Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations

Anand KUMAR
IFOMIS, Faculty of Medicine,
University of Leipzig,
Haertelstrasse 16-18
Leipzig, Germany, D-04107
anand.kumar@ifomis.uni-leipzig.de
URL:
http://www.uni-leipzig.de/~akumar/

Barry SMITH[1,2]
[1]IFOMIS, Faculty of Medicine,
University of Leipzig,
Haertelstrasse 16-18
Leipzig, Germany, D-04107
[2]Department of Philosophy,
SUNY at Buffalo,
Buffalo, NY 14260, USA
phismith@buffalo.edu
URL:
http://ontology.buffalo.edu/smith//

Christian BORGELT
Department of Knowledge
Processing and Language
Engineering, Otto-von-Guericke
University of Magdeburg,
Universitätsplatz 2,
Magdeburg, Germany, D-39106
christian.borgelt@cs.uni-magdeburg.de
URL:
http://fuzzy.cs.uni-magdeburg.de/~borgelt/

Abstract

The Gene Ontology is an important tool for the representation and processing of information about gene products and functions. It provides controlled vocabularies for the designations of cellular components, molecular functions, and biological processes used in the annotation of genes and gene products. These constitute three separate ontologies, of cellular components), molecular functions and biological processes, respectively. The question we address here is: how are the terms in these three separate ontologies related to each other? We use statistical methods and formal ontological principles as a first step towards finding answers to this question.

1 Introduction

(Zhu et al 2004) noted that creating a model of the dynamics of molecular interaction networks offers enormous potential for understanding systems biology. Existing work has led to the development of databases and ontologies which provide classifications and annotations based on a gene product's function, location, structure and so on, as for example in PANTHER (Thomas PD et al 2003), a library of protein families and subfamilies indexed by function, and the Gene Ontology Annotation[1] (GOA) (Camon et al 2003).

Further progress requires a robust formal ontology of structures, locations, functions and processes, linked together via relations such as *is_part_of*, *is_located_at*, *is_realized_by*, and so forth. As a step along this road, we provide a methodology for deriving and representing association rules between the entities present within the separate ontologies of the Gene Ontology.[2] (Gene Ontology Consortium, 2001). Such rules will be able to situate a biological process in relation to a cellular location to an agent. They will be able to relate lower-granularity molecular functions in relation to higher-granularity biological processes, and establish other sorts of relations between entities in different parts of GO.

A preliminary study in this area (Burgun et al 2004) combines ontological, lexical and statistical principles. Their study provides association rules on a selected set of 23 gene products that were potentially involved in enterocyte differentiation and that showed similar levels of expression. (Clelland and Oinn) provide commonly annotated terms based on the CluSTr database (Kriventseva et al 2001), which has recently been incorporated into the QuickGO browser.[3] Association rules have been used for mining gene expression data by (Creighton and Hanash 2003). (Ogren et al 2004) studied the compositional nature of the GO terms and described the dependencies among them.

Our investigation draws on the fact that terms from GO's separate ontologies are often used to annotation the same gene or gene product. We draw on the TIGR database to establish the corresponding patterns of association between terms in GO when taken in its entirety.

In what follows we describe the results of this work We analysed all of TIGR's 84,833

[1] http://www.ebi.ac.uk/GOA/

[2] http://www.geneontology.org/

[3] http://www.ebi.ac.uk/ego/

annotations, pertaining to the 41,502 distinct gene products present within GOA and focusing on the TIGR database within the February 2004 edition of GO. These associations were mined to establish association links between GO terms using standard statistical database techniques based on the so-called apriori algorithm and using a part of speech tagger. The discovered links were then analysed on the basis of methods drawn from foundational ontology.

2 Gene Ontology

2.1 The Cellular Component Ontology

GO's *cellular component (*cc) vocabulary consists of terms such as *flagellum*, *chromosome*, *ferritin*, *extracellular matrix* and *virion*. This ontology is the GO counterpart of anatomy within the medical framework. GO includes in this vocabulary both the extracellular environment of cells and the cells themselves (that is, *cell* is subsumed in GO by *cellular component*).

