Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations using a priori knowledge of gene relationships

Hyunsoo Kim and Haesun Park
College of Computing, Georgia Institute of Technology
801 Atlantic Drive, Atlanta, GA 30332
hskim@cc.gatech.edu, hpark@cc.gatech.edu

ABSTRACT

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. In this paper, we explore the utility of singular value decomposition (SVD) and non-negative matrix factorization (NMF) to extract unrecognized gene relationships from the biomedical literature by taking advantage of known gene relationships. We introduce a way to incorporate a priori knowledge of gene relationships into LSI/SVD and NMF. In addition, we propose a gene retrieval method based on NMF (GR/NMF), which shows comparable performance with latent semantic indexing based on SVD.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Application—*Text processing*; H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval—*Clustering*; G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*Singular value decomposition*

General Terms

Algorithms, Design, Theory

Keywords

Gene relationships, singular value decomposition, non-negative matrix factorization

1. INTRODUCTION

Latent semantic indexing based on singular value decomposition (LSI/SVD) [3, 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 [8]. 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 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, NMF solves the following optimization problem:

$$\min_{W,H} \|A - WH\|_F^2, \;\; s.t. \; W, H \geq 0, \tag{1}$$

where $W \in \mathbb{R}^{m \times k}$ is a basis matrix, $H \in \mathbb{R}^{k \times n}$ is a reduced dimensional representation of A, and $W, H \geq 0$ means that all elements of W and H are non-negative. NMF does not provide us with a unique solution if we can find a full rank square matrix X such that $A = WXX^{-1}H, WX \geq 0, X^{-1}H \geq 0$ [14]. NMF gives us more direct interpretation than PCA due to non-subtractive combinations of non-negative basis vectors. Also, 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 [7]. It has been successfully applied to many problems including text data mining [14, 21] and gene expression data analysis [12, 6]. 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 [14, 15, 9] 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 well capture known gene relationships.

2. GENE-DOCUMENT COLLECTION

To identify unrecognized gene relationships for n genes, a term-by-gene_document matrix A of size $m \times n$ is generated following the scheme proposed in [8]. Each gene-document, which is represented as a column in the matrix A, is generated by concatenation of all titles and abstracts of the PubMed IDs cross-referenced in the human, mouse, and rat Entrez Gene IDs for each gene.

We applied common filtering techniques (*e.g.* removal of com-

TMBIO'06, November 10, 2006, Arlington, Virginia, USA.

mon 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 $m \times n$ term-by-gene_document matrix $A = [a_{ij}]$ was provided by using a log-entropy weighting scheme [2]. The elements of A are often assigned two-part values $a_{ij} = l_{ij} * g_i$, where l_{ij} is the local weight for the i-th term in the j-th gene-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

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

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

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

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 gene-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 $\mathbf{a}_j$ ($1 \leq j \leq n$) 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

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

In this paper, 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 [8]. For each Enrez Gene ID, we downloaded up to 10 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. We built a term-by-gene_document matrix A of size $8,316 \times 50$ in the form of MATLAB sparse arrays generated by Text to Matrix Generator (TMG) [25].

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)		

3. 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 variety 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 SVD. In LSI/SVD, if the matrix A has its SVD,

$$A = U \Sigma V,$$

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

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

is considered, where the columns of U_k are the leading k left singular vectors, Σ_k is an $k \times k$ 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, $\Sigma_k V_k^T$ 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

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

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:

$$\tilde{r} = \frac{|\{\text{known relevant genes}\} \cap \{\text{retrieved genes}\}|}{|\{\text{known relevant genes}\}|},$$

$$\tilde{p} = \frac{|\{\text{known relevant genes}\} \cap \{\text{retrieved genes}\}|}{|\{\text{retrieved genes}\}|},$$

$$\tilde{F}\text{-measure} = \frac{2 * \tilde{r} * \tilde{p}}{\tilde{r} + \tilde{p}}, \quad (3)$$

