

Data and text mining

Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach

Shao Li*, Lijiang Wu and Zhongqi Zhang

Bioinformatics Division, TNLIST and Department of Automation, Tsinghua University, Beijing 100084, China

Received on February 3, 2006; revised on May 16, 2006; accepted on June 29, 2006

Advance Access publication July 4, 2006

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Network reconstruction of biological entities is very important for understanding biological processes and the organizational principles of biological systems. This work focuses on integrating both the literatures and microarray gene-expression data, and a combined literature mining and microarray analysis (LMMA) approach is developed to construct gene networks of a specific biological system.

Results: In the LMMA approach, a global network is first constructed using the literature-based co-occurrence method. It is then refined using microarray data through a multivariate selection procedure. An application of LMMA to the angiogenesis is presented. Our result shows that the LMMA-based network is more reliable than the co-occurrence-based network in dealing with multiple levels of KEGG gene, KEGG Orthology and pathway.

Availability: The LMMA program is available upon request.

Contact: shaoli@mail.tsinghua.edu.cn

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Reconstructing networks of biological entities such as genes, transcription factors, proteins, compounds and other regulatory molecules is very important for understanding the biological processes and the organizational principles of the biological systems (Barabasi and Oltvai, 2004). Rapid progress in the biomedical domain has resulted in enormous amount of biomedical literatures. Along with the booming growth of biomedical researches, literature mining (LM) has become a promising direction for knowledge discovery. Various techniques have been developed which make it possible to reveal putative biological networks hidden in the huge collection of individual literatures (Shatkay and Feldman, 2003). Among them the co-occurrence (co-citation) approach (Stapley and Benoit, 2000; Jenssen *et al.*, 2001) is the simplest and most comprehensive in implementation. It can also be easily adopted to find the association between biological entities, such as the genes relations (Jenssen *et al.*, 2001) and the chemical compound–gene relations (Zhu *et al.*, 2005). In our early study (Zhang and Li, 2004), a subject-oriented literature mining technique has been developed to extract subject-specific knowledge by incorporating prior knowledge from biologists. Such approach can retrieve information contained in abundant literatures regardless of individual experimental conditions. However, the literature-derived

network is relatively crude and redundant. The co-occurrence approach lacks realistic analysis of various types of relations since the literature reservoir is a collection of results from diverse investigations. Moreover, the networks constructed from literatures are usually not specific with respect to a certain biological process and may unavoidably include overlapped relations, resulting in large and densely connected networks lacking in significant biological meaning.

Network reconstruction from the high-throughput microarray data is another active area in the past decade (van Someren *et al.*, 2002; de Jong 2002). Microarray technology that documents large-scale gene expression profiles allows characterizing the states of a specific biological system, providing a powerful platform for assessing global gene regulation and gene function. So far, a number of methods are available on reconstructing gene networks using microarray such as deterministic Boolean networks (Liang *et al.*, 1998) and ordinary differential equations (Zak *et al.*, 2003). However, it is difficult to build a reliable network from a small number of array samples owing to the non-uniform distribution of gene expression levels among thousands of genes. Such a technique is also insufficient for detailed biological investigation when prior knowledge is absent (Le Phillip *et al.*, 2004).

Both literature-based and microarray-based approaches share the common goal of identifying the hidden networks of biological entities. Integrating both the experimental data and the literature knowledge in an iterative fashion seems to be an effective way in biological network modeling (Le Phillip *et al.*, 2004). Various approaches have been developed to identify gene clusters and accompanying literature topics (Küffner *et al.*, 2005), and model some biological process such as neuro-endocrine-immune interactions (Wu and Li, 2005). In this article, we propose a novel approach to reconstruct gene networks through combining literature mining and microarray analysis (LMMA), where a global network is first derived using the literature-based co-occurrence method, and then refined using microarray data. The LMMA approach is applied to build an angiogenesis network. The network and its corresponding biological meaning are evaluated in multiple levels of KEGG Gene, KEGG Orthology and pathway. The results show that the LMMA-based network is more reliable and manageable with more significant biological content than the LM-based network.

2 METHODS

2.1 Co-occurrence-based PubMed literature mining

The first step in the LMMA approach is to derive co-occurrence dataset through literature mining. To find co-citations, a pool of articles and a

*To whom correspondence should be addressed.

dictionary containing gene symbols and their synonyms are required. In LMMA approach, the literature information is mainly obtained from the National Library of Medicine's PubMed database (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Display&DB=PubMed>). In general LM approach, specific subject interactions cannot be highlighted since all interactions tend to have similar co-citation number (Zhang and Li, 2004). We therefore prepare candidate articles and/or biological entities dictionary at two stages to incorporate prior knowledge. First, using a keyword referring to a certain subject, we select PubMed literatures that contain only terms of a biological subject. Next, an authoritative, standard or specific glossary is employed to provide a context for building topic-related gene networks. In the present work, we use the HUGO (Human Genome Organisation, <http://www.gene.ucl.ac.uk>) glossary, which contains ~20000 non-redundant gene symbols, for literature mining.

