BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 1 2005, pages 104–115
doi:10.1093/bioinformatics/bth464

Gene clustering by Latent Semantic Indexing of MEDLINE abstracts

Ramin Homayouni[1,], Kevin Heinrich[2], Lai Wei[1] and Michael W. Berry[2]*

[1]Department of Neurology, University of Tennessee Health Science Center, 855 Monroe Avenue, 416 Link Bldg, Memphis, TN 38163, USA and [2]Department of Computer Science, University of Tennessee, Knoxville, TN 37996-3450, USA

Received on September 30, 2003; revised on July 10, 2004; accepted on August 2, 2004
Advance Access publication August 12, 2004

ABSTRACT

Motivation: A major challenge in the interpretation of high-throughput genomic data is understanding the functional associations between genes. Previously, several approaches have been described to extract gene relationships from various biological databases using term-matching methods. However, more flexible automated methods are needed to identify functional relationships (both explicit and implicit) between genes from the biomedical literature. In this study, we explored the utility of Latent Semantic Indexing (LSI), a vector space model for information retrieval, to automatically identify conceptual gene relationships from titles and abstracts in MEDLINE citations.

Results: We found that LSI identified gene-to-gene and keyword-to-gene relationships with high average precision. In addition, LSI identified implicit gene relationships based on word usage patterns in the gene abstract documents. Finally, we demonstrate here that pairwise distances derived from the vector angles of gene abstract documents can be effectively used to functionally group genes by hierarchical clustering. Our results provide proof-of-principle that LSI is a robust automated method to elucidate both known (explicit) and unknown (implicit) gene relationships from the biomedical literature. These features make LSI particularly useful for the analysis of novel associations discovered in genomic experiments.

Availability: The 50-gene document collection used in this study can be interactively queried at http://shad.cs.utk.edu/sgo/sgo.html

Contact: rhomayouni@utmem.edu

Supplementary information: http://shad.cs.utk.edu/sgo/pubs.html

INTRODUCTION

Recent advances in genomic and proteomic technologies enable investigators to rapidly identify groups of genes that are coordinately regulated in different experimental conditions. However, understanding the functional relationships and the biological effects of co-regulated genes remains to be a time-consuming and arduous task, requiring investigators to manually extract and assemble gene information from various biological databases. Understandably, the efforts to develop data mining tools to extract gene information from the biomedical literature have intensified recently (Shatkay and Feldman, 2003; Wilkinson and Huberman, 2004). As a first step, high-throughput automated methods are needed to rapidly validate genomic data and to identify groups of functionally related genes based on the published literature (Jenssen *et al*., 2001). Once groups of functionally related genes are identified, more computationally intensive text-mining methods such as natural language processing can be used to extract the nature of the relationships among genes (Yandell and Majoros, 2002).

Automated information retrieval (IR) methods have been around since the creation of digital libraries and the World Wide Web (Baeza-Yates and Ribeiro-Neto, 1999). There are three basic models for IR: set theoretic (Boolean), algebraic (vector space) and probabilistic. In the Boolean method, documents are represented by the sets of index terms and are retrieved in a binary fashion, i.e. only if the query contains the index term. In the vector space model, documents are represented by weighted index terms in a multidimensional space. Here, documents are retrieved based on the degree of similarity to the query, even if the query terms do not appear in the document. In the probabilistic model, documents are retrieved based on the probability that they are relevant to the query. Probabilistic models usually require further interaction with the user to improve retrieval performance.

For genomic applications, a number of set theoretic methods have been described in recent years that utilize functional gene annotation in public electronic databases, such as Medical Subject Heading (MeSH) index, LocusLink (Pruitt and Maglott, 2001), Gene Ontology (GO) (Ashburner *et al*., 2000), and numerous protein–protein interaction or biochemical pathway databases such as KEGG (Kanehisa and Goto,

*To whom correspondence should be addressed.

2000). For example, HAPI identifies gene relationships based on co-occurrence of MeSH index terms in representative MEDLINE citations (Masys *et al*., 2001). Other methods such as MAPPFinder (Doniger *et al*., 2003), EASE (Hosack *et al*., 2003) and GoMiner identify gene relationships using the gene function classifications in GO. A major limitation to these methods is from the binary criterion used in indexing. Another limitation is the lack of specificity of controlled vocabularies; since index terms are usually general, specific information about genes are lost. Moreover, a confounding issue arises from the subjectivity of indexers, whereby different index terms may be assigned to the same citation by different indexers (Funk and Reid, 1983).

An alternative approach to retrieve gene relationships would be to query the biomedical literature directly. Hovig and colleagues have developed PubGene, an automated tool to extract gene relationships based on co-occurrence of gene symbols in MEDLINE abstracts (Jenssen *et al*., 2001). PubGene provides a rapid method to identify gene neighbors based on the biomedical literature. However, on average it identifies only 50% of the known gene relationships. This low recall [1] is primarily due to inconsistencies in gene symbol usage in the literature. In IR, these problems are referred to as synonymy (multiple words having same meaning) and polysemy (words having multiple meanings). For example, in addition to the official gene symbol, many genes contain aliases or synonyms that are preferred by different investigators. Often, biochemical or cell biological studies refer to the gene product and not to the gene itself. Because of this inherent noise in the biological literature, relevant information may be overlooked by focusing on the gene symbol or any single word representation of the gene in the literature.

A major issue in the interpretation of genomic data is that often genes/proteins are identified in experiments that have not been previously studied together. Thus using the term co-occurrence methods to extract overlapping abstracts will not be appropriate for the interpretation of discovery-based genomic studies. Ideally, genomic IR methods would classify genes based not only on known or explicit relationships but also on latent or implicit relationships reported in the literature. Several tools such ARROWSMITH and PubMatrix exist that aid in the extraction of implicit textual relationships between distinct sets of MEDLINE abstracts (Smalheiser and Swanson, 1998; Becker *et al*., 2003). Although these tools could be applied for the interpretation of genomic data, they are not at present amenable to high-throughput studies.

Recently, vector space modeling has been explored for gene clustering using functional information in annotated indices or MEDLINE abstracts (Glenisson *et al*., 2003). In this model, the semantic structure of a document is represented as a vector (essentially a bag of words) in word space and the degree of similarity between documents is calculated by the cosine of the angle between document vectors (Berry *et al*., 1995, 1999). The vectors consist of weighted terms, which is a function of the frequency of the terms in and across the documents in the collection. Glenisson *et al*. (2003) demonstrated that the expansion of gene annotation by a vector space strategy resulted in considerable improvement in clustering a subset of genes over term-matching (Boolean) method. A variant of the vector space model, called Latent Semantic Indexing (LSI), improves retrieval by using a classical factorization method from linear algebra (singular value decomposition, SVD) to create a subspace in which text documents are represented as vectors (Deerwester *et al*., 1988, 1990). The components in the subspace may be regarded as a concept derived from the word usage patterns in the document. Hence, the relevant documents are retrieved based on the degree of similarity in the word usage patterns (concepts) in the documents. The mathematical details of this model have been reviewed previously (Berry, 1992; Berry *et al*., 1995, 1999). LSI has been shown to be 30% more effective in finding and ranking relevant items than word-matching methods (Deerwester *et al*., 1990). An important advantage of vector space models to genomic studies is that relationships between genes may be extracted even if they do not co-occur in abstracts.

