Non-Lexical Approaches to Identifying Associative Relations in the Gene Ontology

O. Bodenreider, M Aubry, and A. Burgun

Pacific Symposium on Biocomputing 10:91-102(2005)

NON-LEXICAL APPROACHES TO IDENTIFYING ASSOCIATIVE RELATIONS IN THE GENE ONTOLOGY

OLIVIER BODENREIDER

U.S. National Library of Medicine
8600 Rockville Pike, MS 43, Bethesda, Maryland 20894, USA
E-mail: olivier@nlm.nih.gov

MARC AUBRY

Unité de Génétique Humaine, UMR 6061 CNRS
Avenue du Pr Léon Bernard 35043 Rennes Cedex, France

ANITA BURGUN

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

The Gene Ontology (GO) is a controlled vocabulary widely used for the annotation of gene products. GO is organized in three hierarchies for molecular functions, cellular components, and biological processes but no relations are provided among terms across hierarchies. The objective of this study is to investigate three non-lexical approaches to identifying such associative relations in GO and compare them among themselves and to lexical approaches. The three approaches are: computing similarity in a vector space model, statistical analysis of co-occurrence of GO terms in annotation databases, and association rule mining. Five annotation databases (FlyBase, the Human subset of GOA, MGI, SGD, and WormBase) are used in this study. A total of 7,665 associations were identified by at least one of the three non-lexical approaches. Of these, 12% were identified by more than one approach. While there are almost 6,000 lexical relations among GO terms, only 203 associations were identified by both non-lexical and lexical approaches. The associations identified in this study could serve as the starting point for adding associative relations across hierarchies to GO, but would require manual curation. The application to quality assurance of annotation databases is also discussed.

1. Introduction

The Gene Ontology™ (GO) is an important resource that has transformed the functional annotation of gene products by providing the curators of model organism databases with a controlled vocabulary which has rapidly become a *de facto* standard. GO has over 17,000 terms and is organized in three hierarchies for molecular functions, cellular components, and biological processes. However, if hierarchical relations (*is a*, *part of*) constitute the backbone of ontologies, GO is essentially a skeleton because it completely lacks associative relations across its three hierarchies. Such associative relations would indicate, for example, that a

cellular component is the location of a biological process and that a molecular function is involved in a biological process.

The lack of representation in GO of the relations existing among functions, processes, and components severely limits the power of reasoning based on GO. This issue has been recognized by Bada et al. as they developed the *Gene Ontology Annotation Tool* (GOAT) [1]. One major task in GOAT and its companion project *Gene Ontology Next Generation* (GONG) is the acquisition of such relations and their formal representation in the Ontology Web Language (OWL) [2].

The approach taken in GOAT for acquiring associations between GO terms has been to mine the annotation database *Gene Ontology Annotation* (GOA) for co-occurrence of GO terms. 600,000 associations were obtained by this method, excluding unreliable associations and the hierarchical relations explicitly represented in GO [1]. Another approach to identifying relations among GO terms draws on the compositional structure of these terms. Ogren et al. found that 65% of all GO terms contain another GO term as a proper substring [3]. Finally, in a previous study, we suggested that association rule mining could be applied to identifying dependence relations among GO terms [4]. Kumar et al. successfully applied association rule mining techniques to the annotation databases of six bacterial genomes from The Institute for Genome Research (TIGR) and evaluated their findings in light of formal ontological principles [5].

In this study, rather than identifying all dependence relations, we concentrate specifically on associations among GO terms across ontologies. The primary objective of this study is to investigate three non-lexical approaches to identifying such associative relations in the Gene Ontology (GO) and compare them to lexical approaches. Our three approaches are: computing similarity in a vector space model, statistical analysis of co-occurrence of GO terms in annotation databases, and association rule mining. A secondary objective is to analyze the consistency of the associations discovered across five model organism databases. In other words, the major contribution of this study is not to define novel non-lexical methods for studying term-term associations, but rather to compare multiple existing approaches among themselves and to traditional lexical methods, systematically and across several model organism databases.