We perform LM by sharing assumption with many existing LM systems that when two genes are co-cited in the same text unit, there should be a potential biological relationship between them (Stapley and Benoit, 2000; Jessen *et al.*, 2001). Sentence as a text unit is found to make the good trade-off between precision and recall with high effectiveness (Ding *et al.*, 2002). Accordingly, as our previous work described (Zhang and Li, 2004), we regard two HUGO gene symbols as co-related if they are co-cited in the same sentence. In the HUGO glossary, one gene corresponds to a unique symbol (a one-to-one short mnemonic representation of the gene name) with several aliases. We regard all <alias, symbol> as <key, value> and store them in a hash table, by which many alias co-occurrences are reflected as a corresponding symbol co-occurrence. Capital letters and lowercase are discerned and only the complete words are considered. Next, we count the co-occurrence number of all symbol pairs to form an LM-based network regarded to a special subject.

2.2 Microarray datasets

Microarray datasets related to a biological process are collected from experiments or public repositories such as SMD (Stanford Microarray Database, <http://genome-www5.stanford.edu/>) that stores a large volume of raw and normalized data from public microarray information (Sherlock *et al.*, 2001). The downloaded microarray data are pre-processed following SMD procedures. In the step of Gene Filtering Options, we selected 'center data for each array by mean'. Meanwhile, a K -nearest neighbors method (Troyanskaya *et al.*, 2001) is used to evaluate the missing values in the microarray datasets. Briefly, a Pearson correlation analysis is employed to derive other K genes which have the most similar expression profiles of a missing value of gene x_i in a observation (i.e. a microarray experiment) j, x_{ij} . Then the missing value is retrieved from the weighted mean of the corresponding K genes.

2.3 LMMA-based network construction

The LMMA approach employs a module of statistical multivariate selection for gene interaction analysis. This is based on the hypothesis that if a co-cited gene-pair is positively or negatively co-expressed, they will indeed interact with each other (D'haeseleer *et al.*, 2000; Ge *et al.*, 2001). Taking the values of n genes as variable x_1, x_2, \dots, x_n , the dataset with m observations (i.e. m microarray experiments) and the n variables is denoted by

$$[x_1, x_2, \dots, x_n] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}. \quad (1)$$

Regression approaches are known to be helpful for analyzing microarray data and handling the complex correlations between expression levels of various genes across samples (West *et al.*, 2001; Segal *et al.*, 2003). Assuming the relations of variables in gene expression data follow a linear model (D'haeseleer *et al.*, 1999), a LM-based network can be refined through multiple variables selection, resulting in a network called LMMA network. We define each node coupling with its neighboring nodes as a

'unit-network'. The linear approximation of a sub-model is expressed as, $x_k = \beta_{k0} + \beta_{k1}x_{k1} + \beta_{k2}x_{k2} + \cdots + \beta_{kl}x_{kl} + e$, where the variables $x_{k1}, x_{k2}, \dots, x_{kl}$ denote the neighboring nodes of x_k in the LM-based network, and e is the random error. Subsequently, stepwise multiple variables selection is used to add new variables and eliminate insignificant variables. The significance of a variable is measured by a P -value which is determined from a F -test,

$$F_j = \frac{SSR_j}{SSE/(m-l-1)}, \quad (2)$$

where SSE and SSR_j represent the residual of random error and the residual of variables (except variable j) of the full model, respectively, m is the number of observations and l is the number of variables (including variable j) of the full model.

A cutoff threshold of P -value, named Th_p , is set to determine whether a variable should be added or deleted. Starting from a null model, we add one variable for evaluation at a time. When adding a new variable, some pre-determined variables may exceed the threshold and will be deleted one at a time. The statistical significance of a whole model is also verified using the F -test as follows:

$$F_{Mod} = \frac{SSR/l}{SSE/(m-l-1)}, \quad (3)$$

where SSR represents the residual of random error and the residual of all variables of the full model. The P -value, $P = P(F_{l,m-l-1} > F_{Mod})$. The significance of the LMMA network of one node can be evaluated.

A LMMA-based network is constructed by recombining all the refined sub-networks after the multivariate selection. For a specific interaction between x_i and x_j , there are two regression coefficients. One is when we regress x_i on x_j , the other is when we regress x_j on x_i , and the one with smaller P -value is used in LMMA. Note that the directionality of the LMMA network is currently not considered.

2.4 Network evaluation

First, in a network, a node represents a gene, and a connection between two nodes represents that these two genes are biologically related. The number of connections that a node has in a network is called the degree of the node (Jeong *et al.*, 2000; Han *et al.*, 2004; Song *et al.*, 2005), which indicates how many genes one gene is related with. The distribution of co-citation degrees is analyzed to know the topological property of the network reconstructed. The connectivity of a network is expressed by the shortest path from one node to another through the network. Each pair of nodes (x_i, x_j) has a shortest path $l_{i,j}$. The average path length of the network is defined as

$$\frac{2}{n(n-1)} \sum_i^n \sum_{j \neq i}^n l_{i,j}$$

Second, a permutation test is performed to examine the stability and integrity of the LMMA network. Keeping the total number of connections fixed the same as the LMMA network, we randomly eliminate the connections in the LM sub-network whose nodes (genes) are overlapped with microarray dataset, resulting in the so-called LM-random filtering networks. The cluster sizes between LM-random filtering and LMMA networks are analyzed by Kolmogorov-Smirnov test. Here, a cluster is defined as a group of connected genes separated from other genes. The cluster size denotes the number of genes in a cluster. Next, the average path length of the largest cluster in both LMMA and LM-random filtering networks is normalized (divided by the number of nodes) and then statistically analyzed by t test.