The performance of LSI has been evaluated in many different applications. In some instances, the mathematical conceptualization of the text material has been shown to be similar to information processing exhibited by humans (Landauer *et al*., 1998). For example, after training on 5 million words from an encyclopedia (using 300 factors), LSI scored as well as the average score from a large sample of foreign national students in a multiple-choice vocabulary test used by the Educational Testing Service (ETS) for the Test of English as a Foreign Language (Landauer *et al*., 1998). In addition, in essay grading tasks, LSI performed comparably to human graders; the correlation between two professional graders on a set of 695 opinion essays was 0.86 and the correlation between LSI and the graders was 0.86 (Foltz *et al*., 1999). LSI methods have also been applied in the biological and medical sciences. For example, Landauer *et al*. (2004) recently demonstrated that LSI can be used to visualize themes and relationships from full-text articles in the scientific literature. They postulate that such methods may be useful for understanding the relations among nominal fields of science or help editors with assignment of appropriate reviewers and exploring the scientific impact of articles, and so on. Interestingly, matrix factorization methods similar to those used in LSI were recently shown to be effective in building whole genome bacterial phylogeny using correlated tetrapeptide motifs derived from over 134 000 proteins in 54 different genomes (Stuart and Berry, 2003).

We have developed a new method which utilizes LSI to identify conceptually related genes based on titles and abstracts in MEDLINE citations. A test version of the

[1] Recall refers to the ratio of relevant documents retrieved to the total number of (known) relevant documents.

web-based software environment, called Semantic Gene Organizer© (SGO), is presented here to demonstrate the utility of LSI for genomic studies. SGO is a unique text-mining tool because it allows the identification of relevant genes based on keyword queries as well as gene-abstract queries. Moreover, SGO identifies gene relationships even if the gene names do not co-occur in the same abstracts. We tested the precision [2] of the algorithm by querying for genes in a small and well-defined gene-document collection containing 50 genes. SGO identified genes in specific signaling pathways with high average precision and was not affected by the relative size (number of abstracts) of the gene-documents in the collection. Related genes were identified by rank order or by hierarchical clustering using the gene distance matrices. Remarkably, SGO clustered all of the genes in the document collection accurately and was able to infer relationships that have only recently been directly shown in the published literature. All together, these results clearly demonstrate that LSI-based algorithms such as SGO may provide a powerful tool to rapidly and accurately classify genes based on functional information in the biological literature abstracts.

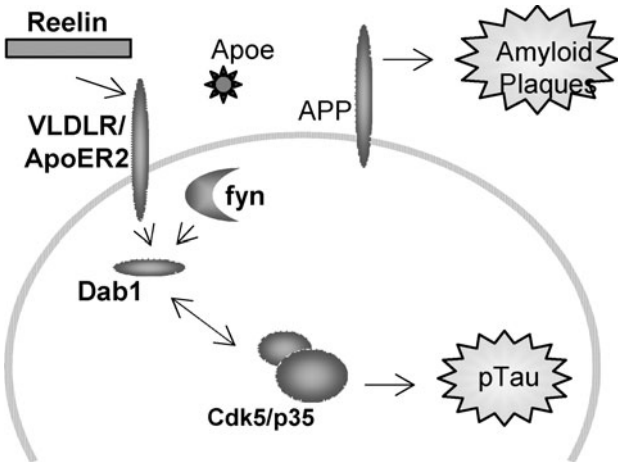

Fig. 1. Schematic summary of the Reelin signaling pathway. Reelin binds directly to lipoprotein receptors, VLDLR and ApoER2, and induces tyrosine phosphorylation of the cytoplasmic adapter protein Dab1 by fyn tyrosine kinase. Dab1 binds to a variety of proteins including amyloid precursor family proteins (APP) and is also phosphorylated on Ser residues by cyclin-dependent protein kinase 5.

SYSTEMS AND METHODS

Gene-document collection

A small 50-gene document collection was constructed by manually selecting genes in three broad categories: (1) development; (2) Alzheimer's disease (AD); (3) cancer biology (Supplementary Table 1). Each gene-document was generated by concatenation of all titles and abstracts of the MEDLINE citations cross-referenced in the mouse, rat and human LocusLink entries for each gene (June 7, 2003; Supplementary Table 2). For evaluation purposes, we chose genes that overlapped between cancer and development but not between cancer and AD. The number of citations in this collection ranged from 8 (APLP1 and CDK5R2) to 361 (TP53), with a median of 30.5 (Supplementary Table 2). As a 'gold standard' for evaluating the performance of SGO, we focused on the small but well-characterized Reelin signaling pathway. The median number of citations for the genes in the Reelin signaling pathway (RELN, VLDLR, LRP8, DAB1 and FYN) was 18.

Reelin is a large extracellular protein that controls neuronal positioning, formation of laminated structures (including the cerebellum) and synapse structure in the developing central nervous system (Rice and Curran, 2001; Tissir and Goffinet, 2003). Recent genetic and biochemical studies have identified several components of the Reelin signaling pathway (Fig. 1 and Table 1). Mice with disruptions in *reelin, disabled-1* or both the *very low-density lipoprotein receptor* (*VLDLR*) and the *apolipoprotein E receptor-2* (*ApoER2*) genes exhibit very similar phenotypes (D'Arcangelo *et al*., 1995, 1999; Howell *et al*., 1997; Sheldon *et al*., 1997; Hiesberger *et al*., 1999). Reelin binds directly to the lipoprotein receptors resulting in tyrosine phosphorylation of Disabled-1 (Dab1). Dab1 is a substrate for the Src family of non-receptor tyrosine kinases. Tyrosine phosphorylated Dab1 interacts with SH2 domains from a number of proteins, including the Src family of kinases (Howell *et al*., 1997). Recent reports indicate that the Src-related family member fyn tyrosine kinase mediates the effect of Reelin on Dab1 (Arnaud *et al*., 2003; Bock and Herz, 2003). Moreover, Dab1 is phosphorylated on serine residues by cyclin-dependent kinase 5 (cdk5) (Keshvara *et al*., 2002). Disruption of *cdk5* gene or its activator p35 results in abnormalities in brain structures very similar, but not identical, to those observed in *reeler* mice (Kwon and Tsai, 1998, 2000).

Accumulating evidence indicates that some components in the Reelin signaling pathway are associated with AD. Dab1 phosphotyrosine binding domain interacts with the NPxY motif in amyloid precursor protein which is closely associated with pathogenesis of AD (Trommsdorff *et al*., 1998; Homayouni *et al*., 1999; Howell *et al*., 1999). Apolipoprotein E blocks the interaction of Reelin with its receptors and is also considered a risk factor for late-onset AD (D'Arcangelo *et al*., 1999; Selkoe, 2001). Finally, cdk5 is deregulated in AD brains and is one of the major kinases that phosphorylates the microtubule-associated protein tau, which is hyperphosphorylated and is a major component of neurofibrillary tangles in AD brains (Lee and Tsai, 2003). Interestingly, tau is also hyperphosphorylated in reeler and Dab1-null mouse brains (Hiesberger *et al*., 1999; Brich *et al*., 2003).