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	$\cos\theta$	Rank	$\cos\theta$	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	0.4770	2	0.9997	2
LRP8	2	2	1	7	0.2981	3	0.9955	7
VLDLR	2	2	9	2	0.2552	4	0.9982	6
FYN	1	5	4	4	0.1760	8	0.9853	8
Genes indirectly associated with the Reelin signaling (six genes)								
CDK5	2	2	3	6	0.1847	7	0.9997	2
CDK5R1	1	5	0	-	0.1972	5	0.9997	2
CDK5R2	1	5	0	-	0.1927	6	0.9997	2
APOE	0	-	0	-	0.1209	14	0.7489	10
SRC	1	5	5	3	0.1256	11	0.8946	9
MAPT	0	-	0	-	0.1209	13	0.6516	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	$\cos\theta$	Rank	$\cos\theta$	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	0.1947	4	0.9977	3
LRP1	0	-	12	8	0.1166	14	0.7699	8
APOE	0	-	0	-	0.1338	8	0.4694	10
A2M	0	-	1	10	0.0890	21	0.4057	11
PSEN1	2	1	37	5	0.3108	2	0.9977	3
PSEN2	0	-	18	6	0.2036	3	0.9971	7
MAPT	0	-	1	10	0.1830	5	0.7012	9
Unrecognized genes associated with the Alzheimer's disease pathway (three genes)								
APLP1	0	-	95	2	0.1623	7	0.9977	3
APLP2	0	-	85	3	0.1644	6	0.9977	3
APBA1	0	-	54	4	0.1338	9	0.9991	2

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.

4. NMF BASED ON ALTERNATING NON-NEGATIVITY-CONSTRAINED LEAST SQUARES

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

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

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

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

which fixes W and solves the optimization with respect to H. Paatero and Tapper [20] originally proposed using a constrained alternating least squares algorithm to solve Eq. (1). Lin [18] discussed about 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. When $k < m$, the non-negative low-rank representation of A is given by H. Here, we adopt a fast algorithm for large scale non-negativity-constrained least squares (NNLS) problems [24] to solve Eqns. (4-5). Bro and de Jong [5] made a substantial speed improvement to Lawson and Hanson's algorithm [13] for large scale NNLS problems. Van Benthem and Keenan [24] devised an algorithm that further improves the performance of NNLS for multivariate data. This algorithm deals with the following NNLS optimization problem given $B \in \mathbb{R}^{m \times k}$ and $A \in \mathbb{R}^{m \times n}$:

$$\min_G \|BG - A\|_F^2, \;\; s.t. \;\; G \geq 0,$$

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 [24].

5. GENE RETRIEVAL VIA NMF (GR/NMF)

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

Table 4: Influence of rank k on the retrieval of genes directly and indirectly associated with the Reelin signal pathway. Recall, precision, and F-measure were computed when 10 genes were retrieved. *Reduced dimension k obtained from the k-selection scheme using only genes directly associated with this pathway.

	k	Recall	Precision	F-measure
LSI/SVD	2	0.5455	0.6000	0.5714
	3*	0.8182	0.9000	0.8571
	4	0.8182	0.9000	0.8571
	5	0.8182	0.9000	0.8571
	6	0.6364	0.7000	0.6667
	10	0.6364	0.7000	0.6667
	20	0.7273	0.8000	0.7619
	30	0.7273	0.8000	0.7619
	40	0.7273	0.8000	0.7619
	50	0.7273	0.8000	0.7619
GR/NMF	2	0.3636	0.4000	0.3810
	3*	0.9091	1.0000	0.9524
	4	0.9091	1.0000	0.9524
	5	0.9091	1.0000	0.9524
	6	0.9091	1.0000	0.9524
	10	0.6364	0.7000	0.6667
	20	0.5455	0.6000	0.5714

5.1 A method for initialization

Most of NMF algorithms require to initialize both W and H, whereas NMF/ANNLS described in this paper needs to initialize only 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 columns are corresponding to a set of known genes $\mathcal{S}_g$ related with one another, otherwise 0. For the other rows of H, the elements are set to 0 only if columns are corresponding to $\mathcal{S}_g$, otherwise random numbers $\in (0.25, 0.75)$.

For example, we know that RELN is related with DAB1, LRP8, VLDLR, and FYN. Thus, the elements of the first row of H have 1 only if columns are corresponding to a set of genes $\mathcal{S}_g$ = {RELN, DAB1, LRP8, VLDLR, FYN}, otherwise 0. The elements of the other rows of H have 0 only if columns are corresponding to $\mathcal{S}_g$, otherwise random numbers $\in (0.25, 0.75)$. Let us assume that the 4th, 5th, 6th, 7th, and 8th columns of H are corresponding to RELN, DAB1, LRP8, VLDLR, and FYN. Then, when $k = 3$, we can build 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 $\bullet$ 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.

5.2 Gene retrieval

NMF/ANNLS 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.

