ISB In Silico Biology An International Journal on Computational Molecular Biology

- Article -

In Silico Biology 7, 0050 (2007); ©2007, Bioinformation Systems e.V.

The use of multiple hierarchically independent Gene Ontology terms in gene function prediction and genome annotation

Yiannis A. I. Kourmpetis[1], Ate van der Burgt[2], Marco C. A. M. Bink[1], Cajo J. F. ter Braak[1]* and Roeland C. H. J. van Ham[2]

[1] Biometris, Wageningen University and Research Centre, 6700 AC Wageningen, The Netherlands
[2] Applied Bioinformatics, Plant Research International, 6708 PB Wageningen, The Netherlands

* Corresponding author

Email: cajo@terbraak.nl

Edited by H. Michael; received June 14, 2007; revised August 23, 2007; accepted August 27, 2007; published October 13, 2007

Abstract

The Gene Ontology (GO) is a widely used controlled vocabulary for the description of gene function. In this study we quantify the usage of multiple and hierarchically independent GO terms in the curated genome annotations of seven well-studied species. In most genomes, significant proportions (6 - 60%) of genes have been annotated with multiple and hierarchically independent terms. This may be necessary to attain adequate specificity of description. One noticeable exception is *Arabidopsis thaliana*, in which genes are much less frequently annotated with multiple terms (6 - 14%). In contrast, an analysis of the occurrence of InterPro hits in the proteomes of the seven species, followed by a mapping of the hits to GO terms, did not reveal an aberrant pattern for the *A. thaliana* genome.

This study shows the widespread usage of multiple hierarchically independent GO terms in the functional annotation of genes. By consequence, probabilistic methods that aim to predict gene function automatically through integration of diverse genomic datasets, and that employ the GO, must be able to predict such multiple terms.

We attribute the low frequency with which multiple GO terms are used in *Arabidopsis* to deviating practices in the genome annotation and curation process between communities of annotators. This may bias genome-scale comparisons of gene function between different species. GO term assignment should therefore be performed according to strictly similar rules and standards.

Keywords: Gene Ontology, genome annotation, annotation strategies, protein function, gene function prediction, multi-label classification, *Arabidopsis* genome

Introduction

The Gene Ontology (GO) [1] provides a controlled vocabulary for the description of gene and gene product attributes in any species. It uses three key domains that provide descriptions of molecular function, biological process and cellular component. Common applications of GO include the functional annotation of genes predicted from whole genome sequences, the functional annotation and comparison of genes assayed in microarray experiments and the analysis of cellular pathways. Our main interest in

GO lies in its application as a classification scheme in automated, probabilistic methods for gene function prediction [2, 3, 4, 5]. Such methods usually employ powerful statistical methods, but they have one built-in restriction: they perform classifications in which terms from at most one functional class of the GO hierarchy are predicted. Recently, the subject of multi-label classification in gene function prediction was addressed in a number of studies. In some of them [6, 7, 8, 9], a single-label classification algorithm was extended for multi-label purposes, but it did not employ the GO. Another study [2] aimed to improve the performance and consistency of gene function prediction by ensuring the True Path Rule, but it could not predict classes that are hierarchically independent. This strongly contrasts with the observation made in this note that genes in the genomes of well-studied model species have often been community-annotated using multiple independent GO terms. The following example illustrates why in principle, it is most relevant for methods of automated gene function prediction to be able to predict multiple independent GO terms.

In the GO datasets (http://www.geneontology.org), the genes *YDL029W* and *YHR107C* from *Saccharomyces cerevisiae* are both annotated with three hierarchically independent terms in the molecular function domain: *YDL029W* is annotated with "ATP binding" (GO: 0005524), "actin binding" (GO: 0003779) and "structural constituent cytoskeleton" (GO: 0005200) and gene *YHR107C* is annotated with the terms "GTPase activity" (GO:0003924), "phosphatidylinositol binding" (GO:0005545) and "structural constituent cytoskeleton" (GO: 0005200). Both genes contribute to the structural integrity of the cytoskeleton, yet they have different molecular functions. It is evident that a single GO term, irrespective of its specificity, will often not suffice to describe the function of a protein in a complete and unique way.

In this study, we investigate and quantify the incidence with which genes from well-studied species (Tab. 1) have been annotated using multiple independent GO terms. We emphasize the importance, not only of developing new methods for automated gene function prediction based on multi-label classification techniques, but also of establishing transparent and standard procedures for GO-term assignment in the curation of gene function annotations.

Table 1: Species and number of annotated genes analyzed.