[2] Precision refers to the ratio of relevant documents retrieved to the total number of documents retrieved (or ranked).

Table 1. Literature relationships between genes directly and indirectly associated with the Reelin signaling pathway

Gene	Experimental evidence[a]		PubMed[b]		LocusLink[c]		SGO rank by query[d]	
	Biochemical	Genetic	co-citation	Rank	Abst. overlap	Rank	Acc. no.	Keyword
Genes directly associated with Reelin signaling (five genes)								
RELN	+	+	n/a	1	n/a	1	**1**	**3**
DAB1	+	+	10	2	4	2	**2**	**1**
LRP8	+	+	0	0	2	3	**3**	**2**
VLDLR	+	+	3	3	1	4	**4**	**4**
FYN	+	+	0	0	0	0	30	47
Genes indirectly associated with Reelin signaling (seven genes)								
CDK5	+	+	0	0	0	0	**5**	**5**
APOE	+		1	4	0	0	**10**	44
SRC	+		0	0	0	0	34	43
MAPT		+	0	0	0	0	**9**	48
APP	+		1	5	0	0	**7**	22
APLP1	+		0	0	0	0	46	**8**
APLP2	+		0	0	0	0	36	**7**

n/a, not applicable.

[a]Direct experimental evidence linking genes to the Reelin signaling pathway is indicated by + (for details see Systems and Methods section).

[b]Gene ranking was calculated based on co-citation frequency of the gene symbols in MEDLINE entries using PubMed (July 7, 2003).

[c]The number of shared citations cross-referenced in the LocusLink entries for each gene (July 7, 2003 build). Citations of sequencing projects were deleted from the list (Okazaki *et al*., 2002; Kawai *et al*., 2001; Shibata *et al*., 2000; Carninci *et al*., 2000; Carninci and Hayashizaki, 1999).

[d]Gene ranking by SGO after Reelin accession number or keyword queries using 50 factors. The top 12 ranked genes are in boldface.

Text representation

The text representation was performed using the object oriented (C++ and Java) software environment called General Text Parser (Giles *et al*., 2003). First, the gene-documents were parsed into keywords (or tokens), and all punctuation (including hyphens) and capitalization were ignored. In addition, articles and other common, non-distinguishing words were discarded using the *stoplist* from Cornell's SMART project repository (ftp://ftp.cs.cornell.edu/pub/smart/english.stop). Next, a term-by-gene matrix was created where the entries of the matrix were the weighted frequencies, a non-negative value used to describe the correlation between that term and the corresponding document. In general, each weight is the product of a local and global component described below. We used a log-entropy weighting scheme for SGO as described previously (Berry and Browne, 1999). The local component l_{ij} and the global component g_i can be computed as

$$l_{ij} = \log_2(1 + f_{ij})$$

$$g_i = 1 + \left(\frac{\sum_j [p_{ij} \log_2(p_{ij})]}{\log_2 n} \right),$$

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}},$$

where f_{ij} is the frequency of the i-th term in the j-th gene-document, p_{ij} is the probability of the i-th term occurring in the j-th gene-document and n is the number of documents in the collection (Berry and Browne, 1999). Several alternative local-global weighting schemes exist which basically award less weight to terms that occur in many gene-documents. Entropy weighting is based on information-theoretic concepts and takes into account the distribution of terms over gene-documents (Dumais, 1991; Baeza-Yates and Ribeiro-Neto, 1999). The log-entropy weighting pair has performed best in most LSI-based retrieval experiments reported (Berry and Browne, 1999).

The weighted frequency for each token is then computed by multiplying its local component by its global component. That is, the term-by-gene document matrix is defined as

$$M = [m_{ij}],$$

$$m_{ij} = l_{ij} * g_{ij}.$$

Once the $m \times n$ term-by-gene document matrix, M, has been created, a truncated SVD of that matrix is performed to create three factor matrices

$$M = U \Sigma V^{\mathrm{T}},$$

where U is the $m \times r$ matrix of eigenvectors of MM^{T}, V^{T} is the $r \times n$ matrix of eigenvectors of $M^{\mathrm{T}}M$ and Σ is the $r \times r$ diagonal matrix containing the r non-negative singular values of M (Golub and Loan, 1996). The size of these factor matrices is determined by r, the rank of the matrix M. By using only the first s columns of the three component submatrices, we can compute M_s, a rank-s approximation to M. In this case, s is

considerably smaller than the rank r. Document-to-document similarity is then computed as

$$M_s^{\mathrm{T}} M_s = (V_s \Sigma_s)\,(V_s \Sigma_s)^{\mathrm{T}}$$

and can be derived from the original formula for the rank-s approximation to M (Berry, 1992). Queries can be treated as *pseudo*-documents and can be computed as

$$q = q_0^{\mathrm{T}} U_s \Sigma_s^{-1},$$

where q_0 is a query vector of associated global term weights, constructed from the user's original input, and the s subscript denotes the first s columns of the corresponding matrix factor.

A given query vector q can be easily compared with all the gene-document vectors of the form $d_j = \Sigma_s V_s^{\mathrm{T}} e_j$, where e_j is the compatible vector of all zeros except the value 1 in position j. Relevance to the query is determined by a ranking of a similarity score, such as the cosine. To be more specific, the score of a gene-document d_j with respect to a query q is defined by the cosine of the angle between the corresponding vectors in the vector space (LSI) model. As discussed by Berry and Brown (1999), the similarity scores were computed as

$$\cos\theta_j = \frac{d_j^{\mathrm{T}}(U_s^{\mathrm{T}} q)}{||d_j||_2 ||q||_2}, \quad j = 1, \ldots, n,$$

and then ranked so that the gene-document vectors having the higher cosine values with the query vector are deemed more relevant to the user's query (Berry *et al*., 1999).

For purposes of performance comparisons, the standard vector space IR model would not exploit the factorization of the term-by-gene matrix M. In this case, the columns of the matrix would be considered as vector representations of the gene-documents and each would have m components (compared to only r components for the LSI model where r is substantially smaller than the number of terms m). User queries would be converted to $m \times 1$ vectors using the global term frequencies for all non-zero components (in much the same manner that columns of M are constructed). Such query vectors of the simpler form q_0 above (for the LSI model) would then be compared with document vectors $d_j = M_j$, where M_j is the j-th column of the matrix M. For larger gene-document collections (scaling toward all of MEDLINE), the standard vector space model will incur significant storage and computational overheads associated with document vectors having on the order of 100 000 or more components (depending on the number of terms parsed).

