OPEN ACCESS Freely available online

PLoS one

Statistical Tests for Associations between Two Directed Acyclic Graphs

Robert Hoehndorf[1,2,3,4]*, Axel-Cyrille Ngonga Ngomo[2], Michael Dannemann[3], Janet Kelso[3]

1 Research Group *Ontologies in Medicine*, Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany, **2** Department of Computer Science, University of Leipzig, Leipzig, Germany, **3** Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, **4** European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Abstract

Biological data, and particularly annotation data, are increasingly being represented in directed acyclic graphs (DAGs). However, while relevant biological information is implicit in the links between multiple domains, annotations from these different domains are usually represented in distinct, unconnected DAGs, making links between the domains represented difficult to determine. We develop a novel family of general statistical tests for the discovery of strong associations between two directed acyclic graphs. Our method takes the topology of the input graphs and the specificity and relevance of associations between nodes into consideration. We apply our method to the extraction of associations between biomedical ontologies in an extensive use-case. Through a manual and an automatic evaluation, we show that our tests discover biologically relevant relations. The suite of statistical tests we develop for this purpose is implemented and freely available for download.

Citation: Hoehndorf R, Ngonga Ngomo A-C, Dannemann M, Kelso J (2010) Statistical Tests for Associations between Two Directed Acyclic Graphs. PLoS ONE 5(6): e10996. doi:10.1371/journal.pone.0010996

Editor: Fabio Rapallo, University of East Piedmont, Italy

Received October 8, 2009; **Accepted** May 12, 2010; **Published** June 16, 2010

Funding: The study was funded by the Max Planck Society and the University of Leipzig. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hoehndor@ebi.ac.uk

Introduction

An increasing number of discoveries, particularly in biomedicine, are facilitated by statistical analyses of data annotated to biomedical ontologies [1]. Biomedical ontologies are generally represented as DAGs, and specific domains are usually represented in distinct, separate DAGs [2–4].

Statistical tests that utilize a single graph can only consider the given domain. However, entities from different domain are linked via biomedical relations [5]. These relations can be vital for the discovery of novel biomedical knowledge. We have designed a family of novel statistical tests to identify strong associations between nodes from two directed acyclic graphs. The tests combine measures of relevance and specificity.

We evaluated our statistical method through an extensive use-case in which we applied our tests to the detection of strong semantic associations between the Gene Ontology [3] and the Celltype Ontology [6] based on co-occurrence in scientific literature. In this use-case, we annotated the ontologies with occurrence and co-occurrence count data of the ontologies category labels in full text scientific articles. The strongest associations identified through our tests are biologically relevant relations.

An implementation of the six novel statistical tests to identify associations between directed acyclic graphs is available as free software from our project webpage at http://bioonto.de/pmwiki.php/Main/ExtractingBiologicalRelations.

State of the art

Our approach to the computation of the strength of the association between two graphs relies on approaches for capturing the semantic similarity between categories in ontologies and for propagating these similarities within DAGs. In the following, we give a brief overview of methods for computing the similarity of categories (a more complete overview can be found in [7]). Most of the existing semantic similarity approaches assume that ontologies contain categories C_i that are annotated with terms $t_{i_1} \ldots t_{i_n} = \phi(C_i)$. Based on this assumption, the computation of the semantic similarity of two categories C_1 and C_2 can be carried out by using the structure of the ontology to which C_1 and C_2 belong (edge-based approaches), the nodes and their properties (e.g., similarity between $\phi(C_1)$ and $\phi(C_2)$) (node-based approaches) or by combining structural knowledge and annotations (hybrid approaches).

The most common edge-based approach consist of using a function of the number of edges between C_1 and C_2 as semantic similarity measure [8,9]. Other approaches combine the previous approach with the lenght of the path from the most specific common ancestor of C_1 and C_2 and the root node [10,11]. Edge-based approaches rely on the nodes being elements of the same graph. Thus, they cannot be utilized when trying to compute the similarity of two nodes from distinct DAGs.

The second category of approaches, the node-based approaches, use the properties of the nodes themselves to compute their similarity. One of the central concept for using annotations to compute similarity is that of information content, which is the negative log-likehood $-log(p(C_i))$ of a term C_i where $p(C_i)$ is the probability of occurrence of the terms in $\phi(C_i)$ in a certain corpus. Based on this value, several similarity metrics have been developed including the information content of the most informative common ancestor used in [12,13] or of the disjoint common ancestors [14].

