An Integrative Approach to Identifying Biologically Relevant Genes

Zheng Zhao[†] Jiangxin Wang[‡] Shashvata Sharma[†]
Nitin Agarwal[†] Huan Liu[†] Yung Chang[‡]
† Department of Computer Science and Engineering, Arizona State University
‡ School of Life Science, CIDV, The Biodesesign Institute, Arizona State University
{zhaozheng, jiangxin.wang, sssharma, agarwal.nitin, huan.liu, yung.chang}@asu.edu

Abstract

Gene selection aims at detecting biologically relevant genes to assist biologists' research. The cDNA Microarray data used in gene selection is usually "wide". With more than several thousand genes, but only less than a hundred of samples, many biologically irrelevant genes can gain their statistical relevance by sheer randomness. Addressing this problem goes beyond what the cDNA Microarray can offer and necessitates the use of additional information. Recent developments in bioinformatics have made various knowledge sources available, such as the KEGG pathway repository and Gene Ontology database. Integrating different types of knowledge could provide more information about genes and samples. In this work, we propose a novel approach to integrate different types of knowledge for identifying biologically relevant genes. The approach converts different types of external knowledge to its internal knowledge, which can be used to rank genes. Upon obtaining the ranking lists, it aggregates them via a probabilistic model and generates a final list. Experimental results from our study on acute lymphoblastic leukemia demonstrate the efficacy of the proposed approach and show that using different types of knowledge together can help detect biologically relevant genes.

1 Introduction

Selecting genes that are critical to a particular biological process has been a major challenge in post-array analysis [14, 16, 32]. Also known as feature selection [10, 21, 22] in machine learning research area, gene selection has attracted intensive research interests and much progress has been made over the last decade in developing effective gene selection algorithms [19, 32, 12, 26]. Given cDNA Microarray data, most existing algorithms try to identify genes that are differentially expressed over the samples. Discriminative genes help classifiers or clustering algorithms to achieve high accuracy [20, 7, 17]. However, does the better accuracy necessarily indicate higher biological relevance of genes? We applied a supervised gene selection algorithm, Fisher score [5] and an unsupervised algorithm, SPEC [44] on the

expression profiling of bone marrow from 18 pediatric patients with acute lymphoblastic leukemia (ALL) [29] to select genes that may provide insight into the pathogenesis of pediatric ALL. The top 20 genes selected by the two algorithms are examined by our biologist collaborators. Table 1 contains a list of the biologically relevant genes identified by the biologists, and the accuracy achieved by the $k$nn classifier on the selected genes. The result shows that a gene list of higher accuracy does not necessarily contain more relevant genes. Hence, selecting genes to achieve high accuracy should not be the sole goal of biological discovery.

Table 1: Biologically relevant genes identified by two gene selection algorithms for childhood ALL.

Unsupervised (ACC: 0.61, REL: 7)			
SFRS5	TM9SF1	WTAP	GPSM3
STAC3	POMP	SLC25A6	
Supervised (ACC: 0.97, REL: 4)			
USP33	IL2RG	SIGIRR	CHCHD2

There could be two sensible explanations. First, a cNDA Microarray data usually contains more than several thousand genes but only fewer than 100 samples. A data set of this kind usually leads to the small sample problem [31]. With so few samples, many genes, which are not biologically relevant, can easily gain their statistical relevance due to randomness [36]. Second, even genes that are related, may have different importance. For instance, to understand a specific biological process, the genes acting as the "trigger" are much more important than the genes acting as the "fire". Therefore, sometimes, the genes that act as the "fire" are not considered as relevant in biologists' study. Addressing these problems goes beyond what the cDNA Microarray data can offer, and necessitates the need for additional information to conduct effective gene selection. Recent developments in bioinformatics have made various knowledge sources available, including the KEEG pathway repository [15], the Gene Ontology database [4] and the NCI Gene-Cancer database [35], etc. Recent work has also revealed the existence of a class of small non-coding RNA species known

as microRNAs, which are surprisingly informative for identifying cancerous tissues [24]. The availability of these various knowledge sources presents unprecedented opportunities to advance research solving previously unsolvable problems. In this work, we propose to develop a platform to study the novel problem of integrating multiple knowledge sources in the process of gene selection for identifying biologically relevant genes. The major challenge in this work is how to address the heterogeneity in different knowledge sources.