Calculation of average precision

Search results can be represented in either graphical or tabular formats. One possible graphical format is that of interpolated precision-recall plots (Baeza-Yates and Ribeiro-Neto, 1999). Specifically, the interpolated precision values at 11 standard recall points (0.0, 0.1, 0.2, ..., 1.0) are plotted. These interpolated precision values are based on the pseudo-precision $\Phi(x)$ defined by $\Phi(x) = \max P(i)$, where $x \leq r_i/r_n$ for $i = 1, 2, \ldots, n$ (Berry and Browne, 1999). Here, r_i denotes the number of relevant documents up to and including position i in the ordered (or returned) list of gene-documents, and $P(i)$ is the precision at the i-th gene-document. In one sense, $P(i)$ is the proportion of gene-documents up to and including position i that are relevant to the given query. The 11-point average precision values (P_{avg}) are calculated by taking an average (mean) of $\Phi(x)$ at the standard recall values for $n = 11$. That is,

$$P_{\mathrm{avg}} = (1/n) \sum_{i=0}^{n-1} \Phi[i/(n-1)].$$

P_{avg} values are considered to be concise representations of their corresponding interpolated precision-recall graphs. As the number of relevant gene-documents per query is rarely constant, it is necessary to average the values of $\Phi(x)$ at prescribed recall levels.

Tree construction

A pair-wise distance matrix was constructed for the 50 genes in the document collection using the distance measure $1 - \cos\theta$, where θ is the vector angle between gene-documents. The hierarchical tree was constructed by implementing PHYLIP version 3.5 (http://evolution.genetics.washington.edu/phylip.html) using the Fitch–Margoliash method, which optimizes the branch lengths by a least squares criterion (Fitch and Margoliash, 1967). The tree was visualized by implementing the Java applet ATV ver. 1.92 (http://www.genetics.wustl.edu/eddy/atv/) (Zmasek and Eddy, 2001).

ALGORITHM

The workflow for Semantic Gene Organizer is shown in Figure 2. First, a gene-document was constructed by concatenating all titles and abstracts for the MEDLINE citations that were cross-referenced in the human, mouse and rat LocusLink entries for a specific gene. The gene-documents were assembled and parsed into a dictionary of terms (tokens) and weighted frequencies that are required for the term-by-gene document (sparse) matrix. In effect, each gene-document is viewed as a bag of words upon which operations can be performed. There are a number of different word weighting schemes that can be used in vector space modeling (Baeza-Yates and Ribeiro-Neto, 1999). The aim of any scheme is to measure similarity within a document while at the same time measuring the dissimilarity of a gene-document from the other gene-documents. In SGO, we use log-entropy weighing scheme to decrease the weight of high-frequency words while giving distinguishing words higher weight (Berry and Browne, 1999). In addition, restrictions on the global and/or document term frequencies can be imposed to control the size of the dictionary. For example, all words that occurred

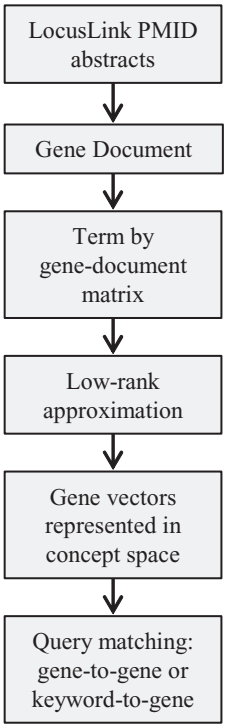

Fig. 2. Work flow for Semantic Gene Organizer. Gene documents are prepared by concatenation of titles and abstracts from citations cross-reference in LocusLink entries. The gene documents are then used to build a term by gene-document matrix upon which SVD is performed to create a low-rank approximation matrix. The gene vectors can be represented with as few as two factors. Queries are represented as pseudo-vectors in this space. Gene-to-gene or gene-to-keyword relationships are derived from the angle between query and gene vectors.

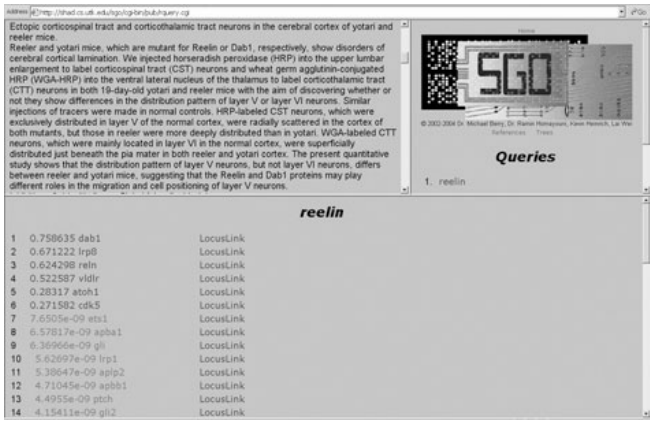

Fig. 3. Screen shot of SGO. This figure illustrates the rank order of genes (lower panel) that are relevant to the query (reelin, upper right panel). The titles and abstracts that comprise the gene documents are shown in the upper left panel. The gene documents that contain the keyword query are in boldface whereas the gene documents that do not contain the keyword query are in gray. SGO is accessible on the Web at http://shad.cs.utk.edu/sgo.

less than two times in one gene-document and in less than two gene-documents were not included in the term by gene-document matrix. This restriction reduced the number of terms in the matrix from 19 780 to 8754.

Term and document vectors for the LSI model deployed by SGO are generated by truncating the SVD of the term-by-gene document matrix to s factors (i.e. only s columns of the orthogonal matrices U and V are used). Thus, LSI produces a rank-reduced space in which to compare two gene-documents at different conceptual levels. In practice, the maximum number of factors is limited by the number of documents in the collection. Fewer factors may be used for broad (more conceptual) comparisons, whereas a larger number of factors may be used for specific (more literal) comparisons. Other studies have demonstrated that for large documents collections the optimal number of factors is approximately 300 (Landauer *et al*., 2004).

Query vectors in SGO are generated by the user and may be in two types: (1) keyword query, which may consist of any number of manually selected terms and (2) accession number query, which consists of all textual information in the abstract document for the given gene. A pseudo-vector is created by using the terms in the keyword query or accession number query and is compared to all other documents vectors in the collection. Since an accession number query vector consists of all the textual information in the document, it will theoretically identify more accurate relationships than a vector consisting of a few keywords. We note that all textual information (words or tokens) contained in the gene abstracts (and not placed in the stoplist) can define valid dictionary entries and hence query terms. Relevance to the query term is determined by ranking a similarity score, defined by the cosine of the vector angles between the query and the gene-documents in the collection. SGO produces a ranked list of genes based on the angle of the gene-abstract documents and the query vectors. For example, Figure 3 shows a screen shot of SGO displaying the ordered list of genes after a Reelin keyword query.

