The 'Open Discovery' Challenge

Jonathan D. Wren

Abstract One of the most exciting goals of literature-based discovery is the inference of new, previously undocumented relationships based upon an analysis of known relationships. Human ability to read and assimilate scientific information has long lagged the rate by which new information is produced, and the rapid accumulation of published literature has exacerbated this problem further. The idea that a computer could begin to take over part of the hypothesis formation process that has long been solely within the domain of human reason has been met with both skepticism and excitement, both of which are fully merited. Conceptually, it has already been demonstrated in several studies that a computational approach to literature analysis can lead to the generation of novel and fruitful hypotheses. The biggest barriers to progress in this field are technical in nature, dealing mostly with the shortcomings that computers have relative to humans in understanding the nature, importance and implications of relationships found in the literature. This chapter will discuss where current efforts have brought us in solving the open-discovery problem, and what barriers are limiting further progress.

1 Introduction

The amount of scientific literature is increasing exponentially,[1] along with most other databases in biomedicine, and there are far more papers published than any individual could ever hope to read. Furthermore, within this vast literature are many

J.D. Wren
Arthritis & Immunology Department, Oklahoma Medical Research Foundation,
825 N.E. 13th Street, Room W313, Oklahoma City, Oklahoma 73104-5005, USA
jdwren@gmail.com

[1] MEDLINE, for example, contained approximately 16 million records at the beginning of 2006, and is growing at a rate of approximately 4%/year, which currently equates to over 2,000 papers published per day.

P. Bruza and M. Weeber (eds.), *Literature-based Discovery,*
Springer Series in Information Science and Knowledge Management 15.

areas of research interest, more than any individual could ever hope to be aware of, leading to increasing specialization of research focus. This narrowing of relative awareness has not been a *barrier* to progress, but one could argue that it *limits* progress. In an age where data is generated faster than knowledge [2], it becomes increasingly important to be able to compile diverse sets of facts to identify high-impact hypotheses [3, 4]. The increasing emphasis on funding and conducting cross-disciplinary research and collaboration is, in part, a consequence of this expansion of information and necessary restriction of individual research focus. Intelligent tools are necessary to navigate, integrate and compile the diversity of available information to better advance all fields of scientific research.

1.1 Text-Based Knowledge Discovery

In 1986 Don Swanson illustrated that two areas of research could be functionally non-interactive, such that discoveries in one field could be relevant to studies in another, yet nonetheless remain unknown by researchers in either field because the fields have little or no overlap. Using a basic approach involving the pairing of keywords between literatures, he demonstrated that regions of overlap could be identified and novel discoveries made [5–9]. Intuitively, we recognize the value of the scientific literature in offering us insight into our own research. Who among us has not, at least once, read an article or attended a talk on a field unrelated to our own and subsequently left inspired with a new insight or direction for our own research? A broad perspective can be extremely valuable.

By enabling a computer to identify potential relationships within the scientific literature, it becomes possible to infer in an automated manner what is *not* known based upon what *is* known. Computers are, after all, perfectly suited to read large amounts of literature, catalog hundreds of thousands of names and synonyms, and simultaneously manipulate and track hundreds of relevant variables. It seems reasonable to stipulate that, for many areas of research with a significant body of associated literature, only a computer could gain the broadest possible perspective. Beyond the technical challenges associated with effective information retrieval (IR), the main challenges to the discovery of new knowledge are enabling a computer to identify *what* is of interest, *why* it is of interest and *how* the information will be conveyed to a human user. The intent of literature based discovery (LBD) is not to bypass the human researcher [10], but to provide a powerful supplement in assisting observation, analysis and inference on a large scale.

Although the LBD approach could be applied to many domains, efforts have thus far focused on the biomedical literature, specifically MEDLINE records. In part this is because MEDLINE records are freely available in electronic format, but also because most LBD efforts identify co-occurring terms as tentative relationships, whether these terms are names or medical subheadings (MeSH). Thus the nature of the association is usually non-specific and is best suited towards associations that are more general in nature. For example, when a gene is mentioned in an abstract

with a disease, there is a good probability that the gene is somehow related to the disease (or suspected to be). Furthermore, if two diseases are frequently mentioned with the same genes, then it not unreasonable to assume that the diseases are related in either their pathogenesis or phenotypic characteristics. The nature of each relationship may not matter as much as the frequency of their association for such inferences. When the nature of the relationship is critical to drawing inferences, then more sophisticated methods will be necessary. For example, if mining a legal/criminal database to find names frequently associated with crimes, the nature of each association is critical to drawing any conclusions – is the person an ordinary citizen, a lawyer prosecuting cases, a judge or a policeman?

During LBD, identifying relationships that are *known* (Fig. 1(1.1)) enables one to infer relationships that are not known, yet potentially *implicit* from the relationships shared by two objects (Fig. 1(1.2)). These shared relationships provide a means to both research and justify the existence of a potentially novel relationship not explicitly contained within the literature. By comparing shared relationship sets identified within the MEDLINE relationship network against what could be expected from a random network model with the same properties, we are able to assign a statistical significance value to any given grouping of relationships (Fig. 1(1.3)).