Third, we employ a leave one out cross validation (LOOCV) (Lachenbruch and Mickey, 1968) approach for evaluating the goodness of fitting in both LM and LMMA networks. According to LOOCV, when the observation (i.e. the microarray experiment) j is omitted for gene i and its neighbors, gene $1(i)$, gene $2(i)$, ..., gene $l(i)$, a new linear network can be constructed based on the remaining observations $x_{j(i)}$ and

$x_{-j1(i)}, x_{-j2(i)}, \dots, x_{-j(i)}$. And the omitted $x_{j(i)}$ can be recovered as $\hat{x}_{j(i)}$ through the corresponding observations of neighboring genes $x_{j1(i)}, x_{j2(i)}, \dots, x_{j(i)}$.

To evaluate the robustness of the network, the mean square error (MSE) for gene i , and the error sum of squares (SSE) for a holistic network, are calculated according to the following equations,

$$\text{MSE}_{(i)} = \frac{1}{m} \sum_{j=1}^m (x_{j(i)} - \hat{x}_{j(i)})^2 \quad (4)$$

$$\text{SSE} = \sum_{i=1}^w \sum_{j=1}^m (x_{j(i)} - \hat{x}_{j(i)})^2, \quad (5)$$

where m is the number of the experiments, and also the times of iteration; $x_{j(i)}$ is the true value and $\hat{x}_{j(i)}$ is the re-evaluated value; w is the number of genes within the network. A lower MSE value refers to good fitting.

A standard-score of MSE, SS_{mse} , can be expressed as

$$\text{SS}_{\text{mse}} = \frac{\sqrt{\text{SSE}/mw}}{\text{STD}\{x_{j(i)}\}}, \quad (6)$$

where $\text{STD}\{x_{j(i)}\}$ represents the standard deviation of the gene expression x_j . The standard-score depicts a relative value of SSE to gene expression variation. Good model exhibits small SS_{mse} value.

2.5 Network validation and pathway extraction

Pathway information is essential for successful quantitative modeling of biological systems (Cary *et al.*, 2005). A well-known pathway database that provides the information of metabolic, regulatory and disease pathways is deposited in KEGG (Kyoto Encyclopedia of Genes and Genomes, <http://www.genome.ad.jp/kegg/>) (Kanehisa and Goto, 2000). The relationship recorded in KEGG database is known to be special on the conception KEGG Orthology (KO, http://www.genome.jp/dbget-bin/get_htext?KO+s+F+f+F), a classification of orthologous genes that links directly to known pathways defined by KEGG. The KO dataset is a single complex flat file containing entries for all of the KO functional terms (the leaf nodes at the fourth level of the KO hierarchy). For more details about KO refer to Mao *et al.* (2005).

In order to take further insights on the underlying biological meanings of our networks, we map the LM- and LMMA-based networks to KEGG pathway database. First, we extract the KO hierarchy and the known associations between genes and their corresponding KO functional terms from the KO dataset. Second, we extract all the annotated genes from the KEGG Genes (KG) dataset. Both the KO hierarchical and the KG hierarchical relations are employed as benchmarks to validate the interactions in the networks. Here, a true positive (TP) defines an entry that is identified in our networks and is also identified in the dataset, a false positive (FP) refers to an entry that is identified in our networks but it does not occur in the dataset, a true negative (TN) represents an entry that is not identified in our networks and it does not occur in the dataset and a false negative (FN) indicates an entry that is not identified in our network but it occurs in the dataset. Here, we consider only KO/KG connections as entries for the definition of true positives, TP, FP, TN and FN. The precision, p , and the recall, r , of a network are derived respectively using the definition $p = \text{TP}/(\text{TP} + \text{FP})$, and $r = \text{TP}/(\text{TP} + \text{FN})$. To validate the effect of the LMMA for the precision of the relations predicted, Fisher Exact test with its online software (<http://www.matforsk.no/ola/fisher.htm>) is used to calculate the exact P -value of comparing the proportions of TP (FP) between LMMA and LM networks.

Moreover, we group the nodes and connections in the LMMA network according to KEGG pathway definitions. Here, a Fisher's Exact Test for KEGG pathway identification described in DAVID (the Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov/helps/functional_annotation.html#fisher) (Dennis *et al.*, 2003) is employed. We perform the KEGG pathway extraction for the LMMA network by statistically evaluating the gene-enrichment in the network, which is compared

Table 1. LM-based and LMMA-based angiogenesis network structures (Thp = 0.150)

	LM-EC	LMMA-EC	LM-ST	LMMA-ST
Common nodes ^a	1257	1031	1258	1162
Connections ^a	6761	2848	6884	3935
Average path length ^a	2.9810	3.6101	2.9741	3.3487
Average degree ^b	5.3738	2.2777	5.4722	3.1375
SSE ^c	522.3206	380.1941	520.2295	479.0745
SS _{mse} ^c	0.0669	0.057	0.0614	0.0589
Microarray size	1257*53	1257*53	1258*119	1258*119

^aIn the largest connected sub-network.

^bIn the whole network.

^cAll nodes except for the isolated ones.

with the random chance. Fisher's Exact Test is adopted to determine whether the proportion of genes of the LMMA network in a KEGG pathway is significantly higher than that for the human genomic background genes.