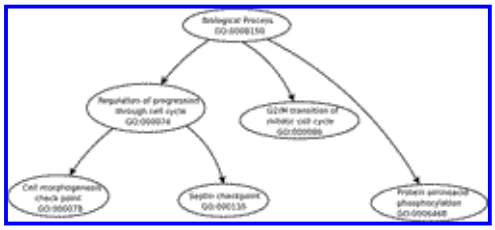
Species name	Biological process	Molecular function	Cellular component	Version
Saccharomyces cerevisiae	6473	6473	6473	1.337
Caenorhabditis elegans	9555	10996	6256	1.83
Drosophila melanogaster	10334	10469	7739	1.95
Homo sapiens	24940	29639	22823	1.48
Mus musculus	15845	17341	16059	1.664
Arabidopsis thaliana	27937	30577	28869	1.1156
Oryza sativa	13994	12669	46942	1.33

Names of the species and number of annotated genes per branch of the Gene Ontology that were used for study. The last column refers to the revision number of the respective gene_association.species files provided by the Gene ontology project [8].

Methods

We studied patterns of GO term assignment in the community-based, manually curated annotations of seven well-studied species (Tab. 1) by analyzing the frequency of multiple term usage in the annotation files provided by the GO project (gene_association.species files; http://www.geneontology.org/GO.current.

annotations.shtmlA). Hierarchical relationships between the terms were checked using the version 1.12 of the GO DAG (http://www.geneontology.org/GO.downloads.ontology.shtml). The primary data files were divided according to the three ontology domains. Multiple annotations of a gene with the same term but with use of different evidence codes were counted as single-term annotations. We counted all genes that were annotated using multiple, hierarchically independent GO terms. This was done by examining the hierarchical relationships between all pairs of annotations for a gene. In the case two terms were hierarchically related (parent-child relationship), the annotation was counted as single-term at the deeper node in the GO DAG (see Fig. 1 for an illustration).



Click on the thumbnail to enlarge the picture

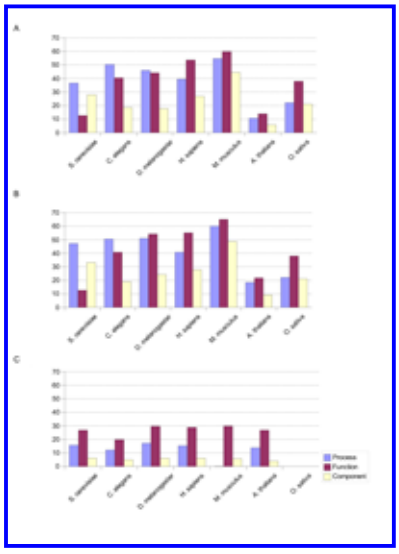
Figure 1: Example of a gene with multiple annotations.
The HSL1 gene from *Saccharomyces* is annotated using five terms from the Biological Process Ontology. Node GO:000074 is the parent of nodes GO:000078 and GO:000135. Thus, four of the five GO terms from the root are hierarchically unrelated, i. e. GO:000078, GO:000135, GO:000086, GO:0006468.

We compared the curated GO project annotations with those that can be automatically derived from the Integr8 database [10]. The Proteome Analysis section of this database provides protein annotations based on InterPro hits to protein families, domains and functional sites. The precomputed files for the species listed in Tab. 1 were downloaded from Integr8 and InterPro hits were mapped to GO terms using the interpro2go map (Mappings of External Classification Systems to GO: http://www.geneontology.org/GO.indices.shtml), available from the GO website. *Oryza sativa* was excluded from this study because it was not available at Integr8 at the time of analysis. Since InterPro entries may correspond to terms from different domains of the GO, the mapped files were again divided according to the ontology. The counting of multiple GO term usage based on InterPro domains was subsequently performed as described for the GO project annotations.

We also investigated the frequency distributions of evidence codes used in GO term assignment for each of the species. The differences in the use of evidence codes between species and between single and multiple annotated genes were analyzed by log-ratio Principal Component Analysis for compositional data [11]. In this analysis we excluded the ND evidence code (designation for "No biological data available"), as it cannot, by definition, result in multiple annotations for a gene.