In recent years, hybrid similarity measures that combine node- and edge-based approaches have been developed. Most of these approaches utilize the information content. For example [15] utilize a combination of edge weights based on node depth and node link density and of the difference of information content of the nodes linked by that edge. Other approaches such as that described in [16] compute edge weights by using a scheme that takes the type of the edge into consideration. The semantic similarity between two terms is set to a function of the maximum of the product of best path between the terms. Again, these approaches can only compute the similarity of terms from the same DAG.

The aim of our approach is to provide a means for the computation of the association between nodes from 2 DAGs, which are, in general, distinct. We do not make similar assumptions about the annotation of edges and nodes as other approaches to semantic similarity. Instead, we go beyong current semantic similarity measures by providing a measure of statistical significance in a distribution of arbitrary node and edge annotations. When applying out method to semantic similarity between ontologies, we can compute initial semantic similarity values for categories which do not belong to the same ontologies.

Methods

Statistics on graphs

Preliminaries of directed acyclic graphs. Our tests take as input two directed acyclic graphs, $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$ that are disjoint ($V_1\cap V_2=\varnothing$). From these two graphs, a graph $G=(V_1\cup V_2,E_1\cup E_2\cup C)$ with $C=V_1\times V_2\cup V_2\times V_1$ is constructed. We denote an edge as an ordered pair of vertices. If an edge connects v_1 and v_2, $e=(v_1,v_2)$, we call v_2 the child of v_1 and v_1 the parent of v_2. If there is a path from v_1 to v_2, we call v_1 a predecessor of v_2 and v_2 a successor of v_1.

In addition to the two graphs, two functions d_1' and d_2' are given as input such that $d_1':V_1\cup V_2\mapsto\mathbb{R}$ and $d_2':C\mapsto\mathbb{R}$. From these two functions, a graph decoration for G is constructed based on the assumption that the two input functions are transitive over the DAG: the decoration $d_1(v)$ of a vertex $v\in V_1\cup V_2$ is the union of $d_1'(v)$ and the values of $d_1'(u)$ for all successors u of v. Similarly, the decoration $d_2(e)$ of an edge $e=(v_1,v_2)$ for $e\in C$ is the union of $d_2'(e)$ and the values of $d_2'(f)$ for all edges f between the successors of v_1 and v_2.

The third component of the input is a score function $score: V_1\times V_2\cup V_2\times V_1\mapsto\mathbb{R}$. We assume that the value of the score function between the vertices v_1 and v_2 depends only on the graph decorations $d_1(v_1)$ of v_1 and $d_1(v_2)$ of v_2 as well as the decoration $d_2(e)$ of the edge $e=(v_1,v_2)$.

The score function is not symmetric, i.e., it is not necessary that $score(x,y)=score(y,x)$. It is intended to measure the association strength between two vertices from the input graphs. Our method identifies whether the score between two vertices is significantly high. A graphical overview of our test method is shown in Figure 1.

Determining the Random Distribution. The score between two vertices v_1 and v_2 is influenced by the topology of the input DAGs: a vertex v that is more general has a larger decoration set $d_1(v)$ due to our basic assumption about transitivity of input graph decorations. Similarily, the cardinality of the decoration set of the edges between nodes from the two input DAGs is larger when the edges connect more general vertices. Therefore, it is insufficient to test for a high score between vertices to consider the score between two vertices as significantly high. A random distribution of the scores of each pair of vertices v_1 and v_2 provides a means for determining the significance of the score between v_1 and v_2. This random distribution depends on the functions d_1' and d_2', the score function and the topology of the input graphs. Hence, we cannot assume any statistical distribution of scores *ab initio*. Instead, we simulate the random distribution of the scores between each vertex pair through multiple random permutations: the d_1'-values that are given as input for our method are randomly swapped with the d_1'-values of vertices in the input DAG from which they originate. There are two options for permutating the d_2'-values for edges: either they are, *mutatis mutandis*, permutated similarily to the d_1'-values of the vertices, or they are permutated depending on the permutation of d'_1-values; in the latter case, when the d_1'-values of v_1 and v_2 are swapped, so are the values of $d_2'(v_1,x)$ and $d_2'(v_2,x)$ for any vertex x.