Researchers have tried to use various types of knowledge to assist gene selection. For instance, the authors in [1] propose to use different types of knowledge about genes to calculate gene similarity, which is then used to identify genes that are closest to the given example genes. In [40] the authors focus on using gene sets, which are groups of genes that share common biological functions, chromosomal locations, or regulations to interpret the gene selection outputs. In [28], gene annotation are used for choosing gene ranking criterion. In [2], protein interaction, gene-disease association and gene function annotation are used for choosing cancer related genes. Gene selection approaches using gene regulatory network and gene ontology are also studied in [18] and [30, 38], respectively. Since most existing work is designed for specific research purposes, they can only handle one or limited types of knowledge of the same category. For instance, the models proposed in [40, 1, 2] can only handle knowledge about genes, but not knowledge about samples. To address this limitation, we propose a integrative approach to systematically incorporate different types of knowledge in gene selection. The approach is based on a probabilistic model for aggregating gene ranking lists, which is obtained by using different types of knowledge. The approach is extensively experimented and tested. Experimental results from our pediatric acute lymphoblastic leukemia (ALL) study show that judiciously using different types knowledge can bring about significant performance improvement to assist biological discovery.

2 An Integrative Approach for Gene Selection

We propose to develop a general approach for systematically integrating different types of knowledge to achieve Knowledge-Oriented Gene Selection, which is named KOGS. Figure 1 presents the major steps in the approach: (1) **Knowledge Conversion** - knowledge understandable for human beings may not be directly applicable in a learning model. Therefore, the first step is to covert different types of human or external knowledge to certain types of internal knowledge that can be used by gene selection algorithms. Assume we have L different external knowledge sources $\mathcal{K}_1^{ext}, \ldots, \mathcal{K}_L^{ext}$. For the ith external knowledge, we can apply a conversion operator $c_i(\cdot)$ to convert the external knowledge $\mathcal{K}_i^{ext}$ to the corresponding internal knowledge $\mathcal{K}_i^{int}$, and this allows us to formalize knowledge conversion
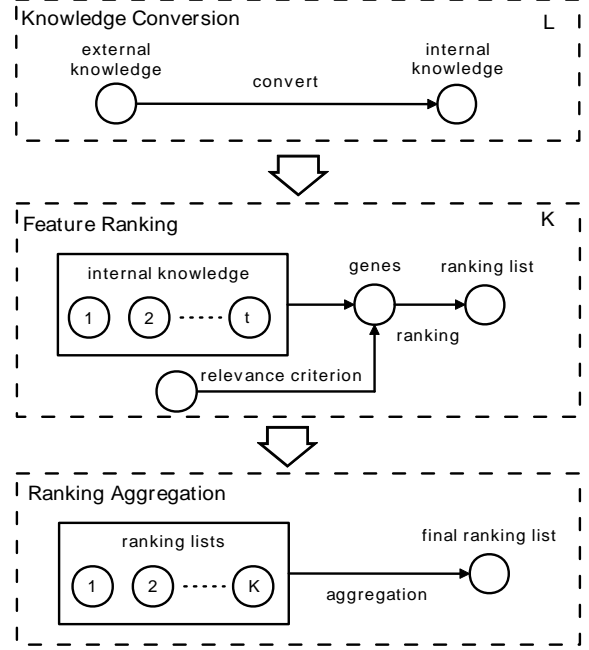

Figure 1: An approach for integrating different types of domain knowledge to assist gene selection.

with the following equation:

$$(2.1) \qquad \mathcal{K}_i^{int} = c_i\left(\mathcal{K}_i^{ext}\right), \; i = 1, \ldots, L$$

