BMC Bioinformatics

BioMed Central

Proceedings

Open Access

Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations

Hyunsoo Kim, Haesun Park* and Barry L Drake

Address: College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

Email: Hyunsoo Kim - hskim@cc.gatech.edu; Haesun Park* - hpark@cc.gatech.edu; Barry L Drake - bldrake@cc.gatech.edu

* Corresponding author

from First International Workshop on Text Mining in Bioinformatics (TMBio) 2006
Arlington, VA, USA. 10 November 2006

Published: 27 November 2007

BMC Bioinformatics 2007, **8**(Suppl 9):S6 doi:10.1186/1471-2105-8-S9-S6

This article is available from: http://www.biomedcentral.com/1471-2105/8/S9/S6

Abstract

Background: The construction of literature-based networks of gene-gene interactions is one of the most important applications of text mining in bioinformatics. Extracting potential gene relationships from the biomedical literature may be helpful in building biological hypotheses that can be explored further experimentally. Recently, latent semantic indexing based on the singular value decomposition (LSI/SVD) has been applied to gene retrieval. However, the determination of the number of factors *k* used in the reduced rank matrix is still an open problem.

Results: In this paper, we introduce a way to incorporate a priori knowledge of gene relationships into LSI/SVD to determine the number of factors. We also explore the utility of the non-negative matrix factorization (NMF) to extract unrecognized gene relationships from the biomedical literature by taking advantage of known gene relationships. A gene retrieval method based on NMF (GR/NMF) showed comparable performance with LSI/SVD.

Conclusion: Using known gene relationships of a given gene, we can determine the number of factors used in the reduced rank matrix and retrieve unrecognized genes related with the given gene by LSI/SVD or GR/NMF.

Background

Latent semantic indexing based on the singular value decomposition (LSI/SVD) [1,2] uses the truncated singular value decomposition as a low-rank approximation of a term-by-document matrix. Recently, LSI/SVD has been applied to gene clustering so as to retrieve genes directly and indirectly associated with the Reelin signaling pathway [3]. This approach may provide us with a powerful tool for the functional relationship analysis of discovery-based genomic experiments. However, this work did not utilize a priori knowledge of gene-gene relationships that are generally available. Moreover, the determination of the number of factors *k* used in the reduced rank matrix is still an open problem even though it is an important parameter that determines a concept space in which gene-documents are projected. In this paper, we suggest a method to estimate the reduced rank *k* in LSI/SVD by taking advantage of known gene relationships. In addition, we propose a gene retrieval method based on the non-negative matrix factorization (GR/NMF), which is a new framework for extracting unrecognized gene relationships from the biomedical literature.

Given a non-negative matrix A of size $m \times n$ and a desired reduced dimension $k < \min\{m, n\}$, NMF finds two non-negative matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ so that $A \approx WH$. A solution to the NMF problem can be obtained by solving the following optimization problem:

$$\min_{W,H} f(W,H) \equiv \frac{1}{2}\left\|A - WH\right\|_F^2, \quad s.t. \quad W,H \geq 0, \tag{1}$$

where $W \in \mathbb{R}^{m \times k}$ is a basis matrix, $H \in \mathbb{R}^{k \times n}$ is a coefficient matrix, $||\cdot||_F$ is the Frobenius norm, and $W, H \geq 0$ means that all elements of W and H are non-negative. Due to $k < m$, dimension reduction is achieved and the lower-dimensional representation is given by H. The NMF gives a direct interpretation due to non-subtractive combinations of non-negative basis vectors. In addition, some practical problems require non-negative basis vectors. For example, pixels in digital images, term frequencies in text mining, and chemical concentrations in bioinformatics are typically non-negative [4]. The NMF has been successfully applied to many problems including text data mining [5,6] and gene expression data analysis [7,8]. Non-negative dimension reduction is desirable for handling the massive quantity of high-dimensional data that require non-negative constraints. The determination of the reduced dimension k and the initialization of W and H are open problems. Some NMF algorithms [5,9,10] require that both W and H be initialized, while NMF based alternating non-negativity-constrained least squares that we describe in this paper only requires the initialization of H. We initialized a part of the matrix H by incorporating a known cluster structure and determined the reduced dimension k that can capture well the known gene relationships.

Results and discussion

For evaluation of our methods, we study two biological pathways: (1) Reelin signaling pathway and (2) Alzheimer's disease pathway. We try to extract unrecognized gene-gene relationships from the biomedical literature by taking advantage of known gene relationships. Table 1 shows 50 genes used for our experiments.

Reelin signaling pathway

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 [11,12]. Reelin binds directly to lipoprotein receptors, the very low-density lipoprotein receptor (VLDLR) and the apolipoprotein E receptor-2 (ApoER2), and induces tyrosine phosphorylation of the cytoplasmic adapter protein Disabled-1 (Dab1) by fyn tyrosine kinase. APOER2 is a gene alias name of LRP8. By using these knowledge, we chose five genes directly associated with Reelin signaling pathway, *i.e.* {RELN, DAB1, LRP8, VLDLR, FYN}.

We will examine if the following indirect gene relationships can be found by using the above knowledge. Dab1 is phosphorylated on serine residues by cyclin-dependent kinase 5 (Cdk5) [13]. The proteins encoded by CDK5R1 (p35) and CDK5R2 (p39) are neuron-specific activators of Cdk5. They associate with Cdk5 to form an active kinase. Apolipoprotein E (ApoE) is a small lipophilic plasma protein and a component of lipoproteins such as chylomicron remnants, very low density lipoprotein (VLDL), and high density lipoprotein (HDL). The ApoER2 is involved in cellular recognition and internalization of these lipoproteins. ApoE blocks the interaction of Reelin with its receptors. The Src related family member fyn tyrosine kinase mediates the effect of Reelin on Dab1 [14,15]. MAPT encodes the microtubule-associated protein tau. Cdk5 is one of the major kinases that phosphorylates tau [16]. MAPT gene mutations have been associated with several neurodegenerative disorders such as Alzheimer's disease, Pick's disease, frontotemporal dementia, corticobasal degeneration and progressive supranuclear palsy. The six genes indirectly associated with the Reelin signaling pathway are CDK5, CDK5R1, CDK5R2, APOE, SRC, and MAPT.

