Discovering implicit associations among critical biological entities

Seki, Kazuhiro

Mostafa, Javed

(Citation)
International Journal of Data Mining and Bioinformatics,3(2):105-123

(Issue Date)
2009-05-01

(Resource Type)
journal article

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/90000946

Discovering Implicit Associations among Critical Biological Entities

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

Tel: +81-78-803-6480

Fax: +81-78-803-6316

Javed Mostafa

Laboratory of Applied Informatics Research

University of North Carolina at Chapel Hill

216 Lenoir Dr., CB#3360, 100 Manning Hall, Chapel Hill, NC 27599-3360, USA

E-mail: jm@unc.edu

Tel: +1-919-962-2182

Abstracts: 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.

Keywords: Text data mining, Literature-based discovery, Information retrieval,

Inference network, Free text, Full text, Domain ontology, Gene Ontology, MeSH, Implicit associations, Gene-disease associations

Biographies

Dr. Kazuhiro Seki received his M.A. and Ph.D. both in information science from University of Library and Information Science, Japan, and Indiana University, Bloomington, respectively. His research interests include natural language processing, information retrieval, machine learning, and their applications to intelligent information processing and management systems. He is an assistant professor at Kobe University, Japan.

Dr. Javed Mostafa is an associate professor in the University of North Carolina at Chapel Hill. He teaches and conducts research in the information science area, specializing in information retrieval and user modeling. He has joint faculty affiliations in the information science program and biomedical research and imaging center (a Medical School entity) at UNC. Recently, his research focused on personalized health information delivery while preserving privacy (a National Science Foundation funded project). He is also the PI on an ongoing educational grant from the Institute of Museum and Library Services focusing on training next generation digital librarians.

1 Introduction

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/advance 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 (Hearst, 1999). The pioneering work in TDM (or literature-based discovery) was conducted by Swanson in the 1980's. He argued that there were 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 demonstrate the validity of 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 (Swanson, 1986). The hypothesis was later supported by experimental evidence (DiGiacomo et al., 1989).

Based on his original work, Swanson and other researchers have developed computer programs to aid hypothesis discovery (Gordon and Lindsay, 1996; Srinivasan, 2004; Swanson and Smalheiser, 1997; Weeber et al., 2001). 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 approaches, models, and evaluation methodologies. For instance, 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 and his colleague had proposed (e.g., Swan-

son, 1988; Smalheiser and Swanson, 1996).

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 (Turtle and Croft, 1991) in IR and conduct various evaluative experiments on a realistic benchmark data set.

2 Scope of this Study

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 this kind of genetic associations has tremendous importance for prevention, prediction, and treatment of hereditary diseases. In this context, predicting or ranking candidate genes for a given disease is crucial to select more plausible ones to speed up genetic association studies.

In developing a framework to discover implicit 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, the system attempts to predict a (unknown) causative gene by way of producing a ranked list of candidate genes.

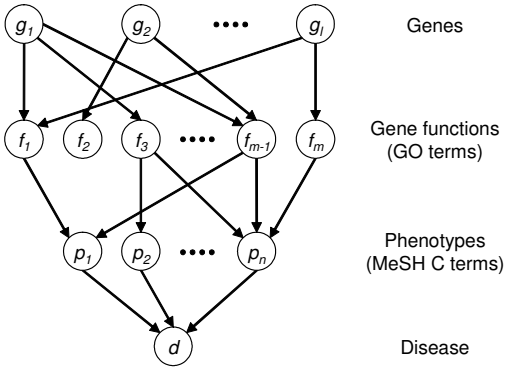

Figure 1: Inference network modeling gene-disease associations.

3 Proposed Approach

Focusing on gene-disease associations, we explored the use of a formal IR model, specifically, the inference network (Turtle and Croft, 1991) for this related but different problem targeting implicit associations. The following details the proposed model and how to estimate the probabilistic parameters 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 IR model to represent gene-disease associations, we treat a disease as a 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 (Salton and McGill, 1983), are not readily applicable as they are not designed to have different sets of concepts (intermediate nodes) to represent documents and queries.

