Proceedings of the Workshop on The Formal Architecture of the Gene Ontology
Leipzig, Germany, May 28-29, 2004.

# DEPENDENCE RELATIONS IN GENE ONTOLOGY: A PRELIMINARY STUDY

ANITA BURGUN[1], OLIVIER BODENREIDER[2], MARC AUBRY[3], JEAN MOSSER[3]

[1]*Laboratoire d'Informatique Médicale, Université de Rennes I*
*Avenue du Pr Léon Bernard 35043 Rennes Cedex, France*

[2]*U.S. National Library of Medicine*
*8600 Rockville Pike, MS 43, Bethesda, Maryland 20894, USA*

[3]*Unité de Génétique Humaine, UMR 6061*
*Avenue du Pr Léon Bernard 35043 Rennes Cedex, France*

Abstract. *The functional interpretation of microarray experiments requires computerized methods that exploit similarities between gene products with respect to their Gene Ontology (GO) annotations. While GO already represents taxonomic and meronomic relations, our objective is to identify associative relations across Gene Ontology hierarchies. The expected benefit of this effort is that these additional relations can be used to produce more consistent annotations. As a first step toward this goal, we analyze dependence relations between GO terms, first in a set of 23 gene products involved in enterocyte differentiation and later in GOA. Our approach takes into account ontological, lexical, and statistical aspects of dependence. Preliminary results suggest the interest of combining these three approaches for accurately identifying ontological and functional dependence relations in GO.*

## 1 Introduction

The biological analysis of pathological conditions should rapidly evolve with the development of micro-array technologies. By exploring the transcription profiles of genes, expression array data provide better diagnosis, prognosis and treatment of diseases. For example, Van de Vijver et al have shown that the gene-expression profile they studied was a more accurate predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histologic criteria [van de Vijver]. In this example, genotype data will improve the selection of patients for adjuvant systemic therapy, previously based on phenotype data.

The results of an expression array experiment consist of large collections of expression profiles that give information on the levels of expression of each gene under various conditions. The first step in expression array analysis is to cluster genes

according to their levels of expression. The underlying assumption is that the genes in a given cluster are related by their participation in common biological mechanisms [Lockhart]. Sets of co-expressed genes can encode products that are involved in a common biological process, and may be localized to the same cellular component.

The second step consists of exploring *functional* similarities in genes that share similar expression patterns (e.g., genes overexpressed or underexpressed under pathological conditions). Valuable sources of information about gene function that can be used to assess the functional coherence of a gene group include the published literature (e.g. [Raychaudhuri], [Blaschke]) and the functional annotations of gene products based on the GeneOntology™ (GO) available in public databanks (e.g. [Tanoue]). The inclusion of GO annotation in microarray datasets provides information about molecular functions, biological processes, and cellular components associated with the gene products.

GO addresses the need for consistent descriptions of gene products in different databases and proposes a controlled vocabulary to represent functional information about gene products [1]. The use of GO terms by many collaborating databases including several of the world's major repositories, facilitates effective searching for all of the available information by computers as well as people. One step further, *ontological properties* of GO are crucial determinants of functional interpretation. One aspect of the use of GO annotations is the ability to group gene products to some high level GO term. For example, while gene products may be precisely annotated as having particular functions such as 'ferric ion binding', biologists may want to get all gene products functioning in iron metabolism could be grouped together as being involved in the more general phenomena 'iron homeostasis'. Empirical grouping exists and sets of high-level GO terms used in genome annotations are published at the GO site[2]. However, computational methods are desirable. Those methods mainly rely on the ontological properties of GO. For example, they will use subsumption relations in GO to group terms that are descendants of 'metal ion binding' (e.g. 'ferric ion binding') under the term 'metal ion binding'. Moreover, there exist biological notions that *depend* on other ones. Intuitively, a dependence relation holds between 'ferric ion binding' (which is a molecular function) and 'iron homeostasis' (which is a biological process). Therefore, given a gene product G, one could expect G to be annotated by 'iron homeostasis' or his children when G has been annotated by 'ferric ion binding'.

We are particularly interested in the phenomenon of inferential annotation based on dependence between GO terms. It can be formulated as follows: *given a gene product G, if A is dependent on B and A annotates G then B is also valid to annotate G.* The The expected benefit of identifying associative relations in GO is that these relations

[1] http://www.geneontology.org/GO.doc.html
[2] ftp://ftp.geneontology.org/pub/go/GO_slims/

can be used to produce more consistent annotations, and, thus, more accurate retrieval of gene products across databases.

