From Terms to Categories: Testing the Significance of Co-occurrences between Ontological Categories

Robert Hoehndorf
Institute for Medical Informatics, Statistics and Epidemiology and
Department of Computer Science, University of Leipzig and
Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Anthropology
hoehndorf@eva.mpg.de

Axel-Cyrille Ngonga Ngomo
Department of Computer Science, University of Leipzig
ngonga@informatik.uni-leipzig.de

Michael Dannemann and **Janet Kelso**
Department of Evolutionary Genetics
Max-Planck-Institute for Evolutionary Anthropology
{michael_dannemann,kelso}@eva.mpg.de

Abstract

The co-occurrence of terms in a text corpus may indicate the presence of a relation between the referents of these terms. We expect co-occurrence-based methods to identify association relations that cannot be found using static patterns. We developed a new method to identify associations between ontological categories in text using the co-occurrence of terms that designate these categories. We use the taxonomic structure of the ontologies to cumulate the number of co-occurrences of terms designating categories. Based on these cumulated values, we designed a novel family of statistical tests to identify associated categories. These tests take both co-occurrence specificity and relevance into consideration. We applied our method to a 2.2 GB text corpus containing fulltext articles and used Gene Ontology's biological process ontology and the Celltype Ontology. The software and results can be found on http://bioonto.de/pmwiki.php/Main/ExtractingBiologicalRelations.

1 Introduction

An increasing number of biomedical ontologies address the problem of biological data integration. Ontologies are a means for organizing and representing basic categories and relationships pertaining to the conceptualization of a domain. Many biomedical ontologies have been developed according to a common set of criteria based on the Open Biomedical Ontologies (OBO) or the OBO Foundry. A common property of these ontologies is their focus on a single domain. This particular property provides an easy means for applying an ontology to a domain-specific application. However, knowledge bridging multiple domains remains hidden and not explicit.

To address this problem, so-called "cross-products" have been created. They define categories from one ontology using categories from other ontologies and relations from the OBO Relationship Ontology (RO) (Smith et al., 2005). Due to the large number of categories in the OBO ontologies, few of these cross-products exist and are maintained. For example, parts of the Gene Ontology (GO) (Ashburner et al., 2000) are defined using categories of cells from the Celltype ontology (CL) (Bard et al., 2005) and relations like **has-participant** from the RO. While many of these cross-products have been created in a manual curation effort, some were created using automated information extraction methods (Bada and Hunter, 2007), which exploit the compositional nature of many terms in these ontologies.

Methods based on term decomposition can provide high quality logical definitions suitable for inclusion in a stable version of the ontology. Yet, they miss several more intricate relations between categories that are not reflected in their names. For example, the relation between *cardiac muscle cells* (CL:0000746) and *heart looping* (GO:0001947) cannot be uncovered using basic pattern matching. Other approaches have been

used to extract relations between categories in ontologies. Among them are association rule mining and statistical analysis of term co-occurrences (Bodenreider et al., 2005).

In this paper, we present a novel method for extracting association relations between categories defined in distinct biomedical ontologies. This method takes as input a set of ontologies and a text corpus. It then detects associations between categories of the input ontologies based on the co-occurrence of terms that designate ontological categories. The data obtained by analyzing co-occurrences is further refined according to the structure of the input ontologies. The resulting association relations can either be considered by human curators, used as input for automated relationship extraction methods or exploited by question answering systems.

2 System and Methods

2.1 Ontologies

An ontology is the specification of a conceptualization of a domain (Herre et al., 2006). Many biological ontologies are represented as directed acyclic graphs (DAGs) and are available in the OBO flatfile format [1]. In these DAGs, nodes represent *categories* and edges represent *relations* between these categories. A category, also called *kind*, *class* or *universal*, is an entity that is general in reality. Examples are *dog*, *apoptosis*) or *to transport sugar*. Categories may have instances, of which some may not be further instantiated. These are called *individuals*. We call the set of all categories in an ontology O $Cat(O)$.

