BIOINFORMATICS

Vol. 20 no. 3 2004, pages 389–398
DOI: 10.1093/bioinformatics/btg421

Knowledge discovery by automated identification and ranking of implicit relationships

Jonathan D. Wren[1,*], Raffi Bekeredjian[2,3], Jelena A. Stewart[2,3], Ralph V. Shohet[2,3] and Harold R. Garner[2,4]

[1]Advanced Center for Genome Technology, Department of Botany and Microbiology, The University of Oklahoma, 620 Parrington Oval Rm. 106, Norman, OK 73019, USA, [2]Department of Internal Medicine, [3]Division of Cardiology and [4]McDermott Center for Human Growth and Development, Department of Biochemistry, Center for Biomedical Inventions, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA

Received on December 11, 2002; revised on March 11, 2003; accepted on August 10, 2003
Advance Access Publication January 22, 2004

ABSTRACT

Motivation: New relationships are often implicit from existing information, but the amount and growth of published literature limits the scope of analysis an individual can accomplish. Our goal was to develop and test a computational method to identify relationships within scientific reports, such that large sets of relationships between unrelated items could be sought out and statistically ranked for their potential relevance as a set.

Results: We first construct a network of tentative relationships between 'objects' of biomedical research interest (e.g. genes, diseases, phenotypes, chemicals) by identifying their co-occurrences within all electronically available MEDLINE records. Relationships shared by two unrelated objects are then ranked against a random network model to estimate the statistical significance of any given grouping. When compared against known relationships, we find that this ranking correlates with both the probability and frequency of object co-occurrence, demonstrating the method is well suited to discover novel relationships based upon existing shared relationships. To test this, we identified compounds whose shared relationships predicted they might affect the development and/or progression of cardiac hypertrophy. When laboratory tests were performed in a rodent model, chlorpromazine was found to reduce the progression of cardiac hypertrophy.

Contact: Jonathan.Wren@ou.edu

Supplementary information: http://innovation.swmed.edu/IRIDESCENT/Supplemental_Info.htm

INTRODUCTION

There is a large difference between what *is* known and what *we* know as individuals. We are only aware of a relatively small fraction of the collective scientific knowledge within any given field. As increasing amounts of information and observations are compiled from different areas of research as individual reports, they can contribute towards a greater understanding in apparently unrelated areas when considered collectively. For example, it has been demonstrated that the useful implications of scientific discoveries can go unnoticed or unutilized because they exist only implicitly from information scattered among different areas of research (Swanson, 1986). By using software to identify words shared between article titles, Swanson and Smalheiser were able to identify common intermediates between Raynaud's Disease (a circulatory disorder restricting blood-flow to the extremities) and the dietary effects of fish oil, leading to the hypothesis and subsequent proof (DiGiacomo *et al*., 1989) that compounds within dietary fish oil could ameliorate the symptoms of Raynaud's Disease (Swanson, 1986; Smalheiser and Swanson, 1998). The term 'non-interactive literatures' was coined to explain why such a reasonable hypothesis had gone unnoticed by researchers in either field alone. Finding methods to utilize greater the biomedical literature in an automated manner to aid scientific discovery is becoming a topic of increasing interest (Yandell and Majoros, 2002).

While innovative, a keyword-based method such as that of Swanson and Smalheiser is both limiting and highly cumbersome, especially where a significant body of literature is concerned, for several reasons: First, only titles are used; second, word phrases such as 'Interleukin 6' are not taken into account, being reduced to 'Interleukin' and '6'; third, synonyms (e.g. 'IL-6') are not considered; finally, and perhaps most importantly, the number of unique keywords grows rapidly per record analyzed, providing an impractically large amount of output for any user to examine (additional

[*] To whom correspondence should be addressed.

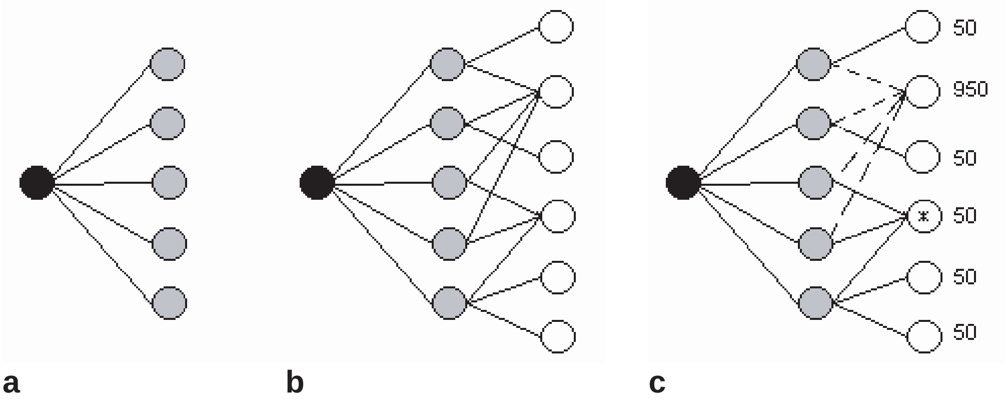

Fig. 1. Using literature-based relationships to engage in the discovery of new knowledge. (**a**) Beginning with an object of interest (black node), tentative relationships are assigned to other objects (gray nodes) when they are co-mentioned within MEDLINE abstracts. (**b**) Each related object (gray) is then queried for its relationships with other objects (white nodes). The white nodes are not directly related to the primary node and are thus only *implicitly* related, through intermediates. (**c**) The relationships shared by white and black nodes are ranked against a random network model to establish how many would be expected by chance alone, given the connectivity of each node in the set. Suppose the entire network consists of 1000 nodes and the numbers to the right represent how many connections each of the implicit nodes have to other nodes in the network. We can then assign a statistical weighting that reflects how exceptional any given implied relationship is based upon the shared intermediates. In this example, a node with 950 relationships may share many relationships with the primary node, but this is nothing exceptional because it is related to most objects in the network. It is thus down-weighted in importance (dashed lines). The connections of the gray nodes must also be taken into account, but is not shown here for simplicity.

discussion in online supplement). Others have designed literature exploration systems that overcome the first three barriers mentioned by looking for the co-occurrences of major Medical Subject Headings (MeSH) descriptors (Hristovski *et al*., 2001) or by mapping text to UMLS concepts (Weeber *et al*., 2000, 2003). The size of the domain to be analyzed remains the most significant problem, however, and has been dealt with thus far by user intervention in the selection of intermediates for analysis.