The approach outlined in Fig. 1 is what has become known as the "open discovery" model [11, 12]. It is also sometimes referred to as "Swanson's ABC discovery model", named because the first input node (black) is referred to as the "A" node, the direct relationships (gray) are referred to as the "B" nodes and the implicit relationships (white) are referred to as the "C" nodes. These implicit relationships have also

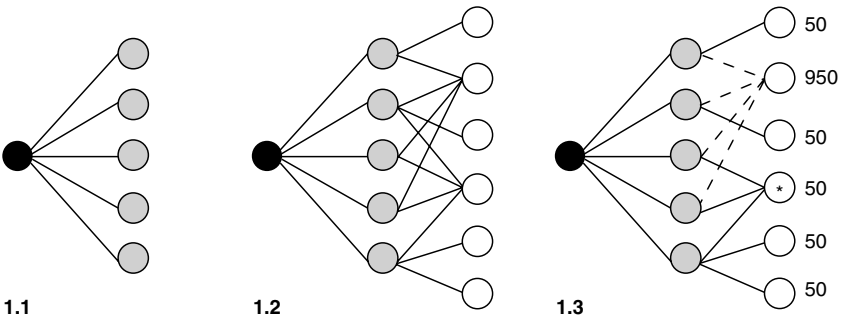

Fig. 1 Using literature-based relationships to engage in the discovery of new knowledge. (**1.1**) Beginning with an object of interest (*black node*), tentative relationships are assigned to other objects (*gray nodes*) when they are co-mentioned within MEDLINE records. (**1.2**) Each related object is then analyzed for its relationships with other objects (*white nodes*). These nodes are not directly related to the primary node, thus they are *implicitly* related. (**1.3**) These shared relationships are ranked against a random network model to establish how many would be expected by chance alone, given the connectivity of each object in the set. In this figure a hypothetical network with 1,000 nodes is analyzed. The node with the most shared relationships (four) is itself a highly connected node (connected to 95% of the network), and thus is less noteworthy from a statistical perspective than another node that shares three relationships and is connected to only 5% of the network (marked with an *asterisk*). A statistical score must be assigned in some manner to rank each of these implicit relationships for their potential significance, such as an observed to expected (Obs/Exp) ratio. Figure reproduced from [1]

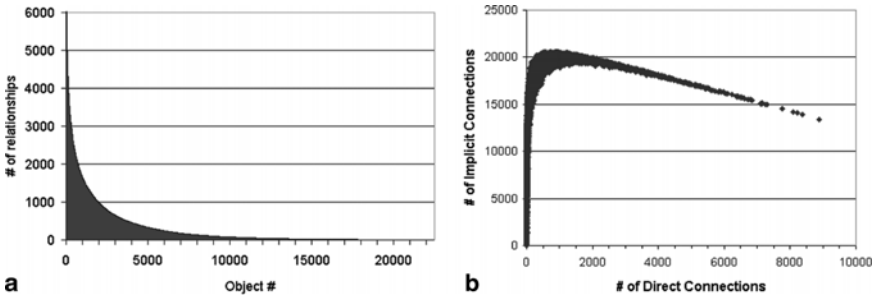

Fig. 2 Structure of the literature-based network. (**a**) The objects in a literature-based network have a disproportionate number of relationships, following a scale-free distribution. (**b**) In the case of the scientific literature, this leads to "extremely small world" network behavior by which most objects in the network are related by at least one intermediate. Figure reproduced from [1]

been referred to as "indirect" and "transitive" relationships. Similarly, the relationships themselves have also been referred to as "associations" and "connections".

Swanson outlined the open-discovery approach conceptually [6], but did not actually engage in it for most of his research because of the problems it posed. Rather, he usually began with the A and C nodes already known and focused upon exploration of the B nodes. However, because the number of relationships per object follows a scale-free distribution (Fig. 2a), the number of implicit connections found by an unbounded search increases rapidly for every direct connection. Figure 2b shows how the number of implicit connections rapidly approaches the maximum number possible (the upper asymptote) given a relatively small number of direct connections [1]. Thus, everything in the database quickly becomes related to the query object and the problem quickly shifts from *finding* implicit connections to *ranking* their potential relevance.

1.1.1 Evaluating Results

One means of quantifying performance when ranking implicit relationships is to score known relationships as if they were not known. In Fig. 1(1.3), for example, the A (black) and C (white) nodes are shown as unconnected. This is because direct relationships (the B nodes) are deliberately screened out from this set. However, if they are not screened out, they too will share relationships with the A node and can be evaluated just as any other C node in the implicit list. A previous study showed that weighting shared nodes (the B nodes) by how unlikely such a set would be shared by chance between two nodes correlated with the probability a relationship was known as well as with the strength of the relationship (Fig. 3).

This problem has been addressed by ranking implicit relationships by their connectivity within a network [1], then by attempting to extend mutual information measure (MIM) calculations from direct relationships to implicit relationships [13] and also by using fuzzy set theory (FST) to identify conceptual domains shared by

