ontology, the associations are considered to "match." This allows us to separate the associations into two types: "matched" and "unmatched." For molecular function, most of the associations match, but for biological process, most of the associations do not match. As with the overlap of assigned proteins (Fig. 2), this highlights the incompleteness of the biological process associations. The matched associations were not analyzed further, as they indicate agreement between PANTHER and FlyBase, and thus represent either correct predictions or systematic errors.

The unmatched associations were reviewed manually, one at a time, by all of the authors together. During manual review of the unmatched biological process associations in PANTHER and FlyBase, we found that in nearly every case, associations from both methods were correct, but different. For example, PANTHER associates a muscarinic acetylcholine receptor with the processes *neuromuscular synaptic transmission* and *G-protein-mediated signaling*, whereas FlyBase associates it with *acetylcholine receptor signaling, muscarinic pathway*. As a result, we found molecular function to be much more informative for analysis of disagreement and error rates, and we focus our discussion accordingly. We manually reviewed all 2083 unmatched molecular function associations, painstakingly if necessary, classifying them as (1) matched (but not detected automatically), (2) correct, (3) incorrect, or (4) inconclusive, given present evidence (see Methods for more detail on these categories). To better understand the sources of error, we report the number of associations in each of these four categories for FlyBase and PANTHER separately (Fig. 3).

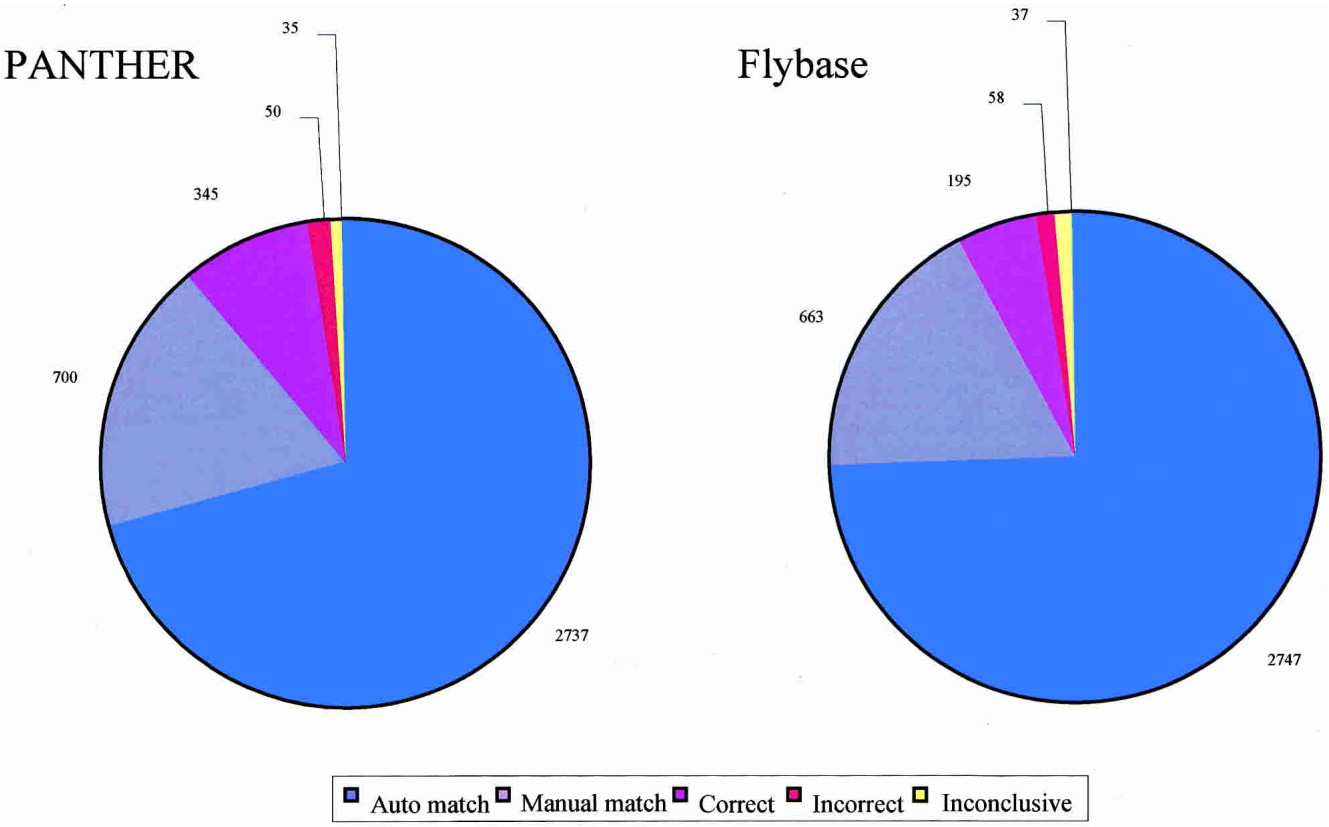


Figure 3 Assessment of molecular function associations of proteins that were classified by both FlyBase and PANTHER. This subset of proteins corresponds to the purple area in Figure 2C. The majority of the molecular function association matches between FlyBase and PANTHER were determined by an automated process (blue slices). The remaining unmatched associations were manually reviewed, and classified as either matched (gray blue), correct (purple), incorrect (red), or inconclusive (yellow).

Types of Errors

Each of the associations classified as "incorrect" was then analyzed further to determine its most likely cause. Table 1 summarizes the frequency of the various types of error identified during the manual review process. There are three main categories of error, which we call "human error," "homology error," and "evidence error." If we take the matched associations to be correct (i.e., assume relatively low rates of systematic error), we find the overall error rate to be <2%. Because the automatically matched associations can only differ in their specificity between the two methods, the only errors that may be missed would be those caused by annotations that are too specific, given the evidence. In the worst case, even if we disregard the automatic matches, the manually reviewed associations alone have an error rate of <5%, although we believe that this would be an overly pessimistic estimate.