3 CONSTRUCTING ANGIOGENESIS NETWORK: AN APPLICATION

Angiogenesis is the process of generating new capillary blood vessels, and a key issue for various disorders especially for a variety of solid tumors, vascular and rheumatoid diseases (Folkman, 1995). Few other processes have such a significant impact as angiogenesis on so many people worldwide. So far, the underlying biological rules of angiogenesis remain unclear. It is therefore critical to understand the molecular basis and biological pathways of angiogenesis (Carmeliet, 2003).

3.1 Reconstruction of LM- and LMMA-based angiogenesis networks

We have successfully reconstructed angiogenesis-oriented networks using both LM and LMMA approaches. First, we collect all the angiogenesis-related PubMed abstracts (till July 24, 2005) using 'angiogenesis' as a keyword. A total of 23 497 'angiogenesis' related PubMed abstracts are indexed automatically. By putting HUGO glossary into this abstract pool, we obtained 1929 angiogenesis-related genes. A total of 9514 co-citations among these genes are extracted to construct the co-occurrence based angiogenesis network. We construct a LM-based network with a co-occurrence number of at least 1. This results in the network with the maximum gene interactions.

Next, we select the gene expression profiles of endothelial cells (EC) and solid tumors (ST) from SMD. It is believed that EC is responsible for the generation of blood vessels and ST is the majority of angiogenesis-dependent diseases (Carmeliet, 2003; Folkman, 1995). The EC microarray dataset contains 44 639 genes and 53 experiments, while the ST microarray dataset contains 39 726 genes and 119 experiments. The largest connected gene network in LM with its genes identified in the EC microarray dataset is called LM-EC network (1257 genes and 6761 connections). Similarly, the largest connected gene network in LM with its genes identified in the ST microarray dataset is called LM-ST network (1258 genes and 6884 connections). Accordingly, two LMMA-based angiogenesis networks, LMMA-EC and LMMA-ST are built (Table 1). Using

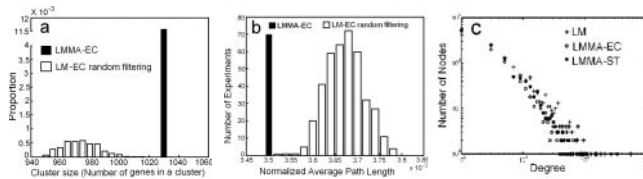


Fig. 1. (a) Comparison of the cluster sizes between the LMMA-EC network and the LM-EC random filtering networks ($P < 0.0001$, by Kolmogorov–Smirnov test). Other clusters are with <10 nodes (data not shown). (b) Comparison of the normalized average path length in the largest cluster between the LMMA-EC and the LM-EC random filtering networks ($P < 0.001$ by t test). (c) Relationship between the number of nodes and the degree of nodes in the whole LM angiogenesis network, LMMA-EC network, and LMMA-ST network ($\text{Thp} = 0.150$). The distribution of degrees in three networks follows a power law, obviously appearing to be scale-free.

the common genes as the baseline, we compare the LM-EC and the LM-ST networks with their corresponding LMMA-EC and LMMA-ST networks respectively.

Table 1 lists the network parameters for LM- and LMMA-based angiogenesis networks. It shows that redundant connections are eliminated after multivariate selection. The connections for LMMA-EC and LMMA-ST networks are much smaller than those of the predominant sub-networks of LM-EC and LM-ST, respectively. The elimination of connections results in a dramatic decrease of the average degrees of genes and a slightly reduction of node number and average path length. Moreover, as shown in Figure 1a and b, when comparing with the LM-random filtering networks derived from the permutation test, the LMMA network results in not only significantly larger cluster size ($P < 0.0001$, by Kolmogorov–Smirnov test), but also smaller path length of the largest cluster ($P < 0.001$ by t -test). The results demonstrate that LMMA is more stable and integrative than that of the LM-random filtering. Similar performance is observed with the LMMA-ST network (Supplementary Fig. S1). Thus, LMMA seems to maintain the backbone of the LM-based angiogenesis network.

Figure 1c shows the relationship between the number of nodes and the degree of nodes in both LM- and LMMA-based angiogenesis networks. Obviously, the profiles follow a power-law distribution, indicating that the topological properties of both networks are scale-free (Jeong *et al.*, 2000; Song *et al.*, 2005). Recent studies (Han *et al.*, 2004; Ozier *et al.*, 2003) show that centrally located, highly connected hub nodes in a scale-free network dominate network operation.

3.2 Comparison of LM- and LMMA-based angiogenesis networks

Top 15 hub genes in both LM-based and LMMA-based angiogenesis networks are listed in Table 2. Vascular endothelial growth factor (VEGF) is identified in both LM and LMMA networks as the hub gene with the highest degree. VEGF is known to be a multifunctional cytokine that plays an important role in vasculogenesis (Mukhopadhyay and Datta, 2004). The activation of endothelial cells by VEGF sets in motion a series of steps towards the creation of new blood vessels (Folkman, 1995).

