BIOINFORMATICS

Vol. 20 no. 2 2004, pages 191–198
DOI: 10.1093/bioinformatics/btg390

Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network

Jonathan D. Wren[1,*] and Harold R. Garner[2]

[1]Advanced Center for Genome Technology, Department of Botany and Microbiology, The University of Oklahoma, 620 Parrington Oval, Rm. 106, Norman, OK 73019 and [2]McDermott Center for Human Growth and Development, Departments of Biochemistry and Internal Medicine, Center for Biomedical Inventions, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75390, USA

Received on April 16, 2003; revised on May 30, 2003; accepted on July 4, 2003

ABSTRACT

Motivation: There is a general scientific need to be able to identify and evaluate what any given set of 'objects' (e.g. genes, phenotypes, chemicals, diseases) has in common. Whether it is to classify, expand upon or identify commonalities and functional groupings, informational needs can be diverse and the best source to identify relationships among a potentially heterogeneous set of objects is the scientific literature.

Results: We first establish a network of related objects by their co-occurrence within MEDLINE records. A set of objects within this network can then be queried to identify shared relationships, and a method is presented to score their statistical relevance by comparing observed frequencies with what would be expected in a random network model. Using Gene Ontology (GO) categories, we demonstrate that this method enables a quantitative ranking of the 'cohesiveness' of a set of objects and, importantly, allows other objects related to this set to be identified and evaluated for their 'cohesion' to it.

Contact: Jonathan.Wren@OU.edu

Supplemental information: A list of ranked genes related to each GO category analyzed can be found at http://innovation.swmed.edu/IRIDESCENT/GO_relationships.htm

INTRODUCTION

The number of articles indexed by MEDLINE is growing exponentially, reflecting an explosion of information driven, in part, by technological improvements. Similarly, new entities of research interest such as genes, diseases, phenotypes and chemical compounds (hereafter simply referred to as 'objects') are discovered regularly. As a consequence, our relative awareness of new research entities and new research discoveries among known entities is decreasing. Complicating this matter is the increasing rate by which data is being gathered and presented to researchers for analysis. Microarrays, for example, can gather tens of thousands of data points to analyze transcriptional response to stimuli. Clustering analyses can group similar response profiles, but says little about similar purpose or function beyond that. When dealing with datasets, especially large ones, there is a need to identify 'cohesiveness' among a set of experimental variables. That is, to identify whether or not a set (or subset) of these variables have been studied together, involved in a common response or pathway, or are sufficiently distinct from random noise or a non-specific response. Also, there is a need to identify what these sets of experimental variables have in common to assist in interpretation of experimental results. Specific research interests can be very diverse, such as identifying genetic pathways affected by a change in experimental conditions, new genes that may be a member of a pathway, or drugs that affect a similar set of genes. As such, there is a need for a general method of approaching the problem and addressing the potential heterogeneity of informational needs.

Searching MEDLINE for information on genes can be a daunting task, as the number of articles published in MEDLINE containing the names of known genes ranges from 0 (unpublished data, yet transcript identified) to over 195 000 (Insulin). MEDLINE contained an estimated 12.6 million records at the time of this writing and is growing at an annual rate of ~500 000 records/year, making manual evaluation of large sets problematic at best. However, awareness of commonalities among experimental variables is central to the process of insight and discovery. As such, efforts to link literature information to experimental data provided by microarrays have recently been the focus of much effort (Masys, 2001; Noordewier and Warren, 2001).

Useful methods of linking genes to informational descriptors have been used in programs like MedMiner (Tanabe *et al.*, 1999) and ARROGANT (Kulkarni *et al.*, 2002), but more sophisticated methods of analysis are needed to tie information found within the scientific literature to a set of

[*]To whom correspondence should be addressed.

variables. Toward this end, methods have been developed such as mapping responding genes to a core set of relevant literature based upon a single best 'kernel' document (Shatkay *et al.*, 2000), giving the user the ability to identify keywords relevant to the retrieved documents. Masys *et al.* developed a method to map keywords within articles regarding a set of genes to a MeSH keyword hierarchy using the UMLS Metathesaurus (Masys *et al.*, 2001). So far, these efforts to identify literature-based shared relationships have predominantly centered upon microarrays, but there is a more global need. For example, clinicians could analyze a set of phenotypes to identify associated diseases or chemicals in the hopes they might provide insight into disease etiology or pharmacology.