The ability to seek out novel, undocumented relationships that are logically implicit from a body of information—yet not explicitly stated within that body—has obvious scientific value. It enables us to use the current state of knowledge to infer possible new relationships that have yet to be studied. Our ability to postulate a potential relationship between two or more things depends principally upon how many common relationships we are aware of between these things, if any, to suggest that a relationship exists where none has been documented. Awareness of relationships, especially sets of relationships, is central to the human process of insight and discovery.

General approach

Herein we describe a method to identify potential relationships within the biomedical literature by defining areas of research interest such as genes, diseases, phenotypes and chemical compounds (hereafter referred to simply as ‘objects’). Beginning with an object of interest (call it ‘A’), we can identify other objects (‘B’) tentatively related to it within the literature (Fig. 1a) by identifying the co-occurrence of A and B objects within MEDLINE records (titles and abstracts). Each B object can then be queried to identify other objects (‘C’) that co-occurred with them in the literature (Fig. 1b). Each of these new objects, C, that are not themselves A or B objects, are related to A only implicitly. That is, they have no documented relationship with A, but share one or more relationships with A. This large list of implicitly related objects will contain potential discoveries of new relationships. However, because of their abundance, it is necessary to prioritize and rank these objects in some manner. To do so, we describe a method to rank relationships shared by two objects within a literature network against a random network model to evaluate how statistically exceptional any given set of shared relationships is (Fig. 1c). We also show that this ranking correlates with the probability that two objects are related as well as the strength (frequency of co-occurrence) by which they are related.

Object co-occurrences exhaustively identify potential relationships

We attempt to identify as many relationships as possible by postulating that a potential relationship exists between two objects when they are observed to co-occur within the same MEDLINE record, an approach also taken by others to identify potential relationships between genes (Stapley and Benoit, 2000; Jenssen *et al*., 2001), proteins (Blaschke *et al*., 1999) and drugs (Rindflesch *et al*., 2000). Some have used the co-occurrence of certain MeSH terms to reflect potential relationships (Hristovski *et al*., 2001; Perez-Iratxeta *et al*., 2002), but MeSH terms for each object within an abstract are not always provided, and in a number of cases (e.g. gene names) are used to reflect the existence of a more general category as opposed to a specific entity. Here, we are interested in associations between any areas of active biomedical research interest. We assemble the primary names and synonyms for genes, diseases, phenotypes and chemical/pharmaceutical compounds into a composite database so that the names can be recognized as they occur within text. These objects were chosen because they are of broad interest in biology and medicine, but the approach we present allows for any 'class' of object to be incorporated that is considered of research interest (e.g. tissue types, protein motifs, cell lines, etc.). It should be noted, however, that addition of a new object class presumes that co-citations with other object classes could be construed as meaningful. For example, country names could be added as an object class, but it is doubtful that any co-occurrences with other objects would be considered biologically meaningful or interesting.

Fuzzy logic is used to weight importance of co-occurrence

The disadvantage of using co-occurrence is that it does not always reflect the existence of a biologically meaningful relationship. To reflect this ambiguity, we borrow from Fuzzy Set Theory and model relationships as probabilistic, that is, ranging from 0 to 1, rather than binary values [for an overview of Fuzzy Set Theory see Steimann (1997) and for a thorough discussion see Klir and Yuan (1995)]. By manually surveying each object co-mentioned within a sample of MEDLINE records, we can estimate the probability that a co-mention reflects the presence of a non-trivial relationship between the two objects. This base probability can then be used to assign a fuzzy score to each relationship, reflecting the probability that one or more co-occurrences are meaningful. Since terms that co-occur more frequently are more likely to represent biologically meaningful relationships (Jenssen *et al*., 2001), each relationship is assigned a score based on the frequency and type (i.e. abstract or sentence) of co-mentions observed and their corresponding error rates (discussed in the section on 'Implementation').

By defining what objects will be recognized rather than using all words reduces the magnitude of analysis and allows a focus on relationships with a higher potential of being considered 'interesting'. Diseases and clinical phenotypes were obtained from Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al*., 2000); chemical compounds and drugs from the MeSH database (Lowe and Barnett, 1994); and genes from Locuslink (Maglott *et al*., 2000) and the Human Gene Nomenclature Committee (HGNC) (Povey *et al*., 2001). As tentative relationships between these objects are identified within text, they are entered into a database. This database enables the construction of a network of relationships, which can be queried to identify relationships shared among a set of objects and to identify novel relationships that are implicit by virtue of shared relationships.

For an implicit relationship—two objects related only through shared intermediates—it is not yet clear what statistical parameters best correlate with the probability of it representing a biologically meaningful relationship. However, we can assume that the probability of an implicit relationship (A ↔ C) being biologically meaningful would not be greater than the least probable of the two individual (A ↔ B or B ↔ C) relationships linking them, where the symbol ↔ is defined as the existence of a non-directional relationship between two objects. This is equivalent to stating that the strength of a chain is no greater than its weakest link.