The network in Fig. 1 consists of four types of nodes: genes (g), gene functions

(f) represented by Gene Ontology (GO) terms,[1] phenotypes (p) represented by MeSH terms under the C category (hereafter referred to as MeSH C terms for short),[2] and disease (d). Each gene node g represents a gene and corresponds to the event that the gene is found in search for the causative genes underlying d. Each gene function node f represents a function of gene products. There 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,[3] 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, causative gene set G for given disease d can be predicted by the probability:

$$P(G|d) = \frac{P(d|G) \times P(G)}{P(d)} \tag{1}$$

where the denominator can be dropped as it is constant for given d. In addition, assuming that $P(G)$ is uniform, $P(G|d)$ can be approximated to $P(d|G)$ below defined as the sum of the probabilities associated with all the paths from G to d.

$$P(d|G) = \sum_i \sum_j P(d|\vec{p_i}) \times P(\vec{p_i}|\vec{f_j}) \times P(\vec{f_j}|G) \tag{2}$$

Eq. (2) quantifies how much a set of candidate genes, G, increases the belief in the development of disease d, where $\vec{p_i}$ (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, we derive the

[1] http://www.geneontology.org

[2] http://www.nlm.nih.gov/mesh

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

following (see Appendix A for more complete derivation):

$$P(d|G) \propto \sum_i \sum_j \left(\frac{P(p_i|d)}{P(\bar{p}_i|d)} \times \frac{P(f_j|p_i)P(\bar{f}_j|\bar{p}_i)}{P(\bar{f}_j|p_i)P(f_j|\bar{p}_i)} \times F(p_i) \times F(f_j) \times P(f_j|G) \right) \quad (3)$$

where p and $\bar{p}$ (f and $\bar{f}$) are used as the shorthand of $p = 1$ and $p = 0$ ($f = 1$ and $f = 0$), respectively, and

$$F(p_i) = \prod_{h=1}^{m} \frac{P(\bar{f}_h|p_i)}{P(\bar{f}_h|\bar{p}_i)}, \; F(f_j) = \prod_{k=1}^{n} \frac{P(\bar{f}_j)P(f_j|\bar{p}_k)}{P(f_j)P(\bar{f}_j|\bar{p}_k)} \; . \quad (4)$$

Note that since p is a binary random variable taking either 0 or 1, $P(\bar{p}|d)$ equals to $1 - P(p|d)$. Similarly, $P(\bar{f}|p)$ is computed as $1 - P(f|p)$, and $P(\bar{f}|\bar{p})$ as $1 - P(f|\bar{p})$. The first factor of the right-hand side of Eq. (3) 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.

3.2 Limitations of the Proposed Model

The inference network described in Section 3.1 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, for example, among phenotypes, such as "phenotype x is often observed with phenotype y," our proposed model does not intend to capture those associations. In other words, 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 disease 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 in our proposed model.

Next, Eq. (2) introduces $\vec{p_i}$ and $\vec{f_j}$ in which only one element of the vectors is positive. While this approximation allows us to assess only the effect of each

factor, p_i and f_j, in isolation from other factors (Baeza-Yates and Ribeiro-Neto, 1999, p. 55), it does not consider the case where strong many-to-one relations exist (e.g., both phenotypes x and y must be observed when disease d is developed). It should be mentioned that, however, even though using only one-to-one relations cannot precisely describe such cases, it could still capture desired relations to a degree because each one-to-one relation would increase the belief in the outcome. For example, if observing phenotype x suggests the development of disease d to some extent and so does observing phenotype y, then it would be more likely to have disease d when observing both x and y than observing only one.

Another limitation is the assumption regarding the uniformity of the probability $P(G)$ since there would be more likely set of genes G that are positive (activated) together. To obtain better estimates for $P(G)$ rather than assuming the uniform distribution, we plan to study the use of known (or predicted) gene-gene interactions (Franke et al., 2006, for example) in future work.

3.3 Probability Estimation

3.3.1 Estimating Conditional Probability $P(p|d)$

The 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 carried out to retrieve articles relevant to d and, within the retrieved articles, we first identify phenotypes (MeSH C terms) strongly associated with the disease based on chi-square statistic. Chi-square is commonly used for testing the independence between two nominal variables and has been used for many applications. Examples include identifying most discriminative features (terms) for text categorization (Yang and Pedersen, 1997) and discovering significant word collocations in a given corpus in statistical natural language processing (Manning and Schütze, 1999).

Given disease d and phenotype p, the chi-square statistic $\chi^2(d, p)$ is computed

as follows.

$$\chi^2(d,p) = \frac{N(n_{11} \cdot n_{22} - n_{21} \cdot n_{12})^2}{(n_{11}+n_{21})\,(n_{12}+n_{22})\,(n_{11}+n_{12})\,(n_{21}+n_{22})} \tag{5}$$

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.