Herein we propose a method to enable object sets to be scored for their 'cohesiveness', as judged by their co-occurrence within the scientific literature (MEDLINE). Importantly, this method also enables other objects to be identified and evaluated for their potential 'cohesiveness' or relevance in relation to the analyzed set. To accomplish this, MEDLINE is exhaustively scanned for potential relationships between objects by identifying co-occurrences within the same record. Term co-occurrence has been used to find tentative relationships between objects such as genes (Stapley and Benoit, 2000; Jenssen *et al.*, 2001), proteins (Blaschke *et al.*, 1999) and drugs (Rindflesch *et al.*, 2000). Since co-occurrence does not *necessarily* reflect the existence of a meaningful relationship, we used Fuzzy Set Theory to assign a weight to the relatedness of two objects based upon their frequency of co-occurrence (see Systems and Methods). By processing all MEDLINE records, a comprehensive network of tentative relationships is created that enables us to evaluate the relatedness of a set of objects based upon the relationships they share (conceptually illustrated in Fig. 1).

Assigning a measure of 'cohesiveness' to a set allows researchers to infer that an experimental grouping is purposeful (assuming the grouped objects are adequately represented within the literature). When used to analyze relationships shared by a set of objects, general 'themes' can be identified (e.g. cancer, apoptosis, diabetes) along with statistically exceptional groupings within the list (e.g. drugs affecting the activity of a group of genes). Finally, it provides a method to identify 'missing members' in a set, by their relatedness to the group as a whole. The method we propose is broadly applicable to a number of scientific analyses, and here we choose the analysis of Gene Ontology (GO) categories to benchmark its performance.

The need for a dynamic controlled vocabulary in biology has been established (Ashburner *et al.*, 2000) and efforts are underway to categorize known genes in terms of their ontology (http://www.geneontology.org/). GO construction is accomplished primarily via a manual, volunteer effort. A large number of ontological classifications have yet to be made, and there is a need to reassign function in response to changing knowledge as well as ascertain which classifications have yet to be assigned based upon current knowledge. Furthermore, it is particularly difficult to evaluate how 'complete' an ontology is with reference to current knowledge when such knowledge spans thousands of articles or more. Thus, the value of automation to aspects of this process has been recognized. In terms of identifying the functional coherence of gene groups, Russ Altman *et al.* have applied and evaluated several approaches that involve comparison of document similarities, of which the method of neighborhood divergence per gene (NPDG) stands out as the most accurate (Raychaudhuri and Altman, 2003), achieving in one study 79% recall at 100% precision (Raychaudhuri *et al.*, 2002b). A minor drawback of this approach, however, is its reliance upon an index relating each object (genes in this case) to a document, which may not always be established—particularly a concern for non-gene objects such as phenotypes and drugs. Furthermore, this method does not explicitly identify which factors make a group 'cohesive'. The author's note: '*Neighbor divergence determines whether a group of genes has a coherent function. It does not tell us the function*' (Raychaudhuri *et al.*, 2002b). Nonetheless, using similar document classification methods, the same authors made significant inroads into automated GO annotation by associating genes with GO codes (Raychaudhuri *et al.*, 2002a). They demonstrated that a supervised machine learning approach can be used to predict gene annotation by assigning GO codes to abstracts and, by extension, the genes associated with the abstracts. The downside is that a document training set must be obtained for

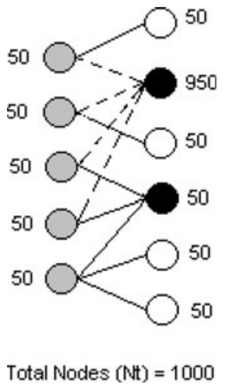

Fig. 1. Identifying shared relationships within a network. A set of objects (gray nodes) is analyzed to find connections shared by members of the set (black nodes). White nodes represent connected but unshared nodes, while numbers adjacent to the nodes represent how many total connections they have within this highly hypothetical network of 1000 nodes. When several members of the set are connected to an object in the network with many connections (black node connected by dashed lines), this is a relatively non-specific relationship since, by definition, sharing such a connection is common to many objects (here, 95% of all objects). Conversely, if another node is connected to more members of this set than would be expected by chance (black node connected by solid lines), then its relatedness to this set can be considered greater despite fewer total connections.

each GO code to be analyzed, and the association of genes to ontology codes is by implication rather than directly.