SYSTEMS AND METHODS

Code was written in Visual Basic 6.0 (SP5) using ODBC extensions to interface with a Microsoft Access 2000 database, with database queries written in SQL. Analyses were run on a Pentium 4–2.4 GHz desktop PC running Windows 2000. Database entries were obtained from the following sources, all downloaded between December 13 and December 25, 2001:

Database	Location
OMIM	ftp://ftp.ncbi.nlm.nih.gov/repository/ OMIM/omim.txt.Z
GDB	http://gdbwww.gdb.org/gdb/ advancedSearch.html
HGNC	http://www.gene.ucl.ac.uk/public-files/nomen/nomeids.txt
LocusLink	ftp://ftp.ncbi.nih.gov/refseq/ LocusLink/LL.out_hs.gz (Human)
MeSH	http://www.nlm.nih.gov/cgi/ request.meshdata (MeSH Trees file)
MEDLINE	National Library of Medicine http://www.nlm.nih.gov
Genome Ontology	http://www.geneontology.org

Values used to ascertain relatedness to one of the two fuzzy sets (i.e. belonging to the category 'related' or 'not related') were based upon the probability that a co-occurrence of objects equated to a non-trivial relationship between the two (see online supplement Fig. S2). Thus, the value of relatedness can range from 0 to 1 and is estimated by the number of times (n) that the two objects were co-mentioned and the error rate (r) associated with the co-mention metric (i.e. sentence or abstract) used to establish the relationship. This formula, $P(\text{related}) = 1 - r^n$, is used to calculate the relatedness of two objects and is referred to as a veracity score. The veracity score can range in value from 0.58 (two objects only co-mentioned once within one abstract) to 1.0. Rather than summing the raw number of co-mentions shared by two objects, summing their veracity scores permits a more accurate estimate of how many relationships are truly shared based upon the known error rate.

We define the ‘strength’ of a relationship as a function of the number of times two objects have been co-mentioned and the probability that each co-mention represents a non-trivial relationship. The term ‘strength’ is used rather than frequency because we record both sentence co-mentions (C_s) and abstract co-mentions (C_a), and need a convenient way to combine the two. The strength score (S) is assigned based upon the individual co-mention error rates, r_s (17% FP) and r_a (42% FP) respectively, by the formula $S = C_s * (1 - r_s) + C_a * (1 - r_a)$.

Materials and methods for the Chlorpromazine–Cardiac Hypertrophy experiments are contained in the online supplement.

ALGORITHM

In this section, we adopt terminology from graph theory and refer to objects as ‘nodes’ and relationships (co-citations) as ‘connections’, also known as the ‘edges’ between nodes. We also define an implicitly related node (C) as one that has no direct connection to the query node (A), yet is connected to one or more intermediate nodes (B) that are simultaneously connected to A. To evaluate the potential significance of an implicitly related node, we compare the set of i nodes (B_i) shared by both the query node A and the implicit node C, against a random network model. Given that we are interested in an node A, and know from processing all literature associated with A that it is related to all nodes in the set B_i, we ask the question ‘Given the number of connections each node in the set B_i has, and the number of connections the target node (C) has, how many connections might we expect between B_i and C by chance alone?’ For example, if C were related to every node in a 1000 node network and A had 100 connections within this network, all of which were shared with C, this would be expected and therefore unexceptional. Thus, dividing the number of observed connections (Obs) between B_i and C by the number of connections we would expect by chance (Exp) provides us with a value reflecting the statistical significance of the shared connections. This value allows us to estimate the potential relevance of a set of connections. For example, if a set of connections linking a disease (A) to a chemical (C) were to encompass highly common nodes such as ‘sodium’ and ‘symptom’, we recognize that—whether true or not—these types of connections are sufficiently vague to be of little use to a scientist in postulating how A and C might have an interesting and specific connection through these intermediates. If the shared connections involve specific transporters or genes, which would not be as frequently mentioned in the literature, it becomes easier to postulate how specific actions of (C) could produce (A).

We derived an expectation value based upon the relative connectivity of each node involved. Assuming nodes are randomly connected in a network with a total of N_t nodes, the probability that a node will be connected to A is given as K_A/N_t where K_A is the total number of connections to A. The probability B will be connected to A [written as $P(B \in A)$] is K_A/N_t and the probability A will be connected to B [written as $P(A \in B)$], is K_B/N_t. Because the formula $P(A \in B)$ or $P(B \in A)$ 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 $P(B \notin A)$], we can define the probability in mathematical terms as:

$$P(A \leftrightarrow B) = 1 - \left(1 - \frac{K_A}{N_t}\right) * \left(1 - \frac{K_B}{N_t}\right) \qquad (1)$$

Intuitively, we expect that if $K_A = N_t$ or $K_B = N_t$ then $P(A \leftrightarrow B) = 1$, 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 confirm the validity of this formula (data not shown). Summing the probability of each individual relationship, we can extend this formula to estimate the expected number of connections a set of nodes, B_i, would share with another object, A, by the equation:

$$\text{Expect}(A \leftrightarrow B_i) = \sum_{i=1}^{n} 1 - \left(1 - \frac{K_A}{N_t}\right) * \left(1 - \frac{K_{B_i}}{N_t}\right) \quad (2)$$

Equation (2) is used to estimate the expected number of shared relationships between B_i and C, given the connectivity of each intermediate (shared) node in the set B_i that A is known to be connected to.