For estimating $P(p|d)$, we reuse the chi-square statistic and normalize $\chi^2(d,p)$ by the maximum to scale it to the range between 0 to 1. It should be mentioned that there are other measures that can be used for identifying phenotypes and estimating $P(p|d)$, including IDF, mutual information, or relative frequencies ($n_{11}/(n_{11}+n_{21})$). However, chi-square produced marginally better predictions in our preliminary experiments on cancer-related diseases and was used in this study.

3.3.2 Estimating Conditional Probability $P(f|p)$

The 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. (2005). Unlike them, however, this study focuses on the use of textual data and domain ontologies and investigates their effects for literature-based discovery. In what follows, we first describe the estimation method for $P(f|p)$ and then for $P(f|\bar{p})$.

For probability estimation, our framework uses MEDLINE records that are assigned any MeSH C terms and are cross-referenced from any gene entry in the Entrez Gene database. For each of such records, we can obtain a set of phenotypes (the MeSH C terms assigned to the record) and a set of gene functions (GO terms) associated with the Entrez Gene entry which cross-references to the MEDLINE record. Considering the fact that the phenotypes and gene functions are associated with the same MEDLINE record, it is possible that some of the phenotypes and

gene functions are associated. A question is what phenotypes and functions are associated and how strong those associations are.

We estimate those possible associations using two alternative 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 C terms) in a sliding, fixed-size window, assuming that associated concepts tend to co-occur more often in the same context than unassociated ones. However, a problem of this strategy is that gene functions and phenotypes are descriptive by nature and may not be expressed in concise GO and MeSH terms (Camon et al., 2005; Schuemie et al., 2004). Schuemie et al. 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. Therefore, 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 user 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 for most IR systems to quantify the specificity of terms in a given document collection (Sparck Jones, 1972). We treat term definitions as pseudo-documents and define IDF for term t as

$$\mathrm{IDF}(t) = \log \frac{N}{Freq(t)} \tag{6}$$

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 original concept, i.e., gene function or phenotype.

Each co-occurrence of the terms from those two proxy term sets (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 estimated 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,

$$S(f, p, a) = \sum_{(t_f, t_p, a)} \frac{w(t_f) \cdot w(t_p)}{|Proxy(f)| \cdot |Proxy(p)|} \tag{7}$$

where t_f and t_p denote any terms in the proxy term sets 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 sizes. As term weight w, this study uses the TF·IDF weighting scheme, also from the IR literature. For term t_p, for instance, we define

$$\mathrm{TF}(t_p) = 1 + \log Freq(t_p, Def(p)) \tag{8}$$

where $Def(p)$ denote p's definition and $Freq(t_p, Def(p))$ denotes the number of occurrences of t_p in $Def(p)$. See Eq. (6) for the definition of IDF.

The association scores, $S(f, p, a)$, are computed for each cross-reference (a pair of a MEDLINE record and a gene) by either *SchemeK* or *SchemeT* and are summed over all cross-references to estimate the association between f and p, denoted as $S(f, p)$. Based on the accumulated associations, we define probability $P(f|p)$ as the relative strength of the association, i.e., $S(f, p)/\sum_{f'} S(f', p)$. In other words, $\sum_{f'} S(f', p)$ is used to normalize $S(f, p)$. This way, f can receive a high probability when the association of f and p stands out among other functions. It is, however, also possible to use other normalization factors that are independent of particular f's, such as the upper bound of the association score (which is the number of training instances in our framework). Further investigation is needed to identify the best strategy.

Then, for estimating $P(f|\bar{p})$ where $p = 0$, we use the association scores between f and $p'(\neq p)$. To be precise, we define $S(f, \bar{p})$ as the sum of those association scores, $\sum_{p' \neq p} S(f, p')$, and estimate $P(f|\bar{p})$ as the relative strength of the association, i.e., $S(f, \bar{p}) / \sum_{f'} S(f', \bar{p})$, in the same way as estimating $P(f|p)$ above.

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 assume that those estimated associations between gene functions and phenotypes are directed from the former to the latter.