Results and discussion

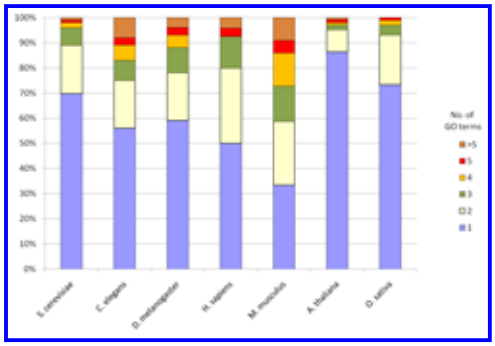
Analysis of the community-based, curated gene annotations of seven well-studied species, as provided by the GO project, show that significant proportions of the genes are annotated with multiple GO terms (Fig. 2a). Proportions range between 6% and 60% among the species and the different domains of the ontology. A large proportion of genes are annotated with three to six independent GO terms and for a considerable number of genes this number exceeds ten terms (Fig. 3). An extreme example is the gene *Notch* from *Drosophila melanogaster* (Flybase ID: Fbgn0004647) which is annotated with 52 independent terms from the Biological Process Ontology.



Click on the thumbnail to enlarge the picture

Figure 2: Multiple annotated genes in seven species. Percentages of genes with multiple hierarchically independent GO terms calculated: (A) using the annotation files provided by the GO project, including all evidence codes; (B) using the same annotation files, but excluding the ND evidence code ("no biological data available"), and; (C) using the InterPro hits as provided by the Integr8 database. Integr8 did not contain data for *O. sativa* at the time of study. The percentage for the Biological Process Ontology for *M. musculus* is missing due to a technical problem. Species designations are as in Tab. 1.

It is commonly accepted that human-curated annotations using the GO provide the as yet most reliable and standardized functional descriptions of genes. For many genes, this involves the use of multiple, hierarchically independent GO terms. Multiple annotations may be required either to describe a single function, process or cellular component as completely and uniquely as possible, or to describe the multiple functions, processes or cellular components in which a single gene product can take part. Most classification methods currently used in automated gene function prediction are not designed to assign multiple GO terms. This implies that such methods will produce incomplete and low-quality functional annotations for many genes.



Click on the thumbnail to enlarge the picture

Figure 3: Proportions of genes annotated with one or more hierarchically independent GO terms used. Each histogram represents the relative composition of genes annotated with the numbers of GO terms indicated.

A wholly unexpected and surprising observation in the comparison of species is the much lower proportion of genes with multiple GO terms in *Arabidopsis thaliana*. Further analysis showed that a relatively large number of genes in the *Arabidopsis* annotation have been assigned the term "unknown" (36.1%, 42% and 35.8% of the annotations in the molecular function, biological process, and cellular component domains, respectively). The term "unknown" is used at the root node of each domain of the GO. In our analysis these nodes were treated as single term annotations. Because this might have biased the comparison, we rescaled the proportions for all the species, excluding genes with the ND evidence code for "No biological data available". The resulting histograms are given in Fig. 2b and persistently show lower proportions of multiple annotations in *Arabidopsis*.

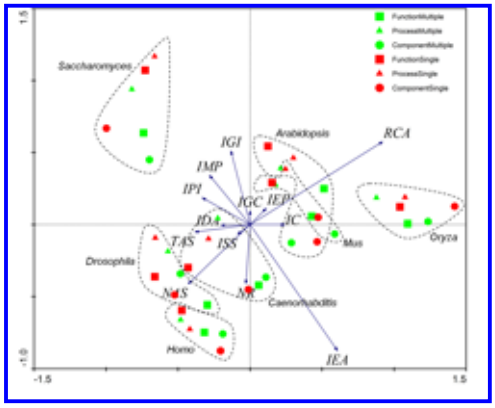
To exclude further the possibility that the deviant pattern in *Arabidopsis* is an artefact resulting from usage of erroneous files in the GO repository, we analyzed two older versions of *Arabidopsis* annotation files (versions 1.949 and 1.959 from March and April 2006, respectively). Both analyses showed similar

and much lower percentages of multiple annotated genes in *Arabidopsis* compared to the other species (data not shown).

To investigate whether the aberrant proportions for *Arabidopsis* can be explained biologically, we performed an independent re-annotation of the gene sets from each species on the basis of their hit lists to InterPro accessions [12]. If the *Arabidopsis* proteome would be biologically different from the other species, as suggested by the GO project annotations (Fig. 2a), an InterPro-based GO annotation can be expected to reveal a similar aberration in the proportion of genes with multiple functional signatures. The results presented in Fig. 2c do not reveal such a difference. It seems obvious therefore that the aberrant proportion of multiple annotations in the GO project must be explained by the different usage of rules for GO term assignment in the annotation process for *Arabidopsis*.

