We are IntechOpen, the world's leading publisher of Open Access books
Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Visual Gene Ontology Based Knowledge Discovery in Functional Genomics

Stefan Götz and Ana Conesa
Bioinformatics and Genomics Department,
Centro de Investigaciones Príncipe Felipe
Spain

1. Introduction

The molecular biology and the study of the genome has been transformed in the last decade by the avalanche of a conspicuous amount of data produced by high-throughput technologies. The accessibility to massive sequencing, microarrays and proteomics methods implies that studies on the genome-wide scale has become feasible for a large number of laboratories throughout the world. The great challenge in current genome research is to transform these data into knowledge, knowledge that will lead to a better understanding of the molecular processes underlying the observed biological phenomena. The discipline of functional genomics has arisen to integrate genomic sequences, experimental genome screening and novel bioinformatics methods to facilitate comprehension of genome-wide data. Within this framework, visualization appears as a helpful component to interpret results from high-throughput experiments and can be indispensable when working with large datasets. Typically, functional genomics makes use of structured functional terms like the Gene Ontology (GO) vocabulary (Ashburner M, 2000) and in this case the "natural" visualization format is a graph of a group of functionally annotated elements. In this graph, nodes represent the functional labels (the GO terms) to which genes or proteins are annotated to and arcs represent the relationships among them. Visualization of genomics data within the GO graph structure is effective in providing a rapid insight into the functional composition of a dataset and for understanding functional information within its biological context. This is of outermost importance in functional genomics research where both the overview and the detail are essential to create hypothesis over the functioning of biological systems.

However, a problem when visualizing GO functional information is that graphs can become extremely large and difficult to navigate if the number of represented elements (genes, proteins or simply GO terms) is high. Moreover, not all nodes are equally important and different kind of information might be associated to them. Therefore, the goal of a good visualization technique is to provide an informative view of the data that is efficient in generating new knowledge.

In this chapter we review current methods to visualize and perform data mining from GO contained genomic data. We start with a brief introduction to the Gene Ontology and present several third-party approaches for the visualization of GO data. The body of this document will focus on visual solutions for GO-based data mining available in the

Blast2GO suite. These include tools for zooming, filtering, highlighting, summarizing, transforming and exporting GO data. Blast2GO is freely available to the scientific community at www.blast2go.org.

## 2. Visualization of genomics data through the Gene Ontology

The **Gene Ontology** is a controlled vocabulary of terms for describing gene product characteristics. It is developed by the GO Consortium (http://www.geneontology.org). The project aims at the standardization of gene product attributes and terms are generally defined to be valid across species. The GO consists of three independent ontologies that describe gene and protein functions from the perspective of the **Biological Process** in which they operate, the **Molecular Function** or role they have and the **Cellular Component** they localize. The GO Consortium also maintains a database of GO term annotated gene products and many laboratories worldwide use this vocabulary to describe their genomic data.

The GO is a hierarchical vocabulary with a Directed Acyclic Graph (DAG) structure, meaning that ancestor (or parent) terms are more general concepts than descendant (or children) terms and that connections between node elements can be many to capture different relationships between biological concepts. The ontology contains more than 30.000 terms distributed in about 20 levels. Another characteristic of the GO is the consistency or "true path rule": given the annotation of a gene to a certain GO level and term, all ancestor terms up to the root must be true for that gene. The Gene Ontology is not only used to annotate gene products with functional labels[1], but also to study their activity from a functional perspective, i.e. to approach genome research as the working of functionally related groups of genes that act coordinately as a *System*.

When considering the visualization of GO contained genomics data, different levels of complexity can be envisaged. First, we can consider the visualization of a single annotated gene or a specific GO term. This typically results in simple graphs that do not pose major representation challenges apart from informing on the relationships between terms and possibly the annotation event for that gene. A more complex situation arises when the annotation of a group of genes has to be displayed. Here large DAGs are produced with many levels and connections between nodes, and information enhancing methods are highly required to identify and quantify the knowledge potential. Finally, one might want to represent a subset of selected GO terms that have been obtained by some data mining or statistical methods, for example, from a functional enrichment analysis of some gene expression data (see below). This represents normally an intermediate situation between the two previous cases with respect to graph size. The visual requirement here would be in the representation of the statistical value associated to the selected terms.

Different GO visualization, browser and query programs can be found that are freely available to the scientific community. A significant number of them are listed at the Gene Ontology website (http://www.geneontology.org/GO.tools.shtml). A common approach to browse the GO is the tree representation, such as provided by the AmiGO (Carbon S, 2009) and QuickGO (Binns D, 2009) engines. Tree browsers show the GO as an indented list, similar to file browsers, where terms can be expanded by mouse click to view their children and navigate down through the hierarchy (see Figure 1). If a term appears in multiple

---

[1] the terms 'functional annotation', 'functional label' and 'GO term' will be used exchangeable throughout the text
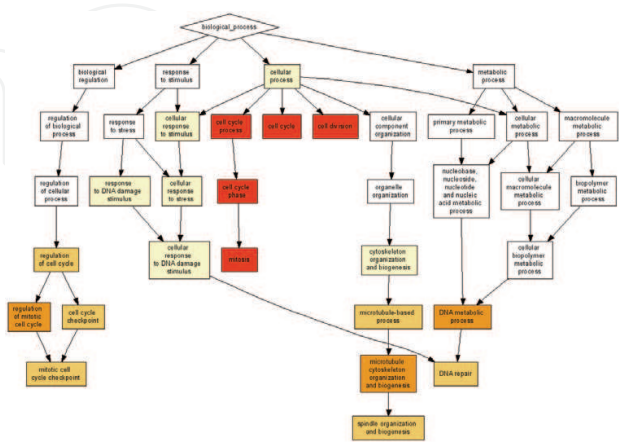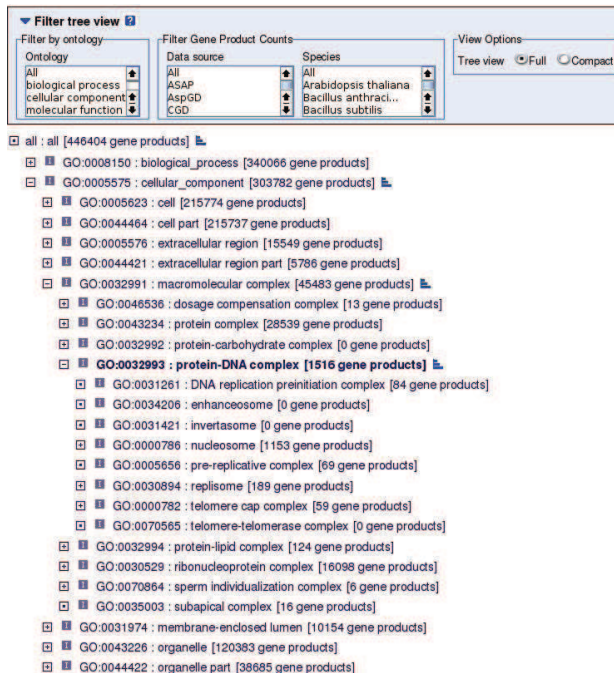
(a) AmiGO browser at the Gene Ontology site, taken from www.geneontology.org