2. Datasets

The three approaches under investigation in this study take advantage of the existing annotation databases created for various model organisms. These databases, made publicly available in a common format by the GO Consortium[*], describe gene products that have been annotated with GO terms by each collabo-

[*] http://www.geneontology.org/GO.current.annotations.shtml

rating group. The annotation databases used in this study correspond to the major model organisms and were downloaded from the GO website[†]:

1. **FlyBase** (*Drosophila melanogaster*)
2. Human subset of **GOA** (*Homo sapiens*)
3. **MGI** (*Mus musculus*)
4. **SGD™** (*Saccharomyces cerevisiae*)
5. **WormBase** (*Caenorhabditis elegans*)

Details about these datasets are provided in Table 1.

Table 1 – Detail of the datasets used in this study

Dataset	Developed by	Web site	Dated
FlyBase	FlyBase Consortium	http://flybase.bio.indiana.edu/	5/22/2004
GOA-Human	European Bioinformatics (EBI)	http://www.ebi.ac.uk/GOA/	6/4/2004
MGI	Jackson Laboratory	http://www.informatics.jax.org/	6/4/2004
SGD	Stanford University	http://www.yeastgenome.org/	6/11/ 2004
WormBase	WormBase Consortium	http://www.wormbase.org/	5/11/2004

Table 2 – Number of unique gene products, GO terms, and gene product-term pairs in the five annotation databases under investigation

Annotation DB	# gene products	# GO terms	# GP-term pairs
FlyBase	9,090	3,597	38,089
GOA-Human	22,720	4,247	92,658
MGI (Mouse)	14,471	3,616	65,571
SGD (Yeast)	6,457	2,412	25,278
WormBase	10,534	1,540	36,695

The version of GO used throughout this study is the June 2004 monthly release, available from the GO website. The GO terms present in the annotation databases but not in the ontology were replaced by current terms whenever possible. For example, the term *amine oxidase (flavin-containing) activity* (GO:0004041) is no longer present and was replaced by *amine oxidase activity* (GO:0008131), with which it is currently asserted to be synonymous. The annotations for which no current GO term existed were ignored. Also ignored were the annotations for which the evidence supporting the association between a gene product and a GO term is insufficient. In practice, we filtered out all annotations inferred from electronic annotation (with 'IEA' as evidence code), because they are not reviewed by curators. We did not include either the negative associations, marked with 'NOT' in the *Qualifier* field of the annotation files. The number of unique gene products, GO terms, and gene product-term pairs in each annotation

[†] http://geneontology.org/

database is given in Table 2. These counts reflect the substitutions and filtering mentioned above.

In addition to GO and the annotation databases, the evaluation relies in part on the Unified Medical Language System® (UMLS®) Metathesaurus®. The UMLS[‡] is a terminology integration project developed at the U.S. National Library of Medicine. The UMLS Metathesaurus integrates many biomedical terminologies, including the Gene Ontology [6]. Although no relations across ontologies are defined in GO, such relations – contributed by other sources – may be present in the Metathesaurus. More specifically, associative relations asserted in other source vocabularies are found in the MRREL table. For example, the GO terms *chloroplast* and *photosynthesis* are also defined in the Medical Subject Headings (MeSH), where they are cross-referenced. This "see also" relationship is recorded in the Metathesaurus between the two concepts. Similarly, the co-occurrence of MeSH descriptors in the MEDLINE database is recorded in the MRCOC table of the Metathesaurus. The edition of UMLS used in this study is 2004AA (April 2004).

3. Methods

The three approaches to identifying associative relations in GO, presented in detail below, can be summarized as follows:

1. A vector space model in which each GO term is described by a vector of gene products corresponding to the annotations of this product in the annotation database for a given organism.
2. Statistical analysis of co-occurrence of GO terms in the annotations of gene product, where the observed frequency of co-occurrence of two GO terms is compared to the frequency expected under the hypothesis of independence of GO terms.
3. Association rules mined from the sets of GO terms extracted from annotation databases, where each transaction corresponds to the annotations of a given gene product in a given annotation database.

In all three cases, the associations identified are restricted to associations across GO ontologies (e.g., molecular function to biological process) by filtering out the association within hierarchies.

Common to the three approaches is the assumption that the dependence relations identified (e.g., among frequently associated GO terms) should reflect *ontological* relations (i.e., among entities whose existence depend on one another), themselves possibly corresponding to *biological* relations. Based on different mathematical principles, the three approaches are expected to identify different sets of dependence relations. While the three methods essentially rely

[‡] http://umlsks.nlm.nih.gov/ (free license required)

on the frequency of association between two GO terms, they use different criteria for determining which associations are significant.