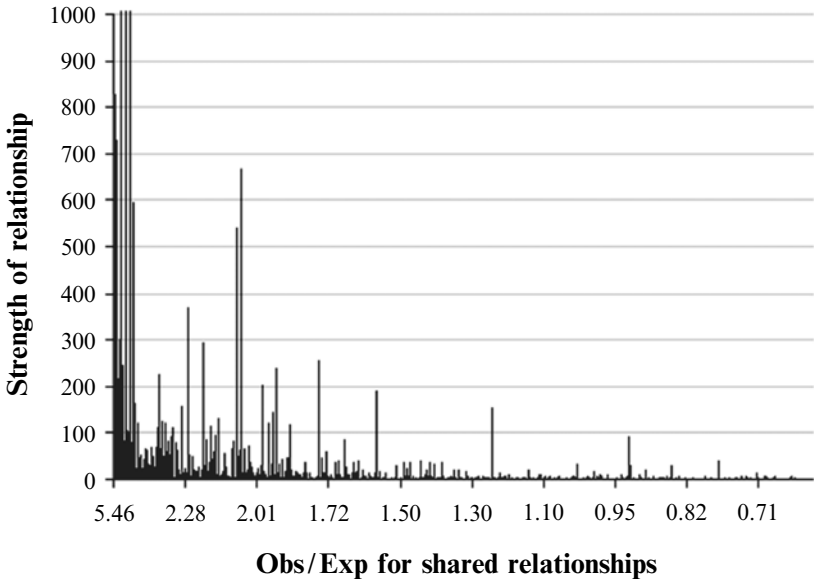

Fig. 3 The object "cardiac hypertrophy" was analyzed to identify all other objects in the database that share literature relationships with it. When a relationship is known (i.e., it has appeared in a MEDLINE title/abstract), a line is plotted on the *y*-axis, which corresponds to how many times the relationship was mentioned in MEDLINE. When the relationship is not known, there is a gap (not all gaps are visible due to *x*-axis compression). Note that frequently mentioned relationships tend to receive high scores when comparing the number of observed relationships shared by two objects to the number of relationships expected by chance (Obs/Exp). Figure reproduced from [1]

two objects [14]. Each approach had its strength and weaknesses in ranking inferences. For example, the FST approach was superior at identifying general concepts (e.g., migraines are associated with pain) whereas the MIM approach was superior at identifying more specific, informative relationships (e.g., migraines are associated with sumatriptan, a medication used to treat migraines). Regardless of the approach used, however, one major problem persisted: The amount of time the user had to spend to identify interesting implied relationships from within the set. This problem is not unique to just the studies mentioned, but rather is a general limitation of LBD in general. Relationships are defined by association and can thus be vague in their nature.

1.2 General Approach

MEDLINE abstracts contain a historical summary of biomedical discovery, and are available in electronic format free of charge from the National Library of Medicine (NLM). Abstracts are typically written without specific format or standardization of content, but are intended to convey the most pertinent aspects of the study being

published. Biomedical interests are broad, yet predominantly focused on several areas of primary interest: Genetics, disease pathology and etiology, study of phenotypes, and the effects and interactions of chemical compounds and small molecules. Recognizing relevant entities or "objects" within these databases such as gene names, diseases, chemical or drug names, and so forth is a challenge in its own right. Using MeSH terms, which are assigned by curators, can bypass nomenclature and ambiguous acronym problems but MeSH terms are limited in their scope (e.g. do not encompass most specific gene names).

As objects are co-cited within a record, LBD approaches assign a tentative relationship, and sometimes a confidence score that reflects some measured probability the relationship is non-trivial. As objects are co-cited more frequently, and/or closer together (e.g. the same sentence), confidence increases that this co-mentioning of objects reflects a meaningful relationship (Fig. 4). All analyses are conducted using this uncertainty measure. This use of co-citations has been adopted in a number of experiments where an automated attempt is made at constructing networks of potential interactions or relationships, mostly between genes or proteins [15–20]. The best known is probably the creation of the PubGene genetic network via co-citation of gene names within MEDLINE [15]. Once all MEDLINE records have been processed, a network of tentative relationships between objects has been constructed and can be analyzed. The method has been applied to MEDLINE, but is extensible to any other domain where discussion is constrained to a focused summary (e.g. an

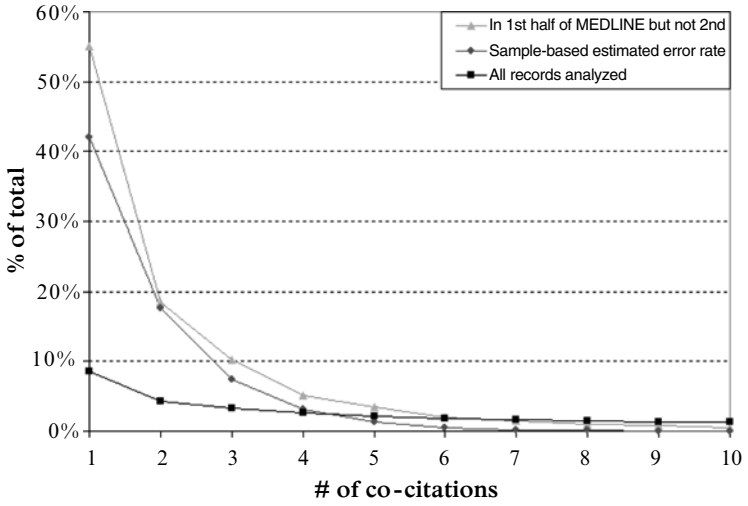

Fig. 4 Analysis of the uncertainty function in assigning tentative "relationships" based upon co-citation. *Top line* represents co-cited objects found within the first half of the 12 million MEDLINE records, but not the second half. Immediately below is the probability the uncertainty function (derived from sample-based error rates) assigns to co-cited relationships based upon the number of co-citations observed. For comparison, the overall distribution in the number of co-citations is shown at *bottom*. Figure reproduced from [1]

abstract) and co-occurrence of terms correlates with the presence or potential presence of a relationship between them (e.g. companies and products, legal precedents and key phrases such as 'workers compensation', etc.).