Because a formal comparison between the annotation strategies for the different species could not be conducted directly, we compared their distributions of usage of evidence codes for both single and multiple annotated genes by log-ratio principal component analysis. The resulting biplot (Fig. 4) shows a widely deviating pattern of evidence code usage. The differences between species appear to be much larger than the differences between single annotated and multiple annotated genes within species.

Arabidopsis does not show any particular deviating pattern relative to the other species and usage of evidence codes appears to be independent of whether genes are annotated with single or multiple hierarchically independent GO terms. *Saccharomyces* appears to be most different in this analysis, which can be attributed to a higher frequency of codes reserved for experimental evidence (IGI, IPI and IMP, see legend for abbreviations). These results further illustrates that systematic differences underlie practices of functional annotation by the annotators and curators of the various species.



Click on the thumbnail to enlarge the picture

Figure 4: Biplot of evidence code distributions used for single and multiple annotated genes.
Each point lies originally in a thirteen dimensional space where each coordinate corresponds to the frequency of use of one evidence code. This two-dimensional representation was achieved by performing a special form of PCA for compositional data (http://www.geneontology.org/GO.current.annotations.shtmlA). A polygon is drawn to cluster the points for each species. Evidence codes: IC: Inferred by Curator; IDA: Inferred from Direct Assay; IEA: Inferred from Electronic Annotation; IEP: Inferred from Expression Pattern; IGC: Inferred from Genomic Context; IGI: Inferred from Genetic Interaction; IMP: Inferred from Mutant Phenotype; IPI: Inferred from Physical Interaction; ISS: Inferred from Sequence or Structural Similarity; NAS: Non-traceable Author Statement; RCA: inferred from Reviewed Computational Analysis; TAS: Traceable Author Statement; NR: Not Recorded. (see http://www.geneontology.org/GO.evidence.shtml).

Conclusion

Assigning function to a gene is an important but also complex operation. Controlled vocabularies like GO provide an excellent infrastructure for the functional description of genes and gene products. As part of

our effort to develop statistical methodology for gene function prediction based on data integration and multi-label classification, we have studied GO term usage in genome annotations. We find that the availability of vocabularies alone does not guarantee that annotators will employ these consistently and in exactly the same manner for each species. This study reveals that the practice of GO term assignment differs considerably between communities of annotators. This will bias genome-scale comparisons of gene function between different species. The quality and comparability of functional annotations will therefore benefit, not only from controlled vocabularies such as the GO, but also from strict application of formal rules for the assignment of GO terms. First and foremost such rules must specify what criteria must be met in order for a GO term to be included in an annotation, irrespective of what evidence code is used for that GO term assignment. A proposal for such rules has been put forward by Clare *et al*. [13]. Finally, the observation of extensive usage of multiple, independent terms to describe a gene function underlines the importance of using multi-label classification methods in the development and application of methods for automated gene function prediction.

Acknowledgements

This work is part of the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

1. The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. Nature Genet. **25**, 25-29.
2. Barutcuoglu, Z., Schapire, R. E. and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. Bioinformatics **22**, 830-836.
3. Pavlidis, P., Weston, J., Cai, J. and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. J. Comput. Biol. **9**, 401-411.
4. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc. Natl. Acad. Sci. USA **100**, 8348-8353.
5. Eisner, R., Poulin, B., Szafron, D., Lu, P. and Greiner, R. (2005). Improving protein function prediction using the hierarchical structure of the gene ontology. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.
6. Clare, A. and King, R. D. (2003). Predicting gene function in *Saccharomyces cerevisiae*. Bioinformatics Suppl. **2**, 42-49.
7. Blockeel, H., Schietgat, L., Struyf, J., Dzeroski, S. and Clare, A. (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics. *In*: Proceedings of PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, vol 4213, LNAI, pp. 18-29.
8. Roth, V. and Fischer, B. (2007). Improved functional prediction of proteins by learning kernel combinations in multilabel settings. BMC Bioinformatics **8** (Suppl. 2), S12.
9. Zhang, M. L. and Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition **40**, 2038-2048.
10. Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U.,

Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I. and Apweiler, R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteome. Nucleic Acids Res. **33**, D297-D302.

11. Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. J. Roy. Stat. Soc. C-App. **51**, 375-392.

12. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J. A., Silventoinen, V., Studholme, D. J., Vaughan, R. and Wu, C. H. (2005). InterPro, progress and status in 2005. Nucleic Acids Res. **33**, 201-205.

13. Clare, A., Karwath, A., Ougham, H. and King, R. D. (2006). Functional bioinformatics for *Arabidopsis thaliana*. Bioinformatics **22**, 1130-1136.