Alzheimer's disease pathway

We obtained the Alzheimer's disease pathway data from the KEGG pathway database [17]. From this pathway, we can obtain an overview of the general picture of the Alzheimer's disease pathway. Amyloid beta precursor protein

Table 1: The genes considered in the data set. The letters 'A', 'C' and 'D' in brackets show the relation with Alzheimer's disease, cancer, development, respectively.

A2M(A)	APBA1(A)	APBB1(A)	APLP1(A)	APLP2(A)	APOE(A)
APP(A)	LRP1(A)	MAPT(A)	PSEN1(A)	PSEN2(A)	ABL1(C)
BRCA1(C)	BRCA2(C)	DNMT1(C)	EGFR(C)	ERBB2(C)	ETS1 (C)
FOS (C)	FYN(C)	KIT(C)	MYC(C)	NRAS(C)	SHC1(C)
SRC(C)	TP53(C)	TGFB1(D)	ATOH1(D)	CDK5(D)	CDK5R1(D)
CDK5R2(D)	DAB1(D)	DLL1(D)	GLI(D)	GLI2(D)	GLI3(D)
JAG1(D)	LRP8(D)	NOTCH1(D)	PAX2(D)	PAX3(D)	PTCH(D)
RELN(D)	ROBO1(D)	SHH(D)	SMO(D)	VLDLR(D)	WNT1(D)
WNT2(D)	WNT3(D)				

(APP) encodes a cell surface receptor and transmembrane precursor protein that is cleaved by secretases to form a number of peptides. The pathway includes {APP, APBB1, LRP1, APOE, A2M, PSEN1, PSEN2, MAPT} among our 50 genes. These eight genes are known genes associated with the Alzheimer's disease pathway.

However, we cannot guarantee that the pathway contains all information regarding the Alzheimer's disease. We will determine whether we can find the following unrecognized knowledge from the above known knowledge. Amyloid beta precursor-like protein 1 (APLP1) affects the endocytosis of APP and makes more APP available for α-secretase cleavage [18]. Site-specific proteolysis of the amyloid-beta precursor protein (APP) by BACE 1 and γ-secretase, a central event in Alzheimer disease, releases a large secreted extracellular fragment (called APP(S)), peptides of 40–43 residues derived from extracellular and transmembrane sequences (Abeta), and a short intracellular fragment (APP intracellular domain) that may function as a transcriptional activator in a complex with the adaptor protein Fe65 and the nuclear protein Tip60. APP is closely related to APP-like protein (APLP) 1 and APLP2, and similar to APP, APLP1 and APLP2 are also cleaved by BACE 1 [19]. Amyloid beta precursor protein-binding, family A, member 1 (APBA1) stabilizes APP and inhibits production of proteolytic APP fragments including the Abeta peptide that is deposited in the brains of Alzheimer's disease patients. Some of the knowledge about genes is from the gene summary entries in the Entrez Gene database. The three unrecognized genes associated with the Alzheimer's disease pathway are APLP1, APLP2, and APBA1.

Performance comparison

For performance comparison, we tested two additional reference methods to identify unrecognized gene-gene relationships. The first method counts the number of shared PubMed citations cross-referenced in the Entrez Gene IDs for each gene. For the Reelin signaling pathway, we counted the number of PubMed co-citations between RELN and other genes. For the Alzheimer's disease pathway, we counted the number of PubMed co-citations between APP and other genes. If a paper is cross-referenced in two genes, the two genes are likely to have a direct or indirect association. The larger number of co-citations provides us with the more probable relationship between two genes, which may be a direct or indirect association. Provided we already knew genes directly associated with a pathway, we can find genes indirectly associated with the pathway. In Table 2, the number of PubMed co-citations between RELN and DAB1 was 9. Even though this method could not find some indirect relationships, *i.e.* (RELN – APOE) and (RELN – MAPT), it could find most of the direct and indirect relationships in the Reelin signaling pathway. However, it found only a known relationship (APP – PSEN1) in the Alzheimer's disease pathway (see Table 3). It cannot suggest potential gene relationships if they do not have co-citations.

The second method counts the frequency of gene symbol 'gene-B' in the gene-document of gene-A, and the frequency of gene symbol 'gene-A' in the gene-document of gene-B to find the relationship between gene-A and gene-B. For example, we searched for a symbol 'DAB1' in the RELN gene-document and a symbol 'RELN' in the DAB1 gene-document in order to find a relationship of (RELN – DAB1). The total frequency of symbol-match for (RELN – DAB1) was 47. Though this method could find all direct

Table 2: Genes directly and indirectly associated with the Reelin signal pathway. The cosine similarities between RELN and genes in the full space and the reduced dimensional space obtained from NMF are also presented. (n/a: not applicable)

Gene	PubMed co-citation		Symbol match		Full space		NMF (k = 3)	
	# co-citation	Rank	# match	Rank	cos θ	Rank	cos θ	Rank
Genes directly associated with the Reelin signaling (five genes)								
RELN	n/a	-	n/a	-	1.0000	1	1.0000	1
DAB1	9	1	47	1	0.7349	2	0.9964	3
LRP8	2	2	1	7	0.4955	4	0.9854	6
FYN	1	5	4	4	0.2887	9	0.9811	7
VLDLR	2	2	9	2	0.5223	3	0.9463	8
Genes indirectly associated with the Reelin signaling (six genes)								
CDK5R1	1	5	0	-	0.2893	7	0.9966	2
CDK5	2	2	3	6	0.2903	6	0.9964	3
CDK5R2	1	5	0	-	0.3228	5	0.9962	5
SRC	1	5	5	3	0.2889	8	0.8433	9
MAPT	0	-	0	-	0.2147	29	0.7630	10
APOE	0	-	0	-	0.2151	28	0.6059	11

Table 3: Known and unrecognized genes associated with the Alzheimer's disease pathway. The cosine similarities between APP and genes in the full space and the reduced dimensional space obtained from NMF are also presented. (n/a: not applicable)