(b) Enriched GO graph by GOrilla, taken from http://cbl-gorilla.cs.technion. ac.il/example.html
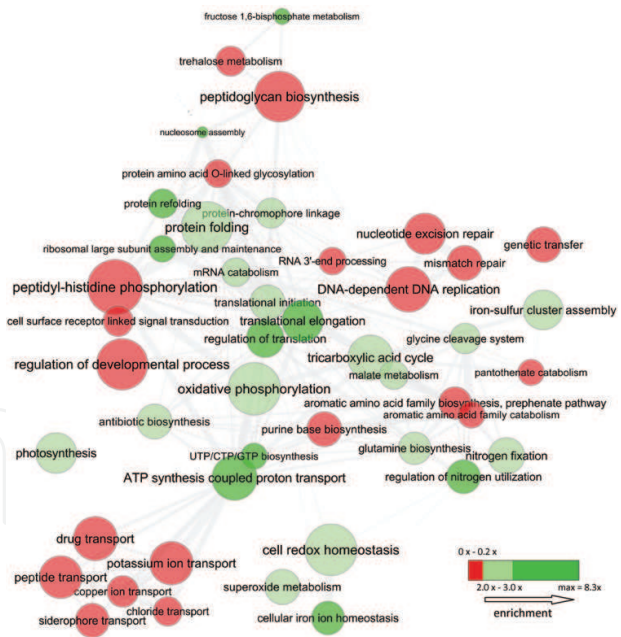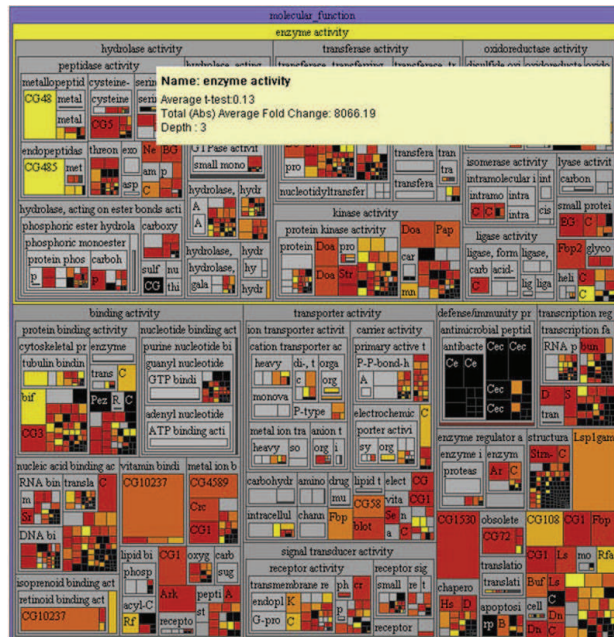
(c) TreeMap view of an annotate dataset, adapted from http://syntheti.cc/ndimensions/ zb/treeviz/treemaps.html

(d) REViGO plot, adapted from (Supek F, 2010)

Fig. 1. Different visualization engines for Gene Ontology data

locations in the tree, all occurrences are displayed. These browsers are indicated for the visualization of single gene or term data and to recover rich information on specific elements of the GO database, but are limited to represent large annotation datasets.

A popular data mining application of the Gene Ontology is the identification of functional categories that are over or underrepresented in a subset of genes selected or ranked for their association to a given phenotype, the so called Pathway Analysis (Al-Shahrour F, 2004). Many bioinformatics tools perform these analyzes, such as Babelomics (Medina I, 2010), DAVID (Dennis G Jr, 2003) or GOrilla (Eran Eden & Yakhini, 2009), to cite a few. DAG based visualization of pathway analysis results typically goes through the display of the subgraph corresponding to the set of enriched GO terms together with a coloring scheme of nodes that indicate the significant value of the enrichment test (see Figure 1).

An interesting development for the visualization of GO based genomics data is the treemap approach (Baehrecke EH, 2004). Treemaps are a space-filling visualization technique for hierarchical structures that show attributes of leaf nodes by size and color-coding. Although, similarly to tree browsers, treemaps impose a flattening of DAG structure, the implementing software offers versatility to code different gene or GO term attributes into the elements of the treemap (see Figure 1). As disadvantage, the compartment-like display of treemaps, far from the familiar graph structure of the Gene Ontology, might be dis-appealing for untrained users and this may explain the lesser use of this solution. Finally, one of the few tools that offer functionalities for reducing the complexity of large GO graphs is REViGO (Supek F, 2010). REViGO is a web server that can take long lists of Gene Ontology terms and summarize them by removing redundant GO terms. The site provides a visualization engine that represent GO graphs by grouping GO terms based on their semantic similarity (see Figure 1).

3. GO visualization with Blast2GO

3.1 Blast2GO

Blast2GO, is a suite of tools and methods for assigning functional labels to novel genome sequences and for knowledge discovery in functional genomics research. The application has, therefore, two main aspects: (i) the generation of functionally annotated sequences according to the Gene Ontology (and other) vocabularies which is based on the similarity between the unknown sequences and known labeled genes (ii) the data-mining of genomics (experimental) datasets making use of this functional information. Blast2GO has been developed to be a user-friendly, biologist-oriented software and employs a large variety of graphical resources to present and analyze annotation data (Götz S, 2008). These features have granted the application great acceptance in the functional genomics scientific community. Blast2GO has been used in hundreds of genomics projects covering a great variety of organisms and it is today the most cited software of its kind (http://www.blast2go.org/b2g_in_papers).

The graphical functions of Blast2GO have been implemented to facilitate browsing and analysis at different levels of complexity of the functional data. In the following sections we will describe in detail all these functionalities and show examples of their application as data mining tools.

3.2 Visualization of one annotated sequence

In the process of assigning functional labels to sequence data by similarity-based automatic procedures a number of steps have to be covered. These typically include:

- comparison of un-characterized sequences to known, annotated genes deposited in public repositories
- identification of primary function candidates as the functional labels belonging to the similar genes
- transfer of selected GO terms to the novel sequences

In Blast2GO this procedure entails the computation of an annotation score (AS) for each node of the reconstructed GO sub-graph encompassed by the set of candidate terms. This annotation-score is used to select transferred labels and to control the intensity of the annotation process. The annotation-score considers the amount of similarity between novel and known sequences: greater similarity implies higher annotation-score, and the quality of the candidate functional term: those candidate terms that are not themselves the result of an automatic procedure are favoured. Additionally, the Blast2GO annotation algorithm takes into account the position of the term within the Ontology: a given term will not be transferred if a descendant is. The reader is referred to (Conesa A, 2005) for a detail explanation of the Blast2GO annotation algorithm.