3.3.3 Estimating Prior Probability $P(f)$

The probability $P(f)$ can be interpreted as the degree of belief that f is in effect when no observation is made (yet). In principle, $P(f)$ should be high if f is in general involved in many diseases and should be low otherwise. To reflect the intuition, we use the number of causative genes with function f, denoted as $N(f)$, in training data. Precisely, we estimate $P(f)$ as the ratio of $N(f)$ to the total number of gene-disease pairs in the training data.

3.3.4 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. For example, 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[4] and those located nearby in the hierarchy are semantically close to each other. Taking advantage of these semantic relations, we enhance probability estimates $P(f|p)$ as follows.

Let us denote by A the matrix whose element a_{ij} is the probability $P(f_j|p_i)$

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

and by A' the other matrix whose element a'_{ij} is updated or enhanced probability $P'(f_j|p_i)$. Then, A' can be 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. Here, we focus only on direct child-to-parent and parent-to-child relations and defines $w_p(i, j)$ as

$$w_p(i,j) = \begin{cases} 1 & \text{if } i = j \\ \dfrac{1}{\#\text{ of children of } p_j} & \text{if } p_i \text{ is a child of } p_j \\ \dfrac{1}{\#\text{ of parents of } p_j} & \text{if } p_i \text{ is a parent of } p_j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Eq. (9) means that the amount of probability is simply 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. (9). Note that this enhancement process can be iteratively applied to take advantage of more distant relationships than children or 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) (Becker et al., 2004). GAD[5] is a manually-curated archive of human genetic studies, containing pairs of a gene and a disease that are reported to have causative relations.

4.1 Creation of Benchmark Data

For evaluation, a benchmark data set was created as follows using the real-world data obtained from GAD.

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

1. Associate each gene-disease pair with the publication date of the article from which the GAD 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. (Since GAD's scope includes complex diseases, one disease may be paired with multiple genes.)

3. For each pair of a disease and its causative gene(s),
   (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 a 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). The particular date was arbitrarily chosen by considering the size of the resulting data and available resources for parameter estimation.
   (b) Remove the most recently reported gene identified above 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 so as to simulate gene-disease association discovery. Tab. 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, and the other classes were used for test.

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 (if given), 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 an evaluation metric, we use *area under the ROC curve* (AUC) for its attractive property as compared to F-score; That is, AUC is not affected by changes in class distribution, i.e., the proportion of positive to negative instances (see Fawcett, 2004, for more detailed discussion).

ROC curves are two dimensional measure for system performance with y axis being true positive proportion (TPP) and x 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 chance that a system ranks a gene randomly picked from the positive set more highly than one from the negative set. To be effective, therefore, AUC must be at least higher than 0.5 which corresponds to a pure guess.

As literature data, this study uses a subset of the MEDLINE data provided for the TREC Genomics Track 2004 (Hersh et al., 2004). 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 a MEDLINE record and a gene) were identified as the training data such that they satisfied all of the following conditions:

1. The MEDLINE records are assigned one or more MeSH C terms to be used as phenotypes,

2. The MEDLINE records are cross-referenced from an Entrez Gene entry so

as to obtain gene functions assigned to the entry,

3. The MEDLINE records have publication dates before 7/1/2003 so as to avoid using new knowledge.

4. The cross-references do not originate from the 257 target genes so as to avoid using possible direct evidence,

Using the literature data, the cross-references, and the tuning data in the cancer class, several parameters were empirically determined for each scheme, *SchemeK* and *SchemeT*, to maximize AUC. Those parameters include the number of MEDLINE articles as the source of phenotypes (n_m), threshold for chi-square statistic to determine phenotypes (t_c), threshold for IDF to determine proxy terms (t_t), and window size for co-occurrences (w_s). For *SchemeT*, for instance, they were set as n_m=700, t_c=2.0, t_t=5.0, and w_s=10 (words) by testing numbers of their possible combinations.

4.3 Results

4.3.1 Overall Performance

With the best parameter settings determined for the cancer class, the system was applied to all the other classes. Tab. 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 infer-

ence network for predicting implicit associations. Comparing the two schemes, it is observed that *SchemeT* does not always outperform *SchemeK* depending on the disease class. However, the overall tendency suggests the advantage of the use of textual data to acquire more precise associations between biological concepts. For the Immune and Metabolic classes, in which AUC dropped by 2.6–9.0%, a close investigation is needed to determine the cause of the problem. Incidentally, without proxy terms described in Section 3.3.2, the overall AUC by *SchemeT* decreased to 0.682 (not shown in Tab. 2), verifying the effectiveness of the use of proxy terms.