1.3 Previous LBD Applications

Open-discovery approaches have been applied to several different research problems, for example to identify compounds implicitly associated with cardiac hypertrophy, a clinically important disease that can develop in response to stress and high blood pressure. By examining the relationships shared by cardiac hypertrophy and one of the highest scoring implicitly associated compounds, chlorpromazine, it was anticipated that chlorpromazine should reduce the development of cardiac hypertrophy. It was tested using a rodent model, by giving mice isoproterenol to induce cardiac hypertrophy, with one group receiving saline injections and the other receiving chlorpromazine. Preliminary experiments suggested that chlorpromazine could significantly reduce the amount of cardiac hypertrophy induced by isoproterenol [1].

1.3.1 Type 2 Diabetes

Another analysis example involved Type 2 Diabetes, also known as Non-Insulin Dependant Diabetes Mellitus (NIDDM), and revealed a line of literature relationships that suggest the pathogenesis of NIDDM is epigenetic (Fig. 5). The analysis furthermore revealed the likely tissue of pathogenic origin (adipocytes), and narrowed the set of potentially causal factors to a general class of compounds (pro-inflammatory cytokines) implicated in the phenotype. Currently, the epigenetic

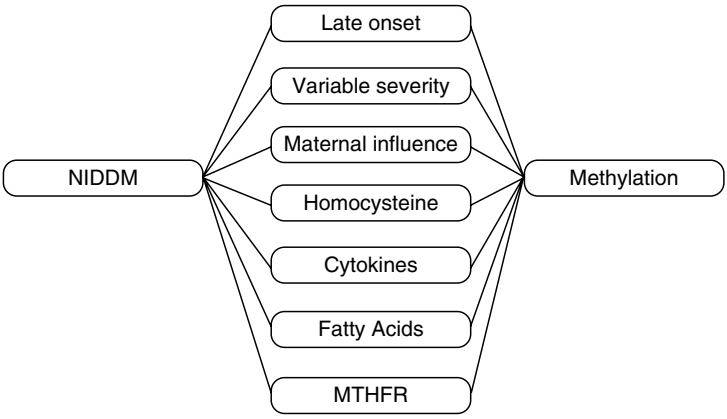

Fig. 5 A program called IRIDESCENT identified critical relationships shared by loss of DNA methylation and NIDDM (not all relationships shown), suggesting a relationship between the two

hypothesis remains untested, but seems to be gaining traction as mutation-based (e.g., single nucleotide polymorphism) models and the "complex disease" hypothesis have difficulty explaining certain observations about the etiology of NIDDM (e.g., why it is on the rise faster than population growth).

Based upon past developments and current research, it seems reasonable to presume that the ultimate goal of LBD research is the development of an intelligent system able to assimilate information in an automated manner, analyze facts and relations therein, and return to the user a set of logical conclusions and suggested courses of action based upon the current state of knowledge.

1.4 Improving on Co-Occurrences

Eventually, to provide a more targeted means of analysis, it will be necessary to expand the open-ended knowledge discovery model to include the nature of relationships in some manner. The general associative model is unfortunately too cumbersome to use, and it is difficult to rigorously test because it makes no predictions as to the nature of relationships. Thus, it is possible that every predicted implicit relationship would be true if one adopted a very lenient definition of the term "relationships". Natural language processing (NLP) provides a means of pinpointing the possible nature of the relationship between co-occurring terms (e.g. A upregulates B, B binds C). Thus it is possible that NLP could be used for the prediction of complementary and antagonistic relationships between unrelated terms.

The current LBD approaches can be summarized as general associative ones – "guilt by association" approaches. Despite their initial successes, there is still room for improvement. Figure 6 shows a general overview of the process assisted by IRIDESCENT as a generic example of how a user would explore potentially novel relationships identified by LBD approaches. First, the user selects an object for analysis. Here, the disease fibromyalgia is chosen. The literature-derived network of relationships is then queried to compile a set of terms related to fibromyalgia and then another set of relationships to each of these related terms (the implicit set). The terms are then displayed to the user for examination. Here, they are sorted in descending order of their observed to expected ratio. Gray rows represent known relationships while white rows represent unknown, implicit relationships. The user then examines the implicit relationships, looking for those that appear interesting – a quality that is highly subjective and usually a function of the examiner (e.g. oncologists would be more interested in cancer-related terms). Once an implicit relationship is chosen for analysis, such as the first implicit relationship on this list, "Parkinson's disease", another window would be opened so that the user could examine what relationships both Parkinson's disease and fibromyalgia share. The user can then examine these shared relationships, once again searching for one or more that look "interesting", and then examining either side of the shared relationship. Here, for example, the user could examine the A–B relationship, which in the window shown would be the relationship between fibromyalgia and females.