2.2 The Molecular Function Ontology

GO's definition of *molecular function* (mf) is: "the action characteristic of a gene product." The mf vocabulary accordingly subsumes terms describing actions, for example: *ice nucleation*, *binding*, or *protein stabilization*.

2.3 The Biological Process Ontology

A *biological process* (bp) is defined in GO as: "A phenomenon marked by changes that lead to a particular result, mediated by one or more gene products". Terms in bp can be quite specific (*glycolysis*) or very general (*death*). GO's mf and bp terms are clearly closely interrelated. The biological process of *anti-apoptosis*, for example, certainly involves the molecular function now labelled *apoptosis inhibitor activity*. Such molecular functions should stand to biological processes in a *part-of* relation. At the same time, however, GO's authors insist that *part-of* holds only between entities within a single vocabulary, and they thus provide no guidance as to the cross-vocabulary relations between the terms. We published a series of papers pointing out these and similar problems in GO as currently constituted (Smith et al 2004; Smith et al 2003; Kumar and Smith 2004; Kumar and Smith 2003).

3 TIGR database annotations

The Institute for Genome Research (TIGR)'s Genome Projects are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans. (http://www.tigr.org/) TIGR has manually curated GO annotation for 6 bacterial genomes (V. cholerae, S. oneidensis, B. anthracis, G. sulfurreducens, P. syringae, and C. burnetii) and two eukaryotes (Arabidopsis thaliana, and Trypanosoma brucei). In addition, automated annotation has effected for Expressed Sequence Tags from several species.

The TIGR database is a rich source of information about gene indices based on genetic sequence. TGICL is a pipeline for the analysis of large Expressed Sequence Tags (EST) and mRNA databases in which the sequences are first clustered on the basis of pairwise sequence similarity and then assembled by individual clusters (Pertia et al 2003). Association rules between GO terms will enable us to determine the clusters of gene expression functions and locations in a way that will add to the knowledge that is contained within representations of such clusters on the basis of the gene indices only.

4 Methods

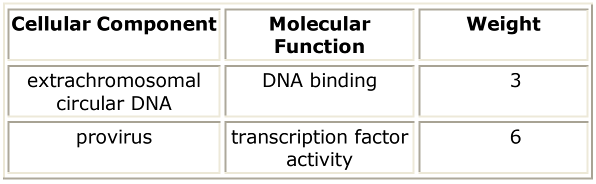
Associations between GO terms were established on the basis of the annotations in the TIGR databases.

4.1 Statistical approach

All the annotations from the TIGR database present within GO's association table were selected and placed into a separate table GO terms were then separated into three separate tables, depending on which of the three GO vocabularies they belong to.

Those GO terms which belong to two different ontologies within GO but are annotated to the same gene products were then separated out for analysis. Three new tables were then created containing those annotations where cc and mf terms, mf and bp terms, and cc and bp terms are annotated together.

The distinct term tuples present were grouped together and their count was used to provide a measure for weighting an association – which is to say how many times two GO terms from two distinct axes are annotated together (Table 1). The co-occurrence of terms within the annotations were then combined together (Table 2).



Cellular Component	Molecular Function	Weight
extrachromosomal circular DNA	DNA binding	3
provirus	transcription factor activity	6

Table 1. Associations between terms belonging to cc and mf, together with an index of how many times such associations occur within the annotation

Cellular Component	Biological Process	Molecular Function	Weight
nucleus	chromosome segregation	ATP binding	19
chloroplast stroma	signal transduction	MAP kinase activity	33

Table 2. Associations between terms belonging all three of GO's ontologies with an index of how many times such associations occur within the annotation

GO's hierarchy has thus far not been considered. Rather we have focused only on the terms themselves to which the annotations are made. For each GO term used for annotation and for each subsuming term in GO's *is_a* hierarchy, we can establish the distance of the former from the latter. (Resnik 1995) has pointed out that the semantic similarity of terms as one traverses the hierarchical tree reduces by a factor of log(p(c)) where p(c) is the probability of finding a child for the term when seeking information. Table 3 thus represents a quantification of this semantic similarity, which can be used to extend the results presented here by using an approach similar to that advanced in (Azuahe and Bodenreider, 2004).