Although computational annotation of high-throughput data is normally an unsupervised process, it might be sometimes convenient to check the annotation event of specific sequences to understand or modify annotation parameters. Blast2GO provides a graphical view on a single sequence annotation through the combination of color, shape and labeling formats. Given the complete set of potential GO term annotation candidates, the corresponding sub-graph is constructed. Primary candidate terms are identified by their GO code while each node in the graph is coloured according to its annotation-score. Finally, the shape of the node is indicative of the effective transference of the term: octagonal-shaped nodes are newly assigned annotations while rectangular nodes are non-transferred terms (see Figure 2). This representation has a number of advantages: on one had it permits the understanding of the relationships among candidate functions and of the potential functional diversity of the novel sequence. On the other hand, it informs on the selection procedure for transferred terms. Finally, it shows the level of specificity (ontology depth) of the novel annotations.

3.3 Combined graph

The major advantage of the GO graph visualization is when representing the information contained in a group of sequences. Two basic features are provided by graph structures: context -biological concepts are placed within the framework of a general biological sense-, and differentiation -by applying different highlighting forms to nodes according to their relevance. In Blast2GO, the graphical function to display multiple sequences is called "Combined Graph" that creates a joint DAG of a set of annotated genes. It is important to mention that Blast2GO supports the true-path consistency rule of the Gene Ontology, meaning that when a given sequence is assigned to a GO term, all parent terms become also annotated with that sequence. Different parameters are available at the Combined Graph function to enhance the visual analysis of a sequences dataset, including coloring schemes, filtering and graphical layout. In most cases, enhancing options rely on the computation of a node associated value that tells something about the relevance of the term within the dataset. These will be described in the following sections together with their use to create customized graphs.
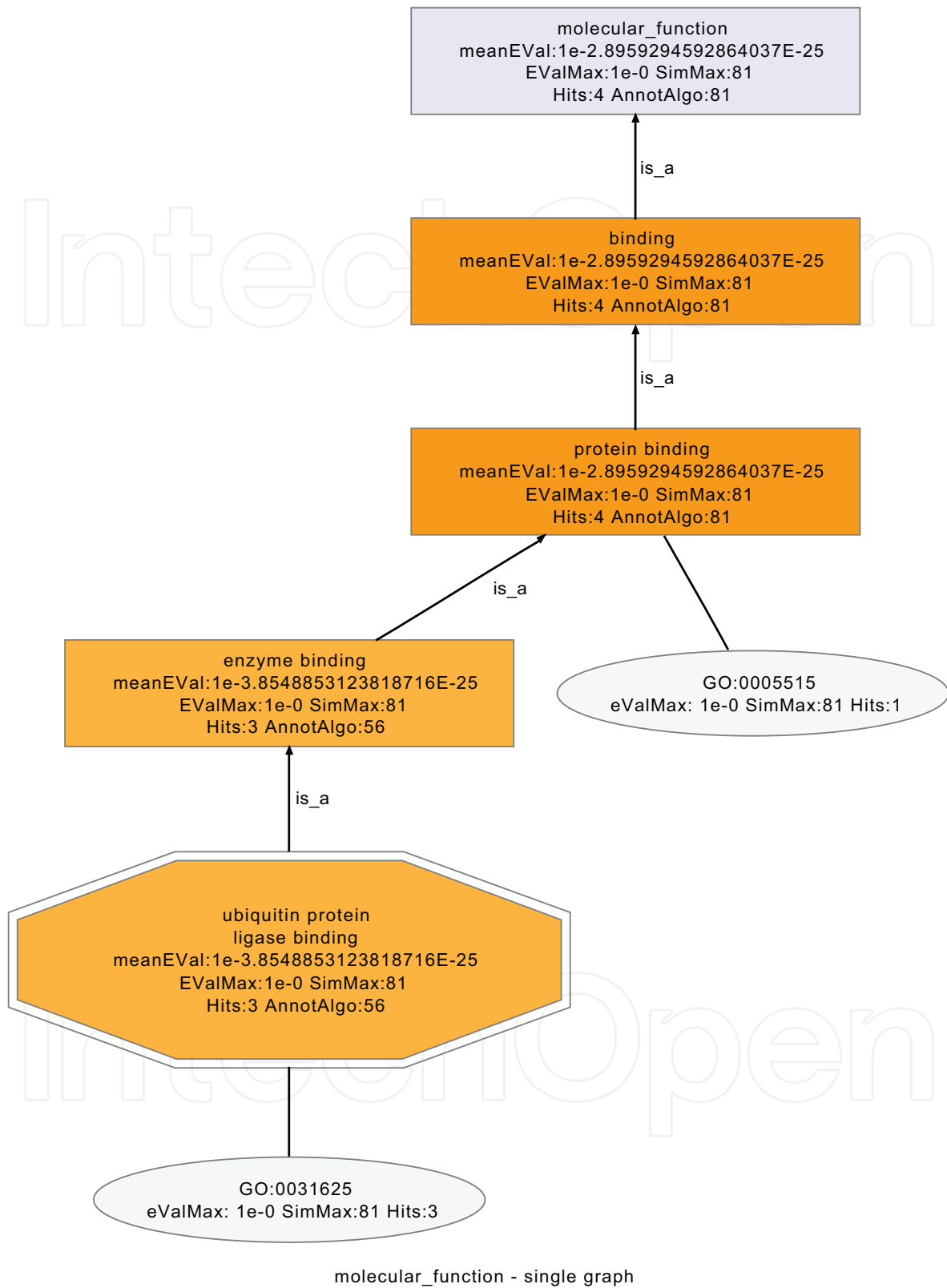
molecular_function
meanEVal:1e-2.8959294592864037E-25
EValMax:1e-0 SimMax:81
Hits:4 AnnotAlgo:81

is_a

binding
meanEVal:1e-2.8959294592864037E-25
EValMax:1e-0 SimMax:81
Hits:4 AnnotAlgo:81

is_a

protein binding
meanEVal:1e-2.8959294592864037E-25
EValMax:1e-0 SimMax:81
Hits:4 AnnotAlgo:81

is_a

enzyme binding
meanEVal:1e-3.8548853123818716E-25
EValMax:1e-0 SimMax:81
Hits:3 AnnotAlgo:56

GO:0005515
eValMax: 1e-0 SimMax:81 Hits:1

is_a

ubiquitin protein
ligase binding
meanEVal:1e-3.8548853123818716E-25
EValMax:1e-0 SimMax:81
Hits:3 AnnotAlgo:56

GO:0031625
eValMax: 1e-0 SimMax:81 Hits:3

molecular_function - single graph

Fig. 2. A GO graph of the molecular function annotation of a single sequence. It shows to possible GO term annotations (oval nodes) for a given sequence and the finally assigned term in octagonal shape. Intermediate nodes are rectangular. Nodes are colored according the Annotation Score ("AnnotAlgo").