Table 1. Analysis of Errors in the GO Associations of PANTHER and FlyBase

	PANTHER	FlyBase
Number of homology errors	8	35
Number of human errors	40	23
Number of evidence errors	2	0
Total number of incorrect associations	50	58
Association error rate (%)	1.3%	1.6%
Number of proteins with at least one incorrect association	49	58
Protein error rate (%)	1.5%	1.8%

Analysis of errors in the GO associations of PANTHER and FlyBase. The "total number of incorrect associations" is the total count of associations judged to be incorrect, regardless of whether the associations are assigned to the same protein or not. The "number of proteins with at least one incorrect association" is the total count of proteins with at least one association judged to be incorrect. Because one protein can have more than one association, this number can be different from "total number of incorrect associations." The "association error rate" is the number of incorrect associations divided by the total number of associations analyzed. The "protein error rate" is the number of proteins with at least one incorrect association divided by the total number of proteins analyzed.

Human Error

The most common error overall is an error on the part of the human curator. The curator's job is to evaluate the evidence, and assign the gene product accordingly. We identified two extremes of curator error: manual error (no evidence at all for the annotation) and judgment error (weak evidence favored incorrectly over strong competing evidence). However, in practice, it is hard to distinguish between these two cases without a detailed description of the curator's decision-making process. A typical example of simple manual error is Trip1 (FBgn0015834), which is a *translation factor* incorrectly labeled by PANTHER as a *transcription factor*—these categories are functionally very different, but similar to the human eye. Another example is FBgn0042083, a *carboxyltransferase* (a ligase) that was incorrectly classified in FlyBase as a *decarboxylase*, another case in which similarity in terms may have caused a manual error. An example of a clear judgment

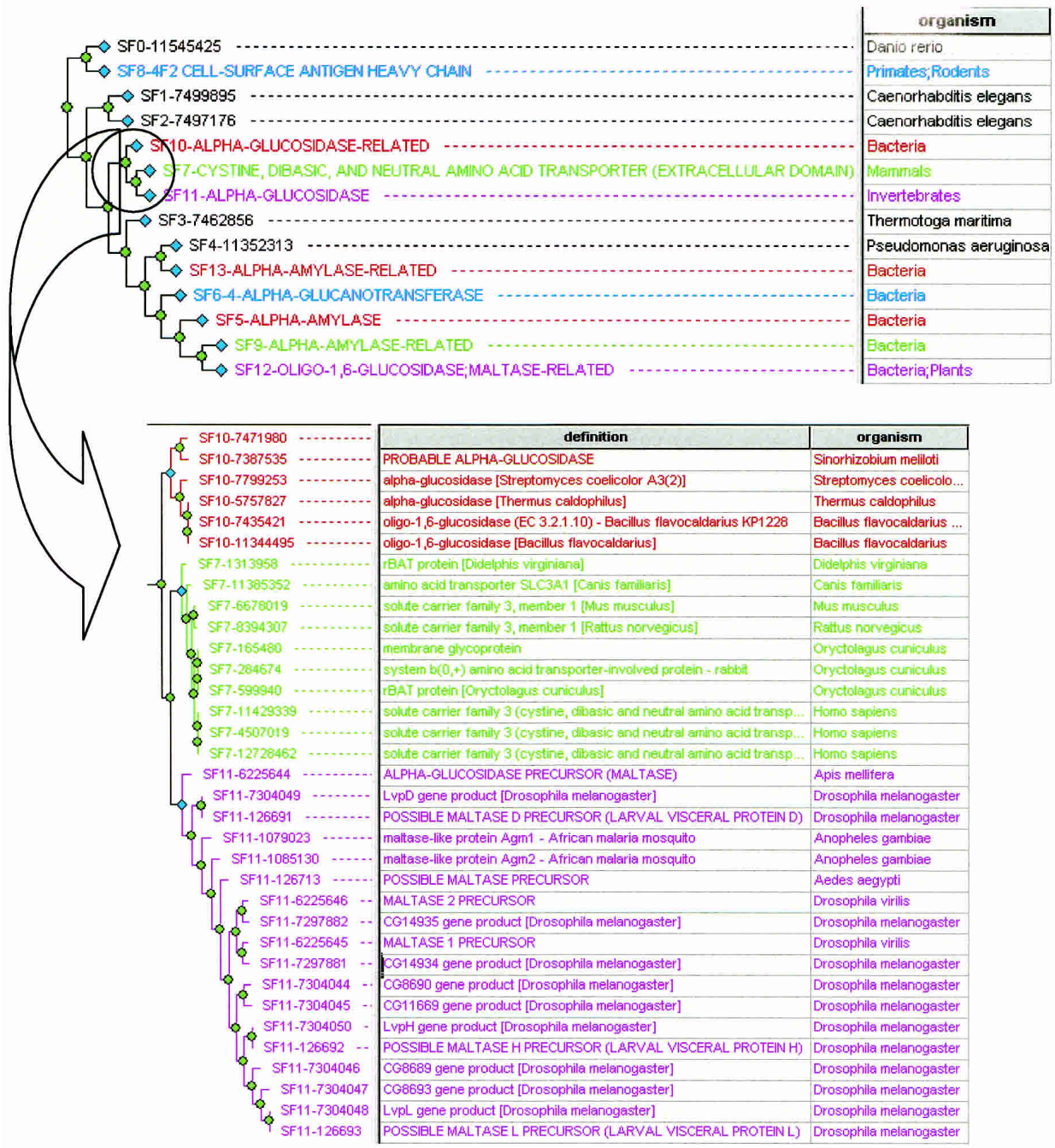


Figure 4 Function inference in the context of a protein sequence tree. This is the PANTHER tree-attribute view, with a sequence-derived tree in the *left* panel, and a table of sequence (or subfamily) attributes in the *right* panel. The *top* figure shows the tree "collapsed" into curator-defined subfamilies. Note that the transporter subfamily (SF7, in green) has been separated by the curator from neighboring groups of proteins that are α-glucosidase-related (SF10 in red, and SF11 in pink). The *bottom* figure shows the "expanded" view with information about each sequence taken from GenBank and SWISS-PROT.