Because our test is intended to identify associations between vertices, we do not assume that the values of d_1' and d_2' are independent. We therefore prefer to use the second option, i.e., that the permutation of the d_2' values depends on the permutation of the d_1'-values.

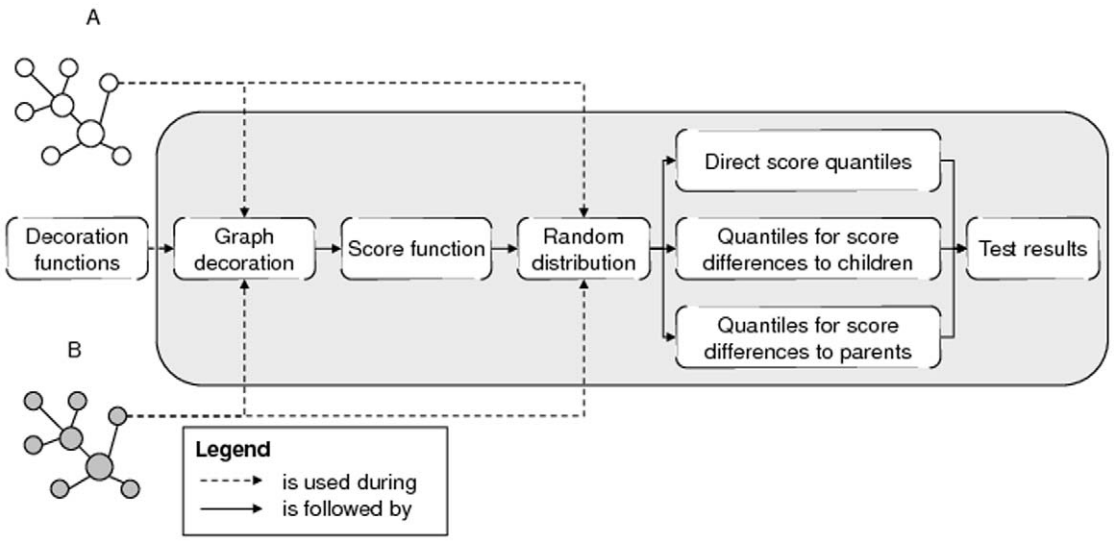

Figure 1. Schematic representation of our method.
doi:10.1371/journal.pone.0010996.g001

Based on these permutations, we first rebuild the graph decorations d_1 and d_2. Then, we calculate and record the values of the score function $score(v_1,v_2)$ for all pairs of vertices v_1 and v_2. In addition, for each vertex u, such that v_1 is a direct successor of u, we calculate and record the score difference $score(v_1,v_2)-score(u,v_2)$. Further, for each vertex w with the direct predecessor v_1, we calculate and record the difference $score(w,v_2)-score(v_1,v_2)$.

Hence, the results of this step are threefold. First, we approximate the random score distribution for each pair of vertices through multiple random permutations. Second, each triple of vertices u, v and $w \in children(u)$ gives rise to a random distribution of score differences between (u,v) and (w,v). Third, each triple u, v and $w \in parents(u)$ yields a random distribution of score differences between (w,v) and (u,v).

Ontologies as graphs

While the tests we develop can be applied to any DAG that satisfies the conditions specified above, their primary application is to test the significance of an association between categories from two ontologies. An ontology is the specification of a conceptualization of a domain [17,18]. Many biological ontologies are represented as directed acyclic graphs (DAGs) and are available in the OBO flatfile format [2]. 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 *red*. 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 *isA* relation, $isA(A,B)$. The relation $isA(A,B)$ can be defined by using the instantiation relation: when $isA(A,B)$, then all instances a of A are instances of B [18]. This definition implies that the *isA* relation is reflexive, transitive and antisymmetric.

A set of categories with the *isA* relation among them form a taxonomy. These taxonomies are often 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 of A in the taxonomy are called children ($children(A) = \{B \mid is\ A\ (B,A) \wedge B \neq A \wedge \forall X\ (is\ A\ (B,X) \wedge is\ A\ (X,A) \rightarrow X = B)\}$), while the direct predecessors are called parents.

In the OBO flatfile format, ontologies are assigned a namespace. Category identifiers are prefixed with the namespace of the ontology to which they belong. Identifiers are therefore 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.