3.3.1 Node relevance and coloring. The Blast2GO Node-Score

A first, straight-forward approach to estimate node relevance is by counting the number of annotated sequences at each GO term and display consequently node color intensity proportionally to this value. This approach is available in Blast2GO under the Combined Graph menu and creates graphs with a gradual coloring from the bottom to the top of the graph, having parent terms darker tints that their children (see Figure 3). This strategy is useful to concentrate attention in general, high populated terms. However, due to GO true-path rule coloring by number of sequences invariably highlights high level nodes and this might not be adequate to find interesting specific terms down in the hierarchy.
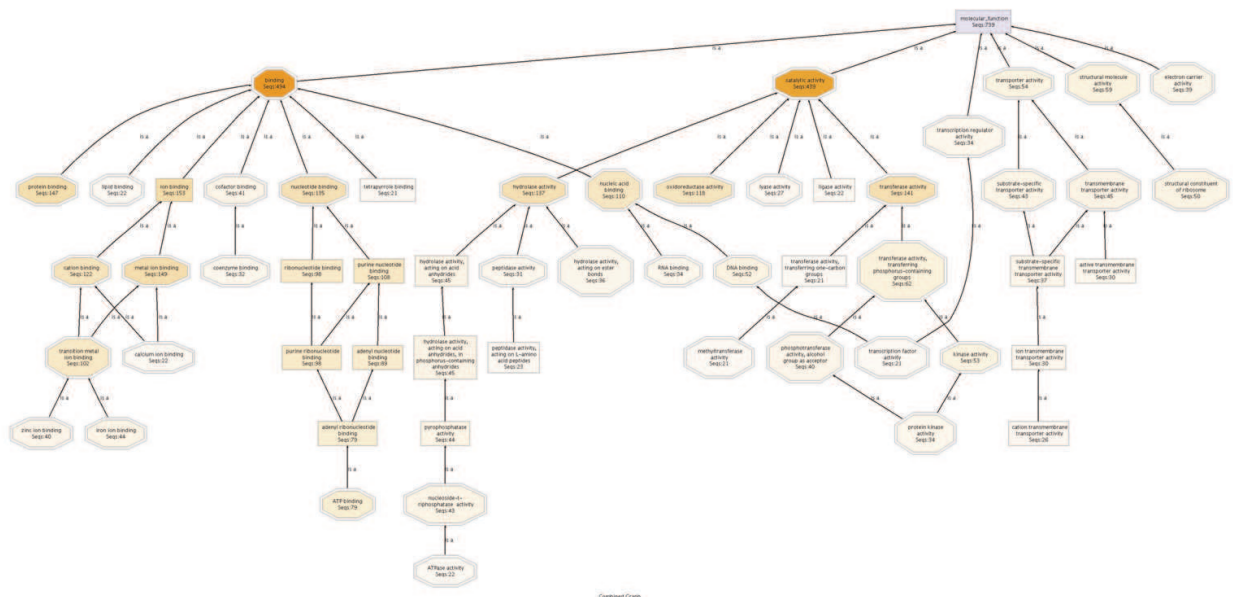

Fig. 3. A GO molecular function DAG of 1000 sequences, representing only terms with at least 20 sequences, colored by the amount of sequences for each term.

Another possibility would be to weight and color nodes by their information content. Within the GO, the information content (IC) (Resnik, 1995) of a node is defined by the inverse of the logarithm of the frequency of annotations: $IC(g) = -log_{10}p(g)$ where g is an element of the GO and GO is the overall set of all GO terms. $p(g)$ is defined as the relative frequency of the occurrence of g. The IC of terms decreases as going up in the graph reaching its minimum at the top level (root) to which all sequences are indirectly assigned ($p(root) = 1$). In a way, this parameter is the opposite at the number of annotated sequences and coloring by this metric results as well in a gradual highlighting scheme. In Blast2GO we propose a novel node metric that combines the information content of a given term and its position withing the Gene Ontology. The node-score (NS) is a trade off between functional assignment and the hierarchical topology of the GO. The node-score is defined as the sum of sequences directly or indirectly associated to a given GO term weighted by the distance between the term and the term of "direct annotation" i.e. the GO term the sequence is originally annotated to. This weighting is achieved by multiplying the sequence number by a factor α [0,∞] to the power of the distance between the term and the term of direct annotation (see Equation 1 for a mathematical expression). In this way, the node-score is accumulative and the information of lower-level GO-terms is considered, but the influence of more distant information (i.e. annotations) is suppressed/decreased depending on the

value of α. This compensates for the drawback of the earlier described method of simply counting the number of different sequences assigned to each GO-term. The α parameter allows this behaviour to be further adjusted. A value of zero means no propagation of information and can be increased by rising α. Figure 4 displays the graph shown in Figure 3 colored according to the node-score metric. Note the difference in the highlighting at the different levels of the DAG.

$$node-score(g) = \sum_{g_a \in desc(g)} gp(g_a) \cdot \alpha^{dist(g,g_a)} \tag{1}$$

where:

- desc(g) represents all the descendant terms for a given GO term g
- dist(g,g_a) is the number of edges between the GO term g and the GO term g_a
- g is an element of the GO where GO is the overall set of all GO terms
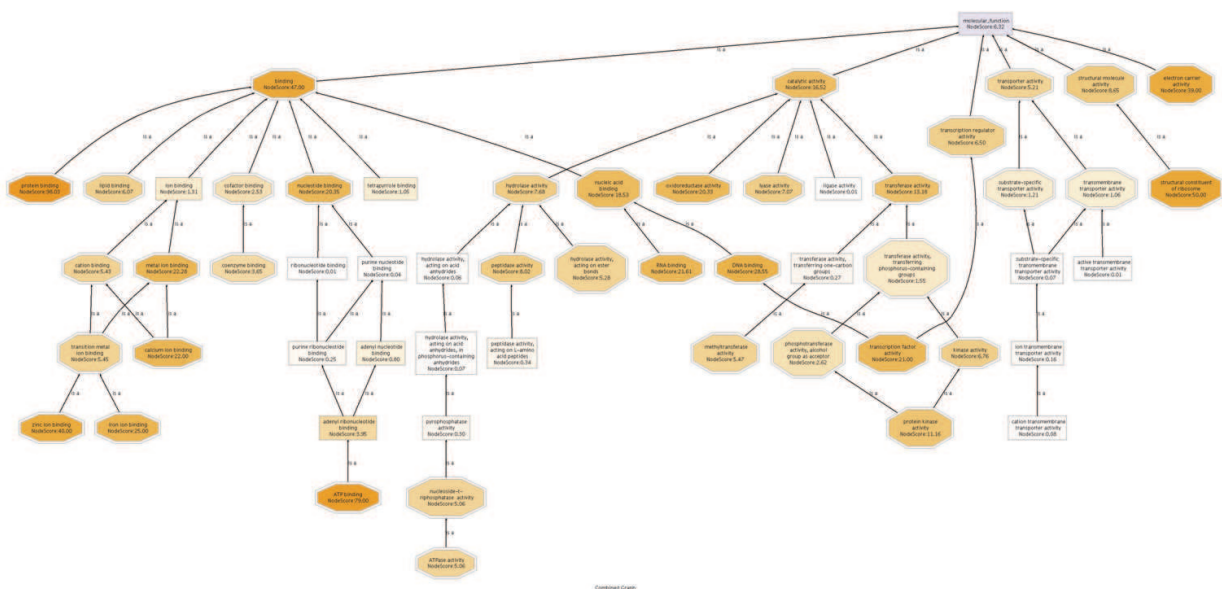- gp(g) is the number of gene products assigned to a given GO term g

Fig. 4. A GO molecular function DAG of 1000 sequences, representing only terms with at least 20 sequences, colored by node score.

3.3.2 Trimming and pruning

Graph coloring, though very helpful to enhance visualization, might not be sufficient to provide handy views of the GO information when many terms have to be displayed. In these cases large and populated graphs are generated and navigation turns tedious or color gradients become too flat to provide clear differences between nodes. In such scenarios it seems more indicated to reduce the size of the graph in a way that relevant information is retained and unimportant terms are hidden. Blast2GO introduces two possibilities for decreasing graph size by using the two node metrics computed in combined graphs. Users can set a filter on the number of sequences annotated to a given node so that only nodes bearing more than a given threshold will be displayed. This results in a "trimming" effect on

the GO DAG. The tree is reduced from the tips and only functional terms at the higher levels of the Ontology are retained. This option can be interesting to summary information and only show functionalities with a sufficient number of members, but at the cost of loosing specific, informativerich terms. Moreover, since different areas of the Gene Ontology are differently developed, trimmed graphs are often uneven in the depth of their different branches. The other option is to filter on the basis of the Blast2GO node-score hiding nodes that do not surpass a specified value. The result here is a pruning effect: the graph is diminished from intermediate nodes that carry redundant information and the resulting tree is enriched in highly informative nodes. Pruning has, however, one potential drawback: since many connecting nodes may get hidden the biological context of the ontology might be affected, having the selected terms a harder task in facilitating the biological interpretability of the genomic dataset. In practice, a balance between the two filtering options has to be achieved to obtain compact and informative graphs which are easy to interpret.

3.3.3 Summarizing, GO slims

Another widely adopted approach for making large GO graph easier to handle is summarizing by a **GO slim**. GO slims are cut-down versions of the GO ontologies containing a subset of the terms of the whole GO. They provide a broad overview of the functional content of a large dataset without the detail of the specific fine grained terms. The process of summarizing a set of GO terms by GO slim implies that terms have to be *mapped* to their corresponding GO slim counterparts and, logically, several GO terms map to the same GO slim term. The result is that a DAG of thousands of nodes can be *condensed* or *slimmed* to a few dozen of key terms, which makes the graph navigable and easy to evaluate. Different GO slim mappings are available that offer different specifications to summarize GO information. Usually these GO slim mappings are specified for organisms or biological domains, such as 'yeast', 'plant' or 'oryzae'. A list of GO slims and a Perl script for mapping can be found at the Gene Ontology site (http://www.geneontology.org/GO.slims.shtml). The **Blast2GO** software offers the functionality to map and represent GO slim summaries of an annotation dataset. Up to 8 different slims are supported and once mapping is completed the regular Blast2GO graphical and enhancing functions (Combined Graph, coloring, filtering, etc) can be applied to the new annotation dataset.

In comparison to the previous graph modification functions, GO slims represent a more static summary of the functional data, since the terms that get displayed are fixed for a giving mapping version. This, which is generally not the most desired option for data mining purposes can be handy when different datasets need to be compared or when a standard summary of the functional content of a dataset needs to be reported.