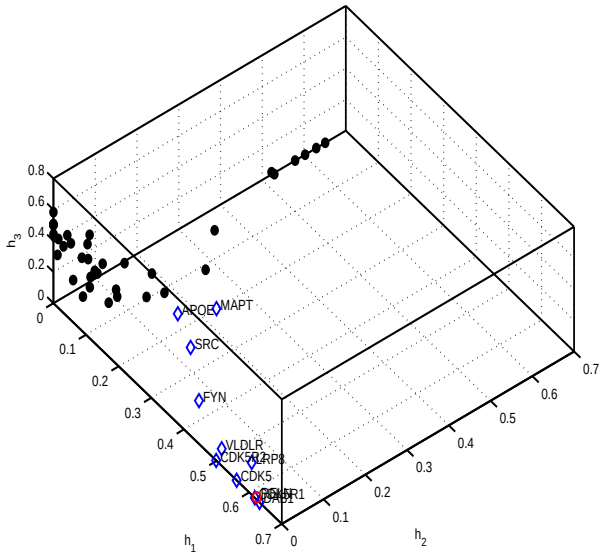

Figure 1: Three-dimensional representations of 50 genes, which were obtained from NMF ($k = 3$) using an initial matrix H built from genes directly associated with Reelin signaling pathway. The j-th gene is located at (h_{1j},h_{2j},h_{3j}), where $H \in \mathbb{R}^{3\times 50} = [h_{ij}]$. (A red circle: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

5.3 Reduced rank estimation for GR/NMF

The determination of the reduced rank k is also an open problem in NMF. As the k-selection scheme for LSI/SVD, we can estimate the reduced rank k for GR/NMF by making use of known gene-gene relationships. The reduced k-dimensional representations of n gene-documents are obtained from NMF. Then, $\tilde{F}$-measure is calculated only from known genes relevant to a query gene, after

Table 5: Influence of rank k on the retrieval of genes associated with the Alzheimer's disease pathway. Recall, precision, and F-measure were computed when 10 genes were retrieved. *Reduced dimension k obtained from the k-selection scheme using only known genes associated with this pathway.

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

retrieving genes by cosine similarities in the reduced k-dimensional space. Even with single k value, NMF can generate different $\tilde{F}$-measure values owing to the random numbers in the initial matrix H. Thus, to decide a $\tilde{F}$-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. It determines $\tilde{F}$-measure values for various k and then chooses the smallest reduced rank k that produces the highest $\tilde{F}$-measure value.

6. RESULTS AND DISCUSSION

For evaluation of various 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.

6.1 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 [22, 23]. 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 we can find the following indirect gene relationships by using above knowledge. Dab1 is phosphorylated on serine residues by cyclin-dependent kinase 5 (Cdk5) [11]. 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 [1, 4]. 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, cortico-basal degeneration and progressive supranuclear palsy. The six genes indirectly associated with the Reelin signaling pathway are CDK5, CDK5R1, CDK5R2, APOE, SRC, and MAPT.

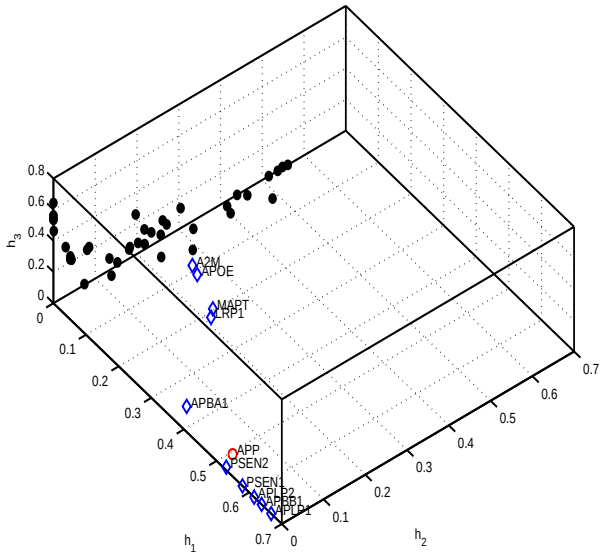

Figure 2: Three-dimensional representations of 50 genes, which were obtained from NMF ($k = 3$) using an initial matrix H built from known genes associated with the Alzheimer's disease pathway. The j-th gene is located at (h_{1j},h_{2j},h_{3j}), where $H \in \mathbb{R}^{3\times 50} = [h_{ij}]$. (A red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

6.2 Alzheimer's disease pathway

We obtained the Alzheimer's disease pathway from KEGG pathway database [10]. From this pathway, we can overview a general picture of the Alzheimer's disease pathway. Amyloid beta precursor protein (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 examine if we can find the following unrecognized knowledge from above known knowledge. Amyloid beta precursor-like protein 1 (APLP1) affects the endocytosis of APP and makes more APP available for α-secretase cleavage [19]. 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 [17]. 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 knowledge about genes is from the gene summary entries in Entrez Gene database. The three unrecognized genes associated with the Alzheimer's disease pathway are APLP1, APLP2, and APBA1.