The objective of this paper is to investigate dependence relations in GO. Our approach is based on ontological, lexical, and statistical aspects of dependence. We selected a set of 23 gene products that were potentially involved in enterocyte differentiation and that showed similar levels of expression. We analyzed which GO terms annotated them and the potential dependence relations between those terms. The preliminary results of a more comprehensive analysis of GO Annotations@EBI (GOA) are also presented. Finally, we discuss the meaning of dependence in this domain, and we suggest potential applications of this approach.

## 2 Background

Dependence is a form of connection between objects or kinds of objects, which may be variously filled. Simons provides several examples of dependence [Simons], including the following:

- Physiological dependence (person a is dependent on drug B iff a cannot survive unless doses of B are regularly administered).
- Causal dependence (detonation a of this mine is dependent on its priming b iff a cannot take place unless b previously takes place).
- Logical dependence (proposition p is dependent on proposition q iff p cannot be true unless q is true).
- Functional dependence (the pressure P of a fixed mass of ideal gas is dependent on its temperature T and volume V iff P cannot vary in value unless at least one of T and V varies in value).
- Practical dependence (skill A is dependent on skill B iff A cannot be mastered unless B is mastered).
- Ontological dependence (accident a, e.g., this whiteness, is dependent on substance b, e.g. this piece of paper, iff a cannot exist unless b exists).

### *2.1 Ontological aspects of dependence*

Aristotle holds that nonsubstances, such as qualities and quantities are ontologically dependent on substances. The general idea of ontological dependence is that ‘x is dependent on y iff x cannot be present unless y is also present’. The dependence of one entity on another is one of *de re* necessity: some entity A ontologically depends on some entity B iff necessarily, if A exists then B exists [Welty].

In the biomedical domain, Kumar gives examples of ontological dependence [Kumar-1], e.g. ‘There is no Cardiac output without a heart and no Cellular motion without some cell which moves’. Biological processes are dependent on on organs, cells and molecules. Kumar also makes a reference to a weaker description of

dependence in ‘the dependence of biomolecular functions upon each other’ [Kumar-2]. Biological functions performed within the cell are such that proteins depend for their functioning on interactions with other molecules and other molecular functions performed by other proteins.

### *2.2 Statistical aspects of dependence*

In probability theory, two events $E_1$ and $E_2$ are independent when the probability of occurrence of the two events simultaneously, $P(E_1 \cap E_2)$, is not greater than the product of the probabilities of occurrence for each event, $P(E_1) \,.\, P(E_2)$. Conversely, when $P(E_1 \cap E_2) > P(E_1) \,.\, P(E_2)$, $E_1$ and $E_2$ are not independent. This non-independence is what we refer to as statistical dependence.

The chi-squared test of association is based on this principle and is used to determine if there is any relationship between two attributes in a sample of data. This test compares the observed frequencies of, say, $P(E_1 \cap E_2)$ to the frequencies that would be expected under the null hypothesis of statistical independence, i.e., $P(E_1) \,.\, P(E_2)$. A large value of the chi-square statistic indicates a deviation from the expected frequencies and the hypothesis of statistical independence is rejected.

The same principle is used for extracting association rules from data (i.e., data mining). These association rules capture the association between two sets of events and are expressed in the form: $A \Rightarrow B$, where B is the set of events that can be predicted from A. Historically, the identification of association rules was applied to analyzing grocery buying patterns, with rules like {bread, milk} $\Rightarrow$ {sugar} expressing that customers buying bread and milk also often buy sugar. More recently, the identification of association rules has been applied to medical databases (e.g. [Brossette]) and bioinformatics (e.g. [Creighton]).

However, when applied to molecular biology databases, association rules have been used to identify links among genes or between genes and experimental conditions rather than among annotations. In this paper, we are interested in identifying associations among annotations in order to reveal additional links in the structure of GO. For example, if genes annotated with $T_1$ are also frequently annotated with $T_2$, we argue that a relationship between $T_1$ and $T_2$ may need to be represented in GO.

### *2.3 Lexical aspects of dependence*

In controlled vocabularies, the terms themselves contain information that is implicit in the names but is not systematically explicitly represented by hierarchical or associative relations in the ontology. For example, given a pair of modifiers such as (‘acute’, ‘chronic’), and a disease name ‘bronchitis’, the two terms ‘acute bronchitis’ and ‘chronic bronchitis’ are co-hyponyms of the term ‘bronchitis’ and may be considered opposed siblings [Bodenreider].