(2) **Feature Ranking** - assume we decide to use K sets of internal knowledge $KNOW_1, \ldots, KNOW_K$ to rank genes, where $KNOW_i$ is defined as: $KNOW_i = \left\{\mathcal{K}_{i_1}^{int} \ldots \mathcal{K}_{i_{t_i}}^{int}\right\}$. Let $\mathcal{C}_i$ be a relevance criterion, $\mathbf{G} = \{g_1, \ldots, g_M\}$ be a set of M genes, and $\mathcal{R}_i(\cdot)$ be a gene ranking function, the task of feature ranking is to use the internal knowledge with the given criterion to rank the relevance of the genes in $\mathbf{G}$, which can be formulated as:

$$(2.2) \qquad R_i^{rank} = \mathcal{R}\left(KNOW_i, \mathcal{C}_i, \mathbf{G}\right)$$

(3) **Rank Aggregation** - after obtained the K ranking lists, they need to be integrated to obtain a final ranking to estimate the relevance of the genes. Let $\mathcal{A}(\cdot)$ be an aggregating operator for ranking lists and $\mathcal{C}$ be an aggregation criterion, we use $\mathcal{A}(\cdot)$ to aggregate the K ranking lists, which can be formulated as:

$$(2.3) \qquad R_F^{rank} = \mathcal{A}\left(R_1^{rank}, \ldots, R_K^{rank}, \mathcal{C}\right)$$

The final gene ranking list can be obtained by considering the ranking lists from all internal knowledge sets in either a supervised or an unsupervised fashion, depending upon how $\mathcal{C}$ is specified. Next, we will study: (1) How to categorize the external knowledge sources; which types of knowledge should be used as the internal knowledge; and how to define

the converting operators $c(\cdot)$ to convert different types of external knowledge to internal knowledge; and (2) Given a set of internal knowledge and a relevance criterion, how to define the ranking operator $\mathcal{R}(\cdot)$ to rank genes; and how to effectively aggregate obtained ranking lists to obtain a final ranking list, in search of biologically relevant genes.

3 Handling Knowledge in KOGS

Different types of external knowledge and internal knowledge need to be handled properly in KOGS to achieve effective gene selection. We now study how to categorize different types of publicly available (i.e., external) knowledge sources and define the types of the internal knowledge that can be used in KOGS. We also show how to convert different types of external knowledge to corresponding internal knowledge.

3.1 External Knowledge Various types of external knowledge sources can be used in gene selection. We categorize them into two groups: the knowledge about genes, and the knowledge about samples. The knowledge about genes usually contains information about the properties of genes or their relationships. Figure 2 presents three different types of knowledge about genes to be used in gene selection: (a) metabolic pathway, which depicts a series of biochemical reactions occurring in cells and reflects how genes interact with each other to accomplish a specific function; (b) gene ontology (GO) annotation [4], which uses a controlled vocabulary to describe the characteristics of genes; and (c) gene sequence, which describes the order of the nucleotide bases of genes. The figure shows that the three types of knowledge have heterogenous representations. The nature of the knowledge determines how it can be used in gene selection. According to the way knowledge is used in gene selection, we further divide different types
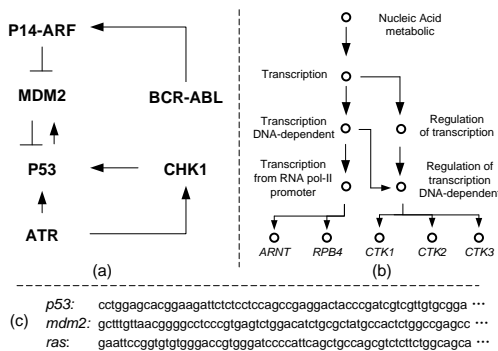