IMPLEMENTATION

Precision and recall rates are estimated

First, we estimated the precision of using co-occurrence as a method of identifying the existence of a non-trivial relationship between two objects by manually evaluating the co-occurring objects within a random set of 25 MEDLINE records (titles and abstracts). We found that two objects co-mentioned within the same sentence were more likely to be related (83%) than objects co-mentioned in the same abstract (58%). Using sentence co-mentions alone, however, would miss 43% of the non-trivial relationships within an abstract. This proportion of correct relationships among abstract co-mentions is similar to the estimates others have obtained (Jenssen *et al*., 2001; Ding *et al*., 2002). Additionally, because judgment of what constitutes a ‘relationship’ and what is ‘non-trivial’ is somewhat subjective, we attempted to estimate this error rate in a more objective way by identifying objects co-mentioned in the first half of MEDLINE (records up until approximately Nov. 1991), but not in the second half. The rationale for this approach comes from the observation that related objects (e.g. insulin–glucose) tend to be

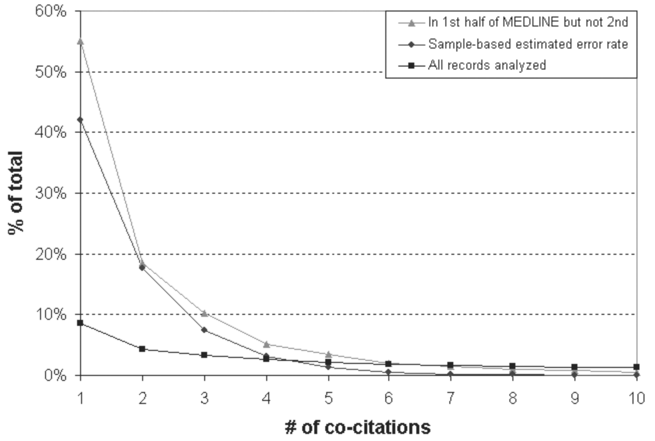

Fig. 2. Analysis of the first 10 000 co-cited objects found within the 1st but not 2nd half of MEDLINE, grouped by the total number of co-citations identified within MEDLINE for the two objects. The fact that these co-citations are non-recurring suggests that the co-citation did not reflect the existence of a relationship studied between the two. As shown here, this distribution is significantly different than the overall distribution in the co-citation frequency of the 63 836 records analyzed.

co-mentioned over the course of many studies after their first co-mention. We reasoned that if two objects are co-mentioned early (establishing their co-existence in the literature), but not again after an equal number of publications, there are several possibilities: the co-mention was the result of two unrelated topics being discussed together (e.g. incidental, broad topical coverage), the objects were once studied for a relationship but none was found or it was in error, or a relationship was established but was not of sufficient interest to warrant further study. Regardless of the exact reason, these represent non-persistent ‘relationships’, and are suggestive of a class of co-mentions (erroneous or uninteresting) that we wish to exclude. We examined the first 10 000 non-persistent co-citations found within MEDLINE and found a similar distribution as would be predicted by the error rate formula (Fig. 2), although by this method the predicted error rate would be slightly higher. This helps to confirm the accuracy of the estimates and to justify the use of a power-law decay function to represent the probability of error.

Resolving ambiguous acronyms

Acronyms were resolved as they occurred within text using an Acronym Resolving General Heuristic (ARGH) to reduce both random and systematic errors in term recognition, which operates with ∼96% precision and 92% recall (Wren and Garner, 2002). A total of 4309 acronyms were flagged as ambiguous (i.e. one definition must be >95% of all identified definitions to be considered unambiguous) and requiring resolution. The ARGH database of MEDLINE acronyms was also used to expand the acronym list for entries within the composite object database, adding 3094 acronyms to database entries that did not have an acronym specified. ARGH also identified 4786 spelling/hyphenation variants observed within MEDLINE for objects within the composite database. It is difficult to assess what impact ARGH has upon the precision or recall when processing records, as the reduction in the false-negative (FN) rate depends upon how common the variant or acronym is and reduction in the false-positive rate depends upon the acronym. For example, the gene calcitonin is associated with the acronym CT, which has a different definition within MEDLINE 96% of the time (Computed Tomography). Gene names like SOCS-3 are unambiguous and unaffected by the use of ARGH to resolve acronyms, but less than half of the definitions of SOCS-3 within MEDLINE would be recognized without the spelling variants provided by ARGH (the ARGH database can be queried at http://lethargy.swmed.edu/argh/argh.asp).

Estimated recall rate of using abstracts versus full-text articles

Abstracts presumably contain the most important findings of a report and important findings are usually reiterated in future abstracts, but it could be argued that some relationships might not be found within abstracts. To estimate this and obtain a recall rate, we calculated the total number of relationships within a domain of knowledge (MEDLINE articles) that are contained in their electronically accessible summary form (MEDLINE titles and abstracts). A set of objects mentioned within review articles was manually compiled and compared to the relationships found within MEDLINE titles and abstracts. The same list was compared to the object database to estimate what percent of object types mentioned in MEDLINE were represented in the databases used. Four objects were randomly chosen from the collective object database, representing one of each object type, with the stipulation that at least two review articles had been written about the object within the past three years. A set of review articles was selected for CTLA-4 (gene) (McCoy and Le Gros, 1999; Green, 2000; Tomer, 2001), Fragile-X Syndrome (disease) (Bardoni *et al.*, 2000; Jin and Warren, 2000; Kooy *et al.*, 2000), cachexia (clinical phenotype) (Barber, 2001; Hasselgren and Fischer, 2001; Tisdale, 2001) and dynorphin (chemical compound) (Steiner and Gerfen, 1998; Caudle and Mannes, 2000). Only objects of the same types (i.e. other genes, diseases, phenotypes and chemicals) were counted.

There were a total of 40 objects mentioned in the literature but not found in the database (2 diseases, 9 phenotypes, 7 genes and 22 chemical compounds). The 2 disease names (Graves’ Opthalamopathy and Relapsing-remitting Experimental Autoimmune Encephalomyelitis) and 9 phenotypes were not mentioned in OMIM. Three of these phenotypes, however, were simply the result of a semantic difference between the OMIM entry and the article (‘rocking’ versus ‘body-rocking’, ‘greater interocular distance’ versus 'increased interocular distance', 'fetal akinesia' versus 'akinesia'). The most problematic category was 'small molecules', for which many chemicals and drugs widely mentioned in the literature (e.g. DAMGO, DADLE, isoprenaline) were simply not found in the MeSH trees database.