error is FBgn0035207, a ubiquitin-protein ligase that shows strong sequence similarity to other ubiquitin-protein ligases, yet was classified by FlyBase as a *small GTPase regulator* on the basis of a much weaker similarity. Another example is cathepsin D (an aspartyl protease, FBgn0038506), labeled by PANTHER as a *cysteine protease*, presumably because most (but not all) cathepsins are cysteine proteases. We find that PANTHER associations have nearly twice the human error rate of FlyBase associations. This is primarily because of the consistency of expert curation: FlyBase associations were made by a few experts trained over a relatively long period of time and performing curation over a period of months to years, whereas the >60 PANTHER curators were temporary consultants who performed their work over a period of days.

Homology Error

Although sequence similarity generally implies shared function, there are exceptions to this rule. Most proteins in the *Drosophila*

Table 2. Comparison of PANTHER/X Abbreviated Ontology and Gene Ontology (GO)

	PANTHER/X	Gene ontology (GO)
Number of categories for molecular function	244	4752
Number of categories for biological process	249	4041
Maximal level of depth for molecular function	3	12
Maximal level of depth for biological process	3	12
Average level of depth for molecular function	2.16	5.33
Average level of depth for biological process	2.15	6.42

The "number of categories for molecular function" and "number of categories for biological process" are the number of distinct terms in each ontology. The "maximal level of depth" is the maximum number of levels separating the most general "top level" terms from the most specific "progeny" terms. The "average level of depth" is calculated by dividing the sum of each category's level by the total number of categories. If the same category appears multiple times in the ontology, all of its occurrences will be calculated. The "top level" molecular function (GO:0003574) and biological process (GO:0008150) terms are the level 0 categories and were not included in the calculation.

genome have not been characterized experimentally, and both of the functional association methods discussed in this paper make inferences based on sequence similarity. Therefore, we cannot assess here the absolute rate of incorrect predictions from homology. However, the two association methods were (1) performed independently of each other, and (2) differ significantly in how sequence similarity information is used to make an inference. To some degree, each method makes different types of "homology errors." Overall, we find that FlyBase associations have more than four times the homology error rate of PANTHER associations.

In the FlyBase method, a curator assigns gene products, one at a time, to functional terms based on BLASTP-inferred homology and known experimental evidence. The most common error in FlyBase associations occurs when an inference is made based on a BLASTP similarity that has high statistical significance, and yet leads to an incorrect functional prediction. We call this a "context error," because careful analysis of sequence interrelationships can resolve whether a prediction is correct or not. PANTHER is less prone to this type of error because a curator associates functions with proteins in the context of a tree-based representation of all the interrelationships between a set of related proteins (a protein family). The family is divided into subtrees (subfamilies) by an expert biologist precisely on the basis of shared function across all subfamily members. Association of molecular function and biological process at the level of the subfamily, rather than the family, minimizes incorrect inferences from sequence similarity. Figure 4 shows an example of how these kinds of errors are avoided by PANTHER. In this example, FBgn0032382 is predicted by FlyBase to be both an α-*glucosidase* and a *neutral amino acid transporter*, but by PANTHER to be an α-*glucosidase*. The BLASTP alignment to the SWISS-PROT (Bairoch and Apweiler 2000) entry Q07837, correctly annotated as a *neutral amino acid transporter*, is highly significant: The *E*-value is 5e-98 and the proteins are 36% identical over 558 amino acids. The α-glucosidase is clearly related to the transporter, but does not share the same function. Inspection of the PANTHER tree representation for this family reveals that the transporter function is confined to a single subfamily that does not contain FBgn0032382 (CG14934). Consistent with this result, the neutral amino acid transporter sequences contain a C-terminal transmembrane segment that is not found in the *Drosophila* protein. The PANTHER subfamily-based analysis makes errors of this kind less common, but does not eradicate them altogether. Particularly challenging are divergent families in which experimental evidence of biochemical function exists for few members. An example of this is the *Drosophila* protein encoded by FBgn0029945. PANTHER/LIB associates it with the term *4-coumarate-CoA ligase*, and FlyBase associates the same protein with *long-chain fatty acid transporter*. In both cases, the function prediction is more specific than warranted by the data (i.e., there is no evidence for specifying coumarate or long-chain fatty acids as a substrate).

PANTHER occasionally makes a different type of homology error: a "low score error." Four of the eight homology errors made by PANTHER were of this type. Sequences are annotated by scoring proteins against a library of ~40,000 Hidden Markov Models (HMMs) that represent different families and subfamilies of proteins. If the best score to any given HMM is not significant enough, there is a chance that the associated functional prediction will be incorrect. Lower scores indicate more remote homology, which in turn implies greater chance of functional divergence. This is a type of a homology error, in that the computational similarity score-based association results in an incorrect prediction. An example of this is the PANTHER association of FBgn0032522 with the function *oxidoreductase*. Although this protein is certainly related to phosphoadenosine phosphosulfate reductase (the closest hit in the PANTHER version 3.0 library), its function is not. The PANTHER/LIB HMM score of −53 indicates evolutionary relatedness but is above the trusted cutoff for functional assignment of −85. A few of the "inconclusive" PANTHER associations derive from apparent homology to dynein and other coiled-coil proteins. These proteins are well-known to present challenges for sequence similarity searching, and more stringent trusted cutoffs should be applied to these cases. PANTHER version 5.0 will support family-specific trusted cutoff values.