Figure 2: An example of three different types of knowledge about genes, (a) Metabolic Pathway, (b) Gene Ontology Annotation, and (c) Gene Sequence.
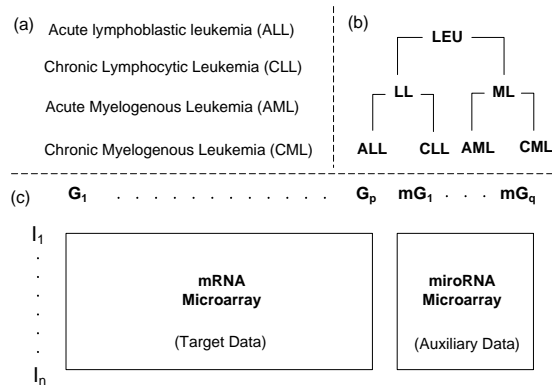

Figure 3: Different types of knowledge about samples, (a) the class label information, (b) sample hierarchy, and (c) an example of the auxiliary data.

of knowledge into three categories: (1) knowledge about gene similarity, $\mathcal{K}_{SIM}^{ext}$, for example, with gene sequence information, gene similarities can be obtained by applying a sequence alignment algorithm. (2) Knowledge of gene functions, $\mathcal{K}_{FUN}^{ext}$, for instance, in a metabolic pathway, a set of genes act together to accomplish particular biological functions; and in gene ontology annotation, the functions of genes are also provided. (3) Knowledge of gene interaction, $\mathcal{K}_{INT}^{ext}$, for example, in the BioGRID [39], over 198000 genetic interactions related to different types of biological functions or processes are recorded. The knowledge of genes is usually accumulated and cross-examined by human researchers in their research by generalizing evidences from multiple experiments, therefore, is relatively reliable, and independent of any specific experiment.

The knowledge of samples usually is about sample categories, $\mathcal{K}_{CAT}^{ext}$, or samples' geometric relationship, $\mathcal{K}_{GEO}^{ext}$. Samples can be categorized with either a flat structure (as shown in Figure 3-(a), which forms the standard class label) or a hierarchical structure, as shown in Figure 3-(b). The geometric relationship among samples, depicted by the pairwise sample similarity, can be derived from a given auxiliary data. Auxiliary data refers to the data collected from the same set of samples that generates the cDNA Microarray, which is the target data for gene selection. The target and the auxiliary data depict the same set of samples, while using different measurements. Auxiliary data may help us get a better understanding of the geometric pattern of the samples. For example, as shown in Figure 3-(c), for gene selection, the microRNA Microarray can serve as auxiliary data, which measures the microRNA expression of samples. cDNA Microarray and microRNA Microarray are collected from the same set of samples. Compared to cDNA Microarray, microRNA Microarray contains only several hundreds of microRNA and are found to be surprisingly informative in separating tissues of cancer and noncancer, as well as differ-

Table 2: The categories and examples of different types of knowledge that can be used in gene selection.

Knowledge	Samples	$\mathcal{K}^{ext}_{CAT}$ - Category	Class Label, Sample Hierarchy
		$\mathcal{K}^{ext}_{GEO}$ - Geometry	miRNA Expression Profile, mRNA Expression Profile
	Genes	$\mathcal{K}^{ext}_{SIM}$ - Similarity	Gene Sequence, Gene Ontology Annotation, Gene Lineage, Gene Locus
		$\mathcal{K}^{ext}_{FUN}$ - Function	Gene Ontology Annotation, Metabolic Pathway, Gene-Disease Association
		$\mathcal{K}^{ext}_{INT}$ - Interaction	Metabolic Pathway, Protein-Protein Interaction

ent types of cancers [13]. Using microRNA Microarray as auxiliary data helps improve our understanding about how cancerous tissues cluster together. Comparing with knowledge about genes, the auxiliary data links to individual experiment, therefore is more specific.

Table 2 summarizes different categories of knowledge that can be used in gene selection. We noticed that some types of knowledge fall into more than one categories. For instance, gene ontology annotation can be used for obtaining the knowledge of both gene similarities, e.g. by comparing shared annotation terms among genes, and gene functions, e.g. by finding out the annotation terms related to specific functions of interest. Different types of knowledge have heterogenous representations and describe genes or samples from different perspectives. The categorization of different types of knowledge helps us generalize the common characteristics of the knowledge from the same category, so that a common approach can be applied on all types of knowledge in that category for knowledge conversion.