Categories may be related to other categories. The most important relation between two categories A and B is the **is-a** relation, $isA(A,B)$. The relation $isA(A,B)$ can be defined using the instantiation relation: when $isA(A,B)$, then all instances a of A are instances of B (Herre et al., 2006). This definition implies that the **is-a** relation is reflexive and transitive.

A set of categories with the **is-a** relation among them form a taxonomy. These taxonomies often are the backbone of the OBO ontologies' DAG structure. We call the set of all successors of a category A the sub-categories $subcat(A) = \{B|isA(B,A)\}$ and its predecessors the super-categories $supcat(A) = \{B|isA(A,B)\}$. The direct successors and predecessors of A in the taxonomy are called children ($child(A) = \{B|isA(B,A) \wedge B \neq A \wedge \forall X(isA(B,X) \wedge isA(X,A) \rightarrow X = B)\}$) and parents, respectively.

In the OBO flatfile format, ontologies are assigned a namespace. Category-identifiers are prefixed with the namespace of the ontology to which they belong. Therefore, they are unique within the OBO ontologies. In addition to a unique identifier, categories are assigned a *name* and a set of *synonyms*. Neither the name nor the set of synonyms must be unique.

2.2 Basic Assumptions

Our method for extracting association relations between categories is based on two main assumptions:

1. Terms can designate ontological categories; the terms that designate the same category are henceforth called the category's synset. Every occurrence of an element of the synset of category C is called an occurrence of C. Every co-occurrence of an element of the synset of the category C with an element of the synset of the category D is called a co-occurrence of C and D.

2. When A is a sub-category of B, then every co-occurrence of A with C is a co-occurrence of B with C. Additionally, every occurrence of A counts as an occurrence of B.

According to our first assumption, we constructed synsets from the synonyms attached to each category in the input ontologies, and counted the occurrences and co-occurrences of these synsets based on two contexts: single sentences and sentences in documents[2]. We used exact matching to identify terms in text. Secondly, we computed the closure of the occurrences and co-occurrences of the categories with respect to the **is-a** relation, as explicated in our second assumption.

Finally, we test for the collocation between categories based on the occurrence and co-occurrence of elements of their synsets. Here, collocation refers to a co-occurrence that is higher

[1] http://www.cs.man.ac.uk/~horrocks/obo/

[2] The second context refers to whole documents, but co-occurrence is based on single sentences. Therefore, when two terms co-occur in two or more sentences within one document, their co-occurrence is only counted once.

than expected by chance. To this end, we designed a family of tests that account for both the ontologies' structure and the term distribution in the text corpus. The tests account for both relevance and specificity of the co-occurrence of categories. In this context, relevance refers to how often the categories co-occur in the text corpus compared to their absolute occurrence. The second aspect of the tests allows the identification of the categories that contain the most information within the ontologies, i.e., they are the most specific categories with respect to the **is-a** relation.

To test our method, we used the biological process (BP) branch of the Gene Ontology (GO) (Ashburner et al., 2000) and the Celltype Ontology (CL) (Bard et al., 2005). Our experiments were conducted using a 2.2 GB text corpus containing 60143 fulltext articles from Open Access journals listed in Pubmed Central.

2.3 Method

We first analyzed the text corpus for the occurrence and co-occurrence of the terms included in the synsets of categories taken from two ontologies. Based on these values, we computed the occurrence and co-occurrence values for the categories. To test the statistical significance of these co-occurrence values, we generated several permutations of the data extracted from the text corpus. These approximate a random distribution of co-occurrence values within the ontologies for the chosen text corpus. We then calculated the p-values for the observed values against this random distribution. Finally, we applied a family of novel tests to these p-values to identify collocated categories from the ontologies. The result of our approach is a list containing pairs of categories that are collocated with respect to a given cutoff.

2.3.1 Text Processing

First, we counted the number of occurrences and co-occurrences of the terms contained in synsets of categories from the input ontologies. We counted the total number of sentences and documents in which at least one element of a synset was found using exact matching. For each pair of categories, we counted the total number of co-occurrences of elements of their respective synsets in sentences. Furthermore, we counted the number of documents in which they co-occured within at least one sentence. We used exact matching and abstained from using any more sophisticated methods for recognizing the ontologies' categories in text at this point in time.