Cellular Component	Molecular Function	Distance
DNA polymerase complex	Epsilon DNA polymerase activity	1
oxoglutarate dehydrogenase complex	oxoglutarate dehydrogenase (lipoamide) activity	1

Table 3. Associations between GO terms belonging to cc and mf taking hierarchy into consideration

4.2 Association rule induction

Association rule induction was originally developed for so-called market basket analysis, which aims at finding regularities in the shopping behaviour of customers of supermarkets, mail-order companies, online stores etc. (Borgelt and Kruse, 2002) Association rules are designed to help in isolating those sets of products that are frequently bought together. This information is expressed in the form of rules like "A customer who buys bread and wine is likely to buy cheese also."

Algorithms for inducing association rules from a set of transactions (the market baskets or shopping carts bought by customers) usually work in two steps: First, the so-called frequent item sets are determined by searching the subset lattice of all items. For this search there are, in principle, two approaches: the *breadth first* search, as employed by the apriori algorithm (Agrawal et al. 1994), and the *depth first* search, on which the eclat algorithm (Zaki et al. 1997) is based. Second, rules are constructed from the frequent item sets and filtered with respect to some quality criterion.

In the GO context, we use association rule induction to discover links between the ontologies in which a gene product is described. That is, we are interested in rules that predict cc from mf or bp, or rules that describe which mfs in which ccs constitute a bp, and so forth.

4.2.1 Assessment of association rules

The number of combinatorially possible rules here is very high. Consequently we need criteria to assess and thus to filter out those association rules which are of serious importance. The standard measures for this purpose are the *support* and the *confidence* of a rule.

The support of an association rule can be defined in two different ways: as the fraction of all transactions to which the rule is applicable (that is, which contain all items in the antecedent of the rule) or the fraction of all transactions for which the rule is correct (that is, which contain *all* items appearing in the rule, regardless of whether in antecedent or consequent). We take the second approach.

The confidence of an association rule is the fraction of cases in which it is correct relative to those in which it is applicable, that is, the ratio of the number of transactions that contain all items in the rule to the number of transactions that contain all items in the antecendent.

A user controls the search for association rules by providing minimum values for support and confidence of the rules to be found. From these values there can be derived a pruning criterion for the search for frequent item sets in the subset lattice as well as filtering criteria for the rules themselves. Only rules that meet both criteria are reported. More intuitively: we are looking for association rules that can be applied often (are above some minimum level of support) and that make reliable predictions (are above some minimum level of confidence).

4.2.2 Application to Gene Ontology

In our application of association rule induction to the analysis of GO we viewed each gene product annotated to a GO term as a transaction. We thus have as many transactions as there are gene products in the data set. We then looked for frequent co-occurrences of terms within such transactions and reported them in the form of association rules. For the search itself we used a well-known implementation of the apriori algorithm by (Borgelt and Kruse 2002).

In addition to filtering based on minimum support and minimum confidence, we also introduced one further selection criterion by requiring that the consequent of a rule must come from a different GO ontology than the terms in the antecedent of the rule, since rules containing only terms from one ontology are likely to recover only the term hierarchy of that ontology, which is not what is needed here. Such filtering can be achieved with the apriori implementation we used via the specification of which part of an association rule an item may appear in.

4.3 Experimental results

We ran the apriori program three times, each time restricting the consequent of the rules to a different ontology. This was required as GO has three orthogonal ontologies and one needed to treat each of them as an antecedent and a consequent with one another. Examples of rules we found are:

membrane [cc]
← oligopeptide transport [bp];
transporter activity [mf]
(0.106%/44, 100.0%)

binding [mf]
← mitochondrial transport [bp];
mitochondrial inner membrane [cc]
(0.106%/44, 100.0%)

protein biosynthesis [bp]
← ribosome [cc];
structural constituent of ribosome [mf]
(0.504%/209, 90.4%)

The two letters in brackets after each term denote the axis the term comes from. The numbers in parentheses at the end of each rule describe the quality of the rule as (S%/A, C%), where S is the support of the rule as a percentage of gene products to which the rule is applicable, A is the absolute number of gene products to which it is applicable (which is designed to complement the information regarding support), and C is the confidence of the rule.

