Pacific Symposium on Biocomputing 12:316-327(2007)

DISCOVERING IMPLICIT ASSOCIATIONS BETWEEN GENES AND HEREDITARY DISEASES

KAZUHIRO SEKI

Graduate School of Science and Technology, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: seki@cs.kobe-u.ac.jp

JAVED MOSTAFA

Laboratory of Applied Informatics Research, Indiana University
1320 E. 10th St., LI 011, Bloomington, Indiana 47405-3907
E-mail: jm@indiana.edu

We propose an approach to predicting implicit gene-disease associations based on the inference network, whereby genes and diseases are represented as nodes and are connected via two types of intermediate nodes: gene functions and phenotypes. To estimate the probabilities involved in the model, two learning schemes are compared; one baseline using co-annotations of keywords and the other taking advantage of free text. Additionally, we explore the use of domain ontologies to complement data sparseness and examine the impact of full text documents. The validity of the proposed framework is demonstrated on the benchmark data set created from real-world data.

1. Introduction

The ever-growing textual data make it increasingly difficult to effectively utilize all the information relevant to our interests. For example, Medline—the most comprehensive bibliographic database in life science—currently indexes approximately 5,000 peer-reviewed journals and contains over 17 million articles. The number of articles is increasing rapidly by 1,500–3,000 per a day. Given the substantial volume of the publications, it is crucial to develop intelligent information processing techniques, such as information retrieval (IR), information extraction (IE), and text data mining (TDM), that could help us manage the information overload.

In contrast to IR and IE, which deal with information explicitly stated in documents, TDM aims to discover heretofore unknown knowledge through an automatic analysis on textual data.[1] A pioneering work in TDM (or literature-based discovery) was conducted by Swanson in the 1980's. He argued that there were

Pacific Symposium on Biocomputing 12:316-327(2007)

two premises logically connected but the connection had been unnoticed due to overwhelming publications and/or over-specialization. For instance, given two premises $A \rightarrow B$ and $B \rightarrow C$, one could deduce a possible relation $A \rightarrow C$. To prove the idea, he manually analyzed numbers of articles and identified logical connections implying a hypothesis that fish oil was effective for clinical treatment of Raynaud's disease.[2] The hypothesis was later supported by experimental evidence.

Based on his original work, Swanson and other researchers have developed computer programs to aid hypothesis discovery (e.g., see Refs. 3 and 4). Despite the prolonged efforts, however, the research in literature-based discovery can be seen to be at an early stage of development in terms of the models, approaches, and evaluation methodologies. Most of the previous work was largely heuristic without a formal model and their evaluation was limited only on a small number of hypotheses that Swanson had proposed.

This study is also motivated by Swanson's and attempts to advance the research in literature-based discovery. Specifically, we will examine the effectiveness of the models and techniques developed for IR, the benefit of free- and full-text data, and the use of domain ontologies for more robust system predictions. Focusing on associations between genes and hereditary diseases, we develop a discovery framework adapting the inference network model[5] in IR, and we conduct various evaluative experiments on realistic benchmark data.

2. Task Definition

Among many types of information that are of potential interest to biomedical researchers, this study targets associations between genes and hereditary diseases as a test bed. Gene-disease associations are the links between genetic variants and diseases to which the genetic variants influence the susceptibility. For example, BRCA1 is a human gene encoding a protein that suppresses tumor formation. A mutation of this gene increases a risk of breast cancer. Identification of these genetic associations has tremendous importance for prevention, prediction, and treatment of diseases. In this context, predicting or ranking candidate genes for a given disease is crucial to select more plausible ones for genetic association studies.

Focusing on gene-disease associations, we assume a disease name and known causative genes, if any, as system input. In addition, a target region in the human genome may be specified to limit the search space. Given such input, we attempt to predict a (unknown) causative gene and produce a ranked list of candidate genes.

3. Proposed Approach

Focusing on gene-disease associations, we explored the use of a formal IR model, specifically, the inference network[5] for this related but different problem targeting implicit associations. The following details the proposed model and how to estimate probabilities involved in the model.

3.1. *Inference Network for Gene-Disease Associations*

In the original IR model, a user query and documents are represented as nodes in a network and are connected via intermediate nodes representing keywords that compose the query and documents. To adapt the model to represent gene-disease associations, we treat disease as query and genes as documents and use two types of intermediate nodes: gene functions and phenotypes which characterize genes and disease, respectively (Fig. 1). An advantage of using this particular IR model is that it is essentially capable of incorporating multiple intermediate nodes. Other popular IR models, such as the vector space models, are not easily applicable as they are not designed to have different sets of concepts to represent documents and queries.