The text processing yielded, for each category C, both its frequency $f(C)$ (total number of occurrence of terms from $syn(C)$ in sentences) and the total number of documents in which an element from $syn(C)$ appeared, $d(C)$. Furthermore, for each pair of categories C_1 and C_2, we obtained both the total number of co-occurrences in sentences $f(C_1,C_2)$ and the total number of documents containing these co-occurrences $d(C_1,C_2)$.

2.3.2 Co-occurrence Cumulation Using Ontologies

The second step in our method implemented our second assumption, i.e., occurrence and co-occurrence between categories is transitive over the **is-a** relation. We assumed that when two categories C and C' stand in the **is-a** relation, C **is-a** C', then every occurrence of C is also an occurrence of C'. This means that the synset-closure $synclos(C)$ of a category C can be constructed as follows:

$$syn(C) \subseteq synclos(C) \quad (1)$$

$$isA(C,C') \rightarrow (syn(C) \subseteq synclos(C')) \quad (2)$$

For all categories C, the values $f_t(C)$ and $d_t(C)$ represent the sum of the values $f(C')$ and $d(C')$ over all of C's sub-categories C'. For all categories C_1 and C_2, we computed the cumulated f- and d-values dubbed $f_t(C_1,C_2)$ and $d_t(C_1,C_2)$:

$$f_t(C_1,C_2) := \sum_{a \in subcat(C_1)} \sum_{b \in subcat(C_2)} f(a,b), \quad (3)$$

$$d_t(C_1,C_2) := \sum_{a \in subcat(C_1)} \sum_{b \in subcat(C_2)} d(a,b), \quad (4)$$

For all categories C_1 and C_2, we defined the following score function:

$$score(C_1,C_2) = \frac{\log f_t(C_1,C_2)}{\log(1+f_t(C_1)) + \log(1+f_t(C_2))} \cdot \frac{\log(d_t(C_1,C_2))}{\log(1+\max(d_t(C_1),d_t(C_2)))} \quad (5)$$

The first component of the score function implements the natural logarithm of the Pointwise Mutual Information (PMI) (Manning and Schütze, 1999) score achieved by the categories with respect to their co-occurrence within sentences. In order to avoid divisions by 0, the denominators

of all members of the score function were incremented. The second component measures a similar value using documents as context. The aim of the score function is to ensure that categories that co-occur relatively often are assigned a high score. The range of the score function is between 0 and 1 and categories with overlapping synsets will have a score of 1.

2.3.3 Determining the Random Distribution

The score of two categories C and D is influenced by the topology of the ontology: categories that are more general occur and co-occur more often, due to our definition of occurrence and co-occurrence of categories. Therefore, it is insufficient to test for a high score to consider the co-occurrence of two categories as significant. A random distribution for the scores of each pair of categories C and D provides a means for determining the significance of a co-occurrence. This random distribution depends on the text corpus, the method for identifying categories, the score function and the topology of the ontologies. Hence, we did not assume any statistical distribution of scores.

We simulate the random distribution of the scores of each category pair through multiple random permutations: the f- and d-values that were measured for each synset during the first step of our method were randomly assigned to categories in the ontology from which they originated. We then calculated and recorded co-occurrence scores for all pairs of categories. In addition, for each category D, such that $isA(D,C_1)$, the score difference $score(C_1,C_2) - score(D,C_2)$ was recorded. Further, for each category E with $isA(C_1,E)$, the score difference $score(E,C_2) - score(C_1,C_2)$ was recorded.

Hence, the results of this step were threefold. First, we approximated the random score distribution for each pair of categories. Second, each triple of categories C, D and $E \in child(C)$ gave rise to a random distribution of score differences between (C,D) and (E,D). Third, each triple C, D and $E \in parent(C)$ yielded a random distribution of score differences between (E,D) and (C,D).