RESULTS

We first evaluated the performance of SGO by comparing the ranking of genes after either an accession number or keyword query on the Reelin signaling pathway (Table 1). The molecular pathway of Reelin signaling has been previously established using both biochemical and genetic approaches (see Gene Document Collection under Systems and Methods section). There are five primary genes that are directly involved in the Reelin signaling pathway: RELN, VLDLR, LRP8, DAB1 and FYN. Some components in the Reelin signaling pathway also interact with other genes, which are refered to as secondary Reelin genes because of their indirect association with Reelin. Table 1 shows a comparison of

the SGO rankings for genes associated with Reelin signaling to the rankings obtained by tabulating the number of shared citations in the LocusLink gene entries or the co-citation frequencies of the gene symbols in MEDLINE abstracts. We found four out of the top five ranked genes (80% recall) identified by SGO were directly associated with Reelin signaling. This result was comparable to that obtained by ranking the overlapping abstracts in LocusLink, but were better than the results obtained by ranking the co-citation frequencies of gene symbols in MEDLINE abstracts (60% recall). Another way to evaluate the performance of SGO is by calculating the precision by which gene-documents are retrieved. As shown in Figure 4A, to obtain a recall value of 100%, precision dropped to 11%. This low precision is due to the low ranking of fyn (47 out of 50). It is noteworthy that a direct association between Reelin and fyn kinase, which is largely associated with cancer biology, has only recently been demonstrated (Arnaud *et al*., 2003; Bock and Herz, 2003). However, these citations were not included in LocusLink at the time of collection. Consequently, fyn ranked 30 and 47 with Reelin accession number and keyword queries, respectively (Table 1).

We examined the performance of SGO on the 50-gene test collection by querying with two additional types of keywords. One set of keywords included GO molecular function classification terms and the other set of keywords included human disease names. Relevant genes for each query were determined using the human GO classification index or the information in RefSeq summaries in human LocusLink entries. As shown in Table 2, different keyword queries had different number of relevant genes. We used a single value summary, called Average Precision, to compare the performance of SGO across different queries as well as for different factorization results (see below). To evaluate the significance of the results, we calculated an expected AP for each relevant set by calculating the AP for 1000 randomly generated rankings of the 50 genes in the document collection. We found that in all cases AP (using 50 factors) was above that which was expected by chance. Furthermore, in preliminary experiments using a much larger collection (>20 000 gene-documents), we found that SGO identified the five primary Reelin genes with an AP of 0.474 compared to an expected AP of 0.0008 (data not shown).

Influence of factorization on retrieval performance

An important distinction between LSI and other vector space models is that LSI uses matrix factorization (SVD) to produce a concept space in which gene-document and query vectors are projected. By changing the number of factors used in the reduced rank matrix, one essentially changes the concepts by which documents are compared. To explore this feature of LSI, we examined the performance of SGO for different keyword queries using 5, 25 and 50 factor space (Fig. 4 and Table 2). For Reelin keyword queries, we found no significant difference in AP (0.84) when using 25 or 50 factors. However, AP dropped to 0.61 when only 5 factors were used. On the other hand, for GO classification and human disease keyword queries, we found that 50 factors produced higher AP in all cases. For example, average precision for neurogenesis and axon guidance improved from 0.27 to 0.37 and 0.10 to 1.0, respectively.

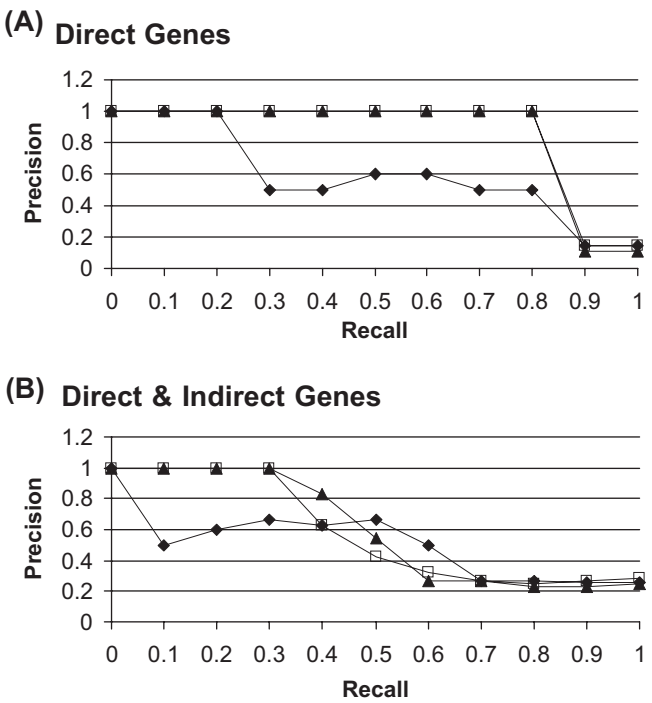

Fig. 4. SGO performance on Reelin keyword query using different factors. (**A**) Direct genes. (**B**) Direct and indirect genes. Precision at 11 recall values (0–1.0) for the Reelin keyword query having 5 directly and 7 indirectly related genes using either 5 (diamonds), 25 (open squares) or 50 (triangles) factors of the reduced rank matrix.

Influence of abstract representation on retrieval performance

Scientific literature is naturally biased toward the well-characterized genes, usually those that are linked to a human disease. For example, there are over 28 000 citations in MEDLINE that refer to the tumor suppressor gene TP53. In contrast, the vast majority of genes have relatively few citations in MEDLINE. In this study, we used only MEDLINE citations assigned to each gene by LocusLink professional curators. Although very limited in number, these citations presumably contain more relevant and functionally important abstracts. As a result, the total number of citations for TP53 used in this collection was 361 while the median number of abstracts for the five primary Reelin genes was 18. Importantly, despite the disparity in the number of abstracts

Table 2. SGO performance for different keyword queries

Query	Relevant genes[a]	SGO performance Rank dis.[b]	AP-25[c]	AP-50	Exp AP[d]
GO classifications					
Apoptosis	7	1–19	0.34	0.45	0.23
Axon guidance	1	1	0.10	1.00	0.09
Cell fate	2	1–10	0.59	0.64	0.11
Kinase	8	1–15	0.72	0.80	0.25
Neurogenesis	10	1–48	0.27	0.37	0.30
Patterning	5	1–10	0.71	0.68	0.19
Transcription	10	1–24	0.40	0.75	0.30
Tyrosine kinase	3	5–10	0.19	0.30	0.14
Human disease					
Alzheimer's Disease	8	3–15	0.72	0.70	0.25
Breast cancer	3	1–5	0.74	0.85	0.14
Lissencephaly	1	1	1.00	1.00	0.09

[a]The number of relevant genes in the 50-gene collection was determined using Gene Ontology classifications or by information in the RefSeq summary in LocusLink (human diseases).
[b]Rank distribution of relevant genes by SGO.
[c]Calculated Average Precision using 25 factors (AP-25) or 50 factors (AP-50) in the term by gene-document matrix.
[d]The expected AP for the number of relevant genes was calculated from 1000 random samplings of the 50 gene-document collection.

assigned to the Reelin signaling genes and other genes in the collection, SGO identified the Reelin signaling genes with high AP. This suggested that LSI performance was not influenced by the biases in the literature for well-studied genes. To explore this issue further, we tested the retrieval performance of SGO when abstract representation for the primary Reelin genes was systematically reduced further while the abstract representation of the remaining genes were unchanged. Three randomly generated gene-documents were created for each gene in the Reelin signaling pathway consisting of 75, 50, 25 and 5% of the original number of abstracts in the collection. We found that the average precision of SGO with a Reelin keyword query was not significantly affected when the number of abstracts for Reelin signaling pathway was reduced by 50% (Fig. 5).