3.1. *Similarity in the vector space model*

Vector space models (VSMs) are frequently used in information retrieval for computing the similarity between documents described as vectors of keywords [7]. A collection of gene products annotated with the controlled vocabulary provided by GO is in fact analogous to a collection of scientific articles indexed with the MeSH controlled vocabulary. Although the primary use of a collection of indexing terms for documents (or annotation terms for gene products) is to compute the similarity among documents (or gene products), our interest here is to compute the similarity among terms. Therefore, we have to transpose the matrix of gene products by GO terms in order to obtain a matrix of GO terms by gene products. As usual in the VSM paradigm, the similarity between two vectors is represented by the angle between these vectors, measured by the dot product of the two (normalized) vectors.

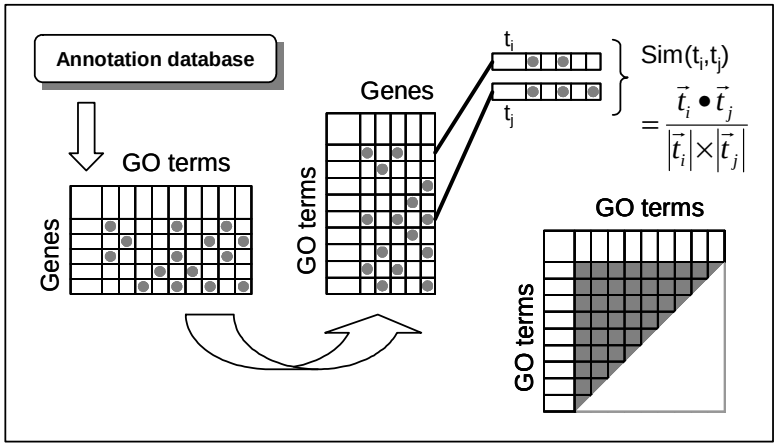

Figure 1 – Similarity in the vector space model from a given annotation database

As illustrated in Figure 1, the original matrix (gene products by GO terms) consists of binary values indicating the presence (1) or absence (0) of an association between a gene product and a GO term in a given annotation database. One such matrix is created for each model organism. The matrix is then transposed. Not represented in the figure is the weighting step. A weight is applied to each binary association in order to lower the importance of an association between a GO term and a gene product when a given gene product is associated with many

GO terms[§]. This weighting scheme, shown in Eq. (1), is known as inverse document frequency (idf) in information retrieval. Here, the weight of each association between a GO term and gene product *j* is inversely proportional to the ratio of the number of annotations for this gene product (n_j) to the total number of distinct gene products in the corresponding annotation database (N). Then, each vector is normalized in order to compensate for differences in the number of genes associated with GO terms. Once the vectors are normalized, their dot product varies between 0 and 1 and measures the similarity between them (see Figure 1). A value of 0 corresponds to no similarity, while 1 indicates complete similarity. Term-term similarity is computed pairwise for all GO terms present, resulting in a half-matrix for each model organism database. We use an arbitrary threshold of .5 for the dot product in order to select the pairs of terms exhibiting a high drgree of similarity.

$$idf_j = \log \frac{N}{n_j} \tag{1}$$

3.2. *Co-occurrence in annotation databases*

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, $P(E_1) \cdot P(E_2)$. Conversely, when $P(E_1 \cap E_2) > P(E_1) \cdot P(E_2)$, E_1 and E_2 are not independent. What we are interested in identifying here are pairs of "non-independent" GO terms, whose frequency of co-occurrence (i.e., simultaneous presence in the annotation of a gene product) is higher than would be expected if the two terms had been used independently by the curators. For a given pair of GO terms (*A*,*B*), information about their association in gene product annotations can be summarized in a two-way contingency table and analyzed statistically [8]:

- n_{AB}, the number of gene products annotated with both term *A* and term *B*
- n_{Ab}, the number of gene products annotated with term *A* but not term *B*
- n_{aB}, the number of gene products annotated with term *B* but not term *A*
- n_{ab}, the number of gene products annotated with neither term *A* or term *B*

The chi-square test of independence (or Pearson's chi-square) is often used to test independence between two categorical variables (here, the presence or absence of a given term in the annotations of genes). The chi-square (χ^2) statistic

[§] The weighting step could probably be omitted in the case of a matrix of GO terms by gene products, because each gene has a limited number of annotations. It would, however, be crucial for computing gene-gene similarity in a matrix of gene-products by GO terms.

relies on the difference between observed frequencies (n_{ij}) for the four events listed above and the frequencies expected under the hypothesis of independence. The χ^2 statistic has a chi-square distribution, specified by its degrees of freedom. There is one degree of freedom in the case of two-way contingency tables for binary variables. A large value of the χ^2 statistic indicates a deviation from the expected frequencies. In this case, i.e., when the corresponding P-value is lower than the usual .05 threshold, the hypothesis of statistical independence is rejected and the association is considered statistically significant. One limitation of the chi-square test is that all expected frequencies are required to be 5 or more. In practice, this condition cannot be met if the frequency of the terms is small.