In this sample, there were 181 objects found within the review articles, 141 of which were also in the composite database (78%). From the 40 objects mentioned in the reviews but not found in the database, 2 were diseases, 9 phenotypes, 7 genes and 22 chemical compounds. From these 141 database objects mentioned within the full-text of the reviews, 138 of them (98%) could be found within the body of a MEDLINE title or abstract, suggesting that most objects pertinent to a review can also be found within an abstract or title. Semantically, 124 of these 138 objects were spelled in the literature the same way they were found in the database, giving a recall rate of 90% (124/138) in terms of identifying the conceptual occurrence of database objects within textual input, and 69% (124/181) in terms of identifying relevant relationships within its domain of knowledge (MEDLINE).

Some of the FN failures to identify objects within text were systematic (e.g. the MeSH entry 5,8,11,14,17-Eicosapentaenoic Acid is almost always referred to in MEDLINE simply as eicosapentaenoic acid) while other failures varied in their rates (e.g. JNK was found to be spelled 81 different ways including 'c-Jun N-terminal kinase' 605 times, 'c-Jun NH2-terminal kinase' 154 times and 'c-Jun amino-terminal kinase' 62 times).

Creating a network of relationships using MEDLINE

A total of 12 037 763 MEDLINE records recorded from 1967 to January 2002 were processed to create a network of 3 482 204 unique relationships between objects. Approximately two-third of the objects in the database found exact literal matches within the literature, identifying at least one relationship for 22 482 of the 33 539 unique objects (85 234 total terms when including synonyms) within the database. As expected, we find a highly disproportionate distribution in the number of relationships per object (Fig. 3a), indicating the network is scale-free in nature. As such, this connectivity contributes to a rapid explosion in the number of implicitly related objects as the number of direct relationships increases (Fig. 3b). Thus, identifying implicitly related objects becomes secondary to being able to rank their potential significance. Furthermore, this also shows that the search for implicit relationships more than one level removed (i.e. $A \leftrightarrow B \leftrightarrow C \leftrightarrow D$) would likely be fruitless in the absence of any further constraints, since all non-circular connections from A are reached relatively rapidly as the domain grows to a modest size.

Ranking all implicit relationships

We evaluated whether this observed to expected ratio (Obs/Exp) we are calculating could be used to estimate the 'relatedness' of two objects solely by examining the relationships they share. To establish this, an Obs/Exp was calculated for all relationship sets (of at least 100 objects) shared by a central query object and any other object in the database, regardless of whether a direct relationship was known or not. The Obs/Exp scores were sorted from highest to lowest on the x-axis, and the strength of the relationship, if known, was plotted on the y-axis. If the strength was not known (i.e. it was an implicit relationship), then no bar was plotted. For the object 'Cardiac Hypertrophy', we see that the higher the Obs/Exp ratio, the more likely the relationship is known (Fig. 4). Furthermore, we note that the higher the Obs/Exp ratio, the stronger the relationship tends to be (i.e. the more frequently they are co-cited).

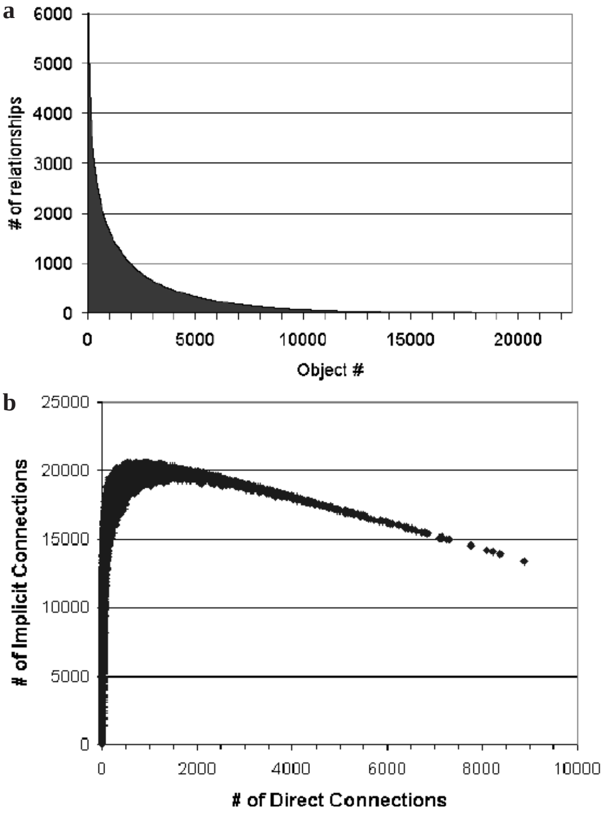

Fig. 3. The distribution in the number of relationships per object follows a scale-free power law distribution. (**a**) A relatively small fraction of the objects in the database are directly related to a large percentage of the total, contributing to a rapid explosion in the number of implicitly related objects. (**b**) As the number of direct relationships increases, the number of implicit relationships rapidly approaches the theoretical maximum, which is the total number of nodes in the network, and then decreases linearly with the number of possible implicit relationships.

To confirm that the trend observed in Figure 4 is not specific to the analysis of cardiac hypertrophy, but rather is a general trend, we randomly picked 100 objects from the database that had between 500 and 1000 relationships within the network

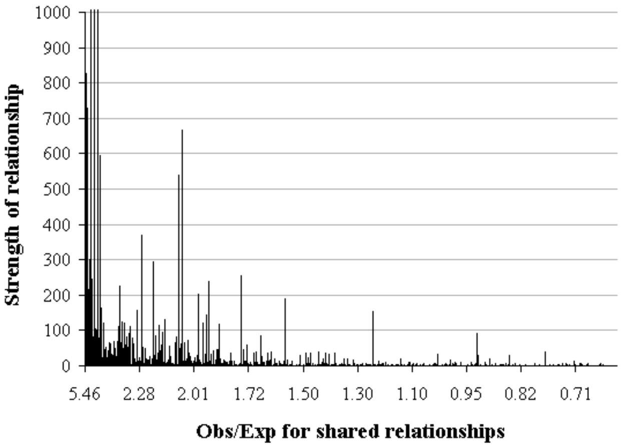

Fig. 4. Objects were ranked for their implicit 'relatedness' to cardiac hypertrophy solely on the basis of the relationships they shared (Obs/Exp). If a relationship in MEDLINE has been established, its strength (based upon frequency of co-occurrence within MEDLINE) is plotted on the y-axis, otherwise it will appear as a gap (meaning no relationship has been established). Shown is a subset of 4887 objects sharing at least 100 relationships with cardiac hypertrophy, sorted by their calculated observed to expected ratio. Due to x-axis compression, not all gaps will be visible on this graph.