Results

Statistics on graphs

To identify strong associations, we designed a family of tests for the score of each edge between the two input DAGs that considers a fragment of the path in the DAG. The tests are designed to measure the significance of the score between vertices v_1 and v_2 based on three criteria: (1) the score $score(v_1,v_2)$ for the association should be higher than expected; (2) for each child u of v_1, $score(v_1,v_2)-score(u,v_2)$ should be higher than expected; and (3) for each parent w of v_1, $score(w,v_2)-score(v_1,v_2)$ should be lower than expected.

The first criterion of our tests identifies hypothetical associations between nodes from two graphs. The second and third criteria are used to verify whether the pair is the best selection, or whether a more specific or more general association is preferable. For this purpose, the second and third criteria test for novelty of the association (compared to the child and parent nodes).

Within this section, let u and v be fixed vertices from the DAGs G_1 and G_2, respectively. Furthermore, let N be the number of permutations that were used to determine the random distributions. The first test we designed, Θ^1, depends on the vertices u and v, the DAG structure and the number of permutations N. It tests for the following properties:

- the score between u and v is high,
- the difference between $score(u,v)$ and $score(u',v)$ for every child u' of v is high,
- the difference between $score(u,v)$ and $score(u'',v)$ for every parent u'' of v is low.

"Being high" and "being low" are 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 u and v, one function for each triple of categories u, v and u' where u' is a child of u, and one for each triple u, v and u'' where u'' is a parent of u. We combine 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 u is carried out: here, the combined value is the geometric mean of $1-x$, where x is the p-value in the corresponding CDF.

Formally, let u and v be fixed vertices from the directed acyclic graphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$, respectively, and let

- N be the number of permutations,
- $score^n(u,v)$ be the score between u and v in the n^{th} permutation,
- $NQ(x,u,v)=P(score^n(u,v)\leq x)$, $1\leq n\leq N$, be the cumulative distribution function (CDF) of $score(u,v)$.
- $DQ^{u_j}(x,u,v)=P(score^n(u,v)-score^n(u_j,v)\leq x)$, $1\leq n\leq N$, be the CDF of the difference between the vertex u and its j^{th} child vertex,
- $DQ(x,u,v)=\{DQ^{u_j}(x,u,v)|u_j \in child(u)\}$,
- $MQ^{u_k}(x,u,v)=P(score^n(u_k,v)-score^n(u,v)\leq x)$, $1\leq n\leq N$, be the CDF of the score difference between the vertex u and its k^{th} parent vertex,
- $MQ(x,u,v)=\{MQ^{u_k}(x,u,v)|u_k \in parent(u)\}$,
- $VQ_{NQ}(x)=P(Var(NQ(x,x_1,x_2))\leq x)$, for all $x_1 \in V_1$ and $x_2 \in V_2$, be the CDF of the variances Var of the distribution $NQ(x,x_1,x_2)$, and VQ_{DQ} and VQ_{MQ} for the distributions $DQ(x,x_1,x_2)$ and $MQ(x,x_1,x_2)$, respectively.

For each child u_j of u, we calculate the difference in scores $\delta_d(u_j)=score(u,v)-score(u_j,v)$. Then, we compute the geometric mean ξ of all values $DQ(\delta_d(u_j),u,v)$. Similarly, we calculate $\delta_m(u_k)=score(u_k,v)-score(u,v)$ for each parent u_k of u, and the geometric mean ψ of all values $MQ(1-\delta_m(u_k),u,v)$. Then we define as our first test

$$\Theta^1(u,v)=NQ(score(u,v),u,v)\cdot\xi\cdot\psi \qquad (1)$$

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 extend 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 define three new distributions for the variances and choose the p-value in the respective CDF as a weight in our tests. We compute 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 define 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-sided, i.e., they are not symmetric. We define two-sided, symmetric tests $\tau^i(u,v)$ for all vertices u and v as

$$\tau^i(u,v) = \Theta^i(u,v) \cdot \Theta^i(v,u) \tag{2}$$

Table 1 lists the combination of properties for all tests. The precise formulation of all six tests can be found in the supplement S1.