Gene	PubMed co-citation		Symbol match		Full space		NMF (k = 3)	
	# co-citation	Rank	# match	Rank	cos θ	Rank	cos θ	Rank
Known genes associated with the Alzheimer's disease pathway (eight genes)								
APP	n/a	-	n/a	-	1.0000	1	1.0000	1
APBB1	0	-	97	1	0.4428	5	0.9989	3
PSEN1	2	1	37	5	0.5018	2	0.9989	3
PSEN2	0	-	18	6	0.4172	6	0.9978	7
LRP1	0	-	12	8	0.2674	10	0.8433	8
A2M	0	-	1	10	0.2350	18	0.7479	9
APOE	0	-	0	-	0.2480	13	0.7142	10
MAPT	0	-	1	10	0.3084	8	0.6649	11
Unrecognized genes associated with the Alzheimer's disease pathway (three genes)								
APBA1	0	-	54	4	0.3468	7	1.0000	1
APLP1	0	-	95	2	0.4490	4	0.9989	3
APLP2	0	-	85	3	0.4824	3	0.9989	3

relationships, it could not find most of the indirect relationships except (RELN – CDK5) and (RELN – SRC) in the Reelin signaling pathway. This low recall problem is primarily due to inconsistencies in gene symbol usage in the literature. It could not find a known relationship (APP – APOE) in the Alzheimer's disease pathway. In contrast to these two reference methods, our gene retrieval method based on NMF could extract most of the direct and indirect gene-gene relationships by using the cosine similarity measure in the reduced dimensional space. We ranked genes by cosine similarity with a query gene RELN for the Reelin signaling pathway. A higher cosine similarity indicates a more probable relationship between two genes, which may be a direct or indirect association. In the full dimensional space, the ranks of two indirect relationships ((RELN – APOE) and (RELN – MAPT)) were 28 and 29, which were larger than those obtained from GR/NMF. APP was used as a query gene for the Alzheimer's disease pathway. In the full dimensional space, the ranks of two known relationships ((APP – APOE) and (APP – A2M)) were 13 and 18. In the reduced dimensional space obtained from NMF, the ranks were reduced so that we could capture the relationships. NMF (k = 3) can also be used to visualize genes in three dimensional space when it can retrieve most of the known relationships. Figure 1 shows gene relationships in the Reelin signaling pathway. Figure 2 illustrates gene relationships in the Alzheimer's disease pathway. By using different initializations of H, we were able to focus on the specific gene relationships in the different pathways. The proposed NMF initialization scheme was evaluated by retrieving genes associated with the Alzheimer's disease pathway in the reduced three-dimensional space obtained from NMF with k = 3. After computing the NMF with different initializations, we obtained F-measure values when 10 genes were retrieved. From 50 different random initializations, the NMF produced the maximal F-measure value (0.9524) only 36 times. Typically, the NMF is sensitive to random initialization since it converges only to a critical point.

On the other hand, by using the proposed initialization scheme, the NMF achieved the maximal F-measure value 49 times. Using known biological knowledge, we were able to improve the probability that the NMF converges to a solution which reflects the known knowledge very well. In addition, since the convergence criterion Eq. (6) (below) is sometimes not tight enough for true convergence, guaranteed and faster convergence by the proposed NMF initialization scheme is required.

Tables 4 and 5 show the influence of the reduced dimension k on the LSI/SVD and GR/NMF retrieval performance. Recall, precision, and F-measures were computed when 10 genes were retrieved. Both cases showed that small k was enough to generate high F-measure values. GR/NMF showed comparable performance with LSI/SVD. By using the k-selection scheme in LSI/SVD, k = 3 was chosen for the Reelin signaling pathway and the Alzheimer's disease pathway. As for GR/NMF, we chose k = 3 for both pathways by using the k-selection scheme and the initialization strategy. Tables 4 and 5 show that LSI/SVD and GR/NMF exhibited excellent retrieval for indirect or unrecognized genes with k = 3.

Practical applications

The LSI/SVD and proposed GR/NMF can elucidate unrecognized gene-gene interactions (*i.e.* edges in a gene interaction graph) from some known gene relationships. There are several types of identified gene-gene interactions. Firstly, a gene relationship identified by our methods can