6.3 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 the some indirect relationships, *i.e.* (RELN - APOE) and (RELN - MAPT), it could find most of 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 examples, we searched for a symbol 'DAB1' in RELN gene-document and a symbol 'RELN' in 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 relationships, it could not find most of 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 direct and indirect gene-gene relationships by using cosine similarity measure in the reduced dimensional space. We ranked genes by cosine similarity with a query gene RELN for the Reelin signaling pathway. The higher cosine similarity provides us with the 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 14 and 13, 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 - LRP1) and (APP - A2M)) were 14 and 21. 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 the three dimensional space when it can retrieve most of 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 could 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 NMF with different initializations, we obtained F-measure values when 10 genes were retrieved. From 50 different random initializations, NMF could produce the maximal F-measure value (0.9524) only 15 times. Typically, NMF is sensitive to the random initialization since it converges only to a local minimum. On the other hand, by using the proposed initialization scheme, NMF could achieve the maximal F-measure value 36 times. Using known biological knowledge, we could improve probability that NMF converges to a solution on which well reflects the known knowledge. In addition, since the convergence criterion Eq. (6) is sometimes not enough for true convergence, 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-measure 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, we decided $k = 3$ for the Reelin signaling pathway and $k = 4$ for the Alzheimer's disease pathway. For NMF, we decided $k = 3$ for both pathways. Tables 4 and 5 show that LSI/SVD and GR/NMF could well retrieve indirect or unrecognized genes at k selected by this scheme.

6.4 Practical applications

The proposed LSI/SVD and 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 the identified gene-gene interactions. Firstly, a gene relationship identified by our methods can 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.

7. CONCLUSIONS

In this paper, we have shown the utility of SVD and NMF so as to extract 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 a reduced rank k estimation scheme for LSI/SVD and GR/NMF, which is generally applicable to information retrieval using SVD and NMF when there exists 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.

8. REFERENCES

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

[2] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41:335–362, 1999.

[3] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[4] H. H. Bock and J. Herz. Reelin activates SRC family tyrosine kinases in neurons. *Curr. Biol.*, 13:18–26, 2003.

[5] R. Bro and S. de Jong. A fast non-negativity-constrained least squares algorithm. *J. Chemometrics*, 11:393–401, 1997.

[6] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, 101(12):4164–4169, 2004.

[7] M. Chu and R. J. Plemmons. Nonnegative matrix factorization and applications. *IMAGE*, 34:1–5, 2005.

[8] R. Homayouri, K. Heinrich, L. Wei, and M. W. Berry. Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics*, 1:104–115, 2005.

[9] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

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

[11] L. Keshvara, S. Magdaleno, D. Benhayon, and T. Curran. Cyclin-dependent kinase 5 phosphorylates disabled 1 independently of reelin signaling. *J. Neurosci.*, 22:4869–4877, 2002.

[12] P. M. Kim and B. Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*, 13:1706–1718, 2003.

[13] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ, 1974.

[14] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[15] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of Neural Information Processing Systems*, pages 556–562, 2000.

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

[17] Q. Li and T. C. Sudhof. Cleavage of amyloid-beta precursor protein and amyloid-beta precursor-like protein by BACE 1. *J. Biol. Chem.*, 279(11):10542–10550, 2004.

[18] C. J. Lin. Projected gradient methods for non-negative matrix factorization. Technical Report Information and Support Service ISSTECH-95-013, Department of Computer Science, National Taiwan University, 2005.

[19] S. Neumann, S. Schobel, S. Jager, A. Trautwein, C. Haass, C. U. Pietrzik, and S. F.Lichtenthaler. Amyloid precursor-like protein 1 influences endocytosis and proteolytic processing of the amyloid precursor protein. *J. Biol. Chem.*, 281(11):7583–7594, 2006.

[20] P. Paatero and U. Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[21] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons. Text mining using non-negative matrix factorizations. In *Proc. SIAM Int'l Conf. Data Mining (SDM'04)*, April 2004.

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

[23] F. Tissir and A. M. Goffinet. Reelin and brain development. *Nat. Rev. Neurosci.*, 4:496–505, 2003.

[24] M. H. van Benthem and M. R. Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *J. Chemometrics*, 18:441–450, 2004.

[25] D. Zeimpekis and E. Gallopoulos. Design of a MATLAB toolbox for term-document matrix generation. In I. S. Dhillon, J. Kogan, and J. Ghosh, editors, *Proc. Workshop on Clustering High Dimensional Data and its Applications at the 5th SIAM Int'l Conf. Data Mining (SDM'05)*, pages 38–48, Newport Beach, CA, April 2005.

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

		Enrez 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