Application to biomedical ontologies

Occurrence and co-occurrence count data as graph decoration. To verify whether the tests we designed yield reasonable results, we applied our method to the detection of significant co-occurrences between ontological categories in natural language texts, as a precursor to the detection of relations between ontological categories. For this purpose, we make the following assumptions:

1. A term occurs in a portion of text if it is an exact substring of this portion of text.
2. 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 category C's synset is called an *occurrence of* C. Every co-occurrence of an element of the category C's synset with an element of the category D's synset is called a co-occurrence of C and D.
3. If 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.

Table 1. Elements of the test score of τ^i.

	combining p-values in the CDF's of score differences from parents to children	variance distribution of scores	variance distributions to children and parents
τ^1	geometric mean		
τ^2	minimum		
τ^3	geometric mean	X	
τ^4	minimum	X	
τ^5	geometric mean	X	X
τ^6	minimum	X	X

doi:10.1371/journal.pone.0010996.t001

To test our method, we used the Gene Ontology (GO) [3] and the Celltype Ontology (CL) [6] as input DAGs. The GO is an ontology specifically designed to describe gene products. It contains three separate ontologies: the biological process, molecular function and cellular component ontologies. Gene products can be tagged with ontology categories to describe and classify them. The CL is an ontology for types of cells. It classifies cells based on criteria such as structure or function.

Based on the input requirements of our test, we constructed synsets from the synonyms attached to each category in the input ontologies, and counted the occurrences and co-occurrences of the categories based on two contexts: single sentences and sentences in documents. 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. The functions that assign the occurrence and co-occurrence count values to a synset of a category for each context are called d and f, respectively.

We used exact string matching to identify terms in text. Our evaluation was conducted using a 2.2 GB text corpus containing 60143 fulltext articles from Open Access journals listed in Pubmed Central. The aim of our method is to test for significant co-occurrences between categories.

Text Processing. First, we counted the number of occurrences and co-occurrences of the terms contained in synsets of categories from the input ontologies. Table 2 shows examples for the synsets of categories. We counted the total number of sentences and documents in which at least one element of a synset was found by 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 [19,20] to evaluate our method. Exact matching provides a large dataset for the evaluation of our method. For practical applications such as relationship extraction, more advanced methods should be chosen.

The text processing yielded, for each category C, both its frequency $f(C)$ and the total number of documents in which C occurred, $d(C)$. 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 $d(C_1,C_2)$.

Count data over ontologies. The first component in our method implements the assumption that the input graph decorations are transitive over the DAG structure. In the case of ontologies, this implements the assumption that occurrence and co-occurrence between categories is transitive over the *isA* relation between categories.

We assumed that when two categories C and C' stand in the *isA* relation, $isA(C,C')$, 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:

$$syn(C) \subseteq synclos(C) \tag{3}$$

$$isA(C,C') \rightarrow (syn(C) \subseteq synclos(C')) \tag{4}$$

Table 2. Example synsets taken from the GO and the CL.

ID	Label	Synonyms
GO:0001574	globoside biosynthetic process	ganglioside biosynthesis; ganglioside formation; ganglioside synthesis
CL:0000114	surface ectodermal cell	cell of surface ectoderm; surface ectoderm cell

doi:10.1371/journal.pone.0010996.t002

For count data, the decoration value of a vertex v in the DAG is equal to the sum of the input value pair $d(v)$ and $f(v)$ and the corresponding input values for v's successors. Therefore, for all categories C, we define $f_t(C)$ and $d_t(C)$ to represent the sum of the values $f(C')$ and $d(C')$ over all of C's sub-categories C'. Furthermore, for all categories C_1 and C_2, we compute the cumulated f - and d-values dubbed $f_t(C_1,C_2)$ and $d_t(C_1,C_2)$:

$$f_t(C_1,C_2) := \sum_{a \in subcat(C_1)} \sum_{b \in subcat(C_2)} f(a,b) \tag{5}$$

$$d_t(C_1,C_2) := \sum_{a \in subcat(C_1)} \sum_{b \in subcat(C_2)} d(a,b) \tag{6}$$

Again, for count data, co-occurrence values between nodes v_1 and v_2 can be summed up over the successors of v_1 and v_2 to yield the decoration of the edge between v_1 and v_2.