Both PANTHER and FlyBase occasionally make "protein domain errors": finding a similarity to only one domain of a multidomain protein, but incorrectly inferring a function that depends either on a different domain or a combination of domains. We find these errors to be rare—only four in FlyBase and three in PANTHER, or ~0.1%. For example, FlyBase classifies

Table 3. Mapping the PANTHER/X Ontology to GO

	Map to equivalent GO categories	Map to more general GO categories	No match	Total
Molecular function				
Overall	160 (65.5%)	64 (26.2%)	20 (8.2%)	244
Level 1 categories	20 (66.7%)	4 (13.3%)	6 (20.0%)	30
Level 2 categories	91 (60.3%)	49 (32.4%)	11 (7.3%)	151
Level 3 categories	49 (77.7%)	11 (17.5%)	3 (4.8%)	63
Biological process				
Overall	167 (67.0%)	66 (26.5%)	16 (6.4%)	249
Level 1 categories	23 (63.9%)	4 (11.1%)	9 (25.0%)	36
Level 2 categories	95 (67.3%)	39 (27.7%)	7 (5.0%)	141
Level 3 categories	49 (68%)	23 (32%)	0.00%	72

For both molecular function and biological process, approximately two-thirds of the PANTHER/X terms can be mapped to an exactly equivalent term in GO, and approximately an additional one-quarter of PANTHER/X terms can be mapped to a GO term that is more general than the PANTHER/X term.

Table 4. Depth of the Mapped PANTHER/X Terms in GO

	GO depth with equivalent PANTHER/X-GO mapping		GO depth with a more general GO category mapped to PANTHER/X	
PANTHER/X level	Molecular function	Biological process	Molecular function	Biological process
Level 1	1.90 ± 0.78	2.78 ± 0.85	1.25 ± 0.50	2.5 ± 1.00
Level 2	3.61 ± 1.21	4.22 ± 1.30	2.18 ± 1.12	3.29 ± 1.04
Level 3	4.45 ± 1.11	5.04 ± 1.09	3.36 ± 1.12	3.52 ± 2.27
Overall	3.68 ± 1.36	4.27 ± 1.37	2.32 ± 1.20	3.32 ± 1.58

For each level of PANTHER/X, we calculated the average (standard deviation) depth, in the Gene Ontology, of the GO term to which a PANTHER/X term was mapped. We calculated statistics separately for PANTHER/X terms that could be mapped to an equivalent GO term, versus PANTHER/X terms that could only be mapped to a more general GO term.

FBgn0038379 as an *atrial natriuretic peptide clearance receptor*, whereas the most specific PANTHER/LIB subfamily term is *guanylate cyclase*. This particular fly protein is certainly homologous to ANPCRs (they both contain a guanylate cyclase-like domain), but it does not contain the other domains present in ANPCRs. PANTHER classifies FBgn0024920 (thymidylate synthase, TS) as both an *oxidoreductase* and a *synthase*. This was because of an error in subfamily curation—vertebrate and invertebrate TSs are closely related to the TS domain of plant bifunctional thymidylate synthase and dihydrofolate reductase (DHFR) proteins, and the curator failed to divide these into distinct groups. The *Drosophila* TS was incorrectly inferred to be a bifunctional enzyme, but it only contains the TS domain and not the DHFR domain.

Evidence Error ("Transitivity Error")

This type of error occurs when the inference process is applied correctly, but the underlying evidence is incorrect. It is often referred to in the literature as a "transitivity error": An incorrect annotation is propagated from one sequence to another via sequence similarity. For example, PANTHER/LIB annotates FBgn0010602 as a *ubiqutin conjugating enzyme*, when it is in fact a *SUMO conjugating enzyme*. These proteins are closely related, and in fact differ only in operating on a distinct but closely related substrate (SUMO vs. ubiquitin). The PANTHER annotation was based in part on a SWISS-PROT annotation of a related sequence, O09181, which was incorrect at the time (it was corrected in SWISS-PROT prior to our analysis). As for the process errors, we do not expect to find all instances of evidence errors in this study, because many of the sources used for annotations were similar for the FlyBase and PANTHER classification processes.

DISCUSSION

Overall, assuming that the systematic error across the two methods is relatively small, the protein-to-function association error rate of both the FlyBase and PANTHER methods is <2%. The primary source of error for both methods appears to be simple human error. This is encouraging, as these errors can be corrected over time, especially as the associations are used and reviewed by a wider community. Although homology-based inference can certainly cause errors in annotation, our analysis indicates that the frequency of such errors is relatively small compared with the number of correct (or at least not obviously incorrect) inferences. In the absence of experimental results, these inferences are extremely valuable. Moreover, these homology errors can be minimized by careful tree-based inference, such as that implemented in PANTHER. Although both PANTHER and FlyBase have comparable overall error rates, the increased rate of homology error in FlyBase is balanced by an increased rate of human error in PANTHER. This is primarily caused by differences in the curation processes. In PANTHER, homology inferences are made simultaneously for large numbers of sequences in the context of a tree, by a relatively large number of curators over a short period of time. FlyBase curators are few and retained over a much longer period of time, and make homology inferences for each protein more or less individually.