In Silico Biology 7, 0050 (2007); ©2007, Bioinformation Systems e.V.

The use of multiple hierarchically independent Gene Ontology terms in gene function prediction and genome annotation

Yiannis A. I. Kourmpetis[1], Ate van der Burgt[2], Marco C. A. M. Bink[1], Cajo J. F. ter Braak[1]* and Roeland C. H. J. van Ham[2]

[1] Biometris, Wageningen University and Research Centre, 6700 AC Wageningen, The Netherlands
[2] Applied Bioinformatics, Plant Research International, 6708 PB Wageningen, The Netherlands

* Corresponding author

Email: cajo@terbraak.nl

Edited by H. Michael; received June 14, 2007; revised August 23, 2007; accepted August 27, 2007; published October 13, 2007

Abstract

The Gene Ontology (GO) is a widely used controlled vocabulary for the description of gene function. In this study we quantify the usage of multiple and hierarchically independent GO terms in the curated genome annotations of seven well-studied species. In most genomes, significant proportions (6 - 60%) of genes have been annotated with multiple and hierarchically independent terms. This may be necessary to attain adequate specificity of description. One noticeable exception is *Arabidopsis thaliana*, in which genes are much less frequently annotated with multiple terms (6 - 14%). In contrast, an analysis of the occurrence of InterPro hits in the proteomes of the seven species, followed by a mapping of the hits to GO terms, did not reveal an aberrant pattern for the *A. thaliana* genome.

This study shows the widespread usage of multiple hierarchically independent GO terms in the functional annotation of genes. By consequence, probabilistic methods that aim to predict gene function

automatically through integration of diverse genomic datasets, and that employ the GO, must be able to predict such multiple terms.

We attribute the low frequency with which multiple GO terms are used in *Arabidopsis* to deviating practices in the genome annotation and curation process between communities of annotators. This may bias genome-scale comparisons of gene function between different species. GO term assignment should therefore be performed according to strictly similar rules and standards.

Keywords: Gene Ontology, genome annotation, annotation strategies, protein function, gene function prediction, multi-label classification, *Arabidopsis* genome

Introduction

The Gene Ontology (GO) [1] provides a controlled vocabulary for the description of gene and gene product attributes in any species. It uses three key domains that provide descriptions of molecular function, biological process and cellular component. Common applications of GO include the functional annotation of genes predicted from whole genome sequences, the functional annotation and comparison of genes assayed in microarray experiments and the analysis of cellular pathways. Our main interest in GO lies in its application as a classification scheme in automated, probabilistic methods for gene function prediction [2, 3, 4, 5]. Such methods usually employ powerful statistical methods, but they have one built-in restriction: they perform classifications in which terms from at most one functional class of the GO hierarchy are predicted. Recently, the subject of multi-label classification in gene function prediction was addressed in a number of studies. In some of them [6, 7, 8, 9], a single-label classification algorithm was extended for multi-label purposes, but it did not employ the GO. Another study [2] aimed to improve the performance and consistency of gene function prediction by ensuring the True Path Rule, but it could not predict classes that are hierarchically independent. This strongly contrasts with the observation made in this note that genes in the genomes of well-studied model species have often been community-annotated using multiple independent GO terms. The following example illustrates why in principle, it is most relevant for methods of automated gene function prediction to be able to predict multiple independent GO terms.

In the GO datasets (http://www.geneontology.org), the genes *YDL029W* and *YHR107C* from *Saccharomyces cerevisiae* are both annotated with three hierarchically independent terms in the molecular function domain: *YDL029W* is annotated with "ATP binding" (GO: 0005524), "actin binding" (GO: 0003779) and "structural constituent cytoskeleton" (GO: 0005200) and gene *YHR107C* is annotated with the terms "GTPase activity" (GO:0003924), "phosphatidylinositol binding" (GO:0005545) and "structural constituent cytoskeleton" (GO: 0005200). Both genes contribute to the structural integrity of the cytoskeleton, yet they have different molecular functions. It is evident that a single GO term, irrespective of its specificity, will often not suffice to describe the function of a protein in a complete and unique way.

In this study, we investigate and quantify the incidence with which genes from well-studied species (Tab. 1) have been annotated using multiple independent GO terms. We emphasize the importance, not only of developing new methods for automated gene function prediction based on multi-label classification techniques, but also of establishing transparent and standard procedures for GO-term assignment in the curation of gene function annotations.

Table 1: Species and number of annotated genes analyzed.