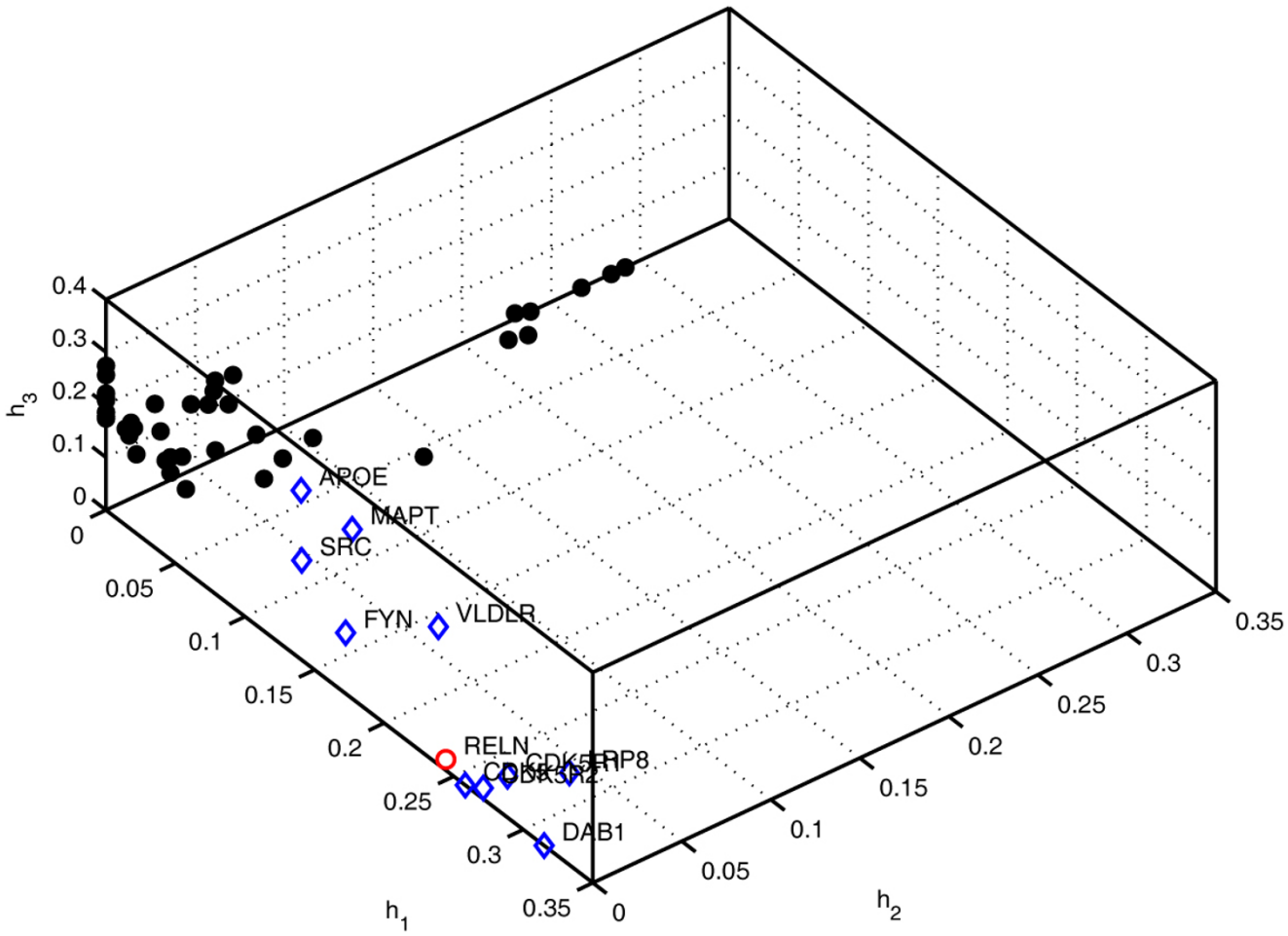

Figure 1

Visualization of genes associated with the Reelin signaling pathway. Three-dimensional representations of 50 genes, which were obtained from NMF (k = 3) using an initial matrix H built from genes directly associated with Reelin signaling pathway. The j-th gene is located at (h_{1j}, h_{2j}, h_{3j}), where $H \in \mathbb{R}^{3\times 50} = [h_{ij}]$. (A red circle: RELN; A red circle and blue diamonds: genes associated with the Reelin signaling pathway; Black dots: other genes)

be a completely novel direct gene-gene interaction so that it needs to be confirmed by wet-laboratory biochemical experiments. Secondly, it can be an indirect gene-gene interaction implicit in a priori knowledge. For instance, if gene-A activates gene-B and gene-B inhibits gene-C, then gene-A and gene-C have an indirect gene-gene interaction. However, one need to be careful since there is still a possibility that gene-A and gene-C have a direct gene-gene interaction. Thirdly, it can be an explicitly known direct gene-gene interaction that is available in public databases although it was not recognized by users in advance.

Conclusion

In this paper, we have shown the utility of the SVD and the NMF extracting unrecognized gene-gene relationships from the biomedical literature. We have introduced a way to incorporate a priori knowledge into LSI/SVD and NMF in order to retrieve unrecognized documents related with a query document. Specifically, we have established the reduced dimension k estimation schemes for LSI/SVD and GR/NMF, which are generally applicable to information retrieval using the SVD and the NMF when there are some known relationships between a query document and other documents. The proposed GR/NMF takes advantage of a priori knowledge of cluster structure in its initialization step. It could retrieve unrecognized genes by using known genes associated with a biological pathway, which showed comparable performance with LSI/SVD. Extracting potential gene relationships from the biomedical literature may be helpful in building biological hypotheses that can be explored further experimentally.