Figure 5, presents three DAGs of the Molecular Function Ontology of the same 1000 sequences. The first graph is unfiltered to show complexity of the information, the second graph shows the functional information after having applied a GO slim reduction. The third graph is filtered and thinned depending on the number of sequences belonging to each GO-term and the node-score. All GO terms with less than 10 sequences were removed (tip nodes) and all the nodes with a node-score smaller than 12 applying an alpha of 0.4 were removed (intermediate nodes).
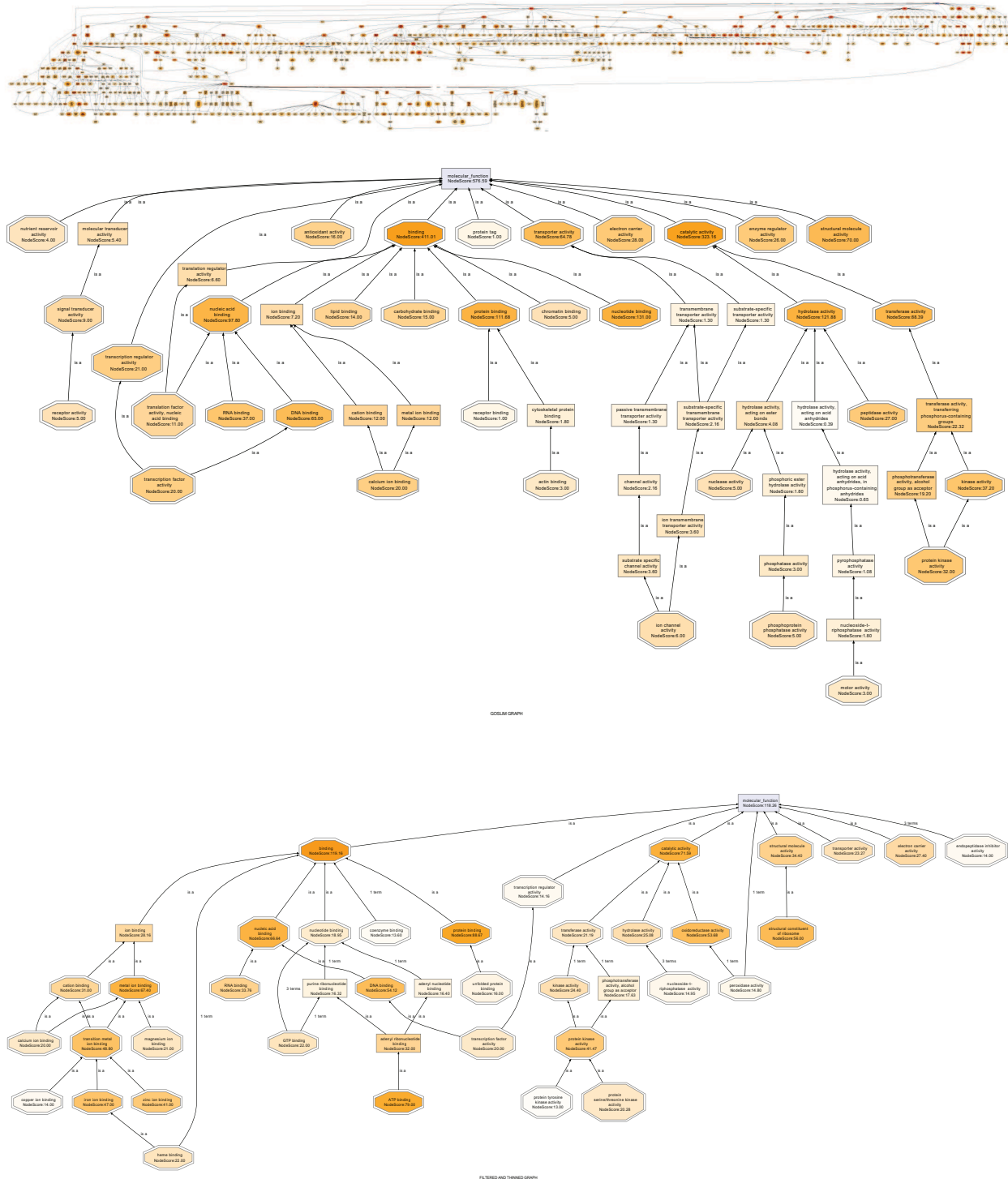
Fig. 5. The molecular functions of 1000 sequences visualized in 3 different ways: The first graph is unfiltered, the second graph shows the functional information after having applied a GO slim reduction and the third graph is filtered and thinned according to the number of sequences belonging to each GO-term and the node-score. All GO terms with less than 10 sequences (tip nodes) and all intermediate terms with a node-score smaller than 12 (with a=0.4) were removed.

3.3.4 Pie charts

Although the GO graph is the most suitable layout for analyzing the functional meaning of an annotated dataset, scientists find some times convenient other kind of representations when presenting and publishing their results. Pie- and less frequently bar-charts are commonly used. Translation of the Combined Graph of a group of sequences into a pie chart, is however, not trivial. As DAGs have multiple hierarchy levels and pies are "flat" representations, some kind of projection or term selection has to be made before creating a pie. One common option is to "cut" the DAG at a selected level and represent terms at this position as a pie chart. The size of the slides will be proportional to the number of annotated genes at each term belonging to the chosen level. This approach poses two main problems. Firstly, it generates size inconsistent diagrams: since one given sequence can be annotated to multiple terms, the sum to the sizes of the different pie slides might be different from the total number of genes in the dataset. This is in slight contradiction with the notion of pie chart (as the distribution of a whole among several classes) and therefore GO-derived pies much be considered as radial flat representations of a graph structure rather than "proper" pie charts. Secondly, the different slides of the pie may not represent terms semantically or informationally equivalent. This is due to the fact that different areas of the GO has different intensity of development. For example, on some branches third level terms may represent very specific notions since in other areas could refer to highly generic concepts. A way to circumvent this problem is to cut different branches of the GO at different levels depending on some value that is indicative of the information content of the node (see Figure 6). Blast2GO offer both options to generate pie charts from Combined Graphs: one-level pie (see Figure 7.a) where the chart contains terms of only one level of the GO and multi-level pie (Figure 7.b). For this last case two different metrics can be use in Blast2GO to indicate the
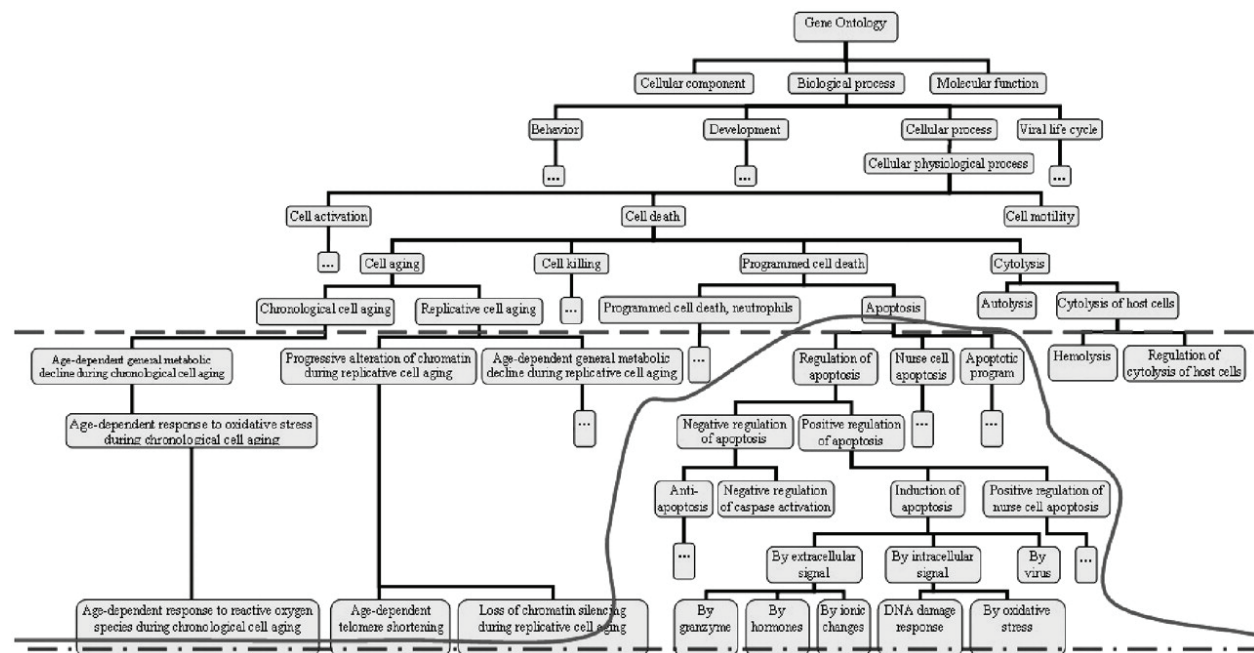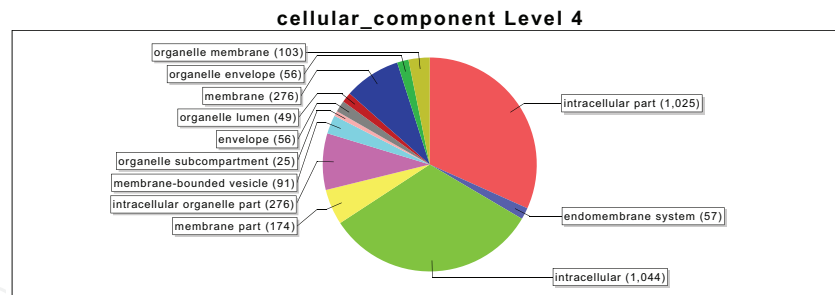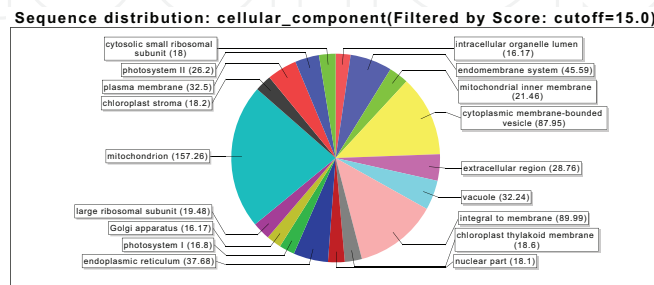