The method presented here is used to analyze a network of related objects and we evaluate how well topical groupings can be quantitatively discerned from random groupings in terms of their cohesiveness. We reasoned that these 'cohesive subnetworks' could be used to identify other objects related to members of a set, and evaluate them in terms of *their* cohesiveness to the set. This type of analysis, among other things, could be used to identify 'missing members' of a set. As such, it could represent a practical way of obtaining a set of unannotated genes that either belong in an ontological category, or are highly related to it. If so, then this method could represent a powerful automated manner of assisting in the ontology development effort. These suggested new members could, after evaluation, be added to the original set. Genes related to an ontology category for reasons other than membership in it (e.g. associated with a closely related process) could be eliminated from the list and the number of remaining members would provide a dynamic estimate of how much curation remains, assuming the relationship network was kept up-to-date with the most current literature. Such an estimate would, of course, be limited by the method itself as well as the availability of literature documenting potential relationships.

SYSTEMS AND METHODS

Code was developed in Visual Basic 6.0 (SP5) using ODBC extensions to interface with a Microsoft Access 2000 database, with database queries written in SQL. The object database consists of primary names and synonyms for genes, diseases, phenotypes and chemical compounds and was constructed using entries obtained from the following databases:

Database (Reference)	Location
OMIM (Hamosh *et al.* 2000)	ftp://ftp.ncbi.nlm.nih.gov/repository/OMIM/omim.txt.Z
GDB	http://gdbwww.gdb.org/gdb/advancedSearch.html
HGNC (Povey *et al.* 2001)	http://www.gene.ucl.ac.uk/public-files/nomen/nomeids.txt
LocusLink (Maglott *et al.* 2000)	ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL.out_hs.gz (Gene names & ID) ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL_tmpl.gz (GOs)
MeSH (Lowe and Barnett 1994)	http://www.nlm.nih.gov/cgi/request.meshdata (MeSH Trees file)
MEDLINE	National Library of Medicine http://www.nlm.nih.gov
GO (Ashburner *et al.* 2000)	http://www.geneontology.org

Over 12 500 000 MEDLINE records were processed (1967—November 2002) to catalog all co-occurrences of these object names (or their synonyms) found within each record, along with their frequency and whether they co-occurred in the same sentence or abstract. Acronyms were resolved with a set of heuristic rules (Wren and Garner, 2002) when defined in text, and undefined acronyms flagged as potentially ambiguous were ignored (a definition must comprise at least 95% of all known definitions to be considered unambiguous). When two objects co-occur within a record, Fuzzy Set Theory was used to assign a value of 'relatedness' to them ranging from 0 (unrelated) to 1 (related) by the formula $P(related) = 1 - r^n$. Here, r represents the probability that assigning a relationship based upon the co-occurrence of objects is an error, and n is the frequency of co-occurrence. When co-occurring in a sentence the probability was estimated at 0.83, while in the same abstract it was 0.58, estimates similar to those obtained by others (Ding *et al.*, 2002). This is calculated for both sentence and abstract co-occurrences and the greater of the two values used. While other, more robust, statistical approaches are available for calculating the strength of association between terms (Wilbur and Yang, 1996), at this point in the algorithm we are more interested in simply whether or not there is a non-trivial relationship between terms.

Details on the GO and Locuslink database versions used are located on the supplemental web page information.

ALGORITHM

In this section, we will adopt terminology from graph theory and refer to objects as 'nodes' and relationships (co-citations) as 'connections', which are equivalent to the 'edges' between nodes. We wish to evaluate the statistical significance of an observation that n nodes within a set, consisting of a total of t nodes (B_t, where $n \leq t$), are all connected to another node (A), which itself may or may not be a member of the set. To do this, we compare the observed number of connections, n, against what would be expected by chance if the same t nodes were connected randomly. This is assuming a random network of the same size as the literature network, and that the nodes being analyzed in the random network have the same number of connections as the set B_t as well as A. The solution will be a function of the connectivity of each node in the set, B_t, as well as the connectivity of A. For example, if A were connected to every node in the network we would expect that B_t would share exactly t connections with A regardless of the connectivity of any or all nodes within B_t (assuming each node in B_t is connected to at least one node in the network).