2.3.4 Significance Testing

To identify strong co-occurrences, we designed a family of tests for each co-occurrence that considers a fragment of the path in the ontology graph. The first kind of tests is assymetrical. At the end of this section, we will introduce a symmetrical form of these tests. The first tests are designed to test the significance of the co-occurrence between C_1 and C_2 based on three criteria: (1) the score $score(C_1,C_2)$ for the co-occurrence should be higher than expected; (2) for each child category D of C_1, $score(C_1,C_2) - score(D,C_2)$ should be higher than expected and (3) for each parent category E of C_1, $score(E,C_2) - score(C_1,C_2)$ should be lower than expected.

The first criterion measures relevance, while criteria (2) and (3) test for specificity. The first criterion establishes high confidence in the co-occurrence strength. The second criterion reflects the assumption, that a collocation must be novel, i.e., it must represent an information gain over the co-occurrences of a sub-category. Therefore, given that $isA(D,C_1)$, we assume that any relevant information gained from the co-occurrence between C_1 and C_2 already appears in the co-occurrence between D and C_2 when the difference between $score(C_1,C_2)$ and $score(D,C_2)$ is low (with respect to the random distribution of scores). We would assume a collocation between D and C_2, because D is more specific than C_1. On the other hand, if the difference between $score(C_1,C_2)$ and $score(E,C_2)$ is high (with respect to the random distribution of scores), and $isA(C_1,E)$, we would assume a collocation between E and C_2. We describe the intuitions behind our tests below. The complete description and formalization of the tests can be found on the project website.

Within this section, let C and D be fixed categories from ontologies O_1 and O_2, respectively. Furthermore, let N be the number of permutations.

The first test we designed depends on the categories C and D, the ontology's structure and the number of permutations N. It tests for the following properties:

- the co-occurrence score between C and D is high,
- the difference between $score(C,D)$ and $score(C',D)$ for every child C' of C is high,
- the difference between $score(C,D)$ and $score(C'',D)$ for every parent C'' of C is low.

"Being high" and "being low" were captured using the values of the cumulative distribution functions (CDFs) obtained by the N permutations performed in the previous step: one function for each pair of categories C and D, one function for each triple of categories C, D and C' where C' is a child of C, and one for each triple C, D and C'' where C'' is a parent of C. We then combined the p-values of the score differences to children in a single value using their geometric mean. A similar combination of the score differences' p-values to the parent categories of C was carried out: here, the combined value is the geometric mean of $1-x$, where x is the p-value in the corresponding CDF.

The geometric mean was used because it has properties that correspond to our intuitions: when the score difference to one of the child categories is very low (the p-value in the CDF is 0), we always prefer the co-occurrence of the child of C and D over the co-occurrence of C and D. The geometric mean would then be 0, and the result of the first test Θ_1 would be 0 as well. Very high differences (the p-value in the CDF is 1) are ignored, i.e., the value of the geometric mean depends solely on the other child categories of C.

The inverse holds for the score differences between C and D and the parents of C and D: when the p-value of the score difference in the CDF is 0, this difference is ignored (because $1-0=1$, and thus does not heavily influence the value of the geometric mean), while a high difference (the p-value in the CDF is 1) results in a final score of 0.

The goal of the test Θ_1 is to find the *most specific* pair of categories which co-occur significantly often. Therefore, the score between the two categories should be high, and provide a significant gain over all the child categories. If there was no such gain, i.e., the score between C and D is high and the score between the children of C and D is high as well, Θ_1 prefers the co-occurrences between children of C and D, because they are more specific and therefore contain more information. The difference to the parents of C should be low, as otherwise there would be a significant gain in the score between a parent of C and D over the score between C and D. Then, Θ_1 prefers the co-occurrence between this parent and D over C and D.

All other tests are extensions of the first test. The second test, Θ_2, uses the minimum function instead of the geometric mean to combine the p-values in the CDFs of the score differences to parents and children.

The first two tests Θ^1 and Θ^2 do not consider the variances of the distributions of scores, differences in scores to children and differences in scores to parents. Therefore, we extended these tests by weighting all three components of the tests with the variances of their corresponding distributions. In these tests, high variance lowers the impact of the result, while lower variance strengthens it.