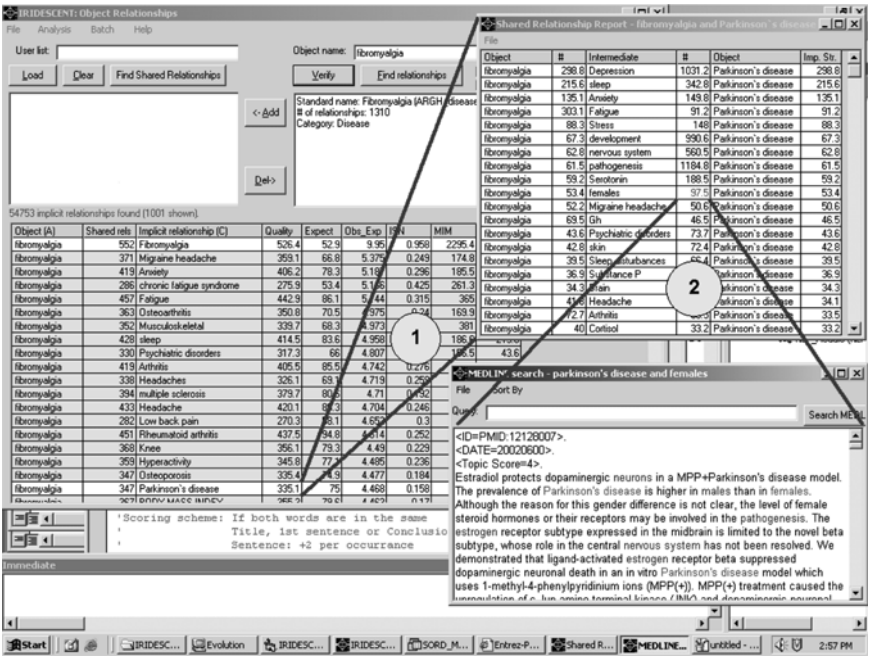

Fig. 6 Using an open-discovery approach to identify implicit relationships and explore shared relationships identified within the literature

The literature associated with this relationship is not shown here, but the nature of this relationship is that approximately 90% of fibromyalgia sufferers are female. Then, examining the corresponding B–C relationships, between females and Parkinson's, shown here in window #2 would pull up something like the next inset window. Examining the literature, with keywords highlighted for convenience, it is apparent that males disproportionately suffer from Parkinson's. Thus, at this point, the user understands one of the aspects of the implied relationship between fibromyalgia and Parkinson's. Users would then examine each of these shared relationships, one by one, to get a better idea of the overall nature of the implied relationship. This last step is the hardest since each individual relationship (e.g., of a disease to gender) may or may not paint a cohesive picture for any overall implied relationship between the two terms. It is entirely possible, if not likely, that many of the bridging B terms may be simple, isolated relationships that do not contribute at all towards an overall relationship between A and C. Thus, it could be confusing for users to try to iteratively construct a picture of a general relationship piece by piece since some of those pieces may only make sense after further analysis while others may not contribute at all towards a general A–C relationship. Subjective interpretation and a limited understanding of the nature of implied relationships are the biggest current barriers to LBD.

In the example shown in Fig. 6, the threshold to declare a relationship as "known" was set to a minimum of four co-mentions. Searching PubMed for "parkinson's and fibromyalgia" in the title or abstract yields two papers, one of which suggests the relationship between the two in terms of the neurotransmitters that are affected in each and the overlap in phenotypes [21]. So, in this case, a relationship is known between the two and was not detected because of the threshold. This also illustrates one of the limitations of the approach – in some cases several abstracts may co-mention two objects, yet examining the text of each one reveals no specific relation between the two. In other cases such as this one, one abstract co-mention may define a relationship. Lower or higher thresholds can be set depending upon user preference for virtually all of these approaches, but this is a persistent caveat. In this specific case, because the co-mentioning article was a review and somewhat speculative in nature, this would tell to the experimentalist interested in validating this connection that empirical work remains to be done. It also provides the experimentalist with many more shared relationships for him/her to better understand the implied relationship prior to experimentation. These shared relationships can be extremely valuable because, aside of these two papers, there is no further research that could be obtained via traditional query methods that would explain how the two diseases are connected.

The first step in better elucidating the nature of relationships might be to enable information extraction (IE) routines to classify directionality in relationships, which could lead to inference of complementary and antagonistic relationships. Figure 7 examines a hypothetical implicit relationship identified by an IE-based open discovery approach, with Fig. 7a showing the current approaches: Commonalities (B_1 through B_6) are identified between two objects (A and C). It is not known what type of relationship is implied by these common relationships until the user examines the text the relationships were identified in (as shown in Fig. 6). This examination can take a significant amount of time. For example, when a tentative relationship between Type 2 Diabetes and Methylation [22] in a previous analysis, although the initial implication was suggested relatively quickly, it took about 2 weeks worth of exploring the connecting relationships to better understand and identify the key components of the implied relationship. Much of this analysis is weaving a growing set of facts into a cohesive summary of what they mean collectively, which includes a willingness to look for both positive and negative evidence as well as judge what weight should be assigned to any observations that appear contradictory given all the other compiled observations. This would not be as much of a problem if it weren't for the fact that many implicit relationships are often examined before one of potential interest is found. A means of summarizing the nature of each implied relationship would be of great assistance.

Where possible, an IE-based approach to LBD would extract the nature of the relationship between objects (e.g., A affects B, but not the other way around). This directionality combined with regulatory information provides a means of inferring the general nature of relationships prior to their examination. For example, in Fig. 7c, we see that A positively affects the intermediates B_1, B_2, and B_5. In turn, these same objects positively affect C. The other intermediate relationships do not immediately