Fig. 6. The multi-level pie function allows instead of a fixed/lowest level of abstraction (dashed line/dash-and-dot line) a custom level of abstraction with different depths in various sub-trees of the GO (continuous line) depending on a user-specified weight criteria (Figure taken from: Khatri & Draghici (2005))

(a) cellular component GO pie at level 4

(b) multi-level pie at node-score 15

Fig. 7. Two different pie charts generated with Blast2GO. (a) shows the functional information of a given annotation set at GO level 4 while (b) shows all the GO terms which surpassed the node-score filter (cutoff=15).

"node-importance" and in this way the point of sectioning: the pie will contain the most specific node of each branch that surpass a user specified cut-off level for the number of sequences or the node-score. So generated pies show a greater diversity of functions that have more similar weights withing the dataset.

3.3.5 Visualization of statistical data

A last scenario where visualization on the DAG has gained popularity is for the display of statistical results of functional genomic data-mining approaches. In this case the relevance of a node is not in the number of annotated sequences, but because some additional experimental data has indicated the association of functions to a certain phenotype. The most popular test is the enrichment analysis of gene expression data. Functional enrichment has been made available at Blast2GO by the integration in the application of the Gossip algorithm (Blüthgen et al., 2005). Gossip computes a Fisher's Exact Test (Fisher, 1922) applying robust FDR (False Discovery Rate) correction for multiple testing and returns a list of significant GO terms ranked by their adjusted p-values. Results are represented in Blast2GO in tabular format or are directly visualized on the DAG, colouring by significance value (see 16 for an example). This kind of "enriched graphs" typically displays highlighted branches of terms semantically related. Analysis of the colored terms within branches permits to identify the level of functional specificity that associated to the phenotype, as well as the level of abstraction where this association is lost. In addition of the Fisher's Exact Test, Blast2GO supports the visualization on the DAG of any kind of statistical result or numeric value directly associated to GO functions. GO terms can be uploaded in a tab-separated format with a column containing the numerical statistical value scaled between 0 and 1. The
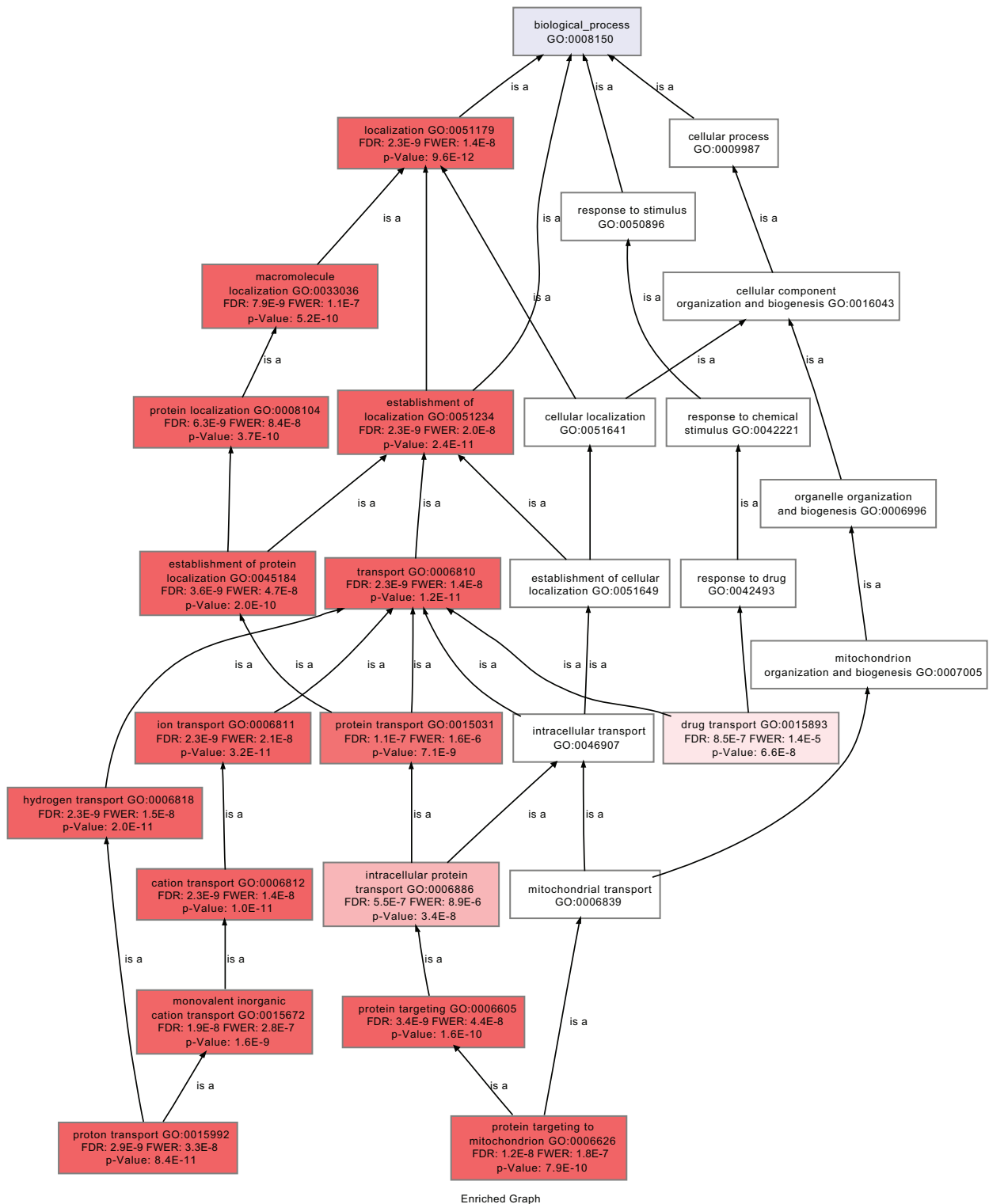
color scale is set form dark red (0) to dark blue (1) passing through white (0.5). A good example of such a use can be found in Conesa et al. (2008). Biological functions related to several clinical measurements were identified through multivariate projection strategies and visualized by a DAG colouring the significant GO terms on a scale from red to blue depending on their significance (see Figure 8).