We defined three new distributions for the variances, and chose the p-value in the respective CDF as a weight in our tests. We computed the scores for each pair of category N times, resulting in one distribution of scores for each pair of categories. Each of these distributions has a variance. The score variance distribution is the finite distribution (containing N elements) of the variances of each of these distributions. We defined the variance distribution for score difference to parent and child analogously.

The tests Θ^3 and Θ^4 use only the variance distribution of scores, while Θ^5 and Θ^6 use all three variance distributions. These tests are one-sides, i.e., they are not symmetric. We define two-sided, symmetric tests $\tau^i(C,D)$ for all categories C and D as

$$\tau^i(C,D) = \Theta^i(C,D) \cdot \Theta^i(D,C) \qquad (6)$$

3 Implementation

The text processing module is implemented in Java. The remaining steps are implemented using a combination of Java classes and Groovy scripts. The source code for all programs is available under the modified BSD license from the project webpage. The implementation uses the functionality of the GNU GetOpt library[3], Java Universal Network/Graph Framework[4] and the Java Colt libraries[5].

4 Discussion

4.1 Results

We applied the method described here to the biological process (BP) branch of the Gene Ontology

[3] http://sourceforge.net/projects/evcgen/

[4] http://jung.sourceforge.net

[5] http://dsd.lbl.gov/~hoschek/colt/

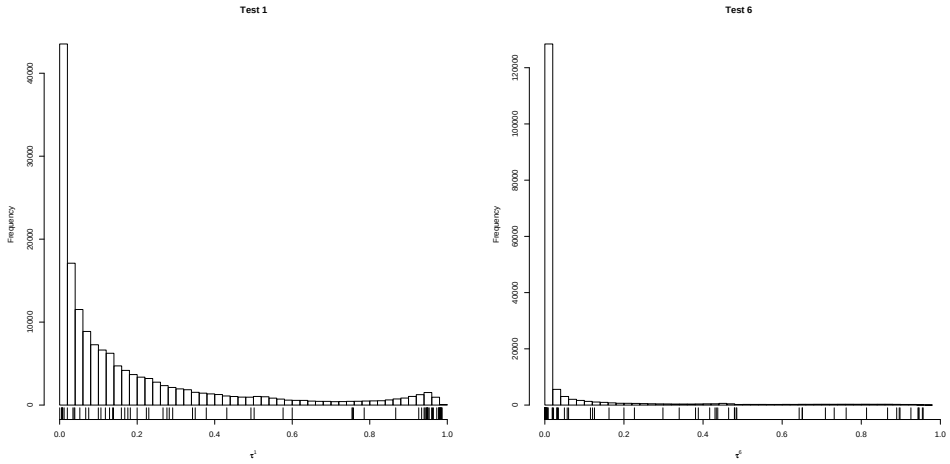

Figure 1: Distribution of test results. The plot on the left shows the distribution of the test results for τ^1. On the right, the same is shown for τ^6. It can be seen that a test using the minimum function (τ^6) is stronger than a test using the geometric mean (τ^1). Furthermore, weighting the tests with the CDFs of the variances (τ^6) produce stronger results than the basic test (τ^1). Below the distributions, the quantiles of the GO-CL dataset for each test are displayed.

p	τ^1	τ^2	τ^3	τ^4	τ^5	τ^6
0.5	0.075	0.017	0.024	0.003	0.007	0.001
0.8	0.288	0.145	0.141	0.047	0.061	0.016
0.9	0.522	0.433	0.298	0.168	0.220	0.120
0.95	0.806	0.790	0.472	0.412	0.456	0.400
0.99	0.952	0.950	0.863	0.826	0.859	0.824

Table 1: The table shows p-quantiles for different p-values for all six tests. Given a p-value (first column), the quantiles show the result of each test for which p-values are below the quantile.

(GO) and the Celltype Ontology (CL). We identified 3,751 out of the 14,542 terms in the GO's biological process ontology in our text corpus. We found 491 of 754 terms from the CL. Terms from the GO's BP branch co-occurred 70,967 times with CL terms.