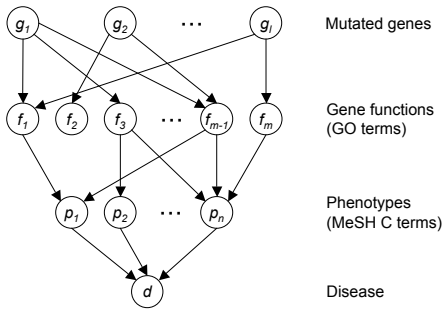

Figure 1. Inference network for gene-disease associations.

The network consists of four types of nodes: genes (g), gene functions (f) represented by Gene Ontology (GO) terms,[a] phenotypes (p) represented by MeSH C terms,[b] and disease (d). Each gene node g represents a gene and corresponds to the event that the gene is found in the search for the causative genes underlying d. Each gene function node f represents a function of gene products. There

[a] http://www.geneontology.org
[b] http://www.nlm.nih.gov/mesh

Pacific Symposium on Biocomputing 12:316-327(2007)

are directed arcs from genes to functions, representing that instantiating a gene increases the belief in its functions. Likewise, each phenotype node p represents a phenotype of d and corresponds to the event that the phenotype is observed. The belief in p is dependent on the belief in f's since phenotypes are (partly) determined by gene functions. Finally, observing certain phenotypes increases the belief in d. As described in the followings, the associations between genes and gene functions ($g \rightarrow f$) are obtained from an existing database, Entrez Gene,[c] whereas both the associations between gene functions and phenotypes ($f \rightarrow p$) and the associations between phenotypes and disease ($p \rightarrow d$) are derived from the biomedical literature.

Given the inference network model, disease-causing genes can be predicted based on the probability defined below.

$$P(d|G) = \sum_i \sum_j P(d|\vec{p}_i) \times P(\vec{p}_i|\vec{f}_j) \times P(\vec{f}_j|G) \tag{1}$$

Equation (1) quantifies how much a set of candidate genes, G, increases the belief in the development of disease d. In the equation, $\vec{p}_i$ (or $\vec{f}_j$) is defined as a vector of random variables with i-th (or j-th) element being positive (1) and all others negative (0). By applying Bayes' theorem and some independence assumptions discussed later, we derive

$$P(d|G) \propto \sum_i \sum_j \left(\frac{P(p_i|d)}{P(\bar{p}_i|d)} \times \frac{P(f_j|p_i)P(\bar{f}_j|\bar{p}_i)}{P(\bar{f}_j|p_i)P(f_j|\bar{p}_i)} \times F(p_i) \times F(f_j) \times P(f_j|G) \right) \tag{2}$$

where

$$F(p_i) = \prod_{h=1}^{m} \frac{P(\bar{f}_h|p_i)}{P(\bar{f}_h|\bar{p}_i)}, \; F(f_j) = \prod_{k=1}^{n} \frac{P(\bar{f}_j)P(f_j|\bar{p}_k)}{P(f_j)P(\bar{f}_j|\bar{p}_k)} \tag{3}$$

The first factor of the right-hand side of Eq. (2) represents the interaction between disease d and phenotype p_i, and the second factor represents the interaction between p_i and gene function f_j, which is equivalent to the odds ratio of $P(f_j|p_i)$ and $P(f_j|\bar{p}_i)$. The third and fourth factors are functions of p_i and f_j, respectively, representing their main effects. The last factor takes either 0 or 1, indicating whether f_j is a function of any gene in G under consideration.

The inference network described above assumes independence among phenotypes, among gene functions, and among genes. We assert that, however, the effects of such associations are minimal in the proposed model. Although there may be strong associations among phenotypes (e.g., phenotype p_x is often observed with phenotype p_y), the model does not intend to capture those associations. That

[c] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene

Pacific Symposium on Biocomputing 12:316-327(2007)

is, phenotypes are attributes of the disease in question and we only need to know those that are frequently observed with disease d so as to characterize d. The same applies to gene functions; they are only attributes of the genes to be examined and are simply used as features to represent the genes under consideration.