First, we want to calculate the probability that a node (B) within the set B_t is connected to A. Assuming nodes are randomly connected in a network with a total of N_t nodes, the probability that B will be a node connected to A (written as $B \rightarrow A$) is given as K_A/N_t where K_A is the total number of nodes connected to A. Similarly, the probability that A will be a node connected to B (written as $A \rightarrow B$) is given as K_B/N_t where K_B is the total number of nodes connected to B. For simplicity, we assume $A \neq B$. In the literature network these events are not independent, such that if $B \rightarrow A$ then $A \rightarrow B$, and this dependency can be written as $A \longleftrightarrow B$. However, we cannot know a priori if a dependency exists and so in the random network model we must assume independence. Then by comparing how many connections between nodes are observed in a given model to what would be expected by chance, assuming independence, we arrive at a statistical measure of how exceptional any given set of connected nodes is. So, in a random network:

$$P(A \longleftrightarrow B) = P(A \rightarrow B) \quad \text{OR} \quad P(B \rightarrow A) \qquad (1)$$

This probability is more easily represented in mathematical terms as the probability B is not related to A and vice versa, written as NOT ($P(A \notin B)$ AND $P(B \notin A)$), where the symbol $\notin$ is used here to mean 'is not connected to'. This probability in mathematical terms is:

$$P(A \longleftrightarrow B) = 1 - \left(1 - \frac{K_A}{N_t - 1}\right) \times \left(1 - \frac{K_B}{N_t - 1}\right) \quad (2)$$

The denominator will be $N_t - 1$ if we assume a node cannot be connected to itself, otherwise it will be N_t. Intuitively, we expect that if $K_A = N_t - 1$ or $K_B = N_t - 1$ then $P(A \longleftrightarrow B) = 1$, since the number of connections to one node does not matter if the other node is connected to all nodes. This formula applies for all non-zero values of K_A and K_B. Random network simulations were conducted to validate this formula (data not shown). Summing the probability of each individual connection, we can extend this formula to estimate the expected number of connections that a set of nodes, B_t, would share with another object, A, by the equation:

$$E(N_{A \leftrightarrow B_t}) = \sum_{i=1}^{t} 1 - \left(1 - \frac{K_A}{N_t - 1}\right) \times \left(1 - \frac{K_{B_i}}{N_t - 1}\right) \quad (3)$$

Dividing the number of observed connections (Obs) between B_t and A by the number of connections we would expect by chance (Exp) provides us with a value reflecting how exceptional the observed number of connections is. We can estimate the statistical significance of this ratio by calculating or approximating (through simulations) the SD associated with the expectation value for each set size. Here, we will consider a number of Obs statistically significant if its value is 2 SD above the expected mean. Assuming a normal (Gaussian) distribution, the area under the curve from $+2\sigma$ to $+\infty$ represents only 2.5% of the total area under the curve and we can thus assert significance at a 97.5% confidence level.

This formula provides a weighting adjustment for each connected node based upon its overall connectivity. When using this formula to evaluate objects related to a set, we are essentially asking the question 'How *specific* is the relationship of this object to the set analyzed?' For example, a widely studied object such as 'cancer' may be related to 4 out of 10 genes in a set, but this result could also be obtained by choosing 10 genes randomly. Thus, the formula evaluates how specific a relationship is to a set.

IMPLEMENTATION AND RESULTS

Sets of objects (3–100 members) were chosen at random from the literature-derived network. Each object in the set had at least one connection within the network. All other objects connected to at least two members of this set were evaluated according to Equation (3), while objects connected to fewer than two of the set were discarded. An average was calculated for all these connecting objects. This was repeated 100 times for each size set to obtain a set size average and SD for the set size average. The same was then done for sets of genes within each ontology category, except the sets were displayed as individual data points rather than averaged so that the distribution can be visualized. We hypothesized that the average Obs/Exp score should be higher due to a larger number of shared relationships within the set and, in fact, this is what we see in Figure 2. As shown in this graph, the range of values is much greater for smaller size sets and converges toward 1.0 as the set size increases. To see why this is necessarily true, imagine a set size increasing to the point where the set encompasses all objects. By definition, there is nothing specific or cohesive about such a set. For smaller sets, it is apparent that there is much more overlap with the random sets than is observed for the larger sets. To identify a good cutoff range for classifying a set of objects as 'cohesive', we compared specificity and sensitivity associated with classifying sets 1, 2 and 3 SD above the average Obs/Exp score for random sets of the same size (Fig. 3). Based upon this analysis, 2 SD (σ) appears to offer the best tradeoff between specificity and sensitivity.

We examined some of the topical entries that scored within 2σ of the random average to see if their low scores might perhaps be in error. We find that a number of ontological categories can have genes that serve a common purpose, yet are sufficiently separate in terms of their genetic associations that they are not frequently mentioned together in the literature (e.g. sensory perception genes, anion transporters). Similarly, some genes have little associated literature and few relations for analysis. This represents a potential limitation of the method.