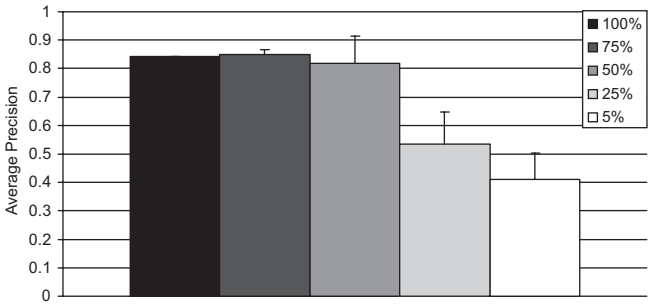

Fig. 5. Effect of abstract representation on SGO performance. Average precision is shown for reelin keyword query of the 50-gene document collection with decreasing representation (100 to 5%) of abstracts for the five primary Reelin genes (RELN, DAB1, VLDLR, LRP8 and FYN). Each bar represents the average AP $\pm$ standard deviation for three randomly generated collections.

Identification of genes indirectly associated with queries

A unique feature of the LSI model is the ability to identify latent (indirect) relationships between the query and documents in the collection. We examined whether SGO could identify genes that were indirectly associated with Reelin, e.g. through association with one of the downstream components. Based on the published literature, we manually assigned seven genes (SRC, CDK5, APOE, MAPT, APP, APLP1 and APLP2) as indirectly associated with Reelin. SGO identified three and two out of seven secondary Reelin genes by the accession number and keyword queries, respectively (Table 1). In contrast, none of the indirectly associated Reelin genes was identified by LocusLink co-citation method and only two indirectly associated genes were identified by gene symbol co-citation method. Interestingly, both APLP1 and APLP2 were identified by SGO even though they were not cited together with Reelin (Table 1). The association between Reelin and APLP1/2 is mediated by Dab1, which was shown to bind directly to the APP family proteins (Homayouni *et al.*, 1999).

We also examined the effect of factorization on the identification of indirectly associated genes. We found no significant difference in the performance of SGO on the identification of indirectly associated genes using either 25 or 50 factors (AP $=$ 0.59 and 0.60, respectively) and a slight decrease in AP when using five factor space (AP $=$ 0.53). The significance of factorization on the identification of indirectly associated genes was more apparent for other keyword queries. For example,

using 25 factors for the keyword query lissencephaly, SGO identified RELN, DAB1, LRP8, VLDLR and ATOH1 in rank order (data not shown). Interestingly, only RELN gene-document contains the word lissencephaly. Therefore, the association of lissencephaly with DAB1, LRP8 and VLDLR is due to the similarity in the text usage patterns in the gene-documents. When 50 factors were used, SGO identified RELN, CDK5, LRP8, APBA1 and DAB1 in rank order. Interestingly, CDK5 gene-document also contains the keyword lissencephaly, indicating that by using more factors the similarities become more literal. Taken together, these results indicate that SGO is useful in extraction of latent relationships compared to currently available co-citation strategies.

Hierarchical clustering

In the previous section we evaluated SGO on pairwise gene-to-gene and gene-to-keyword queries. A more comprehensive way to evaluate SGO would be to examine the relationship of all genes in the document collection to one another. To address this issue, a pairwise distance (derived from the vector angles) for each gene-document was calculated and used to generate a hierarchical tree using the Fitch–Margoliash least squares optimization method for constructing phylogenetic trees (Fitch and Margoliash, 1967). We found that the overall topology of the tree was consistent with our original classification of the genes in that the genes were grouped into two major categories: development/cancer and development/Alzheimer's disease (Fig. 6 and Supplementary Table 1).

Each of these major branches in the tree was subdivided into smaller and more functionally cohesive clusters. For example, the APP, PSEN1 and PSEN2 clusters contain all of the genes found to be mutated in early onset AD. Also, we found that all gene family members (such as APLP1-2, BRCA1-2, PAX2-3 and GLI1-3) were appropriately clustered together. Consistent with the earlier results, RELN, DAB1, LRP8 and VLDLR clustered together. Interestingly, although FYN did not rank highly by Reelin queries (above), hierarchical clustering placed FYN in close proximity of other developmentally important genes and not with the cancer-related genes. Lastly, we were surprised to see clustering of the oncogene SHC1 with Alzheimer genes in spite of only one co-citation among 130 APP abstracts. Interestingly, a recent report (not included in this document collection) demonstrated that SHC1 directly binds with APP and modulates its proteolysis, a process critical in the pathogenesis of AD (Zambrano *et al*., 2004). All together, these results indicate that SGO can provide a robust method for high-throughput method to explore the function of genes based on implicit and explicit information in the biomedical literature.

DISCUSSION

The first step in the interpretation of high-throughput genomic data involves the functional classification of the genes that are co-regulated. Currently, gene classification relies upon the information in manually curated indices and databases. We have developed an automated method using LSI to classify genes based on conceptual modeling of the biological information in the titles and abstracts of the MEDLINE citations. Using a small dataset of 50 gene documents, we demonstrated here that LSI achieves high average precision in identifying gene-to-gene relationships.

The use of concept-based information retrieval methods is particularly appealing for analysis of genomic data which often identify novel associations between genes that are not well-documented in the literature. In contrast to keyword matching (Boolean) methods, vector space methods such as LSI identify gene relationships based on word usage patterns in abstracts even if there are no explicit associations in the literature. In this study, we demonstrated that SGO identified indirectly associated genes with high average precision. For instance, SGO identified a relationship between Reelin and APLP1 although they are not cited together. Also, by hierarchical clustering, SGO grouped FYN with developmental genes such as Reelin and Cdk5 even though a direct association between these pathways was only recently established and was not included in the document collection. Similarly, SHC1 was grouped with AD genes with no direct evidence in the document collection. These results demonstrate the utility of LSI approaches in extracting latent relationships from the literature.

Another unique feature of LSI-based approaches for text mining is the ability to rank order genes based on keyword queries. We demonstrated that SGO identified genes belonging to different GO classification terms or human disease names with high average precision. This feature may be useful in expanding GO gene classifications in an automatic fashion. For example, we found that a keyword query using the GO term 'tyrosine kinase' identified FYN, ABL1 and SRC (data not shown) in addition to the three genes (KIT, EGFR and ERBB2) that were assigned manually by GO curators. Similarly, SGO may be used to automatically assign genes to human diseases for collections such as OMIM. Moreover, because of the ability of LSI to identify latent relationships, SGO may be used as an exploratory tool to identify likely gene candidates for diseases in linkage studies.