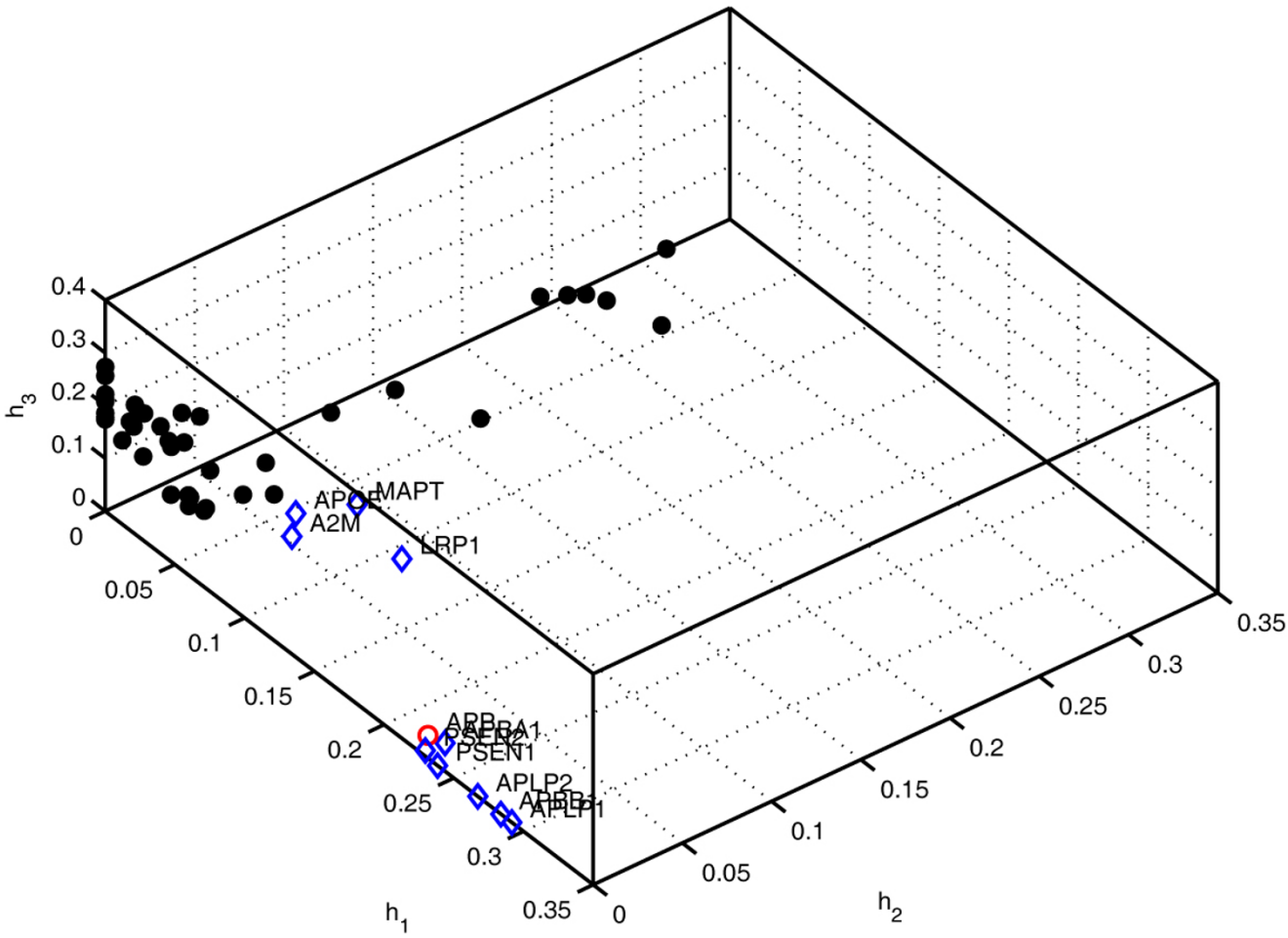

Figure 2

Visualization of genes associated with the Alzheimer's disease pathway. Three-dimensional representations of 50 genes, which were obtained from NMF (k = 3) using an initial matrix H built from known genes associated with the Alzheimer's disease pathway. The j-th gene is located at (h_{1j}, h_{2j}, h_{3j}), where $H \in \mathbb{R}^{3\times 50} = [h_{ij}]$. (A red circle: APP; A red circle and blue diamonds: genes associated with the Alzheimer's disease pathway; Black dots: other genes)

Methods

Gene-document collection

In [3], a gene-document is generated by concatenating all titles and abstracts of the PubMed IDs cross-referenced in the human, mouse, and rat Entrez Gene IDs for each gene. However, this concatenation may cause skewed encoding in favor of a gene that has more PubMed IDs. In our study, to identify unrecognized gene relationships for n genes, a term-by-gene_document matrix A of size $m \times n$ is generated by the following scheme. A total of 50 genes were considered in three broad categories: (1) Alzheimer's disease; (2) cancer; (3) development (see Table 1). These 50 genes are the same genes as those used in [3]. For each Enrez Gene ID, we downloaded up to 10 of the most recent titles and abstracts, which were available as of July, 2006. Table 6 shows Entrez Gene IDs for human, mouse, and rat and the number of PubMed citations for each gene. As an intermediate step, we constructed a term-by-PudMed_document matrix A_p of size 8, 316 × 1, 273 in the form of MATLAB sparse arrays generated by Text to Matrix Generator (TMG) [20], where a PubMed-document is a document generated by concatenation of a title and an abstract for a PubMed ID. Then, from A_p, we built a term-by-gene_document matrix A of size 8, 316 × 50, where 50 gene-documents are centroid vectors for 50 genes.

We applied common filtering techniques (*e.g.* removal of common words, removal of words that are too short or too long, *etc.*) for the purpose of reducing the size of the term dictionary. Stemming was also applied. The $m \times n_p$ term-by-PubMed_document matrix $A_p = [\tilde{a}_{ij}]$ was provided by using a log-entropy weighting scheme [2]. The

Table 4: Influence of the reduced dimension *k* on gene retrieval of the Reelin signal pathway. Recall, precision, and *F*-measure were computed when 10 genes were retrieved.

	k	Recall	Precision	*F*-measure
LSI/SVD	2	0.5455	0.6000	0.5714
	3*	0.9091	1.0000	0.9524
	4	0.7273	0.8000	0.7619
	5	0.6364	0.7000	0.6667
	6	0.6364	0.7000	0.6667
	10	0.6364	0.7000	0.6667
	20	0.5455	0.6000	0.5714
	30	0.8182	0.9000	0.8571
	40	0.7273	0.8000	0.7619
	50	0.8182	0.9000	0.8571
GR/NMF	2	0.4545	0.5000	0.4762
	3*	0.9091	1.0000	0.9524
	4	0.9091	1.0000	0.9524
	5	0.6364	0.7000	0.6667
	6	0.6364	0.7000	0.6667
	10	0.5455	0.6000	0.5714
	20	0.7273	0.8000	0.7619

*The reduced dimension *k* obtained from the *k*-selection scheme using only genes directly associated with this pathway.

Table 5: Influence of the reduced dimension *k* on gene retrieval of the Alzheimer's disease pathway. Recall, precision, and *F*-measure were computed when 10 genes were retrieved.

	k	Recall	Precision	*F*-measure
LSI/SVD	2	0.5455	0.6000	0.5714
	3*	0.9091	1.0000	0.9524
	4	0.9091	1.0000	0.9524
	5	0.8182	0.9000	0.8571
	6	0.8182	0.9000	0.8571
	10	0.7273	0.8000	0.7619
	20	0.8182	0.9000	0.8571
	30	0.8182	0.9000	0.8571
	40	0.8182	0.9000	0.8571
	50	0.8182	0.9000	0.8571
GR/NMF	2	0.4545	0.5000	0.4762
	3*	0.9091	1.0000	0.9524
	4	0.9091	1.0000	0.9524
	5	0.9091	1.0000	0.9524
	6	0.8182	0.9000	0.8571
	10	0.8182	0.9000	0.8571
	20	0.7273	0.8000	0.7619

*The reduced dimension *k* obtained from the *k*-selection scheme using only known genes associated with this pathway.

elements of A_p are often assigned two-part values $\tilde{a}_{ij} = l_{ij} * g_i$, where l_{ij} is the local weight for the *i*-th term in the *j*-th PubMed-document, and g_i is the global weight for the *i*-th term. The local weight l_{ij} and the global weight g_i can be computed as

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

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

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

where f_{ij} is the frequency of the *i*-th term in the *j*-th PubMed-document, p_{ij} is the probability of the *i*-th term occurring in the *j*-th PubMed-document, and n_p is the number of PubMed-documents in the collection.