Ogren et al recently published a study of GO term names [Ogren]. They refer to the notion that derivational phrases encode semantic relations. For example the term 'regulation of cell proliferation' is derived from the term 'cell proliferation' by addition of the phrase 'regulation of'. The associative modifier 'Regulation of' corresponds to a semantic relationship (regulates). Associative relationships in terminologies can be suggested by analyzing the associative modifiers (e.g., 'regulation of') of nested terms (e.g., 'cell proliferation') in complex terms (e.g., 'regulation of cell proliferation'). This phenomenon of creating terms by combining an associative modifier and an existing term is also called reification.

## 3 Material and methods

### *3.1 GO*

The three hierarchies of GO are molecular function (MF), biological process (BP) and cellular component (CC). A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components. Each hierarchy is a directed acyclic graph. According to the documentation, the three GO hierarchies are independent. As a consequence, no relationships are represented across hierarchies. Therefore, annotators are encouraged to annotate to terms from all three hierarchies. In practice, the annotation of a gene product to one GO hierarchy is independent of its annotation to other hierarchies.

However, the notion of dependence is found in GO documentation[3]: "Some GO terms imply the presence of others in the ontology. For example: If X regulation exists, then the process X must also exist. Potentially any process in the ontology can be regulated". Moreover, GO authors recognize that implicit links between GO hierarchies may be conveyed by terms: "there are many cases where component terms are appropriate in the process ontology. For example, Golgi organization and biogenesis is different from lysosome organization and biogenesis, so the anatomical qualifiers 'Golgi' and 'lysosome' are necessary".

We used the Feb. 2004 release of GO. The Molecular Function file contains 7288 terms, the Biological Process file 8337 terms, the Cellular Component file 1390 terms.

### *3.2 Limited in-depth study*

A limited in-depth study was performed on a set of 23 gene products selected by UMR 6061 researchers. This set of 23 gene products corresponds to genes potentially involved in enterocyte differentiation that had similar expression pattern.

[3] http://www.geneontology.org/GO.usage.htm

### 3.2.1 Annotation

GO annotation was performed by submitting this list to SOURCE [Diehn] [4]. SOURCE is a unification tool that dynamically collects and compiles data from many scientific databases to provide, among other information, GO annotations. For each gene product, the resulting file contains a list of GO terms. In the following sections, we will use 'AGED' (Annotated Genes of Enterocyte Differentiation) to refer to our file of gene products annotated by SOURCE.

### 3.2.2 Establishing dependence relations in AGED

The AGED records were analyzed manually. A dependence link was established between two GO terms when both terms annotated the same gene product and exhibited implicit dependence. For example, the dependence relation between 'lipid transporter activity' and 'lipid transport' was added to the AGED file because some gene product is annotated by both 'lipid transporter activity' and 'lipid transport' and the BP 'lipid transport' depends functionally on MFs such as 'lipid transporter activity'

### 3.2.3 Qualitative analysis

In this step, we considered whether dependence relations could be inferred "lexically". For example, the dependence relation between 'lipid transporter activity' (MF) and 'lipid transport' (BP) can be inferred lexically since one string is lexically included in the other.

We also analyzed the status of the dependence relations with respect to ontological aspects, e.g. 'membrane fusion (BP) depends on membrane (CC)' is an instance of typical ontological dependence.

### 3.2.4 Quantitative analysis

A quantitative analysis was performed on the AGED file. Starting from the dependence relations established manually (see 3.2.2) and the initial AGED list, we checked the consistency of annotation in AGED. We searched for missing annotations as well as inconsistent annotations with respect to the set of dependence relations that were listed previously.

## *3.3 Large-scale test in GOA*

Because the result of a lexical analysis of GO terms has already been published ([Ogren]), we elected to focus our large-scale analysis on statistical dependence relations in a database of annotations. We used the Nov 2003 release of GOA-human

[4] http://source.stanford.edu/

downloaded from GO Website. The method we use to investigate the statistical aspects of dependence is based on the notion of co-occurrence of GO terms in GOA. Two GO terms $t_1$ and $t_2$ co-occur if they both annotate the same gene product.

Using the pairs ($t_d$, $t_i$) of GO terms in dependence relation established manually (see 3.2.2), we checked the existence of a co-occurrence relation in GOA between $t_d$ and $t_i$.In addition, we analyzed the whole matrix of co-occurring terms in GOA in order to identify additional significant associations. For each pair of GO terms ($t_1$, $t_2$), the number of times $t_1$ and $t_2$ annotate the same gene product in GOA was computed.