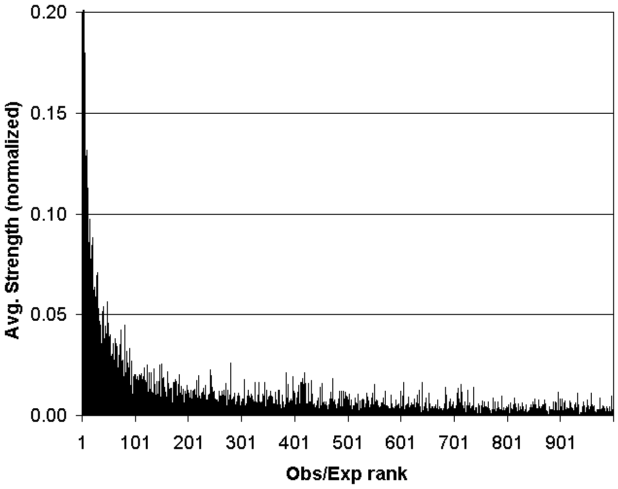

Fig. 5. The observed to expected ratio obtained from identifying and ranking shared relationships correlates with the existence and known strength of a relationship. This enables novel (implicit) relationships to be correlated with the probability they are relevant (as judged by existing relationships) and important (as judged by the strength/frequency of historical reporting).

(this range was chosen simply to ensure that the approximate scale of analysis for each object was similar). Implicit relationships were identified for these objects and ranked by their Obs/Exp values. The top 1000 Obs/Exp scores were taken for each analysis and ranked from 1 (highest Obs/Exp) to 1000 (lowest), and a normalized strength score calculated for each object analyzed, ranging from 1.0 (strongest direct relationship observed) to 0.0 (no relationship observed). Figure 5 shows this average strength plotted against the Obs/Exp rankings, indicating that this is a general trend.

In some ways, the correlation of exceptional groupings with known relationships is not too surprising, as we would expect that two objects with very similar purposes, functions, or involvement in a biological process should interact with and/or be studied with many of the same objects. This does, however, establish that the relatedness of two objects can be correlated with the statistical exceptionality of the relationships they share. More importantly, this provides us with a means to evaluate quantitatively implicit relationships by demonstrating that the Obs/Exp score correlates positively with established relationships. This numeric evaluation enables us to identify new relationships, not found within MEDLINE records, that are more likely to be logically plausible and relevant to the query object because of the relationships they share.

Wet-lab testing of *in-silico* predictions

Cardiac hypertrophy is defined as an increase in the size of myocytes that is associated with detrimental effects on aspects of contractile and electrical function in the heart. It is induced in response to environmental stimuli, such as arterial hypertension, increased cardiac work or hormonal stimuli. It is an intensively studied condition as evidenced by the 4092 articles in MEDLINE containing the key phrase 'cardiac hypertrophy'.[†] A total of 2102 unique objects were co-mentioned within all articles mentioning cardiac hypertrophy and 19 718 unique objects were implicitly related to cardiac hypertrophy through a total of 1 842 599 different paths.

Examining the shared relationships for the implicitly related objects in Table 1, we excluded 'endotoxins' from further study in part because this refers to a class of immuno-inductive compounds rather than a specific compound, and in part because endotoxins are known to have substantial effects on cardiac function that would complicate interpretation of hypertrophy. Morphine was excluded from the study because it would have substantial effects on the behavior of mice (including somnolence and reduced feeding) that would limit the dose used for study. Chlorpromazine (CPZ) was deemed more suitable for further study, in part because it is a commonly used drug and an unrecognized effect on cardiac hypertrophy could have clinical importance. A list of shared relationships between cardiac hypertrophy and CPZ is available in the online web supplement.

Chlorpromazine (CPZ) is an aliphatic phenothiazine compound used principally as an anti-psychotic and anti-emetic drug (Shen, 1999). It has a number of physiological effects

[†] Statistics are as of June 23, 2003, although this analysis was initially conducted in January 2002 when there were fewer articles than this.

Table 1. Chemical compounds within the composite database implicitly related to cardiac hypertrophy

Rank	Implicit relationship	Shared rels	Quality estimate	Expected	Obs/Exp
1	Endotoxin	1301	1025	307	3.34
2	Morphine	1217	939	283	3.32
3	**Chlorpromazine**	**1089**	**824**	**252**	**3.28**
4	Globulin	1130	850	265	3.20
5	Cisplatin	1129	862	274	3.14
6	Neomycin	1105	842	272	3.10
7	Polyethylene glycol	1153	863	279	3.09
8	Phytohemagglutinin	1099	807	266	3.03
9	Methotrexate	1190	897	308	2.91
10	Casein	1165	895	308	2.91
11	Isoleucine	1142	852	293	2.91
12	Galactose	1104	826	284	2.91
13	Progesterone	1448	1132	392	2.89
14	Esterase	1197	908	317	2.86
15	Tetracycline	1066	800	283	2.83
16	Acetone	1075	804	285	2.82
17	Concanavalin A	1317	1002	355	2.82
18	Polysaccharide	1092	829	295	2.81
19	Bromide	1368	1048	381	2.75
20	Methanol	1221	930	354	2.63

The 20 objects with the most implicit connections (shared rels) were extracted and sorted by observed to expected ratio, which is calculated using the probability each direct relationship comprising the implicit relationship is valid (quality estimate). These are compounds that *should not have any documented relationship* with cardiac hypertrophy within the MEDLINE titles and abstracts analyzed yet, at the same time, share many relationships with it.