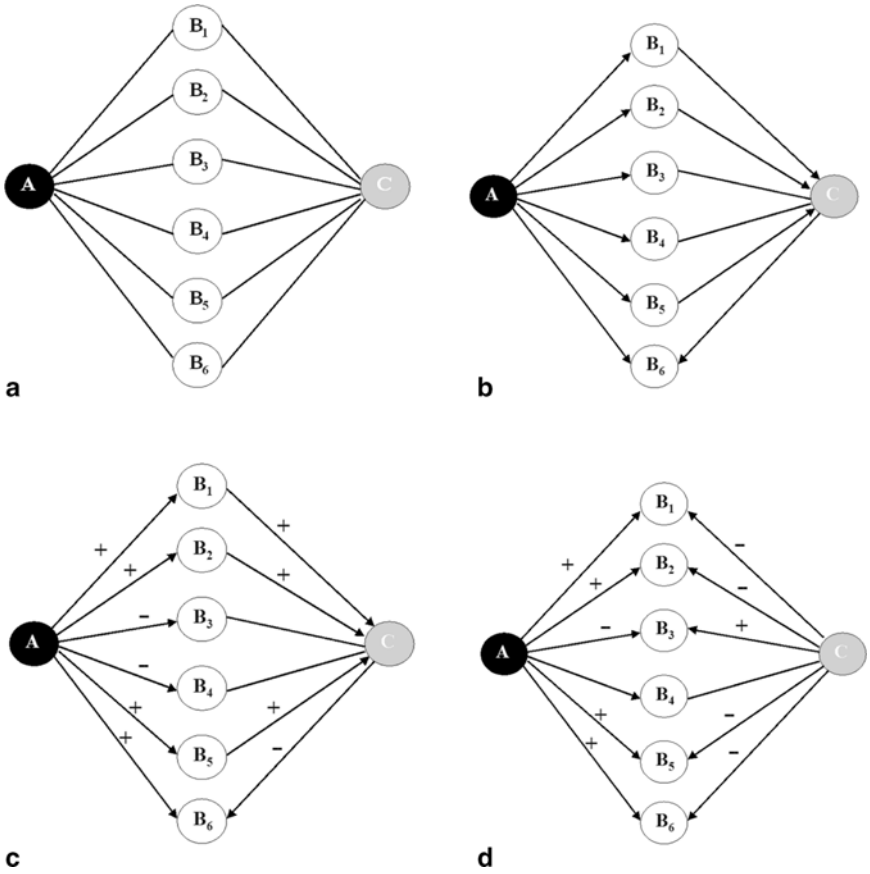

Fig. 7 Relationships identified within text and how IE would change the nature of analysis. (**a**) Current methods of ranking potentially interesting discoveries (i.e., undocumented relationships) rely upon statistical methods that suggest more relationships are shared than would be expected by chance. Here, it is unclear what the nature of the proposed relationship between objects A and C is until a user examines all the A–B and B–C intermediates. (**b**) By incorporating directional information (e.g., A affects B), greater information content is provided to the researcher. Here, for example, A appears to be affecting C through intermediates. (**c**) When information is extracted regarding the nature of relationships (e.g., A increases B), this enables inferences to be made regarding complementary and antagonistic relationships (e.g. A should increase C). (**d**) Multiple types of inferences can be made with this new model, here neither A nor C is predicted to affect the other, but rather they are anticipated to have opposing effects upon their intermediates. Notice that not all relationships necessarily have directionality or information concerning effects

provide information on how A affects C through them, if at all, but neither do they provide any contradictory information. Using this information, and without having to examine the underlying relationships beforehand, we can infer that A positively affects C.

Such a system could potentially be quite an improvement over previous methods, provided certain issues could be resolved. It would provide several ways that

more generalized information could be obtained. In Fig. 7d, for example, A and C are related, but the implied relationship is not between A and C but rather their intermediates. A and C apparently have antagonistic relationships with five out of six of their intermediates. A affects B_1, B_2, B_5 and B_6 positively while C affects them negatively. It also affects B_3 negatively whereas C affects B_3 positively. This type of information would be highly useful for inferring physiological interactions caused by chemicals or pharmaceuticals. B_1, B_2, B_5 and B_5, for example, could be heart rate, sweating, blood pressure and vasoconstriction. A could be a drug that increases them (e.g., isoproterenol) and C could be a drug that reduces them (e.g., valium). This type of system could be very useful for detecting potential drug interactions. If the antagonistic relationships here were positive instead (e.g., C was ephedrine instead of valium), then this would suggest these two drugs should not be given together. Except in a case where neither one alone had sufficient effect or some enhanced effect was deliberately being sought.

Using IE to identify the nature of relationships entails identifying regulatory and associative keywords within text and assigning the appropriate relationship. Several efforts have demonstrated the feasibility and efficacy of this, mostly in terms of protein–protein regulatory interactions [23, 24] but also in more generic terms [16]. The potential for advances in open discovery LBD methods is truly exciting. Eventually, if enough of these intermediate analysis steps could be automated, we may be witnessing the creation of an in silico scientist [25] – software that is able to analyze all electronically available information, draw logical conclusions about what is both possible and plausible and then propose the most logical and efficient course of action to empirically validate hypothesized relationships derived purely in silico. Of course, we are far from that day, but it is not unreasonable to presume it is both possible and perhaps even realizable within a generation or so.

1.5 Using History as a Guide to the Future

The historical discovery of new relationships within MEDLINE abstracts and provides a benchmark dataset for knowledge discovery. It could be argued that any individual experimental validations of relationships predicted by any knowledge discovery method are somewhat anecdotal. That is, a significant amount of user-based decision goes into ascertaining what novel relationships are worth pursuing. Currently, it is not at all clear which LBD approaches are most efficient due to a lack of quantitative methods and gold standard test sets for analysis. One possible way of addressing this might be to turn to a historical analysis. If historical relationship networks could be created, we could study how they have evolved over time, asking the critical question: How many scientific discoveries known today would have been highly ranked inferences in the past – based solely upon what was known at the time? More specifically it can be asked how well any particular approach would have performed historically in predicting the probability an implicit relationship will be of future scientific relevance.