4.3.2 System Performance with Different Size of the Literature

As described in Section 4.2, the MEDLINE data used for this study are about one-third of the entire database, which also restricts the number of cross-references that can be used for parameter estimation. Although the experiment in the previous section has shown the effectiveness of the proposed approach even with the limited amount of data, further improvement might be achieved with larger literature data. To study the potential of the framework, this section examines the relation between the system performance and the size of the literature.

From the available 29,158 cross-references (pairs of a gene and a MEDLINE record), we randomly chose 2,500, 5,000, 7,500, 15,000, and 22,500 pairs for parameter estimation. Based on each of the subsets, probabilities $P(f|p)$ were separately calculated with *SchemeK* and *SchemeT*. Then, they were used to estimate $P(d|G)$ for the test data in all the disease classes (including the cancer class) for comparison. Fig. 2 plots the results where x and y axes represent AUC and the number of cross-references, respectively. Notice that the right-most dots correspond to the case where all the available cross-references are used.

There is an immediate increase when the number of cross references increases from 2,500 to 5,000 irrespective of the schemes. Then, for *SchemeT*, AUC almost linearly increases with the size of training data, while, for *SchemeK*, it slows down after using 15,000 cross-references. This result suggests that, with larger MED-

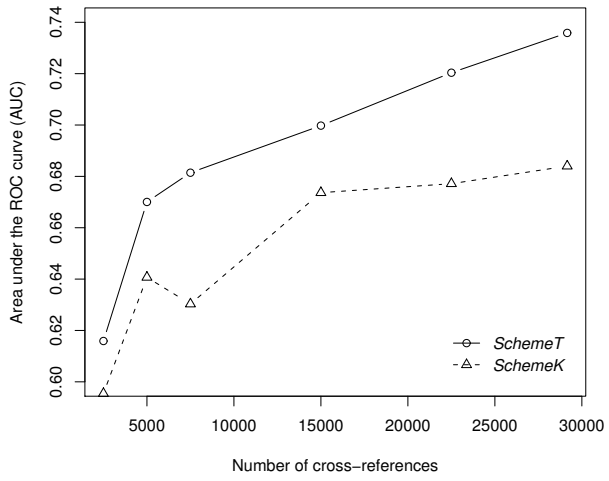

Figure 2: Relation between AUC and the number of cross-references used for estimating $P(f|p)$.

LINE data, we may witness a greater advantage of *SchemeT* over *SchemeK*.

4.3.3 Impact of Full-Text Articles

This section reports our 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 for the text data mining research, such as ours. We use the full-text collection from the TREC Genomics Track 2004 (Hersh et al., 2004), which contains 11,880 full-text articles in the biomedical domain. Among these articles, however, only 679 satisfy the conditions described in Section 4.2 regarding MeSH C terms, cross-references, and publication dates. To make a fair comparison, we conducted experiments using the 679 full-text articles and only the corresponding 679 abstract in estimating $P(f|p)$. Note that, due to the smaller data size used for parameter estimation, the results reported below cannot be directly compared to those described in the previous sections.

Tab. 3 summarizes the results obtained based on only titles and abstracts ("*Abs*") and complete full-text articles ("*Full*"), both using *SchemeT*. (As *SchemeK* does

not use textual information, there is no distinction between *Abs* and *Full* for *SchemeK*.)

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%)

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.

4.3.4 Enhancing Probability Estimates by Domain Ontologies

In order to examine the effectiveness of the use of domain ontologies for enhancing $P(f|p)$, we then applied the proposed method described in Section 3.3.4 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 a small collection of full-text articles instead of abstracts.) Fig. 3 summarizes the results obtained by different number of iterations, where the left and right graphs correspond to *SchemeT* and *Full*, respectively. Incidentally, we used only child-to-parent relations in GO hierarchy for this experiment without using MeSH as it yielded the best results in the cancer class (i.e., the tuning data).

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% after two times of iterations. The difference may be due to the fact that the associations learned by *Full* is more sparse than those by *SchemeT* as the number of available cross-references for *Full* was limited for this experiment.