We also find that transitive errors—propagation of an incorrect annotation of a related sequence—appear to be quite rare. This fact is not as surprising as it seems at first blush: Both PANTHER and FlyBase make use of sources (primarily SWISS-PROT, RefSeq, and OMIM) that have already used manual curation to minimize well-known problems like transitivity error. Another source of error that has been discussed in great detail in the literature is what we term a "domain error"—incorrectly inferring function on the basis of similarity over only a single domain. Although this type of error provides some excellent anecdotes, statistically we find it to be quite rare, occurring at a rate of <0.01% and accounting for ~6% of the total errors. Although we have not attempted to systematically track down every instance of domain error, our findings indicate that domain-based errors in protein function assignment may be more rare than is often assumed. Those that exist will, like the other classes of error, be corrected over time.

One of the most significant findings is the low degree to which the PANTHER and FlyBase associations overlap. In other words, incompleteness (false-negative associations with an ontology term) is a much larger problem than error rate (false-positive associations). Development of function ontologies for biology, and association of these terms with gene products, is in its infancy. Both PANTHER and FlyBase associations are far from complete in even capturing what is known at present or can be inferred, given the present state of biological knowledge. Combining the PANTHER and FlyBase associations into a single resource is an obvious way to increase the completeness of coverage for the *Drosophila* proteome. All of the erroneous associations uncovered by this study have been corrected. Associations made by one method and not the other are now being used to expand the number of curated associations in both FlyBase and PANTHER. Lastly, both sets of associations are available in the AMIGO browser (http://www.godatabase.org).

The full set of associations analyzed for this paper is available as Supplemental Material on the GO and PANTHER Web sites.

METHODS

Our methodology is outlined in Figure 1. Both PANTHER and FlyBase started with a common set of 14,332 predicted proteins from the *Drosophila* genome (Adams et al. 2000): Release 2 from FlyBase (available at http:// www.fruitfly.org/sequence/sequence_db/aa_gadfly.dros.RELEASE2). For the standard ontology, we used GO molecular function Version 2.107 (October 1, 2001) and biological process Version 2.99 (October 2, 2001). To facilitate comparison of associations between proteins and ontology terms, the terms in the two ontologies (GO and