## 4 Results

### *4.1 Limited in-depth study*

The AGED file consists of 23 annotated gene products, corresponding to 139 lines. Each line represents the annotation of one gene product with one GO term, e.g. 'AFP' is annotated by 'extracellular space' (GO:0005615).

#### 4.1.1 Dependence relations

Fourteen different dependence relations can be defined, using the data from the AGED file. Those 14 relations are represented by 18 instances between one gene product and one GO term in AGED. Some of these dependence relationships are given in table 1. Most of the dependence relations hold between a BP and a MF (13/14). Only one relation holds between a MF and a CC.

| Biological Process (BP) | Molecular Function (MF) | #occurrences in AGED | |
|---|---|---|---|
| | | actually present | expected |
| transport | carrier activity | 2 | 2 |
| *bile acid* transport | bile acid transporter activity | 1 | 2 |
| digestion | bile acid transporter activity | 1 | 2 |
| lipid transport | lipid transporter activity | 3 | 4 |
| cholesterol metabolism | high-density lipoprotein binding | 1 | 2 |
| carbohydrate metabolism | oligo-1,6-glucosidase activity | 1 | 1 |
| carbohydrate metabolism | sucrose alpha-glucosidase activity | 1 | 1 |
| iron ion homeostasis | ferric iron binding | 1 | 1 |
| iron ion transport | ferric iron binding | 1 | 1 |

Table 1 : Examples of dependence relations

#### 4.1.2 Qualitative analysis

**Lexical aspects**. Four relations exhibit the lexical pattern of one string contained in the other (e.g. 'lipid transport depends on lipid transporter activity').

**Ontological aspects**. As mentioned before, only one relation between a molecular function and a cellular component was found: cell adhesion depends on membrane components, which corresponds to the classical notion 'a biological function exists iff the substance exists'.

#### 4.1.3 Quantitative results

While the existing annotations in AGED corresponded to 18 occurrences of dependence relations, we found 6 missing annotations. For example, among the 23 gene products in AGED, four are associated with the GO MF term 'lipid transporter activity' while only three of them are annotated with the GO BP term 'lipid transport'. The annotation of one gene product (APOH) with 'lipid transport' is expected but missing in AGED.

### *4.2 Results in GOA*

The co-occurrence between two dependent terms is illustrated in table 2. Among the 19088 gene products in GOA, 21 are annotated with both 'lipid transporter activity' and 'lipid transport', 15 are annotated with 'lipid transporter activity' and not with 'lipid transport', and 30 are annotated with 'lipid transport' and not with 'lipid transporter activity'.

| | | Lipid transporter activity | |
|---|---|---|---|
| | | present | absent |
| Lipid transport | present | 21 | 30 |
| | absent | 15 | 19022 |

Table 2- Number of gene products annotated with 'lipid transporter activity' (MF) and 'lipid transport' (BP) in GOA.

We started analyzing the associations derived from the whole matrix of co-occurrences in GOA. We are now in the process of reviewing some 500 high-support (frequent) and high-confidence (almost systematic) associations. The preliminary results are promising. For example, associations identified by this method include 'voltage-gated sodium channel complex' (CC) with 'cation channel activity' (MF), and 'proton transport' (BP) with 'hydrogen ion transporter activity' (MF).

## 5 Discussion

### *5.1 Functional dependence vs. ontological dependence*

Ontological dependence has been defined as: ‘x is dependent on y iff x cannot be present unless y is also present’ [Smith]. For example, the existence of carbohydrate metabolism depends on that of carbohydrate. Ontological dependence is systematic. Practically, a gene product annotated with a dependent term $t_d$ is also expected to be annotated with the term $t_i$ on which $t_d$ depends. If the annotations were complete and consistent, a statistical analysis of the associations among GO terms would show that $t_d$ is systematically associated with $t_i$. Because it is often reflected in names, ontological dependence may also be suggested by lexical patterns.

On the other hand, biological phenomena rely on functional dependence. For example, carbohydrate metabolism functionally depends on oligo-1,6-glucosidase activity. However, a given gene product may be associated to carbohydrate metabolism without being associated with oligo-1,6-glucosidase activity. Functional dependence can be detected by statistical methods. In contrast with ontological dependence, functional dependence often translates in non-systematic associations. Thresholds for confidence in association rules may be difficult to establish, and dependence rules must be validated by experts.