3.2. *Probability Estimation*

3.2.1. *Conditional Probabilities $P(p|d)$*

Probability $P(p|d)$ can be interpreted as a degree of belief that phenotype p is observed when disease d has developed. To estimate the probability, we take advantage of the literature data. Briefly, given a disease name d, a Medline search is conducted to retrieve articles relevant to d and, within the retrieved articles, we identify phenotypes (MeSH C terms) strongly associated with the disease based on chi-square statistics. Given disease d and phenotype p, the chi-square statistic is computed as

$$\chi^2(d,p) = \frac{N(n_{11}\cdot n_{22} - n_{21}\cdot n_{12})^2}{(n_{11}+n_{21})(n_{12}+n_{22})(n_{11}+n_{12})(n_{21}+n_{22})} \tag{4}$$

where N is the total number of articles in Medline, n_{11} is the number of articles assigned p and included in the retrieved set (denoted as R), n_{22} is the number of articles not assigned p and not included in R, n_{21} is the number of articles not assigned p and included in R, and n_{12} is the number of articles assigned p and not in R. The resulting chi-square statistics are normalized by the maximum to treat them as probabilities $P(p|d)$.

3.2.2. *Conditional Probabilities $P(f|p)$*

Probability $P(f|p)$ indicates the degree of belief that gene function f underlies phenotype p. For probability estimation, this study adopts the framework similar to the one proposed by Perez-Iratxeta *et al.*[6] Unlike them, however, this study focuses on the use of textual data and domain ontologies and investigate their effects for literature-based discovery.

As training data, our framework uses Medline records that are assigned any MeSH C terms and are cross-referenced from any gene entry in Entrez Gene. For each of such records, we can obtain a set of phenotypes (the assigned MeSH C terms) and a set of gene functions (GO terms) associated with the cross-referencing gene from Entrez Gene. Considering the fact that the phenotypes and gene functions are associated with the same Medline record, it is likely that some of the phenotypes and gene functions are associated. A question is, however, what phenotypes and functions are associated and how strong those associations are.

Pacific Symposium on Biocomputing 12:316-327(2007)

We estimate those possible associations using two different schemes: *SchemeK* and *SchemeT*. *SchemeK* simply assumes a link between every pair of the phenotypes and gene functions with equal strength, whereas *SchemeT* seeks for evidence in the textual portion of the Medline record, i.e., title and abstract, to better estimate the strength of associations. Essentially, *SchemeT* searches for co-occurrences of gene functions (GO terms) and phenotypes (MeSH terms) in a sliding window, assuming that associated concepts tend to co-occur more often in the same context than unassociated ones. However, a problem of *SchemeT* is that gene functions and phenotypes are descriptive by nature and may not be expressed in concise GO and MeSH terms. In fact, Schuemie *et al.*[7] analyzed 1,834 articles and reported that less than 30% of MeSH terms assigned to an article actually appear in its abstract and that only 50% even in its full text. It suggests that relying on mere occurrences of MeSH terms would fail to capture many true associations.

To deal with the problem, we apply the idea of query expansion, a technique used in IR to enrich a query by adding related terms. If GO and MeSH terms are somehow expanded, there is more chance that they could co-occur in text. For this purpose, we use the definitions (or scope notes) of GO and MeSH terms and identify representative terms by inverse document frequencies (IDF), which has long been used in IR to quantify the specificity of terms in a given document collection. We treat term definitions as documents and define IDF for term t as $\log(N/Freq(t))$, where N denotes the total number of MeSH C (or GO) terms and $Freq(\cdot)$ denotes the number of MeSH C (or GO) terms whose definitions contain term t. Only the terms with high IDF values are used as the proxy terms to represent the starting concept, i.e., gene function or phenotype.

Each co-occurrence of the two sets of proxy terms (one representing a gene function and the other representing a phenotype) can be seen as evidence that supports the association between the gene function and phenotype, increasing the strength of their association. We define the increased strength by the product of the term weights, w, for the two co-occurring proxy terms. Then, the strength of the association between gene function f and phenotype p within article a, denoted as $S(f, p, a)$, can be defined as the sum of the increases for all co-occurrences of the proxy terms in a. That is,

$$S(f, p, a) = \sum_{(t_f, t_p, a)} \frac{w(t_f) \cdot w(t_p)}{|Proxy(f)| \cdot |Proxy(p)|} \tag{5}$$

where t_f and t_p denote any terms in the proxy terms for f and p, respectively, and (t_f, t_p, a) denotes a set of all co-occurrences of t_f and t_p within a. The product of the term weights is normalized by the proxy size, $|Proxy(\cdot)|$, to eliminate the effect of different proxy sizes. As term weight w, this study used the TF·IDF weighting

scheme. For term t_p for instance, we define TF(t_p) as 1 + $\log Freq(t_p, Def(p))$, where $Def(p)$ denote p's definition and $Freq(t_p, Def(p))$ denotes the number of occurrences of t_p in $Def(p)$.