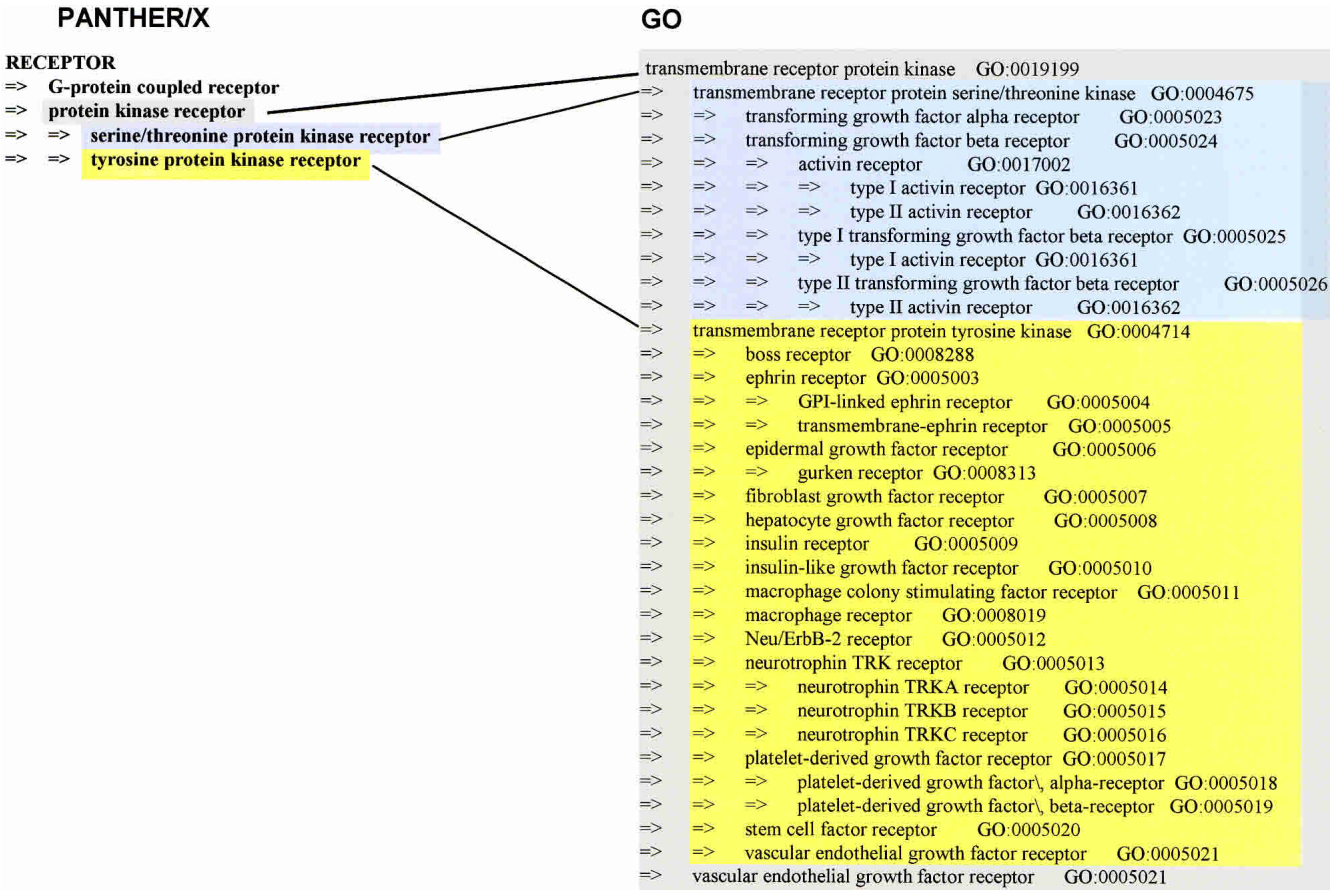


Figure 5 Method for automated comparison of PANTHER and FlyBase assignments. The PANTHER/X ontology was designed as a more lightweight version of GO, and therefore the PANTHER–GO associations will not generally have the same degree of specificity as FlyBase–GO associations. To compare the FlyBase and PANTHER associations directly, it was often necessary to trace up the GO classification to match a given FlyBase association to a PANTHER association. For example, PANTHER may associate a protein with the term *tyrosine kinase receptor* (yellow), which corresponds to the GO category *transmembrane receptor protein tyrosine kinase* (GO:0004714; yellow area designates the term and its children). If this same protein is associated by FlyBase with one or more of the GO categories in yellow, we consider PANTHER and FlyBase assignments as a "match." However, if the protein is associated by FlyBase with *transmembrane receptor protein serine/threonine kinase* (blue area, a sibling but not a child), we consider the PANTHER and FlyBase associations as "unmatched."

PANTHER/X) must be related to each other. For the purposes of this study, we first converted PANTHER associations into GO associations, and then compared the PANTHER–GO and FlyBase–GO associations directly.

GO Associations From FlyBase

FlyBase is a curated database for *Drosophila*. In present versions of FlyBase, Gene Ontology terms are attached as attributes to genes, as surrogates for gene products. There are three major sources of information used to make these attributions. The first is literature curation, information from the primary or review literature being captured by a curator; the second is curation of sequence data, both the primary nucleic acid sequence records from EMBL-Bank/GenBank/DDBJ and the curated protein records of SWISS-PROT; the third is data that originated from the automatic assignment of GO terms to genes by LOVEATFIRSTSIGHT, used during the annotation of Release 1 of the *D. melanogaster* genome sequence (Adams et al. 2000). The automatic assignment of GO terms by LOVEATFIRSTSIGHT resulted in ~20,000 annotations to 50 high-level terms. All of these were reviewed by FlyBase curators, and some 7700 annotations were retained by FlyBase, although many of these were refined to a finer level than originally annotated. The primary basis for accepting, or refining, a LOVEATFIRSTSIGHT annotation was a review of the sequence similarities between computed *Drosophila* genes and SWISS-PROT sequences from other taxa.

GO Associations From PANTHER

Associating PANTHER/X Terms With Proteins

Drosophila proteins (Release 2) were scored against the PANTHER/LIB HMMs (Version 3.0, December 2001). All proteins having an (NLL-NULL) score (Krogh et al. 1994) more significant than −35 were considered to be "hit" by a PANTHER/LIB HMM. This score generally indicates reliable evolutionary relationships, although −85 is the trusted cutoff for high-confidence function prediction. Each *Drosophila* protein was associated with the highest-scoring PANTHER/LIB HMM (either family or subfamily). A total of 8127 (57% of 14,332) *Drosophila* proteins were associated with a PANTHER/LIB HMM. PANTHER/LIB HMMs are, in turn, associated with PANTHER/X molecular function and biological process terms, so each *Drosophila* sequence was associated with functions via the highest-scoring PANTHER/LIB HMM. Of the 8127 proteins that hit a PANTHER/LIB HMM, 4862 (60% of the total) were associated with at least one molecular function term, and 3658 (45%) with at least one biological process term (Fig. 2). For PANTHER Version 3.0, because the curation effort was centered around families containing mammalian proteins, a smaller fraction of *Drosophila* proteins is associated with family and function terms than with human or mouse proteins, for example.