Table 2 lists the P -values for the unit-networks of 15 hub genes derived from the F -test. We calculate the P -values for different

Table 2. The top 15 hub genes identified in LM-based and LMMA-based angiogenesis networks ($\text{Thp} = 0.150$)

Gene	Degree (LM-EC; LM-ST) ^a	Degree (LMMA)		P -value ^b	
		EC	ST	EC	ST
VEGF	554	51	117	0	0
NUDT6	211	51	25	0	0
KDR	182	51	117	0	3.59e–06
SIAT7B	156	51	44	1.19e–07	0
TNF	149	51	46	0	0
IL8	148	26	27	0	0
MVD	126	19	28	0	0
CD34	111	51	22	1.19e–07	0
EGF	104	32	40	1.35e–13	0
IL6	97	31	24	0	0
CDH17	96	30	27	0	0
HIF1A	93	21	38	1.65e–12	0
SOS1	87	14	25	1.54e–11	0
CCM1	83	51	14	6.92e–06	0
PSME3	78	18	34	0	0

^aDegree of these hub genes in both LM-EC and LM-ST networks are the same.

^b P -values are calculated from F -test for the unit-network of each gene (Equation 3).

networks. The results show that the LMMA-based angiogenesis network is more reliable than the LM-based one. Figure 2 illustrates LOOCV gene expression values for the unit-networks of VEGF, EGF, TNF and IL6, respectively. The MSE values of the LMMA unit-networks are smaller than those of the LM, indicating that the LMMA network fits better to the microarray data of angiogenesis. Meanwhile, Table 1 lists the SSE and the SS_{mse} scores resulted from LM- and LMMA-based networks. The reduced errors in LMMA again suggest the improvement of the LMMA-based networks.

Figure 3a and b shows the precision and the recall rates of both the LM- and LMMA-based angiogenesis networks at different threshold Thp . The LMMA-based network exhibits higher precisions and lower recalls than the LM-based one. On the other hand, the recall of LMMA-based network increases gradually with the increasing thresholds. We select a suitable threshold, $\text{Thp} = 0.150$, in the LMMA-based EC and ST networks.

Both LM-EC and LM-ST networks have the same 474 genes corresponding to 355 KO entities covered by KEGG database. When the LM-based network is refined by LMMA, the proportion of the TP rates increases significantly, while the proportion of FP rates decreases evidently. Table 3 shows the statistical results between LM- and LMMA-based angiogenesis networks, which demonstrate that the LMMA approach significantly eliminates the false positive relations.

3.3 Pathway extraction from networks

The statistical significance of pathways in LMMA-based angiogenesis networks is derived from Fisher Exact Test. The results are shown in Table 4 and graphically represented by an example, the EGF (epidermal growth factor) unit-network, in Figure 4. See more in Discussion below. Although many co-occurrence relations are eliminated from the LM-based network, main pathway information, such as the focal adhesion pathway, signaling pathways of TGF-beta, MAPK, Calcium and Wnt, is observed in the

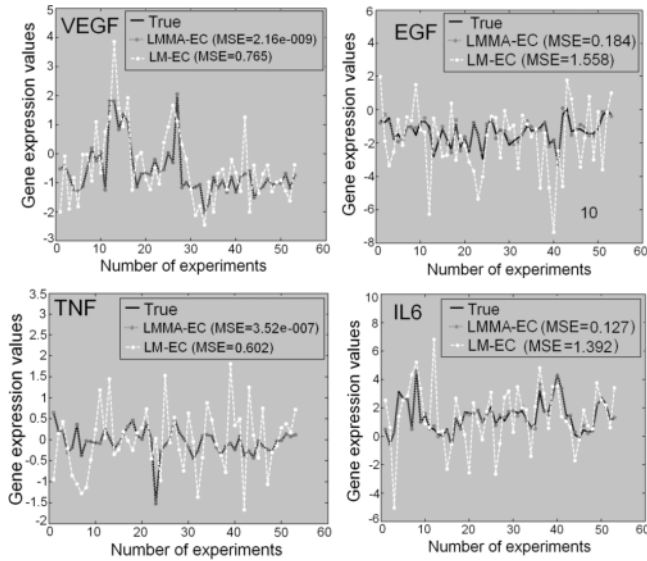


Fig. 2. Gene expression values derived from the leave one out cross validation approach for four hub genes VEGF, EGF, TNF and IL6 in both LM-EC and LMMA-EC networks. A total of 53 experiments in EC microarray dataset are tested.

LMMA-based network with significant P -values. Thus, pathways in LMMA-based network are significantly enriched.

4 DISCUSSION AND CONCLUSION

High false positive rate is a well-known problem in most high-throughput methods for detecting molecular interactions (von Mering *et al.*, 2002). In this work, we developed a LMMA approach to construct networks based on both existing knowledge (literature) and experimental information (microarray). Such approach performs multivariate analysis to modify the literature-derived holistic network using subject-oriented gene expression profiles. To analyze the hidden network buried in microarray datasets, two aspects make it necessary to construct the LM-based network beforehand. First, it is not advisable to construct the network directly from thousands of candidate variables if prior knowledge about the network is not available. Second, the number of variables should not exceed the number of observations (i.e. microarray experiments); otherwise the results will be falsely optimized. Thus, a certain number of arrays are required in LMMA for multivariate selection.

As an application, PubMed literatures and microarray datasets from both the EC and the ST are selected respectively to reconstruct the LMMA network for angiogenesis. The LMMA approach results in a larger cluster size, and a smaller average path length when comparing with a LM-random filtering, while preserves similar topological properties comparing with the LM-based network. Therefore, it indicates that LMMA can eliminate redundant relations while maintain the backbone of the LM-based network.