3.2   Internal Knowledge While defining internal knowledge, the following two issues should be considered. First, the definition should ensure that certain types of external knowledge can be easily converted to its form. Second, it can be effectively used to rank genes. Based on these two considerations, in KOGS, we use the following types of knowledge: knowledge about samples, (1) sample category, $\mathcal{K}^{int}_{CAT}$, (2) sample geometric pattern, $\mathcal{K}^{int}_{GEO}$; and knowledge about genes: (3) gene connection, $\mathcal{K}^{int}_{CON}$, and (4) gene function, $\mathcal{K}^{int}_{FUN}$. Here the gene connection can either refer to the similarity among genes or interaction among genes, since both types of knowledge provides us the information about how genes are connected. Later on, we will show how to propagate gene relevance on the network derived from $\mathcal{K}^{int}_{CON}$. KOGS is not restricted to the four types of internal knowledge defined above. As long as new knowledge can be used to rank genes, it can be treated as a type of internal knowledge. This ensures the extendability of KOGS. While in real applications, we found that most available external knowledge in gene selection can be conveniently converted to one of the four types of internal knowledge. Next we study how to effectively convert various types of external knowledge to internal knowledge.

Table 3: The conversion of different types of external knowledge to internal knowledge.

External Knowledge	Internal Knowledge
$\mathcal{K}^{ext}_{GEO}$, $\mathcal{K}^{ext}_{FUN}$, $\mathcal{K}^{ext}_{SIM}$	$\mathcal{K}^{int}_{GEO}$
$\mathcal{K}^{ext}_{SIM}$, $\mathcal{K}^{ext}_{INT}$	$\mathcal{K}^{int}_{CON}$
$\mathcal{K}^{ext}_{FUN}$	$\mathcal{K}^{int}_{FUN}$
$\mathcal{K}^{ext}_{CAT}$	$\mathcal{K}^{int}_{CAT}$

3.3   Knowledge Conversion We study how to convert external knowledge to internal knowledge. Table 3 contains the information of mapping different types external knowledge to the corresponding internal knowledge. The conversions of $\mathcal{K}^{ext}_{GEO} \rightarrow \mathcal{K}^{int}_{GEO}$, $\mathcal{K}^{ext}_{CAT} \rightarrow \mathcal{K}^{int}_{CAT}$, $\mathcal{K}^{ext}_{SIM} \rightarrow \mathcal{K}^{int}_{CON}$, $\mathcal{K}^{ext}_{FUN} \rightarrow \mathcal{K}^{int}_{FUN}$, and $\mathcal{K}^{ext}_{INT} \rightarrow \mathcal{K}^{int}_{CON}$ are straightforward. For example, $\mathcal{K}^{ext}_{SIM}$, the similarity among genes, and $\mathcal{K}^{ext}_{INT}$, the interaction among genes, can be directly used to construct gene connection graphs, corresponding to $\mathcal{K}^{int}_{CON}$. Below, we show how to perform conversions: $\mathcal{K}^{ext}_{SIM} \rightarrow \mathcal{K}^{int}_{GEO}$ and $\mathcal{K}^{ext}_{FUN} \rightarrow \mathcal{K}^{int}_{GEO}$. The geometric pattern of samples, depicted by the pairwise sample similarity, reflects the structure of the underlying model and is important for building robust learning models [34]. The pairwise distance can also be conveniently used in well studied distance based gene selection algorithms. Figure 4 shows how to convert $\mathcal{K}^{ext}_{SIM}$ and $\mathcal{K}^{ext}_{FUN}$ to $\mathcal{K}^{int}_{GEO}$. The basic idea is to involve the two types of knowledge in the calculation of the pairwise similarity among samples.
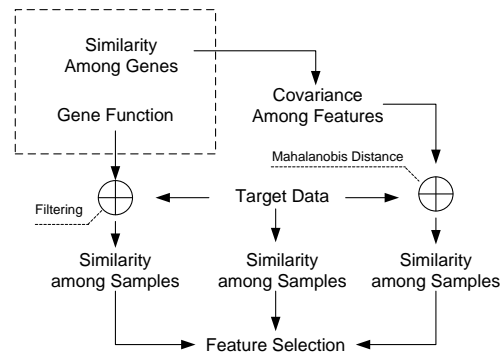