A gene-document vector $\mathbf{a}_i$ (the *i*-th column of *A*) can be easily compared with another gene-document vectors $\mathbf{a}_j$ ($1 \le j \le n$) in the full dimensional space. The similarity scores between two gene-documents ($\mathbf{a}_i$ and $\mathbf{a}_j$) can be computed as

$$\cos(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}. \quad (2)$$

Gene-document vectors having the higher cosine values are deemed more relevant to each other. In this gene retrieval method, a query gene vector is one of column vectors of *A*. This method tries to retrieve genes relevant to the given query gene. In order to compare gene retrieval methods quantitatively, we used the following performance measures. We defined the relevant genes, which include the query gene itself as well as genes related with the query gene. The recall and precision are defined as

$$\text{recall} = \frac{|\{\text{relevant genes}\} \cap \{\text{retrieved genes}\}|}{|\{\text{relevant genes}\}|},$$

$$\text{precision} = \frac{|\{\text{relevant genes}\} \cap \{\text{retrieved genes}\}|}{|\{\text{retrieved genes}\}|}.$$

The weighted harmonic mean of precision and recall, the traditional *F*-measure is defined as

$$F\text{-measure} = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}}.$$

Reduced rank estimation for gene retrieval via LSI/SVD

LSI is based on the assumption that there is some underlying latent semantic structure in the term-by-gene_document matrix that is corrupted by the wide vari-

ety of words used in gene-documents. This is referred to as the problem of polysemy and synonymy. The basic idea is that if two gene-documents represent the same topic, they will share many associating words, and they will have very close semantic structures after dimension reduction via the SVD. In LSI/SVD, if the matrix A has its SVD,

$$A = U\Sigma V,$$

then its rank k approximation for some $k < \text{rank}(A)$,

$$A = U_k \Sigma_k V_k^T$$

is considered, where the columns of U_k are the leading k left singular vectors, Σ_k is a $k \times k$ diagonal matrix with the k largest singular values in nonincreasing order along its diagonal, and the columns of V_k are the leading k right singular vectors. Then, $S_k V_k^T$ is the reduced dimensional representation of A, or equivalently, a gene-document vector $\mathbf{a} \in \mathbb{R}^{m\times 1}$ can be represented in the k-dimensional space as $\hat{\mathbf{a}} = U_k^T \mathbf{a}$. Then, the similarity scores between two gene-documents ($\hat{\mathbf{a}}_i$ and $\hat{\mathbf{a}}_j$) in the k-dimensional space can be computed as

$$\cos(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i^T \mathbf{a}_j}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2} = \frac{(U_k^T \mathbf{a}_i)^T (U_k^T \mathbf{a}_j)}{\|U_k^T \mathbf{a}_i\|_2 \|U_k^T \mathbf{a}_j\|_2}.$$

Gene-documents having the higher cosine values in the reduced k-dimensional space are deemed more relevant to each other.

Here, we suggest a method to estimate the reduced rank k in LSI/SVD in order to retrieve unrecognized genes related with a query gene. If we can capture known gene-gene relationships in the reduced dimensional space obtained from LSI/SVD, we expect that the low-rank representations of gene-document vectors would be reliable to extract other gene relationships as well. This reduced rank k estimation scheme computes the following recall ($\tilde{r}$), precision ($\tilde{p}$), and $\tilde{F}$-measure only from known genes relevant to the given query gene:

$$\tilde{r} = \frac{|\{\text{known relevant genes}\} \cap \{\text{retrieved genes}\}|}{|\{\text{known relevant genes}\}|},$$
$$\tilde{p} = \frac{|\{\text{known relevant genes}\} \cap \{\text{retrieved genes}\}|}{|\{\text{retrieved genes}\}|},$$
$$\tilde{F}\text{-measure} = \frac{2 * \tilde{r} * \tilde{p}}{\tilde{r} + \tilde{p}}, \tag{3}$$

for various k values. It chooses the smallest k that shows the highest $\tilde{F}$-measure value in order to retrieve unrecognized genes related to the given query gene.

Gene retrieval via NMF (GR/NMF)

In this section, we describe a gene retrieval method based on the NMF (GR/NMF) including the initialization of H and the reduced dimension k selection scheme.

NMF based on alternating non-negativity-constrained least squares (NMF/ANLS)

Given a non-negative matrix $A \in \mathbb{R}^{m\times n}$, the NMF based on alternating non-negativity-constrained least squares (NMF/ANLS) starts with the initialization of $H \in \mathbb{R}^{k\times n}$ with non-negative values. Then, it iterates the following ANLS until convergence:

$$\min_W \left\| H^T W^T - A^T \right\|_F^2, \quad s.t. \quad W \geq 0, \tag{4}$$

which fixes H and solves the optimization with respect to W, and

$$\min_H \|WH - A\|_F^2, \quad s.t. \quad H \geq 0, \tag{5}$$

which fixes W and solves the optimization with respect to H. Paatero and Tapper [21] originally proposed using a constrained alternating least squares algorithm to solve Eq. (1). Lin [22] discussed the convergence property of alternating non-negativity-constrained least squares and showed that any limit point of the sequence (W, H) generated by alternating non-negativity-constrained least squares is a stationary point of Eq. (1). After convergence, the columns of the basis matrix W are normalized to unit L_2-norm and the rows of H are adjusted so that the approximation error is not changed. Here, we adopt a fast algorithm for large scale non-negativity-constrained least squares (NLS) problems [23] to solve Eqs. (4–5). Bro and de Jong [24] made a substantial speed improvement to Lawson and Hanson's algorithm [25] for large scale NLS problems. Van Benthem and Keenan [23] devised an algorithm that further improves the performance of NLS. This

Table 6: The number of PubMed citations associated with Entrez Gene IDs for each gene