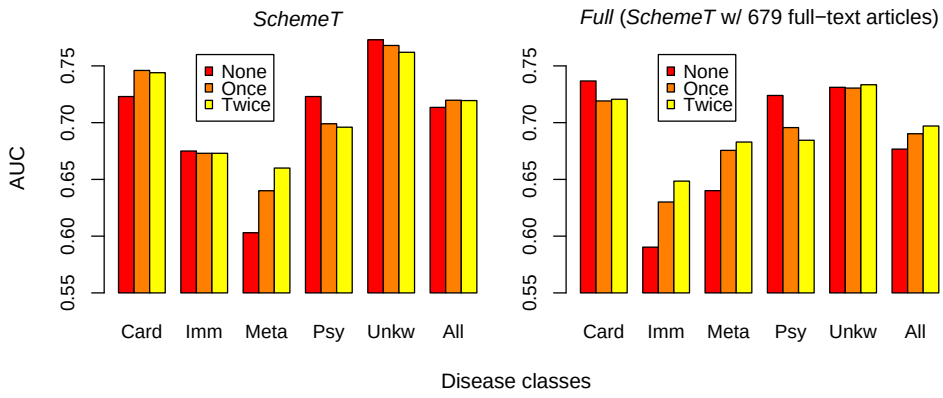

Figure 3: System performance after enhancing associations using GO parent-to-child relations. Three bars in each disease class correspond to # of iterations of enhancement.

The enhancement was intended to uncover the associations not derived from the literature, and thus it may have worked favorably for *Full*. Another possible account is that the quality of the associations obtained by *Full* was higher than that of *SchemeT*. In other words, if the initial associations are noisy, the enhancement process would just propagate false associations through the GO structure, which rather deteriorates system predictions. Our plan is to carry out an experiment with larger full-text data set, so as to determine the effect of domain ontologies.

4.4 Comparative Experiments

To the best of our knowledge, there have been at least a few attempts to develop and evaluate approaches to predicting implicit gene-disease associations (Freudenberg and Propping, 2002; Perez-Iratxeta et al., 2005; Tiffin et al., 2005). This section makes a rough comparison with the related work by using the same data set and by presenting the system performance in the form reported by the others. Specifically, the following focuses on the study by Perez-Iratxeta et al. (2005), which is similar to the present study in the approach and the experimental design.

Perez-Iratxeta et al. (2002, 2005) have developed a system, called G2D, to predict gene-disease associations. G2D is in part similar to *SchemeK* as it utilizes co-annotations of keywords without looking at textual information, although it also employs DNA sequence comparison to find homologous genes similar to those scored highly by keyword co-annotations. As with our system, G2D outputs a ranked list of candidate genes for a given disease. For evaluation, they conducted experiments on 100 known gene-disease associations randomly chosen from Online Mendelian Inheritance in Man (OMIM).[6] Of those 100 associations, they reported that the target genes were successfully found among the 8 best scoring genes in 47 cases and among the 30 best in 62 cases.

To investigate how our approach is compared with that of Perez-Iratxeta et al. (2005), we ran our system with *SchemeK* and *SchemeT* on the same 100 diseases. In the training phase, the following cross-references were excluded to avoid using direct target associations.

- The cross-references originating from the 100 target genes.
- The cross-references pointing to the MEDLINE records which mention any of the input disease names.

Tab. 4 summarizes the results, in which the number of the target genes successfully predicted is indicated at each ranking. Note that, among 100 input diseases, 6 could not be processed either because their target genes were annotated with no GO term or because the disease names could not be associated with any phenotypes.

As shown, 40 and 49 target genes were found among the 8 best scoring genes, and 64 and 75 among the 30 best scoring genes by *SchemeK* and *SchemeT*, respectively (shown in boldface). Comparing with the results obtained by Perez-Iratxeta et al. (2005), i.e., 47 genes among the 8 best and 62 genes among the best 30, *SchemeT* successfully predicted 2 and 13 more true associations among the 8 and 30 best candidates, respectively.

[6] `http://www.ncbi.nlm.nih.gov/omim/`

Table 4: Numbers of target genes correctly predicted above each ranking for 100 monogenic diseases. GAD's cumulative sums are reported only for 8th and 30th (Perez-Iratxeta et al., 2005).