An alternative to the Pearson's chi-square test is the likelihood ratio test (also called G-test or G-square test). The G^2 statistic compares the maximum of the likelihood function under two circumstances: 1) under the hypothesis of independence and 2) under the general, observed conditions. Like the χ^2 statistic, the G^2 statistic has a chi-square distribution (also with one degree of freedom in our setting). Interestingly, the G^2 statistic does not have the minimum expected frequency requirements imposed by the χ^2. However, for the G^2 statistic to be computed, all observed frequencies must be greater than 0.

In practice, for each pair of terms, we first attempt to compute a G^2 statistic. A χ^2 statistic is used instead when the requirements are not met for G^2. Finally, the association is ignored if it fails to meet both G^2 and χ^2 requirements. Because of the low frequency of co-occurrence of the terms in this case, identifying their association is of little interest anyway.

While both χ^2 and G^2 indicate the existence of an association between two variables, neither one describes the strength of the association. Several similarity coefficients have been developed for this purpose [9], which could be used to select the pairs of terms exhibiting a strong association. In this study, however, we simply included all pairs of terms for which the test indicated a statistically significant association, regardless of the strength of the association.

3.3. *Association rule mining*

Association rules capture the association between two sets of events of arbitrary size and are expressed in the form: $A \Rightarrow B$, where B is the set of events that can be predicted from A [10]. Historically, the identification of association rules was applied to analyzing grocery buying patterns, with rules such as $\{bread, milk\} \Rightarrow \{sugar\}$ expressing that customers buying bread and milk also often buy sugar. By applying association rule mining techniques to annotation databases, we expect to discover that genes annotated with the GO term T_1 are also frequently

annotated with T_2. The set of GO terms annotating a gene product is called a transaction in association rule mining parlance.

We used Christian Borgelt's implementation of the *apriori* algorithm[**] to mine association rules. Since our objective is to identify pairs of related GO terms, we restricted the size of the sets under investigation to two. The two major parameters in the algorithm are *support* and *confidence*. Support for the rule $T_1 \Rightarrow T_2$ represents the proportion of genes annotated with both T_1 and T_2. Confidence for the same rule represents the proportion of genes annotated with both T_1 and T_2 among those annotated with T_1. In order to restrict rules to almost systematic associations, we required confidence to be at least 90%. The minimum support was set to a low value (.05%) simply to eliminate "accidental" associations. We use the product of support by confidence to describe the strength of the association.

3.4. *Evaluation*

The first step of the evaluation consists in comparing the results of the three approaches. Associations identified independently by several approaches simultaneously are expected to be stronger and therefore more important. Finally, the presence of the association in the annotation databases of several organisms suggests that this association is stronger than isolated associations. What is evaluated here is essentially the statistical significance of the associations. Evaluating the ontological and biological significance of these associations is beyond the scope of this study.

Additionally, we compared the results of our three approaches to lexical associations and to associations present in the UMLS Metathesaurus.

Lexical relations. Using the method proposed by Ogren et al., we identified all pairs of GO terms where one term is nested as a substring in the other term [3]. In order to reveal additional lexical relations, we did a second run after systematically removing the word 'activity' from terms in the molecular function hierarchy.

UMLS relations. We searched the MRREL table for the presence of associative relations[††] among concepts present in GO. Similarly, we searched the MRCOC table for the presence of co-occurrence relations among GO concepts (co-occurrence of MeSH descriptors in MEDLINE records).

[**] http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html

[††] Their relationship type (REL) is 'RO' for "other related concepts".