A score for occurrences and co-occurrences. For all categories C_1 and C_2, we defined the following score function:

$$score(C_1,C_2) = \frac{\log f_t(C_1,C_2)}{\log(1+f_t(C_1))+\log(1+f_t(C_2))} \cdot \frac{\log(d_t(C_1,C_2))}{\log(1+\max(d_t(C_1),d_t(C_2)))} \tag{7}$$

The first component of the score function implements the natural logarithm of the Pointwise Mutual Information (PMI) [21] score achieved by the categories with respect to their co-occurrence within sentences. PMI has been successfully used in several text mining tools (see, e.g., [22]). 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.

Discussion

Evaluation

We applied the tests to the biological process (BP) branch of the GO and the CL. To recognize the categories in text, we used the identifier of the category, the name and all exact synonyms of the category. On average, every category had 2.1 synonyms. Using exact matching, we identified 3,751 out of BP's 14,542 (26%) categories in our text corpus. We found 491 of 754 (65%) categories from the CL. Categories from the BP co-occurred 70,967 times with CL categories.

Using our method, we identified a total number of 202,627 co-occurrences between categories. After applying our tests, 157,894 co-occurrences produced test values distinct from 0. The remainder obtained a test value of 0 due to numerical restrictions. They were subsequently excluded, because they were indistinguishable from the absence of co-occurrence. We illustrate the quantiles obtained for different p-values in our six tests, τ^i, in Table 3. The distribution of scores for τ^1 and τ^6 are shown in Figure 2. The remaining plots are included in the supplement S1.

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 more restrictive, i.e., they include fewer co-occurrences for a given cutoff. Similarly, tests including the variance for scores are generally more restrictive than tests that are not weighted by the variance of score distributions. In this sense, the tests τ^5 and τ^6 are the most restrictive.

Table 4 shows example associations, and Table 5 shows the kind of relationship between categories that our tests identified for the 100 top-scoring results with respect to the test τ^1. The *has-participant* relation is defined in the OBO Relationship Ontology (RO) [5] as a relation that holds between two categories, where every instance of one category participate in some instance of the other. We define the *Participates-in* relation as a relation between two categories: C_1 *Participates-in* $C_2 \Leftrightarrow \forall x,t_1(instanceOf(x,C_1,t_1) \rightarrow \exists t_2,y\ (instanceOf\ (y,C_2,t_2) \wedge participates\text{-}in\ (x,y,t_2)))$, where *participates-in* is the primitive participation relation between individuals as defined in the RO. We extend the definition of *located-in* in the RO 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 associations 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 sometimes identify too specific or too generic cell types as in the case of the association

Table 3. p-quantiles for different p-values for all tests.

p-value	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

Given a p-value (first column), the quantiles show the result of each test for which p-values are below the quantile.