	SchemeK		*SchemeT*		GAD
Ranking	Number of target genes	Cumulative sum	Number of target genes	Cumulative sum	Cumulative sum
1	10	10	18	18	—
2	6	16	4	22	—
3	7	23	6	28	—
4	5	28	7	35	—
5	4	32	2	37	—
⋮	⋮	⋮	⋮	⋮	⋮
8	3	40	4	49	47
⋮	⋮	⋮	⋮	⋮	⋮
30	1	64	0	75	62

It should be noted that their results and this study are not directly comparable because the experimental settings are not exactly the same. For example, the candidate genes considered in their study and in this study are different; Perez-Iratxeta et al. examined genes located near the target gene (around 30 mega bases), whereas we examined genes in the same sub-band of the same chromosome as the target gene. In addition, Perez-Iratxeta et al. used the entire MEDLINE database for prediction, whereas we used only one-third of it. Although there are such differences in the specifics of the experiments, our results are promising and suggest that it be worthwhile further exploration.

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 the Entrez Gene database and Gene Ontology. The key findings of the present work can be summarized as follows.

- The model and techniques developed for IR, which targets *explicit* information, were shown both applicable and effective for predicting *implicit* gene-disease associations.

- The consideration of textual information (i.e., *SchemeT*) improved system prediction by 4.6% in AUC over simply relying on co-annotations of MeSH C terms and GO terms (i.e., *SchemeK*).

- Using full text improved overall AUC by 5.1% as compared to using only abstracts, although the database cross-references used for parameter estimation in the experiment was limited to a relatively small size.

- The hierarchical structure of GO could be leveraged to enhance probability estimates, especially for those learned from full-text articles.

- The comparative experiment on the 100 diseases from the OMIM database demonstrated that our approach successfully found more true gene-disease associations than that of Perez-Iratxeta et al. (2005) by up to 21%.

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 re-examine the effectiveness of full-text articles in the proposed framework using a larger data set from the TREC 2006 Genomics Track (Hersh et al., 2006). Also, we would like to investigate the better use of domain ontologies. One direction would be to employ different weighting functions (w_p and w_f) in propagating the probability $P(f|p)$, such as the semantic distance (Lord et al., 2003; Resnik, 1999) which better reflects the structure of domain ontology. Furthermore, we are interested in comparing the proposed framework with other IR models in order to study the properties and advantages of our proposed model.

Acknowledgments

This project is partially supported by KAKENHI #19700147, the Nakajima Foundation, the Artificial Intelligence Research Promotion Foundation grant #18AI-255, and the NSF grant ENABLE #0333623. We would like to thank the anonymous reviewers for their helpful comments.

A Derivation of Formula

This section demonstrates step by step how Eq. (3) is derived from Eq. (2). To begin with, the first factor in the right-hand side of Eq. (2) can be transformed as follows by Bayes' theorem and the independence assumption among phenotypes p.

$$\begin{aligned} P(d|\vec{p_i}) &= \frac{P(d)P(\vec{p_i}|d)}{P(\vec{p_i})} \\ &\approx P(d)\prod_{k=1}^{n}\frac{P(p_k|d)}{P(p_k)} \end{aligned} \tag{10}$$

Because $p_i = 1$ and $p_{k\neq i} = 0$ for any $\vec{p_i}$ by definition, the product in the right-hand of Eq. (10) can be written as follows, where p and $\bar{p}$ are used as the shorthand of $p = 1$ and $p = 0$, respectively.

$$\begin{aligned} \prod_{k=1}^{n}\frac{P(p_k|d)}{P(p_k)} &= \frac{P(p_i \mid d)}{P(p_i)}\prod_{k\neq i}\frac{P(\bar{p_k} \mid d)}{P(\bar{p_k})} \\ &= \frac{P(p_i|d)}{P(p_i)}\cdot\frac{P(\bar{p_i})}{P(\bar{p_i}|d)}\prod_{k=1}^{n}\frac{P(\bar{p_k}|d)}{P(\bar{p_k})} \end{aligned} \tag{11}$$

Then, the second factor in the right-hand side of Eq. (2), $P(\vec{p_i}|\vec{f_j})$, can be transformed as follows, assuming the independence among phenotypes and among gene functions.

$$\begin{aligned} P(\vec{p_i}|\vec{f_j}) &\approx \prod_{k=1}^{n} P(p_k|\vec{f_j}) \\ &= \prod_{k=1}^{n}\frac{P(p_k)P(\vec{f_j}|p_k)}{P(\vec{f_j})} \\ &\approx \prod_{k=1}^{n}\prod_{h=1}^{m} P(p_k)\frac{P(f_h|p_k)}{P(f_h)} \\ &= \prod_{k=1}^{n} P(p_k)\times\prod_{k=1}^{n}\prod_{h=1}^{m}\frac{P(f_h|p_k)}{P(f_h)} \end{aligned} \tag{12}$$