In the GO context S is the percentage of gene product IDs to which the relevant rule is applicable (i.e., the percentage of gene product IDs which are annotated with all the terms in the antecedent in the rule), A is the absolute number of gene product IDs to which the rule is applicable, and C is the percentage of gene product IDs for which the rule makes the correct prediction relative to those to which the rule is applicable. Thus for example:

ribosome [cc]
← ribosome biogenesis [bp];
protein biosynthesis [mf]
(0.212%/88, 93.2%)

tells us that 88 gene product IDs (~0.2%) are annotated with the terms *ribosome biogenesis* and *protein biosynthesis*, of which 93.2% (i.e. 82) are also annotated with the term *ribosome*.

4.4 Dependencies based on Part-of-speech Tagging:

The terms within GO were tagged with a part-of-speech tagger Qtag[4] (Mason, 2004) in order to understand the linguistic dependencies between the terms (for example, between adjective and noun or between adverb and verb).

We tagged all the GO terms and also their definitions. The definition tags supplement the GO term tags since definitions are more elaborate and the specificity of the tagger also increases when it deals with complete sentences rather than with the collections of words by which GO terms are constituted.

Among the terms, we find 1813 adjectives, 29 adverbs, 2808 nouns, 11 prepositions and 57 verbs. Among the term definitions, we found 3460 adjectives, 252 adverbs, 4837 nouns, 21 prepositions and 595 verbs. The lowest specificity is found for chemical names.

Unfortunately very few of the interesting part-of-speech generated dependencies can be captured on the basis of an analysis of GO terms alone. This is because there are very few cases of such dependencies where both terms involved are present within GO. Examples are: *heme transport* and *heme transporter activity* or *growth* and *invasive growth.* In many other cases, however, we have complex GO terms whose constituents are not themselves present in GO. Thus *photoreactive repair* is present, but not *repair* or *photoreaction. Hypersensitive response* is present but not *response* and *hypersensitivity*. Terms like *during*, *within* and *without* play an important role in GO's compositional structure, but they are not themselves present within GO. (For more examples see (Smith et al. 2004).) To rectify this defect and to make the corresponding information accessible to software applications ways must be found to link GO to third-party ontologies in which the corresponding constituent terms are themselves subjected to formal treatment.

[4] http://web.bham.ac.uk/O.Mason/software/tagger/

Our statistical approach, apriori-generated association rules and lexical tagging yielded a range of different sorts of associations, for example between a location and a process, a process and a function, a function and another function, a process and another process, an agent and a function and so on. One needs a formal ontology, too, in order to express those relations within a single framework and thus to create a robust representation scheme that can serve as the basis for automated reasoning.

5 Basic Formal Ontology

To do full justice to the information content of GO and to the rules we isolated requires a formal ontological scheme which encompasses both continuants and endurants, and both processes and functions, and which further has the facility to deal with entities found on different levels of granularity (here on the molecular, cellular, and whole-organism levels). Basic Foundational Ontology (BFO) is a framework of this type, the essentials of which can be summarized as follows.

BFO consists of two complementary ontologies, called SNAP (a snapshot ontology of continuant entities existing at a time) and SPAN (a four-dimensional ontology of processes unfolding themselves through time. (Smith and Grenon, forthcoming)

The entities recognized by SNAP ontologies have continuous existence in time, preserve their identity through change and exist in toto at every moment at which they exist at all. They include: independent SNAP entities (substances and their aggregates, parts, and boundaries), and dependent SNAP entities such as qualities, roles, conditions, functions, dispositions, powers, etc.

SPAN entities, in contrast, have temporal parts which means that they unfold themselves in successive phases and can be segmented via segmentation of the temporal intervals which they occupy. SPAN entities include processes in the narrow sense, as well as the instantaneous temporal boundaries of processes, the temporal extents of processes, and so on.

5.1 Formal-Ontological Relations

Formal relations are those types of relations which can traverse the SNAP-SPAN divide; thus they are relations which glue SNAP and SPAN entities together.

A number of parameters can then be used in the construction of sub-ontologies within the wider BFO framework:

– the ontologies from which the relata derive, expressed as an ordered list, called the *signature* of the relation

– the *directionality* of the relation

The principal signatures in the binary case are as follows:

– <SNAP, SNAP>
– <SPAN, SPAN>
– <SNAP, SPAN>
– <SPAN, SNAP>

The first two signatures comprehend relations between ontologies with different domains or granularities. The latter comprehend the relations of realization and participation for example between a function and the process which is its functioning, or between an activity and its agent.