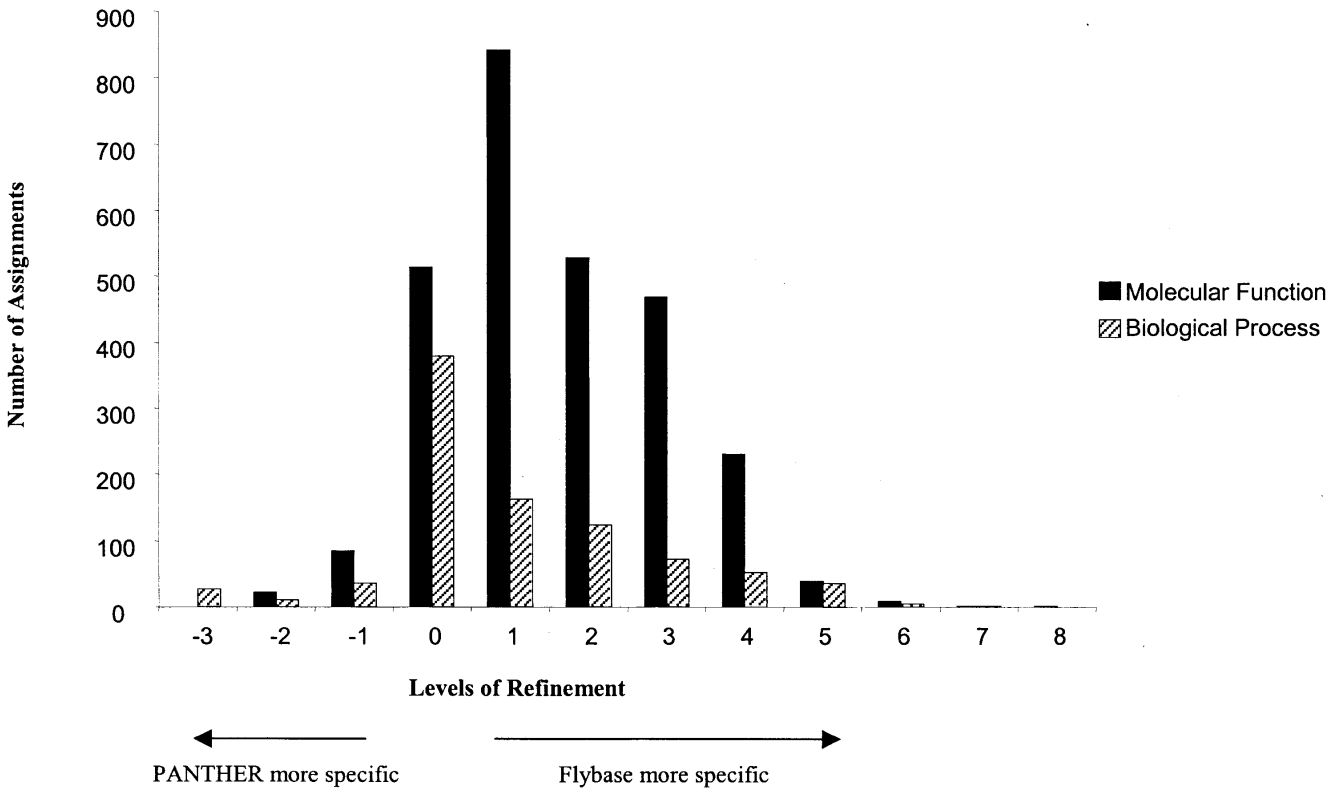


Figure 6 Specificity comparison between matched FlyBase–GO and PANTHER–GO associations. "Levels of refinement" is the number of levels in the GO schema that separate the FlyBase–GO association from the matched PANTHER–GO association. Positive numbers indicate that the FlyBase association was more specific than the matched PANTHER association, whereas negative numbers indicate that the FlyBase association was less specific. In general, the PANTHER–GO associations are less specific because only the PANTHER/X abbreviated ontology was mapped to GO, and not the more specific PANTHER/LIB.

Mapping the PANTHER/X Ontology to GO

It is important to note that PANTHER consists of two separate ontologies: one describing protein function (PANTHER/X) and another describing protein sequence relationships (PANTHER/LIB). These ontologies have been mapped to each other by curators having expertise in a relevant field of biology. The more detailed PANTHER/LIB contains >11,000 curated terms in two levels, family and subfamily. Because many, but not all, of the PANTHER/LIB terms describe protein function, not all can be mapped to function and process ontologies. The PANTHER/X schema contains roughly 500 terms, and was designed as a concise function ontology to allow high-level navigation and analysis of large sets of proteins. For the purposes of this study, we mapped PANTHER/X but not PANTHER/LIB terms to GO terms. Table 2 compares the properties of the PANTHER/X schema and GO: GO contains nearly 20 times as many categories and therefore represents protein function in greater detail. This means that in our comparison, the PANTHER–GO associations are generally less specific than the FlyBase–GO associations, primarily because we did not map the more specific PANTHER/LIB terms as well.

Each term in PANTHER/X was mapped individually to the GO term that was the closest match. Most PANTHER/X terms could be assigned to an exactly equivalent GO term (Table 3). Of the remaining terms, some could be mapped to more general "parent" terms in GO; that is, the PANTHER/X term is more specific. For example, there is no GO term equivalent to PANTHER/X's *synthetase*, and therefore this term was mapped to the more general GO term *enzyme*. There were several PANTHER/X terms that could not be mapped to GO. To decrease the number of unmapped terms, as well as to improve GO, 30 additional terms were added to GO (cf. molecular function Version 2.89 and biological process Version 2.75 at http://www.geneontology.org). The final mapping left only 8.2% of PANTHER/X molecular function terms, and 6.4% of biological process terms, unmapped to GO. The unmapped PANTHER/X terms are mostly higher-level categories that facilitate navigation, such as *select regulatory molecule*, or more pragmatic "functional" categories, such as *extracellular matrix protein* and *cytoskeletal protein*, that are arguably as much a location as a function. Table 4 summarizes the depth, in GO, of the mapped PANTHER/X terms. In general, the deeper the level in PANTHER/X, the deeper in GO, with the PANTHER/X terms mapped to relatively general GO terms. Occasionally, a PANTHER/X term was mapped to more than one GO term by virtue of the relationships in the PANTHER/X ontology. These cases represent parent–child relationships that exist in PANTHER/X but are missing from this version of GO. For example, in PANTHER/X, *ligand-gated ion channel* is a child of both *ion channel* and *receptor*, whereas in GO, *extracellular ligand-gated ion channel* is a child of *ion channel* but not *receptor*. Therefore, we mapped PANTHER/X *ligand-gated ion channel* to two unrelated GO terms: *extracellular ligand-gated ion channel* (equivalent) and *receptor* (parent). The mapping of PANTHER/X to GO is available on the PANTHER and GO Web sites.

Associating GO Terms With Drosophila *Proteins*

PANTHER/X associations for each protein were converted to GO associations using the mapping of PANTHER/X to GO. For each PANTHER/X association, the most specific term was translated into its corresponding GO term (or terms).

Assessing the Agreement Between PANTHER and FlyBase Associations

Automated Assessment

For a given method (FlyBase and PANTHER), each protein was either associated with one or more GO terms, or it was not (i.e., it was classified or unclassified). The overlap of the sets of proteins classified by each method is shown in Figure 2, C and F. For the purposes of comparison, we focus on the proteins that were classified by both methods. For GO molecular function, this represents 3283 proteins (23% of the total), but for GO biological process only 1159 proteins (8%).

For the proteins classified by both methods, we compared the functional associations. To do this, we first needed to define the set of unique functional assignments in each method. For example, FlyBase associates FBgn0031631 with the molecular functions *helicase* (GO:0004386) and *RNA helicase* (GO:0004004). However, *RNA helicase* is a child of *helicase*, that is, the association with *RNA helicase* already implies *helicase*. In these cases, we regard the less specific association (*helicase* in this example) as redundant, and we do not include it in our analysis. (Redundant annotation in FlyBase occurs as a result of literature-based assertions that differ.)