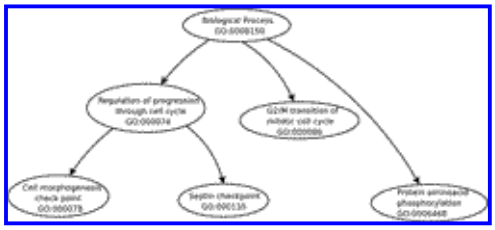
Species name	Biological process	Molecular function	Cellular component	Version
Saccharomyces cerevisiae	6473	6473	6473	1.337

Caenorhabditis elegans	9555	10996	6256	1.83
Drosophila melanogaster	10334	10469	7739	1.95
Homo sapiens	24940	29639	22823	1.48
Mus musculus	15845	17341	16059	1.664
Arabidopsis thaliana	27937	30577	28869	1.1156
Oryza sativa	13994	12669	46942	1.33

Names of the species and number of annotated genes per branch of the Gene Ontology that were used for study. The last column refers to the revision number of the respective gene_association.species files provided by the Gene ontology project [8].

Methods

We studied patterns of GO term assignment in the community-based, manually curated annotations of seven well-studied species (Tab. 1) by analyzing the frequency of multiple term usage in the annotation files provided by the GO project (gene_association.species files; http://www.geneontology.org/GO.current.annotations.shtmlA). Hierarchical relationships between the terms were checked using the version 1.12 of the GO DAG (http://www.geneontology.org/GO.downloads.ontology.shtml). The primary data files were divided according to the three ontology domains. Multiple annotations of a gene with the same term but with use of different evidence codes were counted as single-term annotations. We counted all genes that were annotated using multiple, hierarchically independent GO terms. This was done by examining the hierarchical relationships between all pairs of annotations for a gene. In the case two terms were hierarchically related (parent-child relationship), the annotation was counted as single-term at the deeper node in the GO DAG (see Fig. 1 for an illustration).



Click on the thumbnail to enlarge the picture

Figure 1: Example of a gene with multiple annotations.
The HSL1 gene from *Saccharomyces* is annotated using five terms from the Biological Process Ontology. Node GO:000074 is the parent of nodes GO:000078 and GO:000135. Thus, four of the five GO terms from the root are hierarchically unrelated, i. e. GO:000078, GO:000135, GO:000086, GO:0006468.

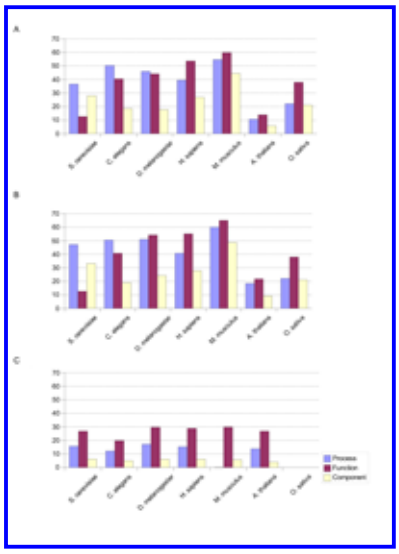
We compared the curated GO project annotations with those that can be automatically derived from the Integr8 database [10]. The Proteome Analysis section of this database provides protein annotations based on InterPro hits to protein families, domains and functional sites. The precomputed files for the species listed in Tab. 1 were downloaded from Integr8 and InterPro hits were mapped to GO terms using the interpro2go map (Mappings of External Classification Systems to GO: http://www.geneontology.org/GO.indices.shtml), available from the GO website. *Oryza sativa* was excluded from this study because it was not available at Integr8 at the time of analysis. Since InterPro entries may correspond to terms from different domains of the GO, the mapped files were again divided according to the ontology. The counting of multiple GO term usage based on InterPro domains was subsequently performed as described for the GO project annotations.

We also investigated the frequency distributions of evidence codes used in GO term assignment for each of the species. The differences in the use of evidence codes between species and between single and

multiple annotated genes were analyzed by log-ratio Principal Component Analysis for compositional data [11]. In this analysis we excluded the ND evidence code (designation for "No biological data available"), as it cannot, by definition, result in multiple annotations for a gene.