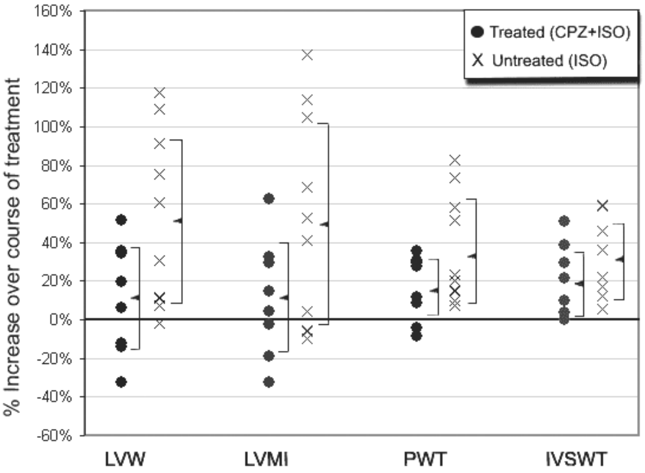

Fig. 6. Chlorpromazine protects against the development of cardiac hypertrophy. Several parameters of ventricular hypertrophy were determined by echocardiography. One group of mice received isoproterenol only (ISO, $n = 10$) and the other received both isoproterenol and chlorpromazine (CPZ + ISO, $n = 8$). Symbols represent individual mice, brackets denote mean (center triangle) and standard deviation for group. LVW = left ventricle weight (CPZ + ISO 11 ± 27%, ISO 51 ± 43%, $P < 0.02$), LVMI = left ventricular mass index (CPZ + ISO 11 ± 28%, ISO 50 ± 52%, $P < 0.04$), PWT = posterior wall thickness (CPZ + ISO 16 ± 16%, ISO 36 ± 27%, $P < 0.05$), IVSWT = intraventricular septum wall thickness (CPZ + ISO 19 ± 18%, ISO 31 ± 20%, $P < 0.12$).

and molecular targets that suggest it might provide an antihypertrophic effect in the heart, one of which is its alpha-adrenergic blocking activity (Morgan and Van Maanen, 1980). Hypertrophy can be induced through over-stimulation of alpha-adrenergic receptors by agonists and this effect can be blocked by alpha-adrenergic antagonists (Colucci, 1982). It has recently been recognized that the calmodulin-dependent phosphatase calcineurin plays an important role in some forms of hypertrophy (Molkentin *et al*., 1998). CPZ has been reported to interact with calmodulin (Marshak *et al*., 1981) as an antagonist, suggesting a potential role beyond that of an alpha-adrenergic receptor. Despite the potential mechanistic connections between cardiac hypertrophy and CPZ, there is no indication within MEDLINE that any relationship between the two has been suggested.

In-silico predicted effect confirmed in rodent model

We looked for an association between CPZ and cardiac hypertrophy in a rodent model. Two groups of mice were given 20 mg/kg/day isoproterenol by osmotic minipump, with one group additionally receiving 10 mg/kg/day CPZ. This dose of CPZ did not perceptibly alter feeding behavior or physical activity. Echocardiograms were obtained before treatment and again before the mice were sacrificed to allow estimation of their left ventricular weight (LVW), left ventricular mass index (LVMI), posterior wall thickness (PWT) and intraventricular septum wall thickness (IVSWT). In all four of the parameters measured, we found that the amount of cardiac hypertrophy was significantly reduced in the isoproterenol (ISO) plus CPZ treated mice ($n = 8$) in comparison to the control group given only ISO ($n = 10$), as evaluated by 1-tailed Student's t-test with unequal variance (Fig. 6).

DISCUSSION

A relationship between CPZ and cardiac hypertrophy has not been previously suggested in the literature. The application of implicit relationship analysis was required for generating the underlying hypothesis of this study. It was previously known that CPZ had modest alpha-blocking activity, but the finding that it interferes with ISO-induced hypertrophy, a pure beta-adrenergic effect, is surprising and provocative. Possible mechanisms include: (1) a previously unsuspected activity of CPZ on beta receptors, either directly or through cross-talk between different classes of receptors; (2) an effect of CPZ on downstream signaling from the beta receptor in the cardiac cells; or (3) a 'pseudo-hypertrophic', non-cellular, effect related to increased myocardial edema or matrix deposition.

Such an investigation could have clinical implications. If this drug exerts a similar effect against common precipitants of hypertrophy, it could provide a clue to molecular structures that should be explored for therapeutic benefit. Moreover, many tens of thousands of patients already receive CPZ, and it may be contributing to cardiac pathology, or protection, in a previously unsuspected fashion. Confirmation and evaluation of the mechanism of CPZ in cardiac hypertrophy will require further work beyond these preliminary investigations.