		Entrez Gene ID			The number of PubMed citations			
Symbol	Gene description	Human	Mouse	Rat	Human	Mouse	Rat	Total
A2M	alpha-2-macroglobulin	2	232345	24153	10	10	10	30
ABL1	v-abl Abelson murine leukemia viral oncogene homolog 1	25	11350	311860	10	10	4	24
APBA1	amyloid beta precursor protein-binding, family A, member 1	320	108119	83589	10	2	6	18
APBB1	amyloid beta precursor protein-binding, family B, member 1	322	11785	29722	10	10	7	27
APLP1	amyloid beta precursor-like protein 1	333	11803	29572	10	10	2	22
APLP2	amyloid beta precursor-like protein 2	334	11804	25382	10	10	5	25
APOE	apolipoprotein E	348	11816	25728	10	10	8	28
APP	amyloid beta precursor protein	351	11820	54226	10	10	10	30
ATOH1	atonal homolog 1 (Drosophila)	474	11921	-	6	10	-	16
BRCA1	breast cancer 1, early onset	672	12189	24227	10	10	6	26
BRCA2	breast cancer 2, early onset	675	-	25082	10	-	3	13
CDK5	cyclin-dependent kinase 5	1020	12568	140908	10	10	10	30
CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	8851	12569	116671	10	10	10	30
CDK5R2	cyclin-dependent kinase 5, regulatory subunit 2 (p39)	8941	12570	-	10	10	-	20
DAB1	disabled homolog 1 (Drosophila)	1600	13131	266729	10	10	4	24
DLL1	delta-like 1 (Drosophila)	28514	13388	84010	10	10	2	22
DNMT1	DNA (cytosine-5-)-methyltransferase 1	1786	13433	84350	10	10	3	23
EGFR	epidermal growth factor receptor	1956	13649	24329	10	10	10	30
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2	2064	13866	24337	10	10	10	30
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1	2113	23871	24356	10	10	10	30
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	2353	14281	24371	10	10	10	30
FYN	FYN oncogene related to SRC, FGR, YES	2534	14360	25150	10	10	10	30
GLI	glioma-associated oncogene homolog 1	2735	14632	140589	10	10	1	21
GLI2	GLI-Kruppel family member GLI2	2736	14633	-	10	10	-	20
GLI3	GLI-Kruppel family member GLI3	2737	14634	140588	10	10	1	21
JAG1	jagged 1 (Alagille syndrome)	182	16449	29146	10	10	5	25
KIT	feline sarcoma viral oncogene homolog	3815	16590	64030	10	10	8	28
LRP1	low density lipoprotein-related protein 1	4035	16971	-	10	10	-	20
LRP8	low density lipoprotein receptor-related protein 8	7804	16975	-	10	10	-	20
MAPT	microtubule-associated protein tau	4137	17762	29477	10	10	10	30
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	4609	17869	24577	10	10	10	30
NOTCH1	Notch homolog 1, translocation-associated (Drosophila)	4851	18128	25496	10	10	10	30
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	4893	18176	24605	10	10	6	26
PAX2	paired box gene 2	5076	18504	-	10	10	-	20
PAX3	paired box gene 3 (Waardenburg syndrome 1)	5077	18505	114502	10	10	2	22
PSEN1	presenilin 1 (Alzheimer disease 3)	5663	19164	29192	10	10	10	30
PSEN2	presenilin 2 (Alzheimer disease 4)	5664	19165	81751	10	10	10	30
PTCH	patched homolog (Drosophila)	5727	19206	89830	10	10	3	23
RELN	reelin	5649	19699	24718	10	10	10	30
ROBO1	roundabout, axon guidance receptor, homolog 1	6091	19876	58946	10	10	2	22
SHC1	Src homology 2 domain containing transforming protein 1	6464	20416	85385	10	10	10	30
SHH	sonic hedgehog homolog (Drosophila)	6469	20423	29499	10	10	10	30
SMO	smoothened homolog (Drosophila)	6608	20596	-	10	9	-	19
SRC	v-src sarcoma viral oncogene homolog	6714	20779	83805	10	10	10	30
TGFB1	transforming growth factor, beta 1	7040	21803	59086	10	10	10	30
TP53	tumor protein p53 (Li-Fraumeni syndrome)	7157	22059	24842	10	10	10	30
VLDLR	very low density lipoprotein receptor	7436	22359	25696	10	10	5	25
WNT1	wingless-type MMTV integration site family, member 1	7471	22408	24881	10	10	5	25
WNT2	wingless-type MMTV integration site family member 2	7472	22413	114487	10	10	6	26
WNT3	wingless-type MMTV integration site family, member 3	7473	22415	24882	8	10	4	22

algorithm deals with the following NLS optimization problem given $B \in \mathbb{R}^{m\times k}$ and $A \in \mathbb{R}^{m\times n}$:

$$\min_{G} \|BG - A\|_F^2, \quad s.t. \quad G \geq 0,$$

where $G \in \mathbb{R}^{k\times n}$ is a solution. It is based on the active/passive set method. More detailed explanations of this algorithm can be found in [23].

A method for initialization

Most NMF algorithms require initialization of both W and H, whereas NMF/ANLS described in this paper requires only initialization of H. In our approach, we incorporate a priori knowledge of gene relationships into the initialization of H. A gene-document is represented as a linear combination of basis vectors. For gene clustering by NMF, gene-documents that are dominated by the same basis vector belong to the same cluster. Here, we propose the following NMF initialization strategy. The elements of the first row of the initial matrix $H \in \mathbb{R}^{k\times n}$ are set to 1 only if the columns corresponding to a set of known genes $\mathcal{S}_g$ are related with one another, otherwise set the elements to 0. For the other rows of H, the elements are set to 0 only if the columns correspond to $\mathcal{S}_g$, otherwise the elements are set to random numbers $\in$ (0.25, 0.75). For instance, we know that RELN is related to DAB1, LRP8, VLDLR, and FYN. Thus, the elements of the first row of H have 1 only if the columns correspond to a set of genes $\mathcal{S}_g$ = {RELN, DAB1, LRP8, VLDLR, FYN}, otherwise set the element to 0. The elements of the other rows of H have 0 only if the corresponding columns are related to $\mathcal{S}_g$, otherwise set the elements to random numbers $\in$ (0.25, 0.75). Let us assume that the 4th, 5th, 6th, 7th, and 8th columns of H correspond to RELN, DAB1, LRP8, VLDLR, and FYN. Then, when k = 3, we can construct an initial matrix H as

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & \cdots & 0 \\ \bullet & \bullet & \bullet & 0 & 0 & 0 & 0 & 0 & \bullet & \cdots & \bullet \\ \bullet & \bullet & \bullet & 0 & 0 & 0 & 0 & 0 & \bullet & \cdots & \bullet \end{pmatrix},$$

where the values in the location of • are random numbers. The columns of the initial matrix H are normalized to unit L_2-norm. This initial matrix H contains a priori cluster structure.

Gene retrieval

NMF/ANLS is used to obtain the final W and H from the initial matrix H. Convergence is tested at every five iterations. The Frobenius norm of the error, *i.e.* $f = ||A - WH||_F$, is computed at each convergence test. The convergence criterion is

$$\frac{f_{prev} - f_{curr}}{f_{prev}} < 10^{-4}, \tag{6}$$

where f_{prev} and f_{curr} are the Frobenius norms in the previous and current convergence tests respectively. The final matrix $H \in \mathbb{R}^{k\times n}$ contains the low-rank representation of the term-by-gene_document matrix A. Hence, the similarity scores between two genes (i and j) in the k-dimensional space can be computed as

$$\cos(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i^T \mathbf{h}_j}{\|\mathbf{h}_i\|_2 \|\mathbf{h}_j\|_2},$$

where $\mathbf{h}_j$ is the j-th column of the final matrix H. Genes having the higher cosine values in the k-dimensional space are deemed more relevant to each other.