Angiogenesis networks constructed by LM and LMMA are tested for accuracy on confident sets of interactions. Both precision and recall rates are calculated against KEGG, one commonly used benchmark. We show that LMMA significantly improves the precision rate when comparing with LM alone. On the other hand, as Bork *et al.* (2004) reported, the choice of benchmark set is still a

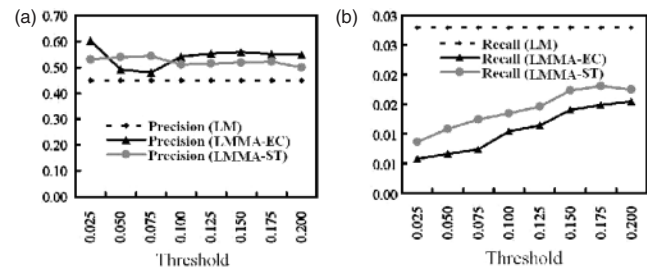


Fig. 3. Comparison of (a) precision and (b) recall in LM, LMMA-EC and LMMA-ST angiogenesis networks at different thresholds. Here LM represents LM-EC and LM-ST since genes in LM-EC and LM-ST are identical when mapping to KEGG. The X axis denotes the P -value thresholds calculated from F -test in the step of statistical multivariate selection. Both the precision and the recall rates are calculated against KEGG.

knotty problem because the agreement among different benchmark sets is surprisingly poor. For example, less than half of all pairs in the KEGG benchmark set are present in the Gene Ontology biological process benchmark set (Bork *et al.*, 2004). Moreover, it is commonly known that co-occurrence in literature often describes or reflects more general relationships between genes. Some of these may be implicit and/or so novel that they have not yet reached the status of common knowledge or accepted fact often required for inclusion in databases such as KEGG. Two aspects mentioned above may be the reason why both the LM and LMMA approaches resulted in a low recall rate (Fig. 3) when calculated against KEGG. Even so, we still show that the integration with microarray data can significantly increase the reliability of gene co-occurrence networks extracted from the literature.

To demonstrate how LMMA reduces the false positive rate and improves the precision, we select, EGF (epidermal growth factor), a key player in angiogenesis as an example. As shown in Figure 4, LMMA totally removes 21 EGF false related genes from LM-based EGF unit-network. First, LMMA deletes five mis-matched genes in LM: SC, SF, AA and PC are abbreviations of stem cells, scatter factor, arachidonic acid or anaplastic astrocytoma, and prostate cancer respectively; IL8RA (interleukin 8 receptor, alpha) is misinterpreted by IL8 and EGF receptor in the lexical order. Second, LMMA cancels eight genes with unknown relations (few co-citation) to EGF in LM: CCR6, FGF16, MAP3K8 and EGF are co-cited in only one PubMed sentence recorded in a gene expression experiment (Gerritsen *et al.*, 2003); the same as IL11, IL10, IL3, IL4 and CCR2. Third, LMMA removes eight genes that seldom have co-occurrences with EGF even by using their alias: NRG2, Scube1, NPY6R, ZNF78L2, IFI44, RNU106, AXPC1 and ANGPTL6. Thus, our results indicate that common errors, which lead to the false relations in LM, can be effectively removed by the LMMA approach.

Moreover, there are 11 most statistically significant KEGG pathways in the LMMA-based angiogenesis networks. See Table 4 for the detailed P -value of each pathway calculated by Fisher Exact Test. Among them the focal adhesion pathway, the adherens junction pathway and the regulation of actin cytoskeleton pathway contribute to the complex processes such as endothelial cell migration, morphogenesis and angiogenesis (Bix *et al.*, 2004). TGF-beta regulates angiogenesis by affecting proliferation, differentiation and

Table 3. The true positive (TP), false positive (FP) and the statistical *P*-values of TP/FP ratio (by Fisher Exact Test) between LM and LMMA networks^a

KEGG	Network	Thp	0.025	0.050	0.075	0.100	0.125	0.150	0.175	0.200
KG	LM	TP	237	237	237	237	237	237	237	237
	LM	FP	1048	1048	1048	1048	1048	1048	1048	1048
	LMMA-EC	TP	39	49	56	76	83	98	108	111
	LMMA-EC	FP	121	175	241	267	303	349	417	458
	TP/FP (LMMA-EC versus LM)	<i>P</i> -value	0.017004	0.034679	0.064785	0.01837	0.02372	0.01526	0.03012	0.04408
	LMMA-ST	TP	71	83	93	101	111	130	137	135
KO	LMMA-ST	FP	223	300	350	392	436	471	499	513
	TP/FP (LMMA-ST versus LM)	<i>P</i> -value	0.0056928	0.021672	0.02762	0.032833	0.033552	0.013196	0.013274	0.021953
	LM	TP	139	139	139	139	139	139	139	139
	LM	FP	170	170	170	170	170	170	170	170
	LMMA-EC	TP	29	33	37	52	57	70	74	77
	LMMA-EC	FP	19	34	40	44	46	55	60	63
KO	TP/FP (LMMA-EC versus LM)	<i>P</i> -value	0.017279	0.087676	0.090294	0.027146	0.017372	0.0097982	0.011667	0.011803
	LMMA-ST	TP	43	54	62	67	73	86	90	87
	LMMA-ST	FP	38	46	52	64	69	80	82	87
	TP/FP (LMMA-ST versus LM)	<i>P</i> -value	0.042857	0.026912	0.020102	0.041336	0.036214	0.028095	0.023083	0.04316

^aHere LM represents LM-EC and LM-ST since genes in LM-EC and LM-ST are identical when mapping to KEGG database. KG = KEGG Gene; KO = KEGG Orthology.