Below, we present various relations which can exist between entities in GO and which were found on the basis of our analysis of the TIGR database. Some of the relations require extending the association rules to include those entities which are absent within GO but included within its compositional structure along the lines described above.

5.1.1 Relations with Signature <SNAP, SNAP> and <SPAN, SPAN>

Transgranular Part-Whole Relations: The relations crossing ontologies of different levels of granularity are pre-eminently relations of part and whole. The <SNAP, SNAP> relations between independent SNAP entities are already present within GO. Example: *nuclear inclusion body* part-of *nucleus*.

Our association rules uncovered transgranular part-whole relations between dependent SNAP entities. For example: *transposase activity* ← DNA *transposition* (1.2%/505, 91.5%). Unlike BFO however GO does not clearly distinguish between functions and processes; hence the above example should be interpreted as involving both a transgranular relation between SNAP dependent entities:

transposase function
← DNA *transposition function*

and a transgranular relation between SPAN entities:

transposase activity
← DNA *transposition function*

This does justice to the fact that DNA transposition activity consists of various activities as its parts, of which is transposase activity.

5.1.2 Relations with Signature <SNAP Independent, SPAN>

Participation: The relation of participation, for example of a runner in a race, is a species of dependence. There are different kinds of participation, which we can order along the following dimensions:

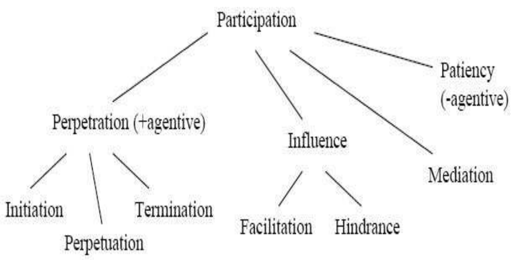


Figure 1. Modes of Participation

Perpetration: A substance perpetrates an action (direct and agentive participation in a process).
Initiation: A substance initiates a process. Example:.

electron transport ← *electron transporter activity* (0.3%/137, 96.4%)

needs to be extended (by drawing on those terms which form part of GO's compositional structure but are not themselves included in GO) by:

electron transport ← *electron transporter*

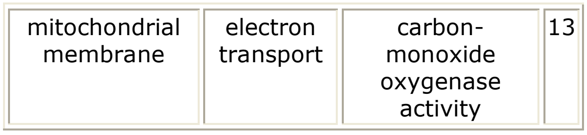
Perpetuation: A substance sustains a process. Perpetuation normally presupposes that at some earlier time an entity entered into the relation of initiation with the process in question. However, perpetuators are of course not by necessity themselves initiators. Example,

proteolysis and peptidolysis ← *subtilase activity (0.2%/63, 96.8%),*

which again needs to be extended by:

proteolysis and peptidolysis ← *subtilase*

Termination: A substance terminates a process (for example an operator terminates the projection of a film). Termination normally presupposes that there has obtained in the immediately prior interval of time a relation of perpetuation. Processes cannot pass from initiation to termination instantaneously. Example:



mitochondrial membrane	electron transport	carbon-monoxide oxygenase activity	13

In the above example, the association between carbon-monoxide oxygenase and electron transport are that of termination. There is an additional association to the location of this process in the mitochondrial membranes.

Facilitation: A substance plays a secondary role in a process. Example:

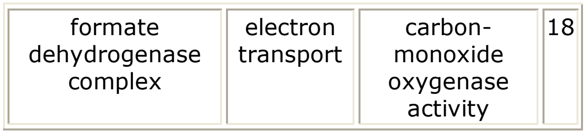
protein biosynthesis ← *large ribosomal subunit (0.1%/56, 98.2%)*

Hindrance, prevention: A substance has a negative effect on the unfolding of a process. Example:

DNA transposition ← *transposase activity (1.2%/505, 91.5%).*

While transposase initiates DNA transposition, it also causes an auto-inhibition later.