As before, given that $p_i = 1$ and $p_{k\neq i} = 0$ for any $\vec{p_i}$ and $f_j = 1$ and $f_{h\neq j} = 0$ for any $\vec{f_j}$, the first factor of the right-hand side of Eq. (12) becomes

$$\prod_{k=1}^{n} P(p_k) = \frac{P(p_i)}{P(\bar{p}_i)} \prod_{k=1}^{n} P(\bar{p}_k) \tag{13}$$

and the second factor becomes

$$\begin{aligned} \prod_{k=1}^{n}\prod_{h=1}^{m} \frac{P(f_h|p_k)}{P(f_h)} &= \frac{P(f_j|p_i)}{P(f_j)} \times \prod_{h\neq j} \frac{P(\bar{f}_h|p_i)}{P(\bar{f}_h)} \times \prod_{k\neq i} \frac{P(f_j|\bar{p}_k)}{P(f_j)} \times \prod_{k\neq i}\prod_{h\neq j} \frac{P(\bar{f}_h|\bar{p}_k)}{P(\bar{f}_h)} \\ &= \frac{P(f_j|p_i)P(\bar{f}_j|\bar{p}_i)}{P(\bar{f}_j|p_i)P(f_j|\bar{p}_i)} \times \prod_{h=1}^{m} \frac{P(\bar{f}_h|p_i)}{P(\bar{f}_h|\bar{p}_i)} \times \prod_{k=1}^{n} \frac{P(\bar{f}_j)P(f_j|\bar{p}_k)}{P(f_j)P(\bar{f}_j|\bar{p}_k)} \times \prod_{k=1}^{n}\prod_{h=1}^{m} \frac{P(\bar{f}_h|\bar{p}_k)}{P(\bar{f}_h)} . \end{aligned} \tag{14}$$

Next, the third factor of Eq. (2), $P(\vec{f_j}|G)$, can be derived from external knowledge source, specifically, Entrez Gene, where each gene is annotated with GO terms (i.e., f's). Based on the database, we define $P(\vec{f_j}|G)$ to be 1 if any gene $g \in G$ is annotated with f_j in the database and 0 otherwise. That is,

$$P(\vec{f_j}|G) = P(f_j|G) = \begin{cases} 1 & \text{if } \exists g \in G \text{ is associated with } f_j \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where we use $P(f_j|G)$ as a simpler representation of $P(\vec{f_j}|G)$.

Lastly, by applying Eqs. (11), (12), (13), (14), and (15) to Eq. (2) and removing constants for given d and any G, one could derive Eq. (3).

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley Longman.

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nature Genetics*, 36:431–432.

Camon, E., Barrell, D., Dimmer, E., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R. (2005). An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 6(Suppl 1):S17.

DiGiacomo, R. A., Kremer, J., and Shah, D. (1989). Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine*, 86(2):158–164.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Laboratories.

Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*, 78(6):1011–1025.

Freudenberg, J. and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18(Supplement 2):s110–s115.

Gordon, M. D. and Lindsay, R. K. (1996). Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, 47(2):116–128.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10.

Hersh, W., Bhuptiraju, R. T., Ross, L., Ross, L., Cohen, A. M., and Kraemer, D. F. (2004). TREC 2004 genomics track overview. In *Proceedings of the 13th Text REtrieval Conference (TREC)*.

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006). TREC 2006 genomics track overview. In *TREC Notebook*. NIST.

Lord, P., Stevens, R., Brass, A., and Goble, C. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing*, 8:601–612.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316–319.

Perez-Iratxeta, C., Wjst, M., Bork, P., and Andrade, M. (2005). G2D: a tool for mining genes associated with disease. *BMC Genetics*, 6(1):45.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Schuemie, M. J., Weeber, M., Schijvenaars, B. J. A., van Mulligen, E. M., van der Eijk,

C. C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604.

Smalheiser, N. R. and Swanson, D. R. (1996). Indomethacin and Alzheimer's disease. *Neurology*, 46(2):583.

Sparck Jones, K. (1972). Statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.

Srinivasan, P. (2004). Text mining: generating hypotheses from Medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18.

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557.

Swanson, D. R. and Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., and Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, 33(5):1544–1552.

Turtle, H. and Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222.

Weeber, M., Klein, H., de Jong-van den Berg, L. T. W., and Vos, R. (2001). Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420.