Using our method, we identified the total number of 202,627 co-occurrences between categories. After applying our tests, 157,894 co-occurrences produced p-values distinct from 0.[6] We illustrate the quantiles obtained for different p-values in our six tests, τ^i, in table 1. The distribution of scores for τ^1 and τ^6 are shown in figure 1. The remaining plots are available on the project webpage.

[6]The remainder obtained a score of 0 due only to numerical restrictions. They were subsequently excluded, because they were indistinguishable from the absence of co-occurrence.

We found that the tests using the minimum instead of the geometric mean of p-values of score differences to parent and child categories are generally stronger, i.e., they include fewer co-occurrences as significant for a given cutoff. Similarly, tests including the variance for scores are generally stronger than tests that are not weighted by the variance of score distributions. In this sense, the tests τ^5 and τ^6 are the strongest.

Relation	Number of occurrences
has-participant	62
Participates-in	13
Located-in	2
unclassified	38

Table 2: Manually identified ontological relations in the 100 top-scoring association results (with respect to τ^1).

Table 2 shows the kind of relationship between categories that our tests identified for the 100 top-scoring results with respect to test τ^1. The *has-participant* relation is defined in (Smith et al., 2005). We define the *Participates-in* relation as: C_1 *Participates-in* $C_2 \iff \forall x,t_1(instanceOf(x,C_1,t_1) \rightarrow \exists t_2,y(instanceOf(y,C_2,t_2) \wedge participates\text{-}in(x,y,t_2)))$, where *participates-in* is the primitive participation relation between individuals as defined in (Smith et al., 2005). We

extend the definition of *located-in* in (Smith et al., 2005) to a relation *Located-in* between processes and objects, which holds when all participants of a process are *located-in* a structure during the entire duration of the process.

In our sample, 38 association relations do not fall under one of the three relations that we investigated. We discovered several kinds of unclassified relations. First, mismatches in granularity lead to strong associations for unrelated categories. For example, *xanthine transport* and *erythrocyte* are closely related according to τ^1. Erythrocytes are involved in the transport of xanthine. However, the GO category *xanthine transport* refers to the inter- and intracellular level of granularity, while erythrocytes transport nutrients between organs. Second, some categories are indirectly related via another category. For example, osteoclasts and lymph node development are related via the protein RANK. Third, when cells have closely related functions, we identify too specific or too generic cell types as in the case of the association between *basophil degranulation* and *mast cell*. Finally, 6 out of 100 associations in our sample seem erroneous.

4.2 Comparison with Other Approaches

We did not compute precision or recall for our method, due to the absence of a gold standard. However, we compared our method with the GO-CL crossproducts available[7] from the OBO Foundry[8]. The dataset contains manually verified relations between categories from the GO and the CL that have been extracted using the method described in (Bada and Hunter, 2007). Because this method is based on the compositional nature of terms in the GO, it exclusively identifies relations in which one category name (usually a type of cell) is a substring of another category name (usually a GO category).

The GO-CL crossproduct contains 396 relations between GO and CL categories. From these 396, we identified 73 that co-occurred in our text corpus. Table 3 shows the percentage of significant co-occurrences within these 73 relations for different cutoffs in our six tests. Figure 1 shows the distribution of the 73 pairs with respect to τ^1 and τ^6.

[7] `http://obofoundry.org/cgi-bin/detail.cgi?id=go_xp_cell`, accessed on January 23rd, 2008.

[8] `http://obofoundry.org`

As our method relies exclusively on the distribution of terms and not on their syntactic structure, it permits the recognition of association relations between categories that could not be recognized using patterns. An example of such an association is *myoepithelial cell* (cells located in the mammary gland) and *milk ejection*.

However, while (Bada and Hunter, 2007) identified well-defined, ontological relations, our approach is designed to identify strongly associated categories that can be further refined using complementary approaches for identifying relationships from text, such as abductive reasoning (Hobbs et al., 1988).