Table 4. KEGG pathways with significant *P*-values in LMMA-based angiogenesis networks (Thp = 0.150)^a

	LMMA-EC (KG)	LMMA-EC (KO)	LMMA-ST (KG)	LMMA-ST (KO)
Focal adhesion pathway	0.00087	1.09e-07	0.00084	2.84e-08
MAPK signaling pathway	0.02825	0.013338	0.015779	0.00910
Adherens junction	2.14e-22	1.31e-13	3.08e-24	2.19e-14
TGF-beta signaling pathway	0.00010	0.00540	8.76e-06	0.00585
Insulin signaling pathway	1.27e-06	0.00264	1.33e-07	0.00225
Calcium signaling pathway	0.00011	0.00373	1.86e-08	6.66e-05
Wnt signaling pathway	—	0.03010	—	0.00548
Regulation of actin cytoskeleton	0.01100	—	0.00020	—
Cytokine-cytokine receptor interaction	5.13E-09	—	9.52E-16	—
Apoptosis	0.00127	—	0.03001	—
Cell cycle	—	0.04594	—	0.02220

^a*P*-values are calculated from Fisher Exact Test. KG = KEGG Gene. KO = KEGG Orthology.

migration of endothelial cells (Lomnyska *et al.*, 2004). Insulin signaling pathway is implicated in cellular mitogenesis, angiogenesis, tumor cell survival and tumorigenesis (Cohen *et al.*, 2005). Many Wnt proteins act through a canonical, beta-catenin signaling pathway (Masckauchan *et al.*, 2005) and are able to control diverse biological processes, such as cell differentiation, proliferation (Masckauchan *et al.*, 2005) and vasculature (Goodwin and D'Amore, 2002). Among the intracellular kinases implicated in angiogenesis, p38 MAPK has been shown to transduce signals critical for vascular remodeling and maturation (Zhu *et al.*, 2003). Ca(2+) signaling is involved in virtually all cellular processes (Munaron *et al.*, 2004). In addition, a variety of stimulatory cytokines, such as tumor necrosis factor (TNF)-alpha, interleukin (IL)-1, -6 and interferon (IFN)-gamma, and growth factors can promote the development of functional and structural vascular changes (Kofler *et al.*, 2005). Therefore, pathway information in the LMMA-based angiogenesis network suggests that multiple pathway interactions boost the activity of either EC or ST, which are in accordance with recent reports (Mukhopadhyay and Datta, 2004; McCarty,

2004). Since multiple pathways are dysfunctional in angiogenesis related disorders such as cancers, a multifocal signal modulation therapy is proposed recently (McCarty, 2004). And LMMA network will be helpful for analyzing the interactions of multiple pathways in such complex biological processes.

As for the usability of LMMA, this system is flexible in application to any biological topic if the related literature and microarray data are available. Note that to construct a LMMA network, the number of all candidate variables (genes) should be controlled in a proper size, and the accuracy of the LMMA approach increases with the increasing number of candidate variables in a certain scope. For the LMMA-based angiogenesis network, it summarizes large amounts of angiogenesis related literatures and high-throughput microarray data. The LMMA approach enables researchers not only to keep up-to-date with all the relevant literature on specialized biological topics, but also to make sense of the relevant large-scale microarray dataset. Meanwhile, the LMMA approach serves as a useful tool for constructing specific biological network and experimental design. Thus, LMMA acts as a valuable computer

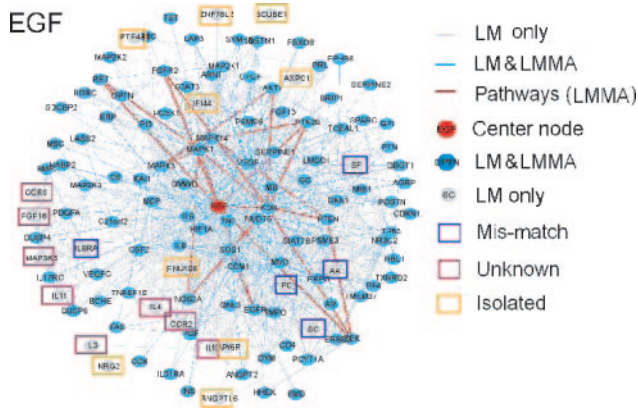


Fig. 4. An EGF (epidermal growth factor) unit-network derived respectively from the co-occurrence literature mining and the LMMA approaches. A total of 21 genes co-cited with EGF in LM are removed by LMMA. By manually revisiting the PubMed records, these 21 genes are found in false relations with EGF resulted from homonymic mis-matches and confused lexical orders (in the blue pane), unknown relations in the Graphviz software (AT&T; <http://www.research.att.com/sw/tools/graphviz/>) is adopted to visualize the constructed network.

representation of the known angiogenesis-related pathways, as well as the interactions among multiple pathways. Such representation will enable a systemic recognition for angiogenesis in the context of complex gene interactions, which is also helpful for studying the regulation of various complex biological, physiological and pathological systems. In the ‘omics’ field, the LMMA approach can be further explored to study protein–protein and other interactions.

ACKNOWLEDGEMENTS

The authors would like to express their great appreciation to B. Li (Boston University, USA), X. G. Zhang and C. Zhang in their lab for helpful discussions and comments. The authors would like to acknowledge the financial support from FANEDD (No. 200366), the Key Project of Chinese MOE (No. 104009) and the Basic Research Foundation of TNLIST.