doi:10.1371/journal.pone.0010996.t003

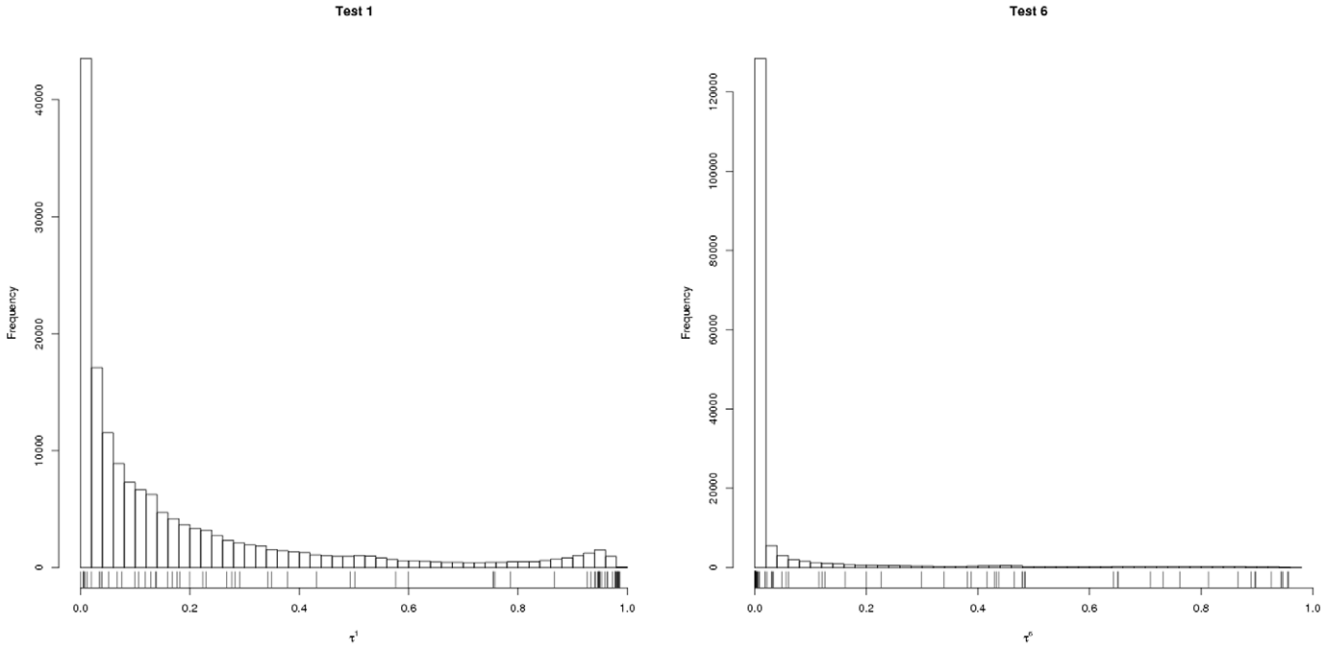

Figure 2. Distribution of test results. The plot on the left shows the distribution of the test results for τ^1. On the right, the same is shown for τ^6. It can be seen that a test using the minimum function (τ^6) is more restrictive than a test using the geometric mean (τ^1). Furthermore, weighting the tests with the CDFs of the variances (τ^6) produces stronger results than the basic test (τ^1). The test results of the GO-CL dataset for each test are displayed below the distributions.
doi:10.1371/journal.pone.0010996.g002

between *basophil degranulation* and *mast cell*. Finally, 6 out of 100 associations in our sample seem erroneous.

We were not able to 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 from the OBO Foundry. The dataset contains manually verified relations between categories from the GO and the CL that have been extracted using pattern matching on category names [23]. As 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 6 shows the percentage of significant co-occurrences within these 73 relations for different cutoffs in our six tests. Figure 2 shows the distribution of the 73 pairs with respect to τ^1 and τ^6.

Table 4. Association examples.

CL	GO
Myoepithelial cell	Milk ejection
Oocyte	Meiotic anaphase I
Osteoclast	Protein geranylgeranylation
Neuroblast	Neuron recognition
Keratinocyte	Keratinization
Sensory neuron	Optic nerve formation
Motor neuron	Spinal cord development
Protoplast	Photosynthesis
Lymphocyte	Chloroplast fission

The results in this table were above the quantile 0.9 in all six tests. While the kind of relation between the categories is apparent for most results, some, like the relation between lymphocytes and chloroplast fission, remain dubious.
doi:10.1371/journal.pone.0010996.t004

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

Important potential applications for our tests arise from the fact that annotations of a large set of biomedical ontologies satisfy the conditions for our tests. Annotations satisfy the True Path Rule [3]: if two categories C and D stand in the *is-a* or *part-of* relation, then any annotation of C is also an annotation of D. Therefore, if gene annotations are used as graph decorations for the two input graphs of our method, the conditions for applying our tests are satisfied. For detecting associations between annotations, an appropriate score function must be chosen based on the hypothesis that is to be tested.

Another potential application of our tests lies in the field of relation extraction. The evaluation of our tests with the GO and

Table 5. Manually identified ontological relations in the 100 top-scoring association results with respect to τ^1.

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

doi:10.1371/journal.pone.0010996.t005

Table 6. Evaluation of our approach with respect to the GO-CL dataset [23].

Recall	τ^1	τ^2	τ^3	τ^4	τ^5	τ^6
99%	0.004	0	0	0	0	0
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

The dataset we used for comparison consists of the 73 relations from the GO-CL crossproduct [23] found in our text corpus. Columns two to seven show the cutoff values required to identify the percentage given in column one of associations as significant using tests one to six. For example, at a cutoff of 0.502, 50% of the relations found in the dataset were significant according to test τ^1.

doi:10.1371/journal.pone.0010996.t006

CL reveals that we are able to detect biologically relevant associations between these ontologies. 94 of the best 100 associations retrieved by τ^1 have biological meaning, as shown in Table 5. Although our approach is unable to detect the types of the biological relations, the associations provide a good starting point for an elaborate approach to the extraction of biological relations.