Recall	τ^1	τ^2	τ^3	τ^4	τ^5	τ^6
95%	0.007	0.006	0.003	0	0.002	0
80%	0.102	0.054	0.028	0.003	0.016	0.002
70%	0.173	0.109	0.049	0.008	0.029	0.004
50%	0.502	0.350	0.173	0.063	0.154	0.060

Table 3: Evaluation of our approach with respect to the GO-CL dataset (Bada and Hunter, 2007). The dataset we used for comparison consists of the 73 relations from (Bada and Hunter, 2007) found in our text corpus. Columns two to seven show the cutoff values required to identify the percentage given in column one of relations as significant using tests one to six.

4.3 Future Research

The method presented in this paper can be enhanced by several means. First, our term identification approach could be improved. A large number of variants of the terms included in the synset of each category may occur in scientific texts. Since our term recognition is based on exact matching, we expect to miss a large number of term occurrences and other references to the ontologies' categories. In particular, this affects the recognition of terms from the GO. We expect that the integration of methods such as (Gaudan et al., 2008) for recognizing GO categories in text would improve our results. Further natural language processing techniques such as stemming could improve the identification of categories in text.

Second, we currently estimate the random score distribution throughout the ontologies using multiple permutations. A deeper statistical analysis could provide insights on how to replace the

random distributions obtained through permutations with the exact random distributions. We expect this to improve the accuracy of our method.

The main goal of our future research will be to extract well-defined, ontological relations between categories. The method we propose in this paper serves as the first step in such an effort, because it generates relevant associations according to the scientific literature used. Additional methods that may be based on the manual generation of patterns (Bada and Hunter, 2007), pattern learning (Hao et al., 2005) or the application of methods from logics and ontologies (Schulz et al., 2006; Hobbs et al., 1988) can then be applied.

In the meantime, we plan to apply our method to other ontologies and lexical resources.This is possible because our method makes use of directed graph structures, in which edges represent relations from a less specific to a more specific entity. Such a graph structure can be extracted from a wide variety of biomedical resources.

4.4 Conclusion

We developed a novel method to identify association relations between ontological categories from co-occurrences between terms obtained using text-mining techniques. For this purpose, we have implemented a suite of tools that can be used to extract these association relations from a text corpus and two ontologies represented in the OBO flatfile format. In order to evaluate the strength of the association relations between the ontological categories, we designed a family of novel statistical tests that account for the ontologies' topologies and test for relevance and specificity.

We applied our method to extract several thousands of associated categories from the Gene Ontology and the Celltype Ontology using a text corpus comprised of fulltext scientific articles from PubMed Central. The association relations that we extracted are available for downloadat `http://bioonto.de/pmwiki.php/Main/ExtractingBiologicalRelations`.

Acknowledgement

We thank Leonardo Bubach, Hernán Burbano and Heinrich Herre for helpful discussions and valuable comments, and Christine Green for her help in preparing the manuscript.

References

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May.

M. Bada and L. Hunter. 2007. Enrichment of obo ontologies. *Journal of Biomedical Informatics*, 40(3):300–315, June.

J. Bard, S. Y. Rhee, and M. Ashburner. 2005. An ontology for cell types. *Genome Biology*, 6(2):R21.

O. Bodenreider, M. Aubry, and A. Burgun. 2005. Non-lexical approaches to identifying associative relations in the gene ontology. *Pac Symp Biocomput*, pages 91–102.

S. Gaudan, A. Jimeno Yepes, V. Lee, and D. Rebholz-Schuhmann. 2008. Combining evidence, specificity, and proximity towards the normalization of gene ontology terms in text. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008(3):9.

Y. Hao, X. Zhu, M. Huang, and M. Li. 2005. Discovering patterns to extract protein-protein interactions from the literature: Part ii. *Bioinformatics*, 21(15):3294–3300, August.

H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek. 2006. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. Onto-Med Report 8, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany.

J. R. Hobbs, M. Stickel, P. Martin, and D. D. Edwards. 1988. Interpretation as abduction. In *26th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 95–103, Buffalo, New York.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

S. Schulz, E. Beisswanger, J. Wermter, and U. Hahn. 2006. Towards an upper-level ontology for molecular biology. *AMIA Annu Symp Proc*, 2006:694–698.

B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. 2005. Relations in biomedical ontologies. *Genome Biol*, 6(5).