In general, scientific discovery falls roughly into one of two categories: Fortuitous and logic-based. *Fortuitous discoveries* are those that arise unexpectedly or by accident. Some might argue that, given the benefit of retrospective hindsight, some fortuitous discoveries might have been anticipated. However, the way this term will be used here is to denote discoveries that could not have been reasonably anticipated given the state of knowledge at the time of discovery. Viagra (sildenafil) is an example of a fortuitous discovery, having been originally developed as a potential treatment for angina, but instead had blockbuster success in treating erectile dysfunction (ED) [26]. The new application for ED was originally observed as a side effect during clinical trials, and while it may now make sense in terms of what is now known about sildenafil's physiological/molecular actions, it is not amenable to computational analysis because the alternative use was published before the original, intended use [26]. Rogaine (minoxidil) shares a similar history with Viagra in that it was originally developed to combat high blood pressure [27], but was discovered later to be a successful treatment for baldness [28]. Between its initial reporting in the literature in 1973 and its later use discovered around 1980, studies were published concerning its pharmacological/molecular actions that might have suggested an alternative use was possible.

Logic-based discoveries occur when an expert postulates that a new relationship can be identified (or ruled out) based upon what is currently known. Whether the expert anticipates the exact answer or not, there is a rationale for both choosing and designing the experiment such that more information can be obtained about the system in question. Conceptually, this is what most knowledge discovery approaches attempt to do: To better understand an area of research (Fig. 1(1.3), black node), unknown variables (Fig. 1(1.3), white nodes) are studied in the context of known variables (Fig. 1(1.3), gray nodes). Logic-based discoveries are those that are thought out and justified, at least to an extent, prior to the commitment of time and resources to further investigation. Preliminary results for a proposed research project typically confer a competitive advantage upon it because they imply a greater chance of success. In the absence of such results, researchers typically justify the proposed commitment of resources by extensive citing of research results obtained from others. In either case, future research is predicated upon current understanding.

It is reasonable to postulate that this latter type of scientific discovery, logic-based, is amenable to computational analysis and that there are numerous relationships published in the literature, shared by two unrelated objects, which suggest the existence of a relationship long before one is recognized. If it can be demonstrated that large-scale computational analysis of scientific information can identify important discoveries prior to their experimental validation, this has very important implications for scientific research in general. It would suggest that, to an extent, human awareness of relationships is a limiting factor in discovery and computational assistance would be of broad benefit to the scientific community.

Due to their relative simplicity and lack of reliance upon proprietary and computationally expensive NLP software, construction of co-citation networks have become an increasingly common way of ascertaining relationships among different

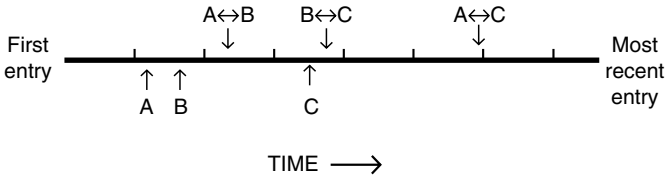

Fig. 8 Entry of objects (below timeline) and relationships (above timeline) into MEDLINE

types of objects within literature-based sources [15–20]. By this method, a relationship is "discovered" when two objects have been co-mentioned in the same abstract. However, this alone does not mean that a relationship has truly been elucidated, proposed, or understood. One could imagine a historical perspective analyzing a co-citation network of random words – one would certainly be able to identify co-citation patterns such as "red↔bird" and "bird↔house" that would predict the eventual "discovery" of the relationship "red↔house", but we can easily recognize that the nature of this relationship and prediction are trivial. This example helps in illustrating the fundamental problem in using co-citation as a metric for identifying a relationship, even when the co-citation has occurred many times: Related objects are almost unavoidably co-mentioned together, but co-mentions do not necessarily reflect a meaningful relationship.

Figure 8 illustrates graphically the variables being analyzed, with MEDLINE depicted as a time-dependant progression of published papers from the first entry to the most recent. At given points in time, the primary object of analysis, A, will first appear within the literature as will other objects such as C which will eventually be discovered to have a relationship with A. A number of intermediate factors such as B (only one is shown here for simplicity) will be related to both A and C prior to the publication of their relationship. Essentially, literature-based discovery methods are predicated upon the assumption that cases such as this exist – that at least a subset of all discoveries could have been predicted prior to their publication.

1.6 Literature Limitations

Electronically available MEDLINE records lack full experimental detail – much sequence information is not published directly in the primary literature but rather deposited into databases. Therefore knowledge discovery methods lack the ability to draw correlations between literature relationships and information contained in genomic and transcriptional (microarray) databases. Integrating experimental data with literature associations should be able to provide experimental insight in several areas.

In Fig. 9, for example, three genes within a genomic region are found to have literature correlations with a disease or phenotype. Once this is known, nearby genomic features such as CpG islands (gray square) or highly repetitive regions

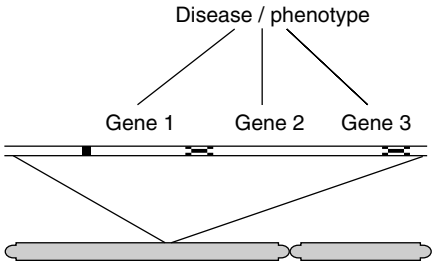

Fig. 9 Correlating literature commonalities with genomic data

(gray X) might offer a hypothesis about how recombination or silencing might contribute towards the etiology of this disease. A similar approach was conducted to identify candidate genes for diseases [29] by association of MeSH phenotypic terms with Gene Ontology (GO) terms through MeSH D terms.