The association scores, $S(f, p, a)$, are computed for each cross reference (a pair of Medline record and gene) by either *SchemeK* or *SchemeT* and are accumulated over all articles to estimate the associations between f's and p's, denoted as $S(f, p)$. Based on the associations, we define probability $P(f|p)$ as $S(f, p)/\sum_p S(f, p)$.

A possible shortcoming of the approach described above is that the obtained associations $S(f, p)$ are symmetric despite the fact that the network presented in Fig. 1 is directional. However, since it is known that an organism's genotype (in part) determines its phenotype—not in the opposite direction, we assumed that the estimated associations between gene functions and phenotypes are directed from the former to the latter.

3.2.3. *Enhancing Probability Estimates $P(f|p)$ by Domain Ontologies*

The proposed framework may not be able to establish true associations between gene functions and phenotypes for various reasons, e.g., the amount of training data may be insufficient. Those true associations may be uncovered using the structure of MeSH and/or GO. MeSH and GO have a hierarchical structure[d] and those located nearby in the hierarchy are semantically close to each other. Taking advantage of these semantic relations, we enhance the learned probabilities $P(f|p)$ as follows.

Let us denote by A the matrix whose element a_{ij} is probability estimate $P(f_j|p_i)$ and by A' the updated or enhanced matrix. Then, A' is formalized as $A' = W_p A W_f$, where W_p denotes an $n \times n$ matrix with element $w_p(i, j)$ indicating a proportion of a probability to be transmitted from phenotypes p_j to p_i. Similarly, W_f is an $m \times m$ matrix with $w_f(i, j)$ indicating a proportion transmitted from gene functions f_i to f_j. This study experimentally uses only direct child-to-parent and parent-to-child relations and defines $w_p(i, j)$ as

$$w_p(i,j) = \begin{cases} 1 & \text{if } i = j \\ \dfrac{1}{\#\text{ of children of } p_j} & \text{if } p_i \text{ is a child of } p_j \\ \dfrac{1}{\#\text{ of parents of } p_j} & \text{if } p_i \text{ is a parent of } p_j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

[d]To be precise, GO's structure is directed acyclic graph, allowing multiple parents.

Equation (6) means that the amount of probability is split equally among its children (or parents). Similarly, $w_p(i, j)$ is defined by replacing i and j in the right-hand side of Eq. (6). Note that the enhancement process can be iteratively applied to take advantage of more distant relationships than children/parents.

4. Evaluation

To evaluate the validity of the proposed approach, we implemented a prototype system and conducted various experiments on the benchmark data sets created from the genetic association database (GAD).[e] GAD is a manually-curated archive of human genetic studies, containing pairs of gene and disease that are known to have causative relations.

4.1. *Creation of Benchmark Data*

For evaluation, benchmark data sets were created as follows using the real-world data obtained from GAD.

(1) Associate each gene-disease pair with the publication date of the article from which the entry was created. The date can be seen as the time when the causative relation became public knowledge.
(2) Group gene-disease pairs based on disease names. As GAD deals with complex diseases, a disease may be paired with multiple genes.
(3) For each pair of a disease and its causative genes,
    (a) Identify the gene whose relation to the disease was most recently reported based on the publication date. If the date is on or after 7/1/2003, the gene will be used as the target (i.e., new knowledge), and the disease and the rest of the causative genes will be used as system input (i.e., old knowledge).
    (b) Remove the most recently reported gene from the set of causative genes and repeat the previous step (3a).

The separation of the data by publication dates ensures that a training phase does not use new knowledge in order to simulate gene-disease association discovery. The particular date was arbitrarily chosen by considering the size of the resulting data and available resources for training. Table 1 shows the number of gene-disease associations in the resulting test data categorized under six disease classes defined in GAD. In the following experiments, the cancer class was used for system development and parameter tuning.

[e] http://geneticassociationdb.nih.gov

Pacific Symposium on Biocomputing 12:316-327(2007)

Table 1. Number of gene-disease associations in the benchmark data.