Conflict of Interest: none declared.

REFERENCES

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
 Bix,G. *et al.* (2004) Endorepellin causes endothelial cell disassembly of actin cytoskeleton and focal adhesions through alpha2beta1 integrin. *J. Cell Biol.*, **166**, 97–109.
 Bork,P. *et al.* (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.*, **14**, 292–299.
 Carmeliet,P. (2003) Angiogenesis in health and disease. *Nat. Med.*, **9**, 653–660.
 Cary,M.P. *et al.* (2005) Pathway information for systems biology. *FEBS Lett.*, **579**, 1815–1820.
 Cohen,B.D. *et al.* (2005) Combination therapy enhances the inhibition of tumor growth with the fully human anti-type 1 insulin-like growth factor receptor monoclonal antibody CP-751,871. *Clin. Cancer Res.*, **11**, 2063–2073.
 Dennis,G.,Jr *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, R60.
 D’haeseleer,P. *et al.* (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.*, 41–52.

D’haeseleer,P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.
 de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.
 Ding,J. *et al.* (2002) Mining Medline: abstracts, sentences, or phrases? *Pac. Symp. Biocomput.*, **7**, 326–337.
 Folkman,J. (1995) Angiogenesis in cancer, vascular, rheumatoid and other diseases. *Nat. Med.*, **1**, 27–31.
 Ge,H. *et al.* (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.*, **29**, 482–486.
 Gerritsen,M.E. *et al.* (2003) Using gene expression profiling to identify the molecular basis of the synergistic actions of hepatocyte growth factor and vascular endothelial growth factor in human endothelial cells. *Br. J. Pharmacol.*, **140**, 595–610.
 Goodwin,A.M. and D’Amore,P.A. (2002) Wnt signaling in the vasculature. *Angiogenesis*, **5**, 1–9.
 Han,J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, **430**, 88–93.
 Jenssen,T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.
 Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
 Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids. Res.*, **28**, 27–30.
 Kofler,S. *et al.* (2005) Role of cytokines in cardiovascular diseases: a focus on endothelial responses to inflammation. *Clin. Sci (Lond.)*, **108**, 205–213.
 Küffner,R. *et al.* (2005) Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics*, **21**, ii259–ii267.
 Lachenbruch,P.A. and Mickey,M.R. (1968) Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.
 Le Phillip,P. *et al.* (2004) Using prior knowledge to improve genetic network reconstruction from microarray data. *In Silico Biol.*, **4**, 335–353.
 Liang,S. *et al.* (1998) Reveal a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, **3**, 18–29.
 Lomnyska,M. *et al.* (2004) Transforming growth factor-beta1-regulated proteins in human endothelial cells identified by two-dimensional gel electrophoresis and mass spectrometry. *Proteomics*, **4**, 995–1006.
 Mao,X. *et al.* (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21**, 3787–3793.
 Masckauchan,T.N. *et al.* (2005) Wnt/beta-catenin signaling induces proliferation, survival and interleukin-8 in human endothelial cells. *Angiogenesis*, **8**, 43–51.
 McCarty,M.F. (2004) Targeting multiple signaling pathways as a strategy for managing prostate cancer: multifocal signal modulation therapy. *Integr. Cancer Ther.*, **3**, 349–380.
 Mukhopadhyay,D. and Datta,K. (2004) Multiple regulatory pathways of vascular permeability factor/vascular endothelial growth factor (VPF/VEGF) expression in tumors. *Semin. Cancer Biol.*, **14**, 123–130.
 Munaron,L. *et al.* (2004) Blocking Ca²⁺ entry: a way to control cell proliferation. *Curr. Med. Chem.*, **11**, 1533–1543.
 Ozier,O. *et al.* (2003) Global architecture of genetic interactions on the protein network. *Nat. Biotechnol.*, **21**, 490–491.
 Segal,M.R. *et al.* (2003) Regression approaches for microarray data analysis. *J. Comput. Biol.*, **10**, 961–980.
 Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.
 Sherlock,G. *et al.* (2001) The Stanford Microarray Database. *Nucleic Acids. Res.*, **29**, 152–155.
 Song,C. *et al.* (2005) Self-similarity of complex networks. *Nature*, **433**, 392–395.
 Stapley,B.J. and Benoit,G. (2000) Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.*, 529–540.
 Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
 van Someren,E.P. *et al.* (2002) Genetic network modeling. *Pharmacogenomics*, **3**, 507–525.
 von Mering,C. *et al.* (2002) Comparative assessment of large scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
 West,M. *et al.* (2001) Predicting the clinical status of human breast cancer using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

- Wu,L.J. and Li,S. (2005) Combined literature mining and gene expression analysis for modeling neuro-endocrine-immune interactions. *Lect. Notes Comput. Sci.*, **3645**, 31–40.
- Zhang,C. and Li,S. (2004) Modeling of neuro-endocrine-immune network via subject oriented literature mining. *Proc. BGRS*, **2**, 167–170.
- Zhu,S. et al. (2005) A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics*, **21**, ii245–ii251.
- Zhu,W.H. et al. (2003) Requisite role of p38 MAPK in mural cell recruitment during angiogenesis in the rat aorta model. *J. Vasc. Res.*, **40**, 140–148.