Fig. 8. GO DAG (part) of the pool of significant terms (biological processes) detected for functional variables by the statistical model. Term color intensity is proportional to the importance of the functional class in the model. Hexagonal nodes are the actual selected GO terms.

Fig. 9. Enriched graph of the subset of genes expressed in membrane in Soybean. Node filter has been set at FDR ≤ 1 · 10⁻⁶. Nodes are colored accordingly to their FDR value in the Fisher's Exact Test against the whole Soybean genome array. Figure taken from (Conesa & Götz, 2008).

4. Concluding remarks

Visualization under the frame of the Gene Ontology Graph is an effective data mining approach for analyzing large functional genomics datasets as it provides functional information within a broad biological context. The different enhancing, summarizing and exploring functionalities described in this chapter are valuable for granting easy access to the most relevant functional information and facilitating the process of knowledge discovery. These options are readily available in the Blast2GO application, an up-to-date bioinformatics tool for advanced GO graph exploring. As the Gene Ontology further expands to include new relationships between terms, links among branches or richer annotation data, and novel metrics become available for assessing and comparing functional datasets, additional graphical features will need to be developed.

5. References

Al-Shahrour F, Díaz-Uriarte R, D. J. (2004). Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes, *Bioinformatics* 20: 578–580.

Ashburner M, Ball CA, B. J. B. D. B. H. C. J. D. A. D. K. D. S.-E. J. H. M. H. D. I.-T. L. K. A. L. S. M. J. R. J. R. M. R. G. S. G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium, *Nat Genet* 25: 25–29.

Baehrecke EH, Dang N, B. K. S. B. (2004). Visualization and analysis of microarray and gene ontology data with treemaps, *BMC Bioinformatics* 5: 84.

Binns D, Dimmer E, H. R. B. D. O. C. A. R. (2009). Quickgo: a web-based tool for gene ontology searching, *Bioinformatics* 25: 3045–3046.

Blüthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H. & Beule, D. (2005). Biological profiling of gene groups utilizing gene ontology., *Genome Informatics* 16(1): 106–115. URL: *http://view.ncbi.nlm.nih.gov/pubmed/16362912*

Carbon S, Ireland A, M. C. S. S. M. B. L. S. A. H. W. P. W. G. (2009). Amigo: online access to ontology and annotation data, *Bioinformatics* 25: 288–289.

Conesa, A., Bro, R., García-García, F., Prats, J. M. M., Götz, S., Kjeldahl, K., Montaner, D. & Dopazo, J. (2008). Direct functional assessment of the composite phenotype through multivariate projection strategies., Genomics 92(6): 373–383. URL: http://dx.doi.org/10.1016/j.ygeno.2008.05.015

Conesa, A. & Götz, S. (2008). Blast2go: A comprehensive suite for functional analysis in plant genomics, *International Journal of Plant Genomics* 2008: 1–13. URL: *http://dx.doi.org/10.1155/2008/619832*

Conesa A, Götz S, G.-G. J. T. J. T. M. R. M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics* 21: 3674–3676.

Dennis G Jr, Sherman BT, H. D. Y. J. G. W. L. H. L. R. (2003). David: Database for annotation, visualization, and integrated discovery, *Genome Biology* 4: P3.

Eran Eden, Roy Navon, I. S. D. L. & Yakhini, Z. (2009). Gorilla: A tool for discovery and visualization of enriched go terms in ranked gene lists, *BMC Bioinformatics* 10: 48.

Fisher, R. A. (1922). On the interpretation of "chi" from contingency tables, and the calculation of p, *Journal of the Royal Statistical Society* 85(1): 87–94.

URL: *http://dx.doi.org/10.2307/2340521*

Götz S, García-Gómez JM, T. J.W. T. N. S. N. M. R. M. T. M. D. J. C. A. (2008). High-throughput functional annotation and data mining with the blast2go suite, *Nucleic Acids Research* 36: 3420–35.

Khatri, P. & Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics* 21(18): 3587–3595.
URL: *http://dx.doi.org/10.1093/bioinformatics/bti565*

Medina I, Carbonell J, P. L. M. S. G. S. C. A. T. J. P.-M. A. N.-C. R. S. J. G. F. M. M. M. D. D. J. (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling, *Nucleic Acids Research* 38 Suppl: W210–3.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy, *IJCAI 1995*, pp. 448–453.
URL: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.5277*

Supek F, Skunca N, R. J. V. K. S. T. (2010). Translational selection is ubiquitous in prokaryotes, *PLoS Genet* 6: e1001004.
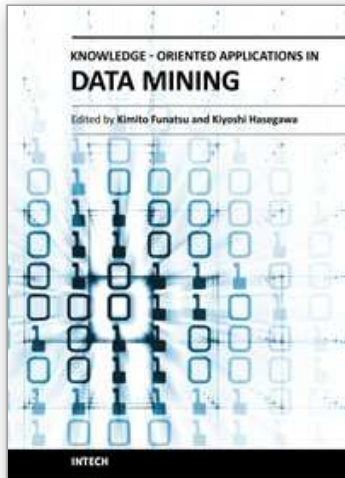
Knowledge-Oriented Applications in Data Mining

Edited by Prof. Kimito Funatsu

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Stefan Götz and Ana Conesa (2011). Visual Gene Ontology Based Knowledge Discovery in Functional Genomics, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech, Available from: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/visual-gene-ontology-based-knowledge-discovery-in-functional-genomics

INTECH
open science | open minds