Using the set averages for randomly assembled sets, we reasoned that genes that share many relationships and have a high Obs/Exp ratio with respect to the genes in a given ontological category, but are not themselves included in the category, might represent an enriched set of candidate genes for possible inclusion in the ontological category analyzed. Table 1 shows an example of a set of genes within

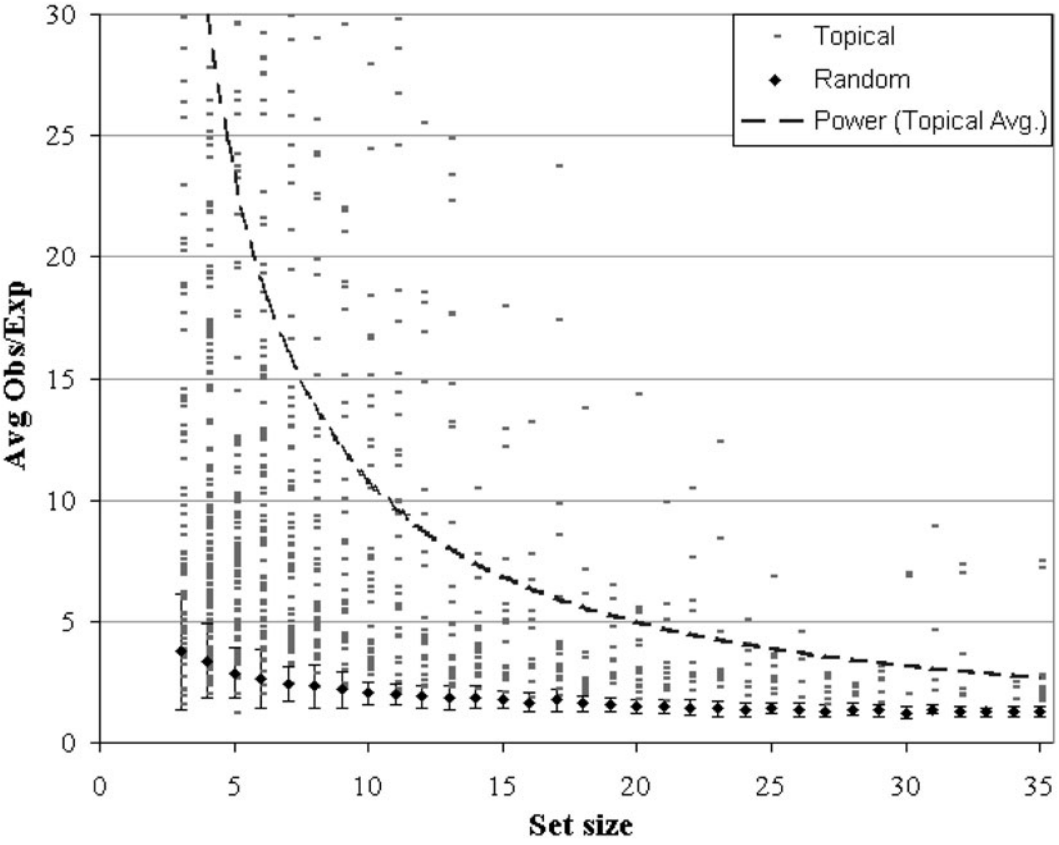

Fig. 2. Comparison of the average observed to expected ratio (Obs/Exp) for topical and random sets. Sets of objects, ranging in size, were either generated randomly or obtained from classifications within the GO database. For the random sets, only the average Obs/Exp value of all sets analyzed is shown, along with bars indicating 1 SD. The average Obs/Exp is shown for each individual topical set analyzed, along with a power-law trend line fit to the topical set size averages. There are 127 data points with an Obs/Exp above 30 that are not shown in this graph.

an ontological category (brain development) along with the number of relationships each gene has within the literature-based network. Table 2 shows the output produced by analysis of this set using the method described. Within this table are a number of object names related to the genes in Table 1, and illustrates in part the nature of the problem we are attempting to address. Some of these relationships, while perhaps true, are not particularly exceptional such as the objects ‘tumor’, ‘nucleus’ and ‘apoptosis’, which are all very highly-related (common) objects within the literature-based network, and their non-specificity to the set is reflected by a low Obs/Exp ratio. Examining the gene names within this list, however, reveals a number of genes also implicated in brain development but not annotated as such within Locuslink’s GO (at the time of this writing), such as engrailed human homologs EN1 and EN2 (Sarnat *et al*., 2002), SHH (Marti and Bovolenta, 2002), as well as BMP-4 and FGF8 (Crossley *et al*., 2001). These genes have a much higher Obs/Exp ratio, suggesting a strong association with this ontological category. Another gene name, caudal, is in the list but scores low because ‘caudal’ is also a word frequently used to describe structures toward the tail end of the body.