A fundamental problem in information retrieval, particularly in context of the biomedical literature, is that document size and representation are not normally distributed. Often, the document collection is biased by a few genes which contain an overwhelming majority of the text material. For example, in the 50 gene-document collection nine genes (TP53, TGFB1, EGFR, BRCA1, APP, APOE, ERBB2 and PSEN1) accounted for more than 50% of all abstracts, whereas the gene-document for Reelin contained only 29 abstracts. Despite this bias in the document collection, SGO identified genes in the Reelin signaling pathway with high average precision (Figs 4 and 5) and accurately grouped genes in the entire document collection (Fig. 6). These results demonstrate the power

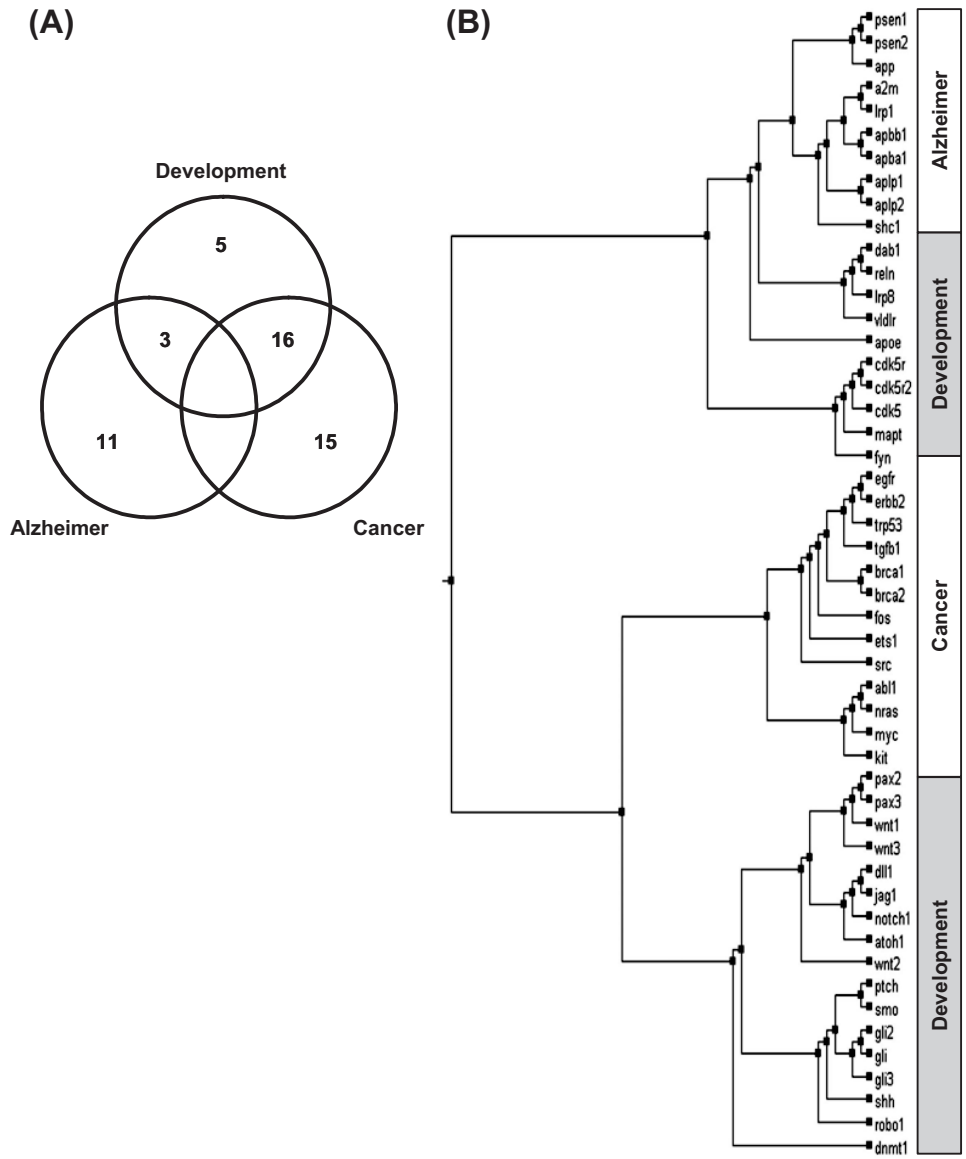

Fig. 6. Gene neighbors deduced from the literature. (**A**) Classification of genes based on manual examination of the biomedical literature. Some developmental genes were associated with AD and others with cancer (for details see Supplementary Table 2). (**B**) Automated classification of the genes in the same collection by SGO. A hierarchical tree was constructed using the PHYLIP algorithm and a distance matrix derived from the cosine of the vector angles between gene-documents.

of LSI-based methods to accurately identify relationships despite biases in the representation of the textual information in the dataset. However, it is important to note that the performance of SGO is entirely dependent on the accuracy of the abstracts assigned to the genes in the document collection. For instance, SGO did not find a recently described relationship between Reelin and Fyn because the citations were not assigned in LocusLink.

In this study, we tested the performance of SGO on a small 50 gene-document collection. However, several features of SGO make it a very amenable tool for genome-wide literature mining. First, it is scalable (Chen *et al*., 2001). Text collections comprising over 150 000 documents and 125 000 terms can be easily parsed for LSI models based on nearly 300 factors. This would only require ∼330 MB for storage of the term and document vectors and ∼300 MB of RAM to construct and factor the term-by-gene document matrix (assuming 0.1% density). Second, it is fast. Query-matching amounts to simple cosine calculations and rankings which can be computed on the order of a few seconds (or reloaded from memory in the case of

common or routing-type queries). Third, it does not require dynamic interaction with MEDLINE through PubMed web interface because the gene-documents and the term-by-gene document matrix is constructed in the front end. In addition, users can index their own gene-document collections from MEDLINE, full-text documents or other sources.

In summary, we have shown proof-of-concept that LSI-based methods may provide a powerful new tool for functional analysis of discovery-based genomic experiments. Importantly, LSI may be used in a number of other applications in the post-genomic era. Moreover, because of the ability for LSI to extract latent relationships from the biomedical literature, it may be useful as a hypothesis-generating tool to identify potential new relationships that can be explored further experimentally.

ACKNOWLEDGEMENTS

We thank the reviewers for their many helpful comments and suggestions. We also thank Lijing Xu, Mi Zhou and Yan Cui for technical help. This work was supported by UT Center for Genomics and Bioinformatics (R.H.), UT Center for Neurobiology of Brain Diseases (R.H.), UT Center for Information Technology Research (K.H. and M.W.B.) and Computational Sciences Initiative at the University of Tennessee and Oak Ridge National Laboratory (K.H.).

REFERENCES

Arnaud,L., Ballif,B.A., Forster,E. and Cooper,J.A. (2003) Fyn tyrosine kinase is a critical regulator of disabled-1 during brain development. *Curr. Biol.*, **13**, 9–17.

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

Baeza-Yates,R. and Ribeiro-Neto,B. (1999) *Modern Information Retrival*. ACM Press, New York.

Becker,K.G., Hosack,D.A., Dennis,G.,Jr, Lempicki,R.A., Bright,T.J., Cheadle,C. and Engel,J. (2003) PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, **4**, 61.

Berry,M.W. (1992) Large scale singular value computations. *Int. J. Supercomputer App.*, **6**, 13–49.

Berry,M.W. and Browne,M. (1999) *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia.

Berry,M.W., Drmac,Z. and Jessup,E. (1999) Matrices, vector spaces, and information retrieval. *SIAM Rev.*, **41**, 335–362.

Berry,M.W., Dumais,S. and O'Brien,G. (1995) Using linear algebra for intelligent information retrieval. *SIAM Rev.*, **37**, 573–595.