Results and discussion

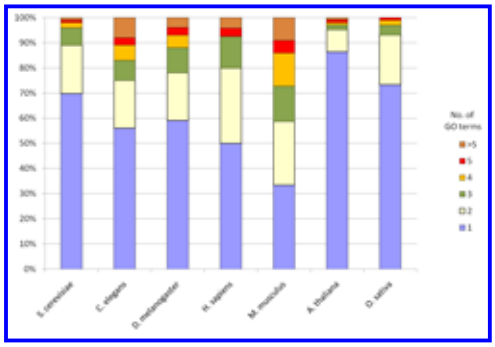
Analysis of the community-based, curated gene annotations of seven well-studied species, as provided by the GO project, show that significant proportions of the genes are annotated with multiple GO terms (Fig. 2a). Proportions range between 6% and 60% among the species and the different domains of the ontology. A large proportion of genes are annotated with three to six independent GO terms and for a considerable number of genes this number exceeds ten terms (Fig. 3). An extreme example is the gene *Notch* from *Drosophila melanogaster* (Flybase ID: Fbgn0004647) which is annotated with 52 independent terms from the Biological Process Ontology.



Click on the thumbnail to enlarge the picture

Figure 2: Multiple annotated genes in seven species. Percentages of genes with multiple hierarchically independent GO terms calculated: (A) using the annotation files provided by the GO project, including all evidence codes; (B) using the same annotation files, but excluding the ND evidence code ("no biological data available"), and; (C) using the InterPro hits as provided by the Integr8 database. Integr8 did not contain data for *O. sativa* at the time of study. The percentage for the Biological Process Ontology for *M. musculus* is missing due to a technical problem. Species designations are as in Tab. 1.

It is commonly accepted that human-curated annotations using the GO provide the as yet most reliable and standardized functional descriptions of genes. For many genes, this involves the use of multiple, hierarchically independent GO terms. Multiple annotations may be required either to describe a single function, process or cellular component as completely and uniquely as possible, or to describe the multiple functions, processes or cellular components in which a single gene product can take part. Most classification methods currently used in automated gene function prediction are not designed to assign multiple GO terms. This implies that such methods will produce incomplete and low-quality functional annotations for many genes.



Click on the thumbnail to enlarge the picture

Figure 3: Proportions of genes annotated with one or more hierarchically independent GO terms used. Each histogram represents the relative composition of genes annotated with the numbers of GO terms indicated.

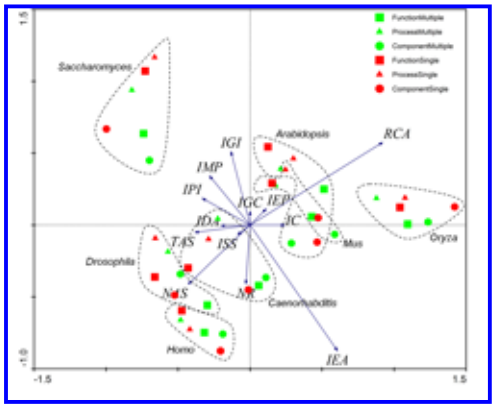
A wholly unexpected and surprising observation in the comparison of species is the much lower proportion of genes with multiple GO terms in *Arabidopsis thaliana*. Further analysis showed that a relatively large number of genes in the *Arabidopsis* annotation have been assigned the term "unknown" (36.1%, 42% and 35.8% of the annotations in the molecular function, biological process, and cellular component domains, respectively). The term "unknown" is used at the root node of each domain of the GO. In our analysis these nodes were treated as single term annotations. Because this might have biased the comparison, we rescaled the proportions for all the species, excluding genes with the ND evidence code for "No biological data available". The resulting histograms are given in Fig. 2b and persistently show lower proportions of multiple annotations in *Arabidopsis*.

To exclude further the possibility that the deviant pattern in *Arabidopsis* is an artefact resulting from usage of erroneous files in the GO repository, we analyzed two older versions of *Arabidopsis* annotation files (versions 1.949 and 1.959 from March and April 2006, respectively). Both analyses showed similar and much lower percentages of multiple annotated genes in *Arabidopsis* compared to the other species (data not shown).

To investigate whether the aberrant proportions for *Arabidopsis* can be explained biologically, we performed an independent re-annotation of the gene sets from each species on the basis of their hit lists to InterPro accessions [12]. If the *Arabidopsis* proteome would be biologically different from the other species, as suggested by the GO project annotations (Fig. 2a), an InterPro-based GO annotation can be expected to reveal a similar aberration in the proportion of genes with multiple functional signatures. The results presented in Fig. 2c do not reveal such a difference. It seems obvious therefore that the aberrant proportion of multiple annotations in the GO project must be explained by the different usage of rules for GO term assignment in the annotation process for *Arabidopsis*.