Cancer	Cardio-vascular	Immune	Metabolic	Psych	Unknown	Total
45	36	61	23	12	80	257

4.2. *Experimental Setup*

Given input (disease name d, known causative genes, and a target region), the system computes the probability $P(d|G)$ as in Eq. (3) for each candidate gene g located in the target region, where G is a set of the known causative genes plus g. The candidate genes are then outputted in a decreasing order of their probabilities as system output.

As evaluation metrics, we use *area under the ROC curve* (AUC) for its attractive property as compared to the F-score measure (see Ref. 8 for more details). ROC curves are two dimensional measure for system performance with x axis being true positive proportion (TPP) and y axis being false positive proportion (FPP). TPP is defined as TP/(TP+FN), and FPP as FP/(FP+TN), where TP, FP, FN, and FP denote the number of true positives, false positives, false negatives, and false positives, respectively. AUC takes a value between 0 and 1 with 1 being the best. Intuitively AUC indicates the probability that a gene randomly picked from positive set is scored more highly by a system than one from negative set.

For data sets, this study used a subset of the Medline data provided for the TREC Genomics Track 2004.[9] The data consist of the records created between the years 1994 and 2003, which account for around one-third of the entire Medline database. Within these data, 29,158 cross-references (pairs of Medline record and gene) were identified as the training data such that they satisfied all of the following conditions: 1) Medline records are assigned one or more MeSH C terms to be used as phenotypes, 2) Medline records are cross-referenced from Entrez Gene to obtain gene functions, 3) cross references are not from the target genes to avoid using possible direct evidence, 4) Medline records have publication dates before 7/1/2003 to avoid using new knowledge.

Using the cross references and the test data in the cancer class, several parameters were empirically determined for each scheme, including the number of Medline articles as the source of phenotypes (n_m), threshold for chi-square statistics to determine phenotypes (t_c), threshold for IDF to determine proxy terms (t_t), and window size for co-occurrences (w_s). For *SchemeT*, they were set as n_m=700, t_c=2.0, t_t=5.0, and w_s=10 (words) by testing a number of combinations of their possible values.

4.3. *Results*

4.3.1. *Overall Performance*

With the best parameter settings learned in the cancer class, the system was applied to all the other classes. Table 2 shows the system performance in AUC.

Table 2. System performance in AUC for each disease class. The figures in the parentheses indicate percent increase/decrease relative to *SchemeK*.

Scheme	Cardio-vascular	Immune	Metabolic	Psych	Unknown	Overall
K	0.677	0.686	0.684	0.514	0.703	0.682
T	0.737 (8.9%)	0.668 (-2.6%)	0.623 (-9.0%)	0.667 (29.8%)	0.786 (11.7%)	0.713 (4.6%)

Both *SchemeK* and *SchemeT* achieved significantly higher AUC than 0.5 (i.e., random guess), indicating the validity of the general framework adapting the inference network for this particular problem. Comparing the two schemes, *SchemeT* does not always outperform *SchemeK* but, overall, AUC improved by 4.6%. The result suggests the advantage of the use of textual data to acquire more precise associations between concepts. Incidentally, without proxy terms described in Section 3.2.2, the overall AUC by *SchemeT* decreased to 0.682 (not shown in Tab. 2), verifying its effectiveness.

4.3.2. *Impact of Full-Text Articles*

This section reports preliminary experiments examining the impact of full text articles for literature-based discovery. Since full-text articles provide more comprehensive information than abstracts, they are thought to be beneficial in the proposed framework. We used the full-text collection from the TREC Genomics Track 2004,[9] which contains 11,880 full-text articles. However, the conditions described in Section 4.2 inevitably decreased the number of usable articles to 679. We conducted comparative experiments using these full-text articles and only the corresponding 679 abstract in estimating $P(f|p)$ for fair comparison. Note that, due to the small data size, these results cannot be directly compared to those reported above.

Table 3 summarizes the results obtained based on only titles and abstracts ("*Abs*") and complete full-text articles ("*Full*") using *SchemeT*.
Examining each disease class, it is observed that the use of full-text articles lead to a large improvement over using abstracts except for the immune class. Overall, the improvement achieved by full texts is 5.1%, indicating the potential advantage of full text articles.

Table 3. System performance in AUC based on 679 articles. The figures in the parentheses indicate percent increase/decrease relative to *Abs*.