Bock,H.H. and Herz,J. (2003) Reelin activates SRC family tyrosine kinases in neurons. *Curr. Biol.*, **13**, 18–26.

Brich,J., Shie,F.S., Howell,B.W., Li,R., Tus,K., Wakeland,E.K., Jin,L.W., Mumby,M., Churchill,G., Herz,J. and Cooper,J.A. (2003) Genetic modulation of tau phosphorylation in the mouse. *J. Neurosci.*, **23**, 187–192.

Chen,C., Stoffel,N., Post,M., Basu,C., Bassu,D. and Behrens,C. (2001) In Aberer,K. and Liu,L. (eds), Telcordia LSI engine: implementation and scalability issues. *Proceedings of the 11th International Workshop on Research Issues in Data Engineering*, IEEE Computer Society, Heidelberg, Germany, 51–58.

D'Arcangelo,G., Homayouni,R., Keshvara,L., Rice,D.S., Sheldon,M. and Curran,T. (1999) Reelin is a ligand for lipoprotein receptors. *Neuron*, **24**, 471–479.

D'Arcangelo,G., Miao,G.G., Chen,S.C., Soares,H.D., Morgan,J.I. and Curran,T. (1995) A protein related to extracellular matrix proteins deleted in the mouse mutant reeler. *Nature*, **374**, 719–723.

Deerwester,S.C., Dumais,S.T., Furnas,G.W., Harshman,R.A., Landauer,T.K., Lochbaum,K.E. and Streeter,L.A. (1988) *Computer Information Retrieval Using Latent Semantic Structure*. Bell Communications Research, Inc., USA.

Deerwester,S.C., Dumais,S.T., Landauer,T.K., Furnas,G.W. and Harshman,R.A. (1990) Indexing by latent semantic analysis. *J. Inform. Sci.*, **41**, 391–407.

Doniger,S.W., Salomonis,N., Dahlquist,K.D., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7.

Dumais,S. (1991) Improving the retrieval of information from external sources. *Behavior Res. Meth. Instr. Comp.*, **23**, 229–236.

Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279–284.

Foltz,P.W., Laham,D. and Landauer,T.K. (1999) Automated essay scoring: applications to educational technology. *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 939–944.

Funk,M.E. and Reid,C.A. (1983) Indexing consistency in MEDLINE. *Bull. Med. Libr. Assoc.*, **71**, 176–183.

Giles,J.T., Wo,L. and Berry,M.W. (2003) GTP (General Text Parser) software for Tex mining. In Bozdogan,H. (ed.), *Statistical Data Mining and Knowledge Discover*. CRC Press, Boca Raton, FL.

Glenisson,P., Antal,P., Mathys,J., Moreau,Y. and De Moor,B. (2003) Evaluation of the vector space representation in text-based gene clustering. *Pac. Symp. Biocomput.*, 391–402.

Golub,G. and Loan,C.V. (1996) *Matrix Computations*. Johns-Hopkins, Baltimore.

Hiesberger,T., Trommsdorff,M., Howell,B.W., Goffinet,A., Mumby,M.C., Cooper,J.A. and Herz,J. (1999) Direct binding of Reelin to VLDL receptor and ApoE receptor 2 induces tyrosine phosphorylation of disabled-1 and modulates tau phosphorylation. *Neuron*, **24**, 481–489.

Homayouni,R., Rice,D.S., Sheldon,M. and Curran,T. (1999) Disabled-1 binds to the cytoplasmic domain of amyloid precursor-like protein 1. *J. Neurosci.*, **19**, 7507–7515.

Hosack,D.A., Dennis,G., Jr, Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Howell,B.W., Gertler,F.B. and Cooper,J.A. (1997) Mouse disabled (mDab1): a Src binding protein implicated in neuronal development. *EMBO J.*, **16**, 121–132.

Howell,B.W., Lanier,L.M., Frank,R., Gertler,F.B. and Cooper,J.A. (1999) The disabled 1 phosphotyrosine-binding domain binds to the internalization signals of transmembrane glycoproteins and to phospholipids. *Mol. Cell. Biol.*, **19**, 5179–5188.

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

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Keshvara,L., Magdaleno,S., Benhayon,D. and Curran,T. (2002) Cyclin-dependent kinase 5 phosphorylates disabled 1 independently of Reelin signaling. *J. Neurosci.*, **22**, 4869–4877.

Kwon,Y.T. and Tsai,L.H. (1998) A novel disruption of cortical development in p35(−/−) mice distinct from reeler. *J. Comput. Neurol.*, **395**, 510–522.

Kwon,Y.T. and Tsai,L.H. (2000) The role of the p35/cdk5 kinase in cortical development. *Results Probl. Cell Differ.*, **30**, 241–253.

Landauer,T.K., Laham,D. and Derr,M. (2004) From paragraph to graph: latent semantic analysis for information visualization. *Proc. Natl Acad. Sci., USA*, **101**, 5214–5219.

Landauer,T.K., Laham,D. and Foltz,P.W. (1998) Learning human-like knowledge by singular value decomposition: a progress report. In Jordan,M.I., Kearns,M.J. and Solla,S.A. (eds), *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, vol. 10, pp. 45–51.

Lee,M.S. and Tsai,L.H. (2003) Cdk5: one of the links between senile plaques and neurofibrillary tangles? *J. Alzheimers Dis.*, **5**, 127–137.

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

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Rice,D.S. and Curran,T. (2001) Role of the reelin signaling pathway in central nervous system development. *Annu. Rev. Neurosci.*, **24**, 1005–1039.

Selkoe,D.J. (2001) Alzheimer's disease: genes, proteins, and therapy. *Physiol. Rev.*, **81**, 741–766.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.

Sheldon,M., Rice,D.S., D'Arcangelo,G., Yoneshima,H., Nakajima,K., Mikoshiba,K., Howell,B.W., Cooper,J.A., Goldowitz,D. and Curran,T. (1997) Scrambler and yotari disrupt the disabled gene and produce a reeler-like phenotype in mice. *Nature*, **389**, 730–733.

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

Stuart,G.W. and Berry,M.W. (2003) A comprehensive whole genome bacterial phylogeny using correlated peptide motifs defined in a high dimensional vector space. *J. Bioinformatics Comput. Biol.*, **1**, 475–493.

Tissir,F. and Goffinet,A.M. (2003) Reelin and brain development. *Nat. Rev. Neurosci.*, **4**, 496–505.

Trommsdorff,M., Borg,J.P., Margolis,B. and Herz,J. (1998) Interaction of cytosolic adaptor proteins with neuronal apolipoprotein E receptors and the amyloid precursor protein. *J. Biol. Chem.*, **273**, 33556–33560.

Wilkinson,D.M. and Huberman,B.A. (2004) A method for finding communities of related genes. *Proc. Natl Acad. Sci., USA*, **101**, 5241–5248.

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

Zambrano,N., Gianni,D., Bruni,P., Passaro,F., Telese,F. and Russo,T. (2004) Fe65 is not involved in the platelet-derived growth factor-induced processing of Alzheimer's amyloid precursor protein, which activates its caspase-directed cleavage. *J. Biol. Chem.*, **279**, 16161–16169.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.