Our method is designed for the detection of associations between two DAGs. However, it can be generalized to test for associations between n graphs. The result of the tests would then be significant n-ary associations between n nodes from n graphs.

Conclusions

We developed a family of novel statistical tests for associations between two directed acyclic graphs. The tests account for the graphs' topologies and test for relevance and specificity of associations. The tests are suitable for the detection of associations between categories from two biomedical ontologies, in particular those which comply with the OBO criteria [24].

In an extensive use-case, we applied our tests to the discovery of associations between categories from the Gene Ontology and the Celltype Ontology that were decorated with the number of occurrences and co-occurrences of the categories' labels in a large corpus of full-text articles. Our results show that a large proportion of the associations discovered by our tests are biologically relevant relations.

The family of tests is implemented in a Java library, which is available as free software from our project webpage at http://bioonto.de/pmwiki.php/Main/ExtractingBiologicalRelations.

Supporting Information

Supplement S1 Statistical tests for associations between two directed acyclic graphs and their application to biomedical ontologies.

Found at: doi:10.1371/journal.pone.0010996.s001 (0.14 MB PDF)

Acknowledgments

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

Author Contributions

Conceived and designed the experiments: RH ACNN MD JK. Performed the experiments: RH ACNN MD. Analyzed the data: RH ACNN MD. Contributed reagents/materials/analysis tools: RH ACNN MD. Wrote the paper: RH ACNN MD JK.

References

1. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech 25: 1251–1255.
2. Golbreich C, Horrocks I (2007) The OBO to OWL mapping, GO to OWL 1.1! In: Golbreich C, Kalyanpur A, Parsia B, eds. Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions, Innsbruck, Austria, Jun 6–7. Aachen, Germany: CEUR-WS.org, volume 258 of *CEUR Workshop Proceedings*.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–29.
4. Beissbarth T, Speed TP (2004) Gostat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics 20: 1464–1465.
5. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, et al. (2005) Relations in biomedical ontologies. Genome Biol 6.
6. Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome Biology 6: R21.
7. Pesquita C, Faria D, Falco AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5: e1000443.
8. Wu Z, Palmer MS (1994) Verb semantics and lexical selection. In: Pustejovsky J, ed. Proceedings of the 32th Annual Meeting on Association for Computational Linguistics (ACL '94), June 27–30, 1994, New Mexico State University, Las Cruces, New Mexico, USA. Morgan-Kaufman Publishers, San Francisco, CA, USA. pp 133–138.
9. Wu H, Su Z, Mao F, Olman V, Xu Y (2005) Prediction of functional modules based on comparative genome analysis and gene ontology application. Nucleic Acids Res 33: 2822–2837.
10. Wu X, Zhu L, Guo J, Zhang DY, Lin K (2006) Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucl Acids Res 34: 2137–2150.
11. del Pozo A, Pazos F, Valencia A (2008) Defining functional distances over gene ontology. BMC bioinformatics 9: 50+.
12. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95). San Mateo, CA: Morgan Kaufmann. pp 448–453. URL citeseer.ist.psu.edu/resnik95using.html.
13. Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine learning (ICML-98). Madison, Wisconsin.
14. Couto FM, Silva MJ, Coutinho PM (2005) Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management. New York, NY, USA: ACM. pp 343–344. doi:http://doi.acm.org/10.1145/1099554.1099658.
15. Othman RM, Deris S, Illias RM (2008) A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. J of Biomedical Informatics 41: 65–81.
16. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of go terms. Bioinformatics 23: 1274–1281.
17. Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5: 199–220.
18. Herre H, Heller B, Burek P, Hoehndorf R, Loebe F, et al. (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.
19. Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 33: 783–786.
20. Gaudan S, Yepes AJ, Lee V, Rebholz-Schuhmann D (2008) Combining evidence, specificity, and proximity towards the normalization of gene ontology terms in text. EURASIP Journal on Bioinformatics and Systems Biology 2008: 9.
21. Manning CD, Schütze H (1999) Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.
22. Pantel P, Lin D (2002) Discovering word senses from text. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Special Interest Group on Knowledge Discovery in Data. New York, NY, USA: ACM Press, ISBN:1-58113-567-X. pp 613–619.
23. Bada M, Hunter L (2007) Enrichment of obo ontologies. Journal of Biomedical Informatics 40: 300–315.
24. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech 25: 1251–1255.