1.7 Integrating Gene Expression Measurements

Within microarray experiments there are groups of genes that respond transcriptionally to changes in experimental conditions. Space limitations prevent more than a few of these genes from being mentioned within MEDLINE abstracts, so this is information that would not be obtainable the way most LBD approaches are currently implemented. However, many microarray datasets are cataloged in NCBI's Gene Expression Omnibus (GEO) [30]. A number of methods are available to cluster transcriptional responders into groups, which could then be cataloged and integrated into the literature-based network. This confers the additional advantage that implicit analyses might point directly to experimental results.

Perhaps the most difficult part of microarray analysis is not so much the cleaning, normalization and clustering of data, but ascertaining the biological relevance of the response. To do this, the researcher must first identify what is already known about the response observed within the experiment to gain confidence that aspects of their experiment correspond with previous observations. Second, and perhaps most important, they must ascertain what their experiment has told them that is not already known. The purpose of integrating microarray response datasets is to be able to answer both these questions.

2 Summary

Natural human limitations of time, expertise, speed of understanding and personal interests prevent researchers from being aware of more than a fraction of the cumulative scientific knowledge gained to date. Computers cannot yet substitute for

human understanding, but can act as a mental "prosthesis" in examining and analyzing this body of knowledge. Vast amounts of time and resources have already been spent to gain this knowledge, but it has not yet been exploited for all the value it holds. Observation and perspective have always been key components in advancing science and medicine, thus we must recognize that limitations in these areas also limit the rate of progress. LBD research will help reduce these barriers and provide a broader perspective. The ability to examine networks of biomedical interactions and infer novel hypotheses holds exciting promise for health-related research.

References

1. Wren, J.D., et al., Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 2004. 20(3): 389–398
2. Valencia, A., Search and retrieve. Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? EMBO Rep, 2002. 3(5): 396–400
3. Blagosklonny, M.V. and A.B. Pardee, Conceptual biology: unearthing the gems. Nature, 2002. 416(6879): 373
4. Bray, D., Reasoning for results. Nature, 2001. 412(6850): 863
5. Swanson, D.R., Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med, 1986. 30(1): 7–18
6. Swanson, D.R., Undiscovered public knowledge. Libr Q, 1986. 56: 103–118
7. Swanson, D.R., Migraine and magnesium: eleven neglected connections. Perspect Biol Med, 1988. 31(4): 526–557
8. Swanson, D.R., Somatomedin C and arginine: implicit connections between mutually isolated literatures. Perspect Biol Med, 1990. 33(2): 157–186
9. Swanson, D.R. and N.R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell, 1997. 91: 183–203
10. Smalheiser, N.R., Informatics and hypothesis-driven research. EMBO Rep, 2002. 3(8): 702
11. Pratt, W. and M. Yetisgen-Yildiz. LitLinker: capturing connections across the biomedical literature. In Proceedings of the International Conference on Knowledge Capture (K-Cap'03), 2003, Florida
12. Srinivasan, P., Text mining: generating hypotheses from MEDLINE. J Am Soc Inf Sci Technol, 2004. 55(5): 396–413
13. Wren, J.D., Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics, 2004. 5(1): 145
14. Wren, J.D., Using fuzzy set theory and scale-free network properties to relate MEDLINE terms. Soft Computing, 2006. 10(4): 374–381
15. Jenssen, T.K., et al., A literature network of human genes for high-throughput analysis of gene expression. Nat Genet, 2001. 28(1): 21–28
16. Rindflesch, T.C., et al., EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput, 2000. 517–528
17. Stapley, B.J. and G. Benoit, Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput, 2000. 529–540
18. Xenarios, I., et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res, 2002. 30(1): 303–305
19. Andrade, M.A. and P. Bork, Automated extraction of information in molecular biology. FEBS Lett, 2000. 476(1–2): 12–17
20. Blaschke, C., et al., Automatic extraction of biological information from scientific text: protein–protein interactions. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, 1999, pp. 60–67

21. Burgunder, J.M., Pathophysiology of akinetic movement disorders: a paradigm for studies in fibromyalgia? Z Rheumatol, 1998. 57(Suppl 2): 27–30
22. Wren, J.D. and H.R. Garner, Data-mining analysis suggests an epigenetic pathogenesis for Type II Diabetes. J Biomed Biotechnol, 2005. 2: 104–112
23. Xenarios, I., et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res, 2002. 30(1): 303–305
24. Zanzoni, A., et al., MINT: a Molecular INTeraction database. FEBS Lett, 2002. 513(1): 135–140
25. Wren, J.D., The emerging in-silico scientist: how text-based bioinformatics is bridging biology and artificial intelligence. IEEE Eng Med Biol Mag, 2004. 23(2): 87–93
26. Boolell, M., et al., Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. Int J Impot Res, 1996. 8(2): 47–52
27. DuCharme, D.W., et al., Pharmacologic properties of minoxidil: a new hypotensive agent. J Pharmacol Exp Ther, 1973. 184(3): 662–670
28. Zappacosta, A.R., Reversal of baldness in patient receiving minoxidil for hypertension. N Engl J Med, 1980. 303(25): 1480–1481
29. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, Association of genes to genetically inherited diseases using data mining. Nat Genet, 2002. 31(3): 316–319
30. Edgar, R., M. Domrachev, and A.E. Lash, Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 2002. 30(1): 207–210