Because a formal comparison between the annotation strategies for the different species could not be conducted directly, we compared their distributions of usage of evidence codes for both single and multiple annotated genes by log-ratio principal component analysis. The resulting biplot (Fig. 4) shows a widely deviating pattern of evidence code usage. The differences between species appear to be much larger than the differences between single annotated and multiple annotated genes within species.

Arabidopsis does not show any particular deviating pattern relative to the other species and usage of evidence codes appears to be independent of whether genes are annotated with single or multiple hierarchically independent GO terms. *Saccharomyces* appears to be most different in this analysis, which can be attributed to a higher frequency of codes reserved for experimental evidence (IGI, IPI and IMP, see legend for abbreviations). These results further illustrates that systematic differences underlie practices of functional annotation by the annotators and curators of the various species.



Click on the thumbnail to enlarge the picture

Figure 4: Biplot of evidence code distributions used for single and multiple annotated genes.
Each point lies originally in a thirteen dimensional space where each coordinate corresponds to the frequency of use of one evidence code. This two-dimensional representation was achieved by performing a special form of PCA for compositional data (http://www.geneontology.org/GO.current.annotations.shtmlA). A polygon is drawn to cluster the points for each species. Evidence codes: IC: Inferred by Curator; IDA: Inferred from Direct Assay; IEA: Inferred from Electronic Annotation; IEP: Inferred from Expression Pattern; IGC: Inferred from Genomic Context; IGI: Inferred from Genetic Interaction; IMP: Inferred from Mutant Phenotype; IPI: Inferred from Physical Interaction; ISS: Inferred from Sequence or Structural Similarity; NAS: Non-traceable Author Statement; RCA: inferred from Reviewed Computational Analysis; TAS: Traceable Author Statement; NR: Not Recorded. (see http://www.geneontology.org/GO.evidence.shtml).

Conclusion

Assigning function to a gene is an important but also complex operation. Controlled vocabularies like GO provide an excellent infrastructure for the functional description of genes and gene products. As part of our effort to develop statistical methodology for gene function prediction based on data integration and multi-label classification, we have studied GO term usage in genome annotations. We find that the availability of vocabularies alone does not guarantee that annotators will employ these consistently and in exactly the same manner for each species. This study reveals that the practice of GO term assignment differs considerably between communities of annotators. This will bias genome-scale comparisons of gene function between different species. The quality and comparability of functional annotations will therefore benefit, not only from controlled vocabularies such as the GO, but also from strict application of formal rules for the assignment of GO terms. First and foremost such rules must specify what criteria must be met in order for a GO term to be included in an annotation, irrespective of what evidence code is used for that GO term assignment. A proposal for such rules has been put forward by Clare *et al*. [13]. Finally, the observation of extensive usage of multiple, independent terms to describe a gene function underlines the importance of using multi-label classification methods in the development and application of methods for automated gene function prediction.

Acknowledgements

This work is part of the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

1. The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. Nature Genet. **25**, 25-29.

2. Barutcuoglu, Z., Schapire, R. E. and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. Bioinformatics **22**, 830-836.

3. Pavlidis, P., Weston, J., Cai, J. and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. J. Comput. Biol. **9**, 401-411.

4. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc. Natl. Acad. Sci. USA **100**, 8348-8353.

5. Eisner, R., Poulin, B., Szafron, D., Lu, P. and Greiner, R. (2005). Improving protein function prediction using the hierarchical structure of the gene ontology. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.

6. Clare, A. and King, R. D. (2003). Predicting gene function in *Saccharomyces cerevisiae*. Bioinformatics Suppl. **2**, 42-49.

7. Blockeel, H., Schietgat, L., Struyf, J., Dzeroski, S. and Clare, A. (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics. *In*: Proceedings of PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, vol 4213, LNAI, pp. 18-29.

8. Roth, V. and Fischer, B. (2007). Improved functional prediction of proteins by learning kernel combinations in multilabel settings. BMC Bioinformatics **8** (Suppl. 2), S12.

9. Zhang, M. L. and Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition **40**, 2038-2048.

10. Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I. and Apweiler, R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteome. Nucleic Acids Res. **33**, D297-D302.

11. Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. J. Roy. Stat. Soc. C-App. **51**, 375-392.

12. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J. A., Silventoinen, V., Studholme, D. J., Vaughan, R. and Wu, C. H. (2005). InterPro, progress and status in 2005. Nucleic Acids Res. **33**, 201-205.

13. Clare, A., Karwath, A., Ougham, H. and King, R. D. (2006). Functional bioinformatics for *Arabidopsis thaliana*. Bioinformatics **22**, 1130-1136.