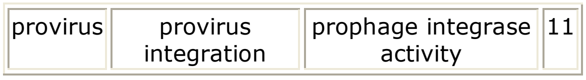
Mediation: A substance plays an indirect role in the unfolding of a process relating other participants. Example:



formate dehydrogenase complex	electron transport	carbon-monoxide oxygenase activity	18

In this relationship, formate dehydrogenase plays the role of a mediator of electron transport.

Patiency: A substance is being acted on by a process. Example:



provirus	provirus integration	prophage integrase activity	11

In this relationship, provirus is being integrated by the process of provirus integration.

5.1.3 Relations with Signature <SNAP Dependent, SPAN>

Realization: There are three main modes of realization, which result by applying the distinctions dealt with in 5.1.2 (initiation, termination and persistence) but substituting "function" for "activity".

Initiation: *electron transport* ← *electron transporter function*

Termination. *electron transport* ← *carbon-monoxide oxygenase function*

5.1.4 Relations with Signature <SPAN, SNAP>

Relations between Processes and Substances

Creation: A process brings into being a substance. Example:

integral to membrane	steroid biosynthesis	3(or 17)beta-hydroxysteroid dehydrogenase activity	2

As in the previous cases, this association can be extended to:

steroid ← *3(or17)beta-hydroxysteroid dehydrogenase activity*

Sustaining in being: A process sustains in being a substance.

membrane	cell growth and/or maintenance	protease inhibitor activity	60

The protease inhibitor activity maintains the membrane structure. Further work needs to be done to understand the granularity of such relations.

Degradation: A process has negative effects upon a substance. Example:

D-amino-acid dehydrogenase complex	amino acid metabolism	lyase activity	2

This relation associates a lyase activity with the degradation of D-amino-acid dehydrogenase complex.

5.1.5 Spatiotemporal projection

Temporal Projection. Processes are directly projectible onto the axis of time. And a substance is indirectly projectible onto a period of time through the mediation of a process in which it is involved. Examples:

Biological process	Molecular function	Weight
acetyl-CoA biosynthesis from pyruvate	dihydrolipoamide dehydrogenase activity	3
acetyl-CoA biosynthesis from pyruvate	dihydrolipoamide S-acetyltransferase activity	5
acetyl-CoA biosynthesis from pyruvate	phosphate acetyltransferase activity	1
acetyl-CoA biosynthesis from pyruvate	pyruvate dehydrogenase (lipoamide) activity	5
acetyl-CoA biosynthesis from pyruvate	pyruvate dehydrogenase activity	5

While we cannot infer such projections directly from a single association rule, we can get a rough approximation if we consider more than one rule together.

Spatial Projection. Processes are projectible also onto the SPAN spatiotemporal regions in which they occur, as also onto the (SNAP) spatial regions where they start and end. Examples:

membrane ←*oligopeptide transport; transporter activity (0.1%/44, 100.0%)*

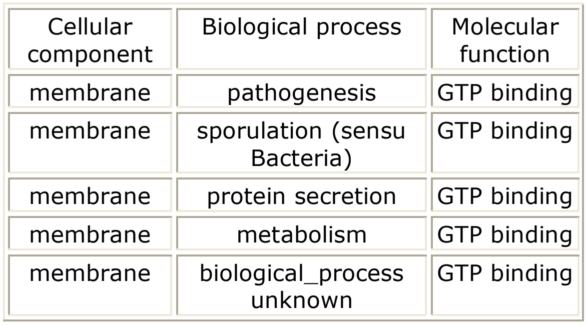
ribosome ←*ribosome biogenesis; protein biosynthesis (0.2%/88, 93.2%)*

6 Discussion

The association rules yielded by our methodology are only the beginning of a process of deciphering the ontological relations across various granularities within the extended GO framework. We need to formally analyse all the rules in order to understand how best they fit within formal ontology and how they can be put together to create larger ontologies within systems biology.

In validating these associations one method is to consider those associations which have been detected on the basis of the annotations present within GO coming from other source databases. Clearly if a rule obtains across a plurality of databases then the corresponding association will be stronger. One disadvantage of this method, however, is that annotations of gene products from different databases to GO are not uniform and so the results cannot be relied upon beyond a certain limit of accuracy. Various other aspects of annotations, for instance, the species under investigation or the pertinent strength of evidence will need to be considered.