Genes associated with the set of genes in each GO category, but not within the category, were output for further analysis. A total of 163 791 annotations were predicted. Associations by co-mention are based solely upon the gene name and no attempt to discern species is made when scanning MEDLINE. However, the ontology annotation for each species within Locuslink is compared to the identified literature association for each gene name. Thus, a number of the predicted associations on the list will be for genes in species in which the ontology association has not been annotated, but may be annotated in a homolog. For example, the gene GRM3 (metabotropic glutamate receptor 3) is currently annotated with the GO term ‘synaptic transmission’ in humans (Locuslink ID# 2913) but not in rats (Locuslink ID# 24416).

We evaluated gene names associated with ontology categories identified by this method for evidence in the literature suggesting they should be included in the category. Such evidence was said to exist if it suggested the gene play a direct

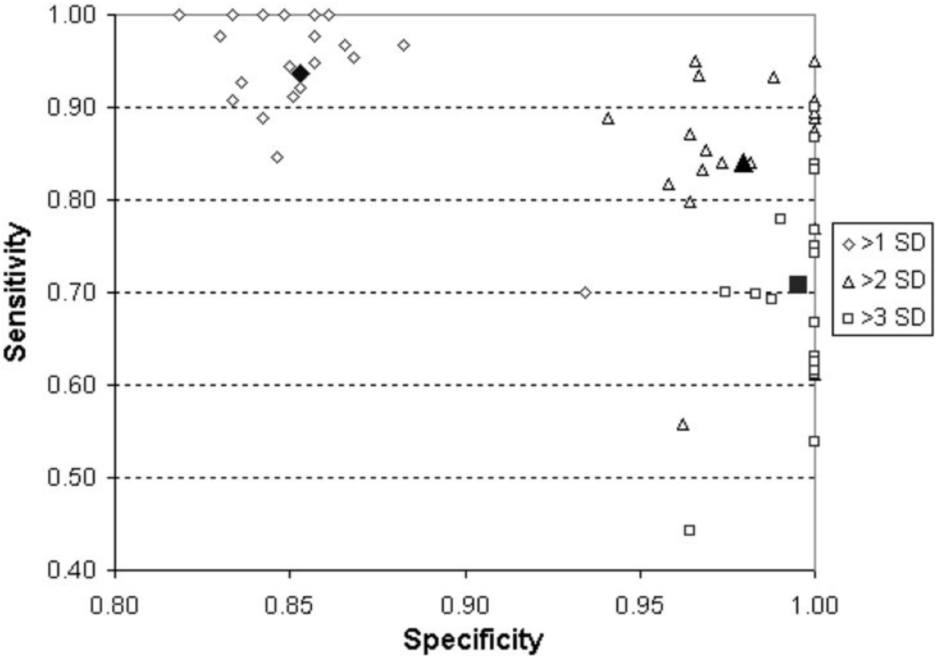

Fig. 3. Specificity versus sensitivity curve associated with the use of cutoff values when classifying a set of objects as 'cohesive'. Cohesiveness is determined by how much higher a set's average Obs/Exp score is from the random average. Each GO set, ranging in size from 3 to 23 members and comprising a total of 1346 categories, was compared with a number of randomly assembled sets equal to the number of categories of the same size available for analysis (average # of random decoys per topical set analyzed = 67.3). Specificity (precision) and sensitivity (recall) were calculated for each set size, when using cutoffs of 1, 2 and 3 SD above the average Obs/Exp score for random sets of the same size. Hollow symbols represent data points for each set size, while larger filled symbols represent an overall average for all set sizes analyzed. While 3 SD offers the highest classification specificity, there is a tradeoff in sensitivity.