Text	Cardio-vascular	Immune	Metabolic	Psych	Unknown	Overall
Abs	0.652	0.612	0.566	0.623	0.693	0.643
Full	0.737 (13.0%)	0.590 (-3.6%)	0.640 (13.0%)	0.724 (16.2%)	0.731 (5.5%)	0.676 (5.1%)

4.3.3. *Enhancing Probability Estimates by Domain Ontologies*

In order to examine the effectiveness of the use of domain ontologies for enhancing $P(f|p)$, we applied the proposed method to *SchemeT* in Tab. 2 and to *Full* in Tab. 3. (Note that *Full* is also based on *SchemeT* for estimating $P(f|p)$ but uses full-text articles instead of abstracts). Figure 2 summarizes the results for different number of iterations, where the left and right plots correspond to *SchemeT* and *Full*, respectively. Incidentally, we used only child-to-parent relations in GO hierarchy for this experiment as it yielded the best results in the cancer class.

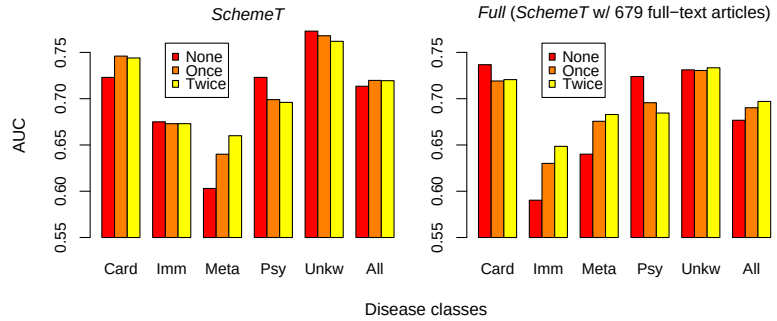

Figure 2. System performance after enhancing associations using GO parent-to-child relations. Three bars in each disease class correspond to # of iterations of enhancement.

For *SchemeT*, the effects were less consistent across the classes and, overall, the improvement was small. For *Full*, on the other hand, we observed clearer improvement except for two classes, *Cardiovascular* and *Psych*, and the overall AUC improved by 4.0%. The difference is presumably due to the fact that the associations learned by *Full* is more sparse than those by *SchemeT* as the amount of the training data for *Full* was limited for this experiment. The enhancement was intended to uncover missed associations and thus worked favorably for *Full*.

5. Conclusion

This study was motivated by Swanson's work in literature-based discovery and investigated the application of IR models and techniques in conjunction with the

use of domain-specific resources, such as gene database and ontology. The key findings of the present work are that a) the consideration of textual information improved system prediction by 4.6% in AUC over simply relying on co-annotations of keywords, b) using full text improved overall AUC by 5.1% as compared to using only abstracts, and c) the hierarchical structure of GO could be leveraged to enhance probability estimates, especially for those learned from small training data. Moreover, we created realistic benchmark data, where old and new knowledge were carefully separated to simulate gene-disease association discovery.

For future work, we plan to investigate the use of semantic distance[10] in propagating the probabilities $P(f|p)$. In addition, we would like to compare the proposed framework with the previous work (e.g., Ref. 6) and with other IR models having one intermediate layer between genes and disease so as to study the characteristics of our model.

Acknowledgments

Dr. Mostafa was funded through the NSF grant #0549313.

References

1. M.A. Hearst. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, 1999.
2. D.R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.
3. P. Srinivasan. Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.
4. M. Weeber, R. Vos, H. Klein, L. Berg, R. Aronson, and G. Molema. Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association*, 10(3):252–259, 2003.
5. H. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.
6. C. Perez-Iratxeta, M. Wjst, P. Bork, and M. Andrade. G2D: a tool for mining genes associated with disease. *BMC Genetics*, 6(1):45, 2005.
7. M.J. Schuemie, M. Weeber, B.J.A. Schijvenaars, E.M. van Mulligen, C.C. van der Eijk, R. Jelier, B. Mons, and J.A. Kors. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604, 2004.
8. T. Fawcett. ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Laboratories, 2004.
9. W. Hersh, R.T. Bhuptiraju, L. Ross, A.M. Cohen, and D.F. Kraemer. TREC 2004 genomics track overview. In *Proceedings of the 13th Text REtrieval Conference*, 2004.
10. P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing*, 8:601–612, 2003.