Overall, we have demonstrated that an analysis of shared relationships scored against a random network model has the potential to elucidate novel and interesting relationships not documented within MEDLINE, but rather based upon information contained therein. Automating the relationship identification process enables us to bypass the monumental time and effort that would be required to record manually every relationship within MEDLINE's 12 million abstracts, and using an object-based model reduces the need to ascertain which relationships are of interest, since the objects within the database are presumably those a user would be interested in. However, there are shortcomings in the use of this method: first, there is the problem of 'uninteresting' relationships. To some extent, this will be user-dependant. Objects that share many relationships may indeed have a relationship themselves, but the nature of their relationship may be such that it would not be considered interesting or worth investigating. Second, ascertaining the nature of the implied relationship by examining the shared relationships is time-consuming. Methods of providing a summary analysis or better evaluating which of the shared relationships are potentially interesting would be highly desirable. Third, comparison to a random network model relies upon the analyzed text to be focused (non-random) in nature. To the extent that writing is random or functionally detached within an analyzed textual unit, trivial connections will be made and fewer groupings will stand out statistically. Finally, work still needs to be done on better establishing relationships, beyond what we have done here in scoring the probability that a co-occurrence was meaningful. For example, the method asserts that a relationship is known when two objects have been mentioned together within the same abstract, and unknown if they have not. While this may be a good generalization, two objects may have been co-mentioned several times and yet the overall nature of their relationship, or certain aspects thereof, remains unknown. Nonetheless, we believe this method will prove to be of utility in a field where the amount of information continues to increase exponentially.

ACKNOWLEDGEMENTS

Thanks to Irene Rombel for a helpful review of the manuscript. This work was funded in part by NSF-EPSCoR grant no. EPS-0132534 (J.D.W.), NIH/NCI grant CA81656, NIH/NHLBI PGA grant 5U01HL6688002, NIH/NHLBI grant P50 CA70907, a Hudson Foundation grant and a National AHA Grant-in-Aid (R.V.S.).

REFERENCES

Barber,M.D. (2001) Cancer cachexia and its treatment with fish-oil-enriched nutritional supplementation. *Nutrition*, **17**, 751–755.

Bardoni,B., Mandel,J.L. and Fisch,G.S. (2000) FMR1 gene and fragile X syndrome. *Am. J. Med. Genet.*, **97**, 153–163.

Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *ISMB*, **7**, 60–67.

Caudle,R.M. and Mannes,A.J. (2000) Dynorphin: friend or foe? *Pain*, **87**, 235–239.

Colucci,W.S. (1982) Alpha-adrenergic receptor blockade with prazosin. Consideration of hypertension, heart failure, and potential new applications. *Ann. Int. Med.*, **97**, 67–77.

DiGiacomo,R.A., Kremer,J.M. and Shah,D.M. (1989) Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am. J. Med.*, **86**, 158–164.

Ding,J., Berleant,D., Nettleton,D. and Wurtele,E. (2002) Mining Medline: abstracts, sentences or phrases? *Pac. Symp. Biocomput.*, Kauau, Hawaii, **7**, 326–337.

Green,J.M. (2000) The B7/CD28/CTLA4 T-cell activation pathway. Implications for inflammatory lung disease. *Am. J. Respir. Cell Mol. Biol.*, **22**, 261–264.

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.

Hasselgren,P.O. and Fischer,J.E. (2001) Muscle cachexia: current concepts of intracellular mechanisms and molecular regulation. *Ann. Surg.*, **233**, 9–17.

Hristovski,D., Stare,J., Peterlin,B. and Dzeroski,S. (2001) Supporting discovery in medicine by association rule mining in Medline and UMLS. *Medinfo*, **10**, 1344–1348.

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.

Jin,P. and Warren,S.T. (2000) Understanding the molecular basis of fragile X syndrome. *Hum. Mol. Genet.*, **9**, 901–908.

Klir,G. and Yuan,B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall.

Kooy,R.F., Willemsen,R. and Oostra,B.A. (2000) Fragile X syndrome at the turn of the century. *Mol. Med. Today*, **6**, 193–198.

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.

Marshak,D.R., Watterson,D.M. and Van Eldik,L.J. (1981) Calcium-dependent interaction of S100b, troponin C, and calmodulin with an immobilized phenothiazine. *Proc. Natl Acad. Sci., USA*, **78**, 6793–6797.

McCoy,K.D. and Le Gros,G. (1999) The role of CTLA-4 in the regulation of T cell immune responses. *Immunol. Cell Biol.*, **77**, 1–10.

Molkentin,J.D., Lu,J.R., Antos,C.L., Markham,B., Richardson,J., Robbins,J., Grant,S.R. and Olson,E.N. (1998) A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell*, **93**, 215–228.

Morgan,J.P. and Van Maanen,E.F. (1980) The role of differential blockade of alpha-adrenergic agonists in chlorpromazine-induced hypotension. *Arch. Int. Pharmacodyn. Ther.,* **247**, 135–144.

Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

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.

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.*, 517–528.

Shen,W.W. (1999) A history of antipsychotic drug development. *Comp. Psychiatry* **40**, 407–414.

Smalheiser,N.R. and Swanson,D.R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comp. Meth. Prog. Biomed.*, **57**, 149–153.

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.

Steimann,F. (1997) Fuzzy set theory in medicine. *Artif. Intell. Med.*, **11**, 1–7.

Steiner,H. and Gerfen,C.R. (1998) Role of dynorphin and enkephalin in the regulation of striatal output pathways and behavior. *Exp. Brain Res.*, **123**, 60–76.

Swanson,D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, **30**, 7–18.

Tisdale,M.J. (2001) Cancer anorexia and cachexia. *Nutrition*, **17**, 438–442.

Tomer,Y. (2001) Unraveling the genetic susceptibility to autoimmune thyroid diseases: CTLA-4 takes the stage. *Thyroid*, **11**, 167–169.

Weeber,M., Klein,H., Aronson,A.R., Mork,J.G., de Jong-van den Berg,L.T. and Vos,R. (2000) Text-based discovery in biomedicine: the architecture of the DAD-system. *Proceedings of AMIA Annual Fall Symposium*, Los Angeles, CA, 903–907.

Weeber,M., Vos,R., Klein,H., De Jong-Van Den Berg,L.T., Aronson,A.R. and Molema,G. (2003) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J. Am. Med. Inform. Assoc.*, **10**, 252–259.

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. *Meth. Inform. Med.*, **41**, 426–434.

Yandell,M.D. and Majoros,W.H. (2002) Genomics and natural language processing. *Nat. Rev. Genet.*, **3**, 601–610.