4. Results

4.1. *Associations identified*

Examples of association identified specifically by each method are presented in Table 3. The first three are the methods under investigation: VSM (vector space model), COC (co-occurrence in annotation databases), and ARM (association rule mining). The others are the methods used in the evaluation: LEX (lexical relations), REL (associative relations in UMLS), and MDL (co-occurrence in MEDLINE). Quantitative results are presented in Table 4 where the number of associations identified by each method is broken down by category of association.

Table 3 – Examples of association identified specifically by each method

Method	Association	
VSM	MF: *ice binding*	[GO:0050825]
	BP: *response to freezing*	[GO:0050826]
COC	MF: *chromatin binding*	[GO:0003682]
	CC: *nuclear chromatin*	[GO:0000790]
ARM	MF: *carboxypeptidase A activity*	[GO:0004182]
	BP: *proteolysis and peptidolysis*	[GO:0006508]
LEX	MF: *mannosyltransferase activity*	[GO:0000030]
	CC: *mannosyltransferase complex*	[GO:0000136]
REL	CC: *cell-matrix junction*	[GO:0030055]
	BP: *cell adhesion*	[GO:0007155]
MDL	CC: *synaptic vesicle*	[GO:0008021]
	BP: *exocytosis*	[GO:0006887]

Table 4 – Number of associations identified by each method for each category of association (MF: molecular function; CC: cellular component; BP: biological process)

	VSM	COC	ARM	LEX	REL	MDL
MF-CC	499	893	362	917	0	0
MF-BP	3057	1628	577	2523	0	1
CC-BP	760	1047	329	2053	22	469
Total	4316	3568	1268	5493	22	470

4.2. *Overlap*

A total of 13,398 associations were identified by at least one method, 7,665 by at least one of the three major methods (VSM, COC, and ARM) and 5,963 by at least one of the evaluation methods (LEX, REL, and MDL). Of these, only 230 associations were identified by both major and evaluation methods. Examples of associations identified independently but simultaneously by several methods are presented in Table 5. As illustrated in Figure 2, 12% of the associations identi-

fied by the three major methods were identified by more than one method. In contrast, only a few lexical associations are also present in the UMLS.

Out of the 7,665 associations identified by at least one method, 5,950 (78%) came from only one annotation database. In 1,116 cases (16%), the association was simultaneously identified in two annotation databases, 6% in three, and 2% in two. Only 41 associations (less than 1%) were present in all five databases. Pairwise, after normalizing by the number of annotations in each database, the highest rates of overlap are between MGD and GOA-Human and MGD and WormBase, the lowest between SGD and GOA-Human and SGD and MGD.

Table 5 – Examples of association identified simultaneously by several methods

Association		**VSM**	**COC**	**ARM**	**LEX**	**REL**	**MDL**
MF: *potassium channel activity* BP: *potassium ion transport*	[GO:0005267] [GO:0006813]	X	X	X			
MF: *chemokine activity* BP: *immune response*	[GO:0008009] [GO:0006955]		X	X			
CC: *hemoglobin complex* BP: *oxygen transport*	[GO:0005833] [GO:0015671]	X	X				
MF: *taste receptor activity* BP: *perception of taste*	[GO:0008527] [GO:0050909]	X		X			
MF: *metal ion transporter activity* BP: *metal ion transport*	[GO:0046873] [GO:0030001]	X		X	X		
CC: *transport vesicle* BP: *transport*	[GO:0030133] [GO:0006810]				X	X	
CC: *gap junction* BP: *cell communication*	[GO:0005921] [GO:0007154]	X	X				X

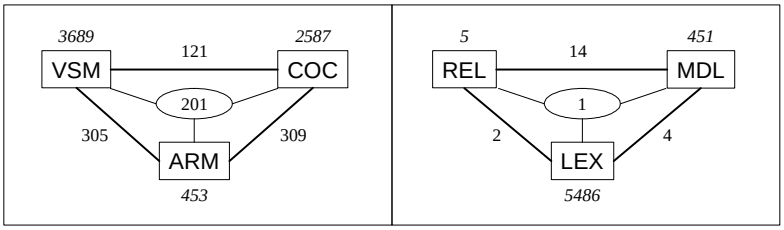

Figure 2 – Number of associations specific to each method (italic) or common to several methods

5. Discussion

Applications. The major application of our methods is of course to help enrich GO with associative relations across ontologies. Ontology creation and extension is a daunting task. However, by automatically extracting candidate relations from annotation databases, the approaches investigated in this study can significantly

reduce the human effort required. We recommend that the associations identified in this study serve as the starting point for adding associative relations across hierarchies to GO. The associations we identified could also be used for quality assurance purposes, i.e., to assess the consistency and completeness of annotation databases. Analogously, knowledge of frequently associated terms could be presented to curators in annotation environments.