After defining the unique set of GO associations for FlyBase and PANTHER, we could automatically "match" a PANTHER association to a FlyBase association using the GO ontology structure. As discussed above, PANTHER/X was designed as a more lightweight version of GO, and therefore the mapped GO associations will not generally have the same degree of specificity as FlyBase–GO associations. To compare the FlyBase and PANTHER associations directly, it was often necessary to trace up the GO classification to match a given FlyBase association to a PANTHER association (Fig. 5). For example, PANTHER may associate a protein with the PANTHER/X term *tyrosine kinase receptor*, which corresponds to the GO category *transmembrane receptor protein tyrosine kinase*. If this same protein is associated by FlyBase with the GO category *ephrin receptor*, which is a child of *transmembrane receptor protein tyrosine kinase*, we consider PANTHER and FlyBase assignments as a "match." However, if the protein is associated by FlyBase with *transmembrane receptor protein serine/threonine kinase* (a sibling but not a child), we consider the PANTHER and FlyBase associations as "unmatched." For the automatic matching, only exactly equivalent terms in the two ontologies were considered. Associations for which PANTHER/X terms mapped to more general GO terms were subjected to manual review to ensure that there was no bias in our analysis caused by having mapped PANTHER/X to GO and not vice versa.

Figure 6 shows the distribution of differences in the specificity of PANTHER and FlyBase assignments. The PANTHER associations are most commonly one to two levels less specific than FlyBase. This is exactly what would be expected because the PANTHER/LIB ontology contains family and subfamily terms that represent one to two additional levels of specificity, but these terms have not yet been mapped to GO.

Manual Review

Every molecular function association not matched by the automated process was reviewed by all of the authors together. There were a total of 1130 PANTHER associations and 953 FlyBase associations, for 1462 different proteins, that were subjected to review. Upon review, we placed each association into one of four categories: (1) a "match" that was not detected automatically, (2) unmatched but "correct," (3) unmatched and "incorrect," or (4) unmatched and "inconclusive." Of the 2083 discrepancies, 1363 were actually matches. This was generally because of discrepancies in the parent–child relationships in the PANTHER/X and GO ontologies. For example, FBgn0000473 is a cytochrome P450. The FlyBase association is *cytochrome P450*, and the PANTHER association is *oxidoreductase*. Because *cytochrome P450* is not a child of *oxidoreductase* in GO (although cytochrome P450s are classified as oxidoreductases by the Enzyme Commission), the automated matching process could not match these two assignments. Upon manual review, we consider PANTHER and FlyBase assignments were actually a "match." Another example is the term *nuclease*. In PANTHER/X, *nuclease* is a child of *nucleic acid binding*, whereas in GO, it is a child of *hydrolase* (both are correct). The associations also do not match when they have sibling rather than parent–child relationships. For example, FBgn0001219 is associated with *HSP70/HSP90 organizing protein* (GO:0008077) by PANTHER, and *heat shock protein* (GO:0003773) by FlyBase. In GO, *HSP70/HSP90 organizing protein* and *heat shock protein* are siblings under the category of *chaperone*, and therefore did not match using the automated process. After manual review, we considered many of these cases to be matches as well.

If the associations were not considered to be a match, then each one was individually assessed using all available information, including the literature, as to whether it was correct, incorrect, or inconclusive. We found that 342 PANTHER associations and 195 FlyBase associations did not match (i.e., the two methods gave different associations), but they were all correct. This is not surprising because both PANTHER and FlyBase efforts are relatively new, and both are far from complete. This incompleteness is particularly evident for biological process classifications. Both the automated comparison and manual review confirmed that the PANTHER and FlyBase biological process associations overlapped very little, yet were largely both correct. We found that 33 PANTHER molecular function associations and 37 FlyBase associations were unresolved (inconclusive) at present, largely owing to the lack of evidence or knowledge. Several associations were judged to be simply wrong, and these were analyzed further to characterize the types of errors that were made.

ACKNOWLEDGMENTS

We thank Richard Mural and Randy Ribaudo for helpful comments on the manuscript; Thomas Hatton and Steven Ladunga for helping to review preliminary classification data; and Mark Yandell for helpful discussion. Work by M.A. was also supported by a Medical Research Council Project Grant.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. The Gene Ontology Consortium. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science* **269:** 496–512.

FlyBase Consortium. 2002. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **30:** 106–108.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 563–567.

Karp, P.D. and Riley, M. 1993. Representations of metabolic knowledge. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1:** 207–215.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235:** 1501–1531.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Mewes, H.W., Albermann, K., Heumann, K., Liebl, S., and Pfeiffer, F. 1997. MIPS: A database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.* **25:** 28–30.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J., et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661–1671.

Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. 2003a. PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31:** 334–341.

Thomas, P.D., Campbell, M.C., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. 2003b. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* (this issue).

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415:** 871–880.

WEB SITE REFERENCES

http://panther.celera.com; PANTHER Classification System.

http://www.ensembl.org/Caenorhabditis_briggsae/; Ensembl *C. briggsae* Genome Server.

http://www.flybase.org; FlyBase@flybase.bio.indiana.edu; FlyBase.

http://www.fruitfly.org/sequence/sequence_db/aa_gadfly.dros.RELEASE2; FlyBase Release 2.

http://www.geneontology.org/; Gene Ontology Consortium.

http://www.godatabase.org/; AMIGO, your friend in the Gene Ontology.

http://www.hgsc.bcm.tmc.edu/projects/rat/; Rat Genome Project.

http://www.ncbi.nih.gov/genome/guide/mouse/; Mouse Genome Resources.

Received September 4, 2002; accepted in revised form June 30, 2003.