### *5.2 Potential applications*

Potential applications of our approach include quality assurance in annotation databases. It provides methods to control annotation quality and to ensure consistency between databases. As mentioned on GO website, the accuracy of GO annotations is a high priority: “Each member organization is responsible for keeping its own annotations accurate and up to date, and for correcting any errors. Users can report errors to the GO mailing list. The GO Consortium is also looking into possible ways to improve quality assurance further, such as manually reviewing selected annotations and developing tools to automate detection of potentially erroneous annotations”.

Furthermore, dependence relations shall be used to complement reasoning based on subsumption and meronomy in various applications. Applications that would take advantage of these reasoning capabilities are of major interest in biomedicine, such as functional interpretation of microarrays and information retrieval.

### *5.3 Limitations*

In this preliminary study, we have purposely not exploited two potentially important sources of additional knowledge: hierarchies and secondary ontologies.

**Hierarchies** can be usefully combined with lexical and statistical tools. While lexical tools can suggest a dependence relation between 'transport' (BP) and 'transporter activity' (MF), this dependence relation could be automatically propagated to 'carrier activity' through the subsumption link 'carrier activity is-a transporter activity'. Similarly, the generalization of association rules by combination with knowledge from hierarchies has been suggested by [Ramakrishnan].

In addition to existing GO hierarchies, **secondary ontologies** would be helpful to represent dependence. For example, we established an association between 'carbohydrate metabolism' (BP) and 'sucrose alpha-glucosidase activity' (MF), and between 'cholesterol metabolism' (BP) and 'high-density lipoprotein binding' (MF). As also mentioned by Wroe et al [Wroe], an ontology of biomolecules in which, for example, sucrose is represented as a kind of carbohydrate is needed to help identify such relations automatically. Additional domain knowledge is needed to represent all the elements that participate in dependence relations between 'digestion' (BP) and 'bile acid transporter activity' (MF).

## 6 Conclusion

Dependence relations in GO are complex. This preliminary study has shown the potential benefits of combining ontological, lexical, and statistical approaches to exploring various aspects of dependence relations. While ontological dependence is expected to be represented in well-formed biomedical ontologies, biologists are also interested in functional dependence. Additional work is needed, in particular an in-depth, manual analysis of the association rules based on the co-occurrence of GO terms in large annotation databases.

## Acknowledgements

Marc Aubry is funded in part by the Conseil Régional de Bretagne.

## References

[Ashburner] Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.Nat Genet. 2000 May;25(1):25-9.
[Blaschke ] Blaschke C, Oliveros JC, Valencia A. Mining functional information associated with expression arrays. Funct Integr Genomics. 2001 Mar; 1(4): 256-68.
[Bodenreider] Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. Int J Med Inf. 2002 Dec 4;67(1-3):85-95.

[Brossette] Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc. 1998 Jul-Aug;5(4):373-81.
[Creighton] Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics. 2003 Jan;19(1):79-86.
[Diehn] Diehn M et al.. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res,2003,31,1:219-23
[Gene Ontology Consortium] Gene Ontology Consortium.Creating the gene ontology resource: design and implementation. Genome Res. 2001 Aug;11(8):1425-33.
[Kumar-1] Kumar A, Smith B. The ontology of blood pressure: a case study in creating ontological partitions in biomedicine; in press
[Kumar-2] Kumar A, Smith B. A framework for Protein Classification. Proc. German Conference on Bioinformatics, Munich, Oct 12-14. 2003.. 2;55-57
[Lockhart] Lockhart DJ et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 1996 Dec; 14(13): 1675-80.
[Ogren] Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms.Pac Symp Biocomput,2004,9:214-225.
[Ramakrishnan] Ramakrishnan Srikant and Rakesh Agrawal. Mining Generalized Association Rules. In Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, September 1995.
[Raychaudhuri ] Raychaudhuri S, Schutze H, Altman RB. Using text analysis to identify functionally coherent gene groups. Genome Res. 2002 Oct;12(10):1582-90.
[Simons] Simons P. Parts. A study in ontology. Oxford University Press, 1987.
[Smith] Smith B, Williams J, Schulze-Kremer S The ontology of the gene ontology. Proc AMIA Symp. 2003; : 609-13.
[Stevens] Stevens R, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000;16(2):184-5
[Tanoue] Tanoue J, Yoshikawa M, Uemura S. The GeneAround GO viewer. Bioinformatics 2002 Dec;18(12):1705-6
[van de Vijver] van de Vijver MJ et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002 Dec 19; 347(25): 1999-2009.
[Welty] Welty C, Guarino N. Supporting ontological analysis of taxonomic relationships. Data & Knowledge Engineering, 39(1), 51-74
[Wroe] Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pac Symp Biocomput 2003:624-35