Table 1. Genes in the ontological category of 'brain development' (GO ID# 7420) at the time of analysis, sorted by the number of related objects identified within MEDLINE

Gene Name	# relationships	LocusLink ID
NT-3	1111	LL:18205
TTF-1	425	LL:21869
BF1	406	LL:2290
DLX2	152	LL:1746
PGDH	152	LL:26227
SIX3	140	LL:6496
Hesx1	128	LL:15209
FMR2	121	LL:2334
ZIC	101	LL:7545
ZIC2	88	LL:7546
BMI1	87	LL:648
Cart-1	69	LL:8092
BF2	68	LL:2291
RB18A	21	LL:5469
Lhx6	20	LL:26468
hyh	14	LL:15591
Vax1	12	LL:22326
PITPNM	11	LL:9600
UNC5C	6	LL:8633
NKX2B	0	LL:4821

role (biological process/molecular function) or be localized in the appropriate cellular compartment (cellular component). We randomly chose 50 of the entries and conducted a literature search using the gene name(s) in concert with ontology keywords/phrases, trying various search combinations. Of the sample surveyed, 26 (52%) played a role in or were a part of the ontological category, 12 were related to the category in some way but did not belong in it, nine genes were not related in any direct or obvious manner, and three genes represented erroneous associations due to ambiguous gene symbols (e.g. CCT2 which stands for 'chaperonin subunit 2' in mammals but 'phosphocholine cytidylyltransferase 2' in *Drosophila*, and MT2 which stands for 'metallothionein 2' in mammals but 'methyltransferase 2' in *Drosophila*).

To aid in user evaluation and hopefully assist in the ontology development effort, the list is electronically available. Each gene in the list has been hyperlinked to its Locuslink ID. The complete list can be accessed at http://innovation.swmed.edu/IRIDESCENT/GO_relationships.htm

DISCUSSION

The method we have developed can be applied to a number of scientific questions concerning the known relatedness of a set.

Table 2. Objects related in the literature to one or more of the genes in Table 1 (only first 21 relationships shown), sorted by the total number of shared relationships identified within the network.

Object name	# shared	Expect	Obs/Exp	Locuslink ID
Nervous system	14	6.28	2.23	
Transcription factor	14	4.20	3.34	
Neurons	13	6.16	2.11	
Tumor	13	9.55	1.36	
Fibroblasts	10	5.51	1.81	
Lymphoma	9	3.81	2.36	
Nucleus	9	6.81	1.32	
SHH	9	0.46	19.48	LL:6469
Alternative splicing	8	2.43	3.29	
Secreted	8	4.40	1.82	
Apoptosis	7	4.65	1.50	
DNA-binding protein	7	1.54	4.54	
Hypoplasia	7	2.32	3.02	
Oncogene	7	2.03	3.45	
Zinc	7	4.02	1.74	
BMP-4	6	0.39	15.22	LL:652
Caudal	6	2.52	2.38	LL:1044
Cysteine	6	3.95	1.52	
Ectodermal	6	1.05	5.70	
Engrailed	6	0.27	22.01	LL:2019
FGF8	6	0.28	21.68	LL:2253

Four out of the five genes on the list have high Obs/Exp ratios, suggesting their presence on the list is due to strong relationships with the specific members of the 'brain development' category

With microarray technology, it represents a potential method of ascertaining whether or not a set of transcriptional responders contains members with documented relationships. In this way, a researcher could decide whether or not the experiment measured a specific response, giving the potential to recognize when a transcriptional response is the result of less stringent hybridization conditions or errors such as cross-hybridization (Wren *et al*., 2002). Importantly, it also allows related non-genetic factors from microarray experiments to be identified and ranked such as phenotypes, diseases, metabolites and chemical compounds.

We note that *Drosophila* names appear to represent a disproportionate number of false positives, in part because the use of gene symbol definitions diverges from mammalian standards, but also because a large number of *Drosophila* gene names are also morphologically identical to common words (e.g. *basket, arrow, red*). Ambiguous gene names have been previously noted as problematic in co-occurrence networks (Jenssen *et al*., 2001). Certainly, a higher quality output could be achieved by identifying an effective way of dealing with this ambiguity. As random objects are added to a set, its average Obs/Exp score will gradually converge toward the random noise level. Thus, the addition of noise (i.e. unrelated or random entries) to any set of related objects will reduce their 'set cohesion' and obscure existing commonalities. This could be problematic in experiments where a number of interrelated subsystems are present within a much larger whole. The quality of the output and reliability of the calculated observed to expected ratio will depend upon the ability of the experimenter to accurately define a set of interest.