Figure 4: Obtaining the knowledge of sample geometry, using different types of knowledge of genes.

3.3.1 $\mathcal{K}_{SIM}^{ext} \rightarrow \mathcal{K}_{GEO}^{int}$ Given similarities among genes, gene covariance can be constructed and used in calculating the pairwise sample similarity via Mahalanobis distance [25], which is defined as:

$$(3.4) \qquad \|\mathbf{x} - \mathbf{y}\|_M^2 = (\mathbf{x} - \mathbf{y})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{y}).$$

In the equation, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$ are two samples with M genes $g_1, \ldots, g_M$, and $\mathbf{C} \in \mathbb{R}^{M \times M}$ is the covariance matrix. In comparison to the standard Euclidian distance, Mahalanobis distance provides a better way to determine the similarities among samples by considering the probability distribution of the underlying model, and the ellipsoid best representing the probability distribution can be estimated from $\mathbf{C}$ [11]. In real applications, $\mathbf{C}$ is usually estimated from the data by the following equation:

$$(3.5) \qquad \mathbf{C} = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T,$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are the N samples of the data, with $\bar{\mathbf{x}}$ being their mean. Although Equation (3.5) specifies an unbiased estimator of the covariance matrix, when sample size is small, it may return a poor estimation. Instead of using the data, the covariance matrix can also be obtained from our knowledge about gene similarities, which may provide another (more stable and reliable) way for estimating $\mathbf{C}$. The following proposition shows how to construct the covariance matrix from $\mathcal{K}_{SIM}^{ext}$, the knowledge of gene similarity.

PROPOSITION 3.1. *Given gene similarity matrix* $W \in \mathbb{R}^{M \times M}$ *of the M genes, with W_{ij} specifying the similarity between genes g_i and g_j. Let D be a diagonal matrix with d_{ii} = $\sum_k w_{ik}$, then $K = (D - W)^+$ specifies a kernel. Using its embedding, the covariance matrix can be obtained by:*

$$(3.6) \qquad \mathbf{C} = K \left(I - \frac{1}{l} U \mathbf{1} \mathbf{1}^T U^T \right) K.$$

In the proposition, l is the number of involved genes, $\mathbf{1}$ is the vector with 1 as its only elements. $(\cdot)^+$ denotes the pseudo-inverse and $K = U \Sigma U^T$ is the SVD [8] of K.

3.3.2 $\mathcal{K}_{FUN}^{ext} \rightarrow \mathcal{K}_{GEO}^{int}$ In a biological study, some particular biological functions may be of special interests according to the research purpose. Given $\mathcal{K}_{FUN}^{ext}$, the knowledge of gene functions, and $\mathcal{F}$, a set of biological functions of interests, data can be filtered by the genes associated with $\mathcal{F}$,

$$(3.7) \qquad X_{\mathcal{F}} = \Pi_{G_{\mathcal{F}}} (X),$$

where $G_{\mathcal{F}}$ is the genes related to $\mathcal{F}$, and $\Pi(\cdot)$ is the projection operator. Using the filtered data $X_{\mathcal{F}}$, the pairwise sample similarity matrix W can be obtained through any similarity measure. Since all genes in $G_{\mathcal{F}}$ are related to the biological functions of interest, geometric distribution specified by W should reflect the distribution under the influence of the functions. In case the functions are closely related to the biological process under study, the distribution will give us an insight of the process, and help us to select biologically relevant genes. Using genes which are known to have a particular function as the seeds