Advantages and limitations. Many associations identified by our approaches cannot be found by lexical methods. The performance of lexical methods could be improved by factoring in term variation (inflection, derivation) and using more rigorous parsing of the terms; it would, however, remain poor due to the limited number of synonyms available in GO for each term. By imposing constraints such as minimum frequency and statistical significance of co-occurrence and minimum confidence for association rules, our approaches are more selective than the unrestricted methods used in GOAT. Limitations – not specific to our approaches – include the fact that what is identified is the presence of associations between GO terms, not their nature. Moreover, the manual curation of the associations identified remains necessary in order to assess their biological significance.

Evaluation. The limited overlap between associations identified by our major methods and the evaluation methods was somewhat unexpected. The lexical relation between, for example, *transport* and *transport vesicle* is ontologically valid but never present in annotations. Although the biomedical literature plays a role in both approaches, the limited overlap between annotation databases and MEDLINE co-occurrences may have the following explanations. Many annotations are derived from sources other than the literature (e.g., inferred from sequence or structural similarity) and MEDLINE co-occurrences are not guaranteed to relate to the same gene when several genes are discussed in an article.

Generalization. As shown in earlier studies [4,5], dependence relations can be found both within and across the three GO ontologies. Although this study is purposely restricted to the identification of associative relations across GO ontologies, our methods actually identified almost as many dependence relations *within* ontologies (not reported on here). The lexical method captures five times as many associations within ontologies than across, including a majority of direct parent-child associations. Because curators are unlikely to use both a parent term and its child in the annotation of a gene, the associations within ontologies captured by our methods are essentially between distinct subtrees of GO hierarchies (e.g., between *metallopeptidase activity* [*catalytic activity* subtree] and *zinc ion binding* [*binding* subtree]). Finally, our approaches could be applied to other domains (e.g., for identifying relations among terms of a clinical terminology using clinical databases indexed with this terminology).

Future directions. Many interesting aspects of the association between GO terms are beyond the scope of this paper. Those issues, which we expect to address in the near future, include the redundancy of associations across species and applications to the functional interpretation of experimental results.

Acknowledgments

Marc Aubry's contribution is funded in part by the Conseil Régional de Bretagne. The authors thank Kelly Zeng who provided technical support for computing similarity in the vector space model.

References

1. Bada, M., Turi, D., McEntire, R. & Stevens, R. Using Reasoning to Guide Annotation with Gene Ontology Terms in GOAT. *SIGMOD Record* **33**(2004). http://www.acm.org/sigmod/record/issues/0406/04.Bada_Turi_McEntire_Stevens.pdf
2. Wroe, C.J., Stevens, R., Goble, C.A. & Ashburner, M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput*, 624-35 (2003)
3. Ogren, P.V., Cohen, K.B., Acquaah-Mensah, G.K., Eberlein, J. & Hunter, L. The compositional structure of Gene Ontology terms. *Pac Symp Biocomput*, 214-25 (2004)
4. Burgun, A., Bodenreider, O., Aubry, M. & Mosser, J. Dependence relations in Gene Ontology: A preliminary study. *Workshop on The Formal Architecture of the Gene Ontology - Leipzig, Germany, May 28-29, 2004* (2004). http://mor.nlm.nih.gov/pubs/pdf/2004-go_workshop-ab.pdf
5. Kumar, A., Smith, B. & Borgelt, C. Dependence relationships between Gene Ontology terms based on TIGR gene product annotations. *Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm 2004)*, 31-38 (2004)
6. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32 Database issue**, D267-70 (2004)
7. Baeza-Yates, R. & Ribeiro-Neto, B. *Modern information retrieval*, 513 p. (ACM Press ; Addison-Wesley, New York; Harlow, England, 1999).
8. Agresti, A. *An introduction to categorical data analysis*, xi, 290 p. (Wiley, New York, 1996).
9. Duarte, J.M., dos Santos, J.B. & Melo, L.C. Comparison of similarity coefficients based on RAPD markers in the common bean. *Genet. Mol. Biol.* **22**, 427-432 (1999)
10. Agrawal, R., Imielinski, T. & Swami, A. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD*, 207-216 (1993)