Raychaudhuri *et al*. (2002a) noted that granularity of the GO codes was problematic for larger categories, where overly general categories such as 'metabolism' yielded less specific results. We observe a similar problem in that genes associated with broad categories frequently have associations with the category, but are not specific to that category alone. In part, this can be adjusted for by increasing the stringency (e.g. random average $+3\sigma$) for larger categories.

Given the volume of genetic information in the literature and the limited amount of time available to curate and develop ontologies, we feel that this type of approach can aid the process. What also may be of potential use, although yet to be determined, are the relationships that ontological categories have with other, non-gene objects such as diseases, phenotypes, chemicals or drugs. These types of relationships could suggest the creation of new ontological categories.

ACKNOWLEDGEMENTS

The authors thank Michael Huebschman and Luther White for their help in manuscript preparation. This work was funded in part by NSF-EPSCoR grant # EPS-0132534 (for JDW), the State of Texas, NIH/NCI grant CA81656 (HRG), NIH/NHLBI grant P50 CA70907 (HRG), the Hudson Foundation (HRG) and by Contract DAAD13-02-C-0079 from the Soldier Biological Chemical Command, APG to the Biological Chemical Countermeasures program of UTSW.

REFERENCES

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Ismb*, **15**, 60–67.

Crossley,P.H., Martinez,S., Ohkubo,Y. and Rubenstein,J.L. (2001) Coordinate expression of Fgf8, Otx2, Bmp4, and Shh in the rostral prosencephalon during development of the telencephalic and optic vesicles. *Neuroscience*, **108**, 183–206.

Ding,J., Berleant,D., Nettleton,D. and Wurtele,E. (2002) Mining Medline: Abstracts, Sentences or Phrases? *Pac. Symp. Biocomp.*, Kauau, Hawaii.

Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian inheritance in man (OMIM). *Hum. Mutat.*, **15**, 57–61.

Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Kulkarni,A., Williams,N., Lian,Y., Wren,J., Mittleman,D., Persemlidis,A. and Garner,H. (2002) ARROGANT: an application to manipulate large gene collections. *Bioinformatics*, **11**, 1410–1417.

Lowe,H.J. and Barnett,G.O. (1994) Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Jama*, **271**, 1103–1108.

Maglott,D.R., Katz,K.S., Sicotte,H. and Pruitt,K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.

Marti,E. and Bovolenta,P. (2002) Sonic hedgehog in CNS development: one signal, multiple outputs. *Trends Neurosci.*, **25**, 89–96.

Masys,D.R. (2001) Linking microarray data to the literature. *Nat. Genet.*, **28**, 9–10.

Masys,D.R., Welsh,J.B., Lynn Fink,J., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **17**, 319–326.

Noordewier,M.O. and Warren,P.V. (2001) Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol.*, **19**, 412–415.

Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.*, **109**, 678–680.

Raychaudhuri,S. and Altman,R.B. (2003) A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, **19**, 396–401.

Raychaudhuri,S., Chang,J.T., Sutphin,P.D. and Altman,R.B. (2002a) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, **12**, 203–214.

Raychaudhuri,S., Schutze,H. and Altman,R.B. (2002b) Using text analysis to identify functionally coherent gene groups. *Genome Res.*, **12**, 1582–1590.

Rindflesch,T.C., Tanabe,L., Weinstein,J.N. and Hunter,L. (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.*, **5**, 517–528.

Sarnat,H.B., Benjamin,D.R., Siebert,J.R., Kletter,G.B. and Cheyette,S.R. (2002) Agenesis of the mesencephalon and metencephalon with cerebellar hypoplasia: putative mutation in the EN2 gene–report of 2 cases in early infancy. *Pediatr. Dev. Pathol.*, **5**, 54–68.

Shatkay,H., Edwards,S., Wilbur,W.J. and Boguski,M. (2000) Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 317–328.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.*, **5**, 529–540.

Tanabe,L., Scherf,U., Smith,L.H., Lee,J.K., Hunter,L. and Weinstein,J.N. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214, 1216–1217.

Wilbur,W.J. and Yang,Y. (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.*, **26**, 209–222.

Wren,J.D. and Garner,H.R. (2002) Heuristics for identification of acronym–definition patterns within text: towards an automated construction of comprehensive acronym–definition dictionaries. *Methods Inf. Med.*, **41**, 426–434.

Wren,J.D., Joslin,J., Kulkarni,A., Butow,R.A. and Garner,H.R. (2002) Cross-hybridization on PCR-spotted microarrays. *IEEE Eng. Med. Biol.*, **21**, 71–75.