NMF can generate different final H matrices because of the random numbers in the initial matrix H. Therefore, it is natural to repeat NMF with different initial matrices to obtain final H matrices. GR/NMF selects one of the final H matrices, which generates the highest $\tilde{F}$-measure value using only known genes related with a query gene (see Eq. (3)). If there are several final H matrices that yield the same maximal $\tilde{F}$-measure value, it chooses one producing the highest average of cosine values between the query gene and other known genes related with the query gene.

Reduced dimension estimation for GR/NMF

The determination of the reduced dimension k is also an open problem in the NMF. As with the k-selection scheme for LSI/SVD, we can estimate the reduced dimension k for GR/NMF by making use of known gene-gene relationships. The reduced k-dimensional representations of n gene-documents are obtained from the NMF. Then, the $\tilde{F}$-measure is calculated only from known genes relevant to a query gene, after retrieving genes by cosine similarities in the reduced k-dimensional space. Even with a single k value, the NMF can generate different $\tilde{F}$-measure values owing to the random numbers in the initial matrix H. Thus, to determine an $\tilde{F}$-measure value for each k, the k-selection scheme selects the highest $\tilde{F}$-measure value after computing $\tilde{F}$-measure values with different initial matrices. This determines $\tilde{F}$-measure values for various k

values and then chooses the smallest reduced dimension k that produces the highest $\tilde{F}$-measure value.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
HK instigated the project, designed methods, performed experiments, and drafted the manuscript. HP and BLD made significant contributions to the design of coding in the term-by-gene_document matrix. All authors were involved in revising it critically for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements
The work of the first two authors is supported in part by the National Science Foundation Grants CCR-0204109, ACI-0305543, and CCF-0621889. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This article has been published as part of *BMC Bioinformatics* Volume 8 Supplement 9, 2007: First International Workshop on Text Mining in Bioinformatics (TMBio) 2006. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S9.

References
1. Berry MW, Dumais ST, O'Brien GW: **Using linear algebra for intelligent information retrieval.** *SIAM Review* 1995, **37:**573-595.
2. Berry MW, Drmac Z, Jessup ER: **Matrices, vector spaces, and information retrieval.** *SIAM Review* 1999, **41:**335-362.
3. Homayouri R, Heinrich K, Wei L, Berry MW: **Gene clustering by latent semantic indexing of MEDLINE abstracts.** *Bioinformatics* 2005, **21:**104-115.
4. Chu M, Plemmons RJ: **Nonnegative matrix factorization and applications.** *IMAGE* 2005, **34:**1-5.
5. Lee DD, Seung HS: **Learning the parts of objects by non-negative matrix factorization.** *Nature* 1999, **401:**788-791.
6. Pauca VP, Shahnaz F, Berry MW, Plemmons RJ: **Text mining using non-negative matrix factorizations.** *Proc SIAM Int'l Conf Data Mining (SDM'04)* 2004.
7. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data.** *Genome Research* 2003, **13:**1706-1718.
8. Brunet JP, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization.** *Proc Natl Acad Sci USA* 2004, **101(12):**4164-4169.
9. Lee DD, Seung HS: **Algorithms for non-negative matrix factorization.** *Proceedings of Neural Information Processing Systems* 2000:556-562 [http://citeseer.ist.psu.edu/lee01algorithms.html].
10. Hoyer PO: **Non-negative matrix factorization with sparseness constraints.** *Journal of Machine Learning Research* 2004, **5:**1457-1469.
11. Rice DS, Curran T: **Role of the reelin signaling pathway in central nervous system development.** *Annu Rev Neurosci* 2001, **24:**1005-1039.
12. Tissir F, Goffinet AM: **Reelin and brain development.** *Nat Rev Neurosci* 2003, **4:**496-505.
13. Keshvara L, Magdaleno S, Benhayon D, Curran T: **Cyclin-dependent kinase 5 phosphorylates disabled 1 independently of Reelin signaling.** *J Neurosci* 2002, **22:**4869-4877.
14. Arnaud L, Ballif BA, Forster E, Cooper JA: **Fyn tyrosine kinase is a critical regulator of disabled-1 during brain development.** *Curr Biol* 2003, **13:**9-17.
15. Bock HH, Herz J: **Reelin activates SRC family tyrosine kinases in neurons.** *Curr Biol* 2003, **13:**18-26.
16. Lee MS, Tsai LH: **Cdk5: one of the links between senile plaques and neurofibrillary tangles.** *J Alzheimers Dis* 2003, **5:**127-137.
17. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
18. Neumann S, Schobel S, Jager S, Trautwein A, Haass C, Pietrzik CU, FLichtenthaler S: **Amyloid precursor-like protein 1 influences endocytosis and proteolytic processing of the amyloid precursor protein.** *J Biol Chem* 2006, **281(11):**7583-7594.
19. Li Q, Sudhof TC: **Cleavage of amyloid-beta precursor protein and amyloid-beta precursor-like protein by BACE 1.** *J Biol Chem* 2004, **279(11):**10542-10550.
20. Zeimpekis D, Gallopoulos E: **Design of a MATLAB toolbox for term-document matrix generation.** *Proc Workshop on Clustering High Dimensional Data and its Applications at the 5th SIAM Int'l Conf Data Mining (SDM'05), Newport Beach, CA* 2005:38-48.
21. Paatero P, Tapper U: **Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values.** *Environmetrics* 1994, **5:**111-126.
22. Lin CJ: **Projected gradient methods for non-negative matrix factorization.** In *Tech Rep Information and Support Service ISSTECH-95-013 Department of Computer Science, National Taiwan University*; 2005.
23. van Benthem MH, Keenan MR: **Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems.** *J Chemometrics* 2004, **18:**441-450.
24. Bro R, de Jong S: **A fast non-negativity-constrained least squares algorithm.** *J Chemometrics* 1997, **11:**393-401.
25. Lawson CL, Hanson RJ: *Solving Least Squares Problems Englewood Cliffs, NJ: Prentice-Hall*; 1974.