A further extension of the work will be to predict "unknown" entities. A large number of annotations is made to the three GO terms *cellular component unknown*, *molecular function unknown* and *biological process unknown*. In those caseswhere there we have association rules without such unknown terms but otherwise relating to the same entities, it could be that the relevant unknown entity will be able to be predicted. This can be done with annotations from a single database source or by putting together annotations from multiple sources. We need to do further work in this area. For example, in the following example, we can establish that in the case where mf is *GTP binding* and cc is *membrane*, the known bf is either *pathogenesis* or *sporulation* (sensu *Bacteria*) or *protein secretion of metabolism*. Thus it is probable that the unknown process here corresponds to one or other of these alternatives.



Cellular component	Biological process	Molecular function
membrane	pathogenesis	GTP binding
membrane	sporulation (sensu Bacteria)	GTP binding
membrane	protein secretion	GTP binding
membrane	metabolism	GTP binding
membrane	biological_process unknown	GTP binding

7 Acknowledgements

Work on this paper was carried out under the auspices of the Wolfgang Paul Program of the Humboldt Foundation and also of the EU Network of Excellence in Semantic Datamining and the project "Forms of Life" sponsored by the Volkswagen Foundation.

References

Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases . *Proc International Conference on Very Large Databases*, Santiage, Chile, (Morgan Kaufmann, 1994) 478-499.

Azuaje FJ, Bodenreider O. Incorporating Ontology-Driven Similarity Knowledge into Functional Genomics: An Exploratory Study. *Proc IEEE Fourth Symposium on Bioinformatics and Bioengineering* 2004. (In press)

Borgelt C, Kruse R. Induction of Association Rules: Apriori Implementation. in: *15th Conference on Computational Statistics* (Compstat 2002, Berlin, Germany) Physica Verlag, Heidelberg, Germany 2002

Burgun A, Bodenreider O, Aubry M, Mosser J. Dependence Relations in Gene Ontology: A Preliminary Study. *Gene Ontology Workshop*, Leipzig, May 2004.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*. 2004 Jan 1;32 Database issue:D262-6.

Clelland S, Oinn T. Comparing protein structure and function: mapping CluSTr into GO and analysis of the results. http://www.ebi.ac.uk/ego/project.pdf?&format=simple

Creighton C, Hanash S. Mining gene expression databases for association rules. *Bioinformatics*. 2003 Jan;19(1):79-86.

Gene Ontology Consortium. Creating the Gene Ontology. *Genome Res*. 2001. 11: 1425-1433.

Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M., Apweiler, R. (2001). CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res* 29: 33-36

Kumar A, Smith B. Towards a Proteomics Metaclassification. Proc *IEEE Fourth Symposium on Bioinformatics and Bioengineering* 2004. (In press)

Kumar A, Smith B. The Universal Medical Language System and the Gene Ontology: Some Critical Reflections. *Lecture Notes in Computer Science*. 2003 Sep; 2821/2003: 135 – 148.

Mason, O. Automatic Processing of Local Grammar Patterns. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, University of Birmingham, 6-7th January 2004, p.166-171.

Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The Compositional Structure of Gene Ontology Terms. Pacific Symposium on Biocomputing 2004;9:214-225

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003 Mar 22;19(5):651-2.

Resnik P. “Using information content to evaluate semantic similarity in a taxonomy”, in Proc. of the *14th International Joint Conference on Artificial Intelligence*, Montreal, pp. 448-453, 1995.

Smith B and Grenon P. The Cornucopia of Formal Relations, forthcoming in DIALECTA

Smith B, Koehler J, Kumar A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. in: Proc DILS 2004. (*Lecture Notes in Bioinformatics* Nr. 2994)

Smith, B., Williams, J., Schulze-Kremer, S.: The Ontology of the Gene Ontology. In: *Proc. Annual Symposium of the American Medical Informatics Association* (2003) 609-613

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*. 2003 Sep;13(9):2129-41.

Zhu H, Huang S, Dhar P. The next step in systems biology: simulating the temporospatial dynamics of molecular network. *Bioessays*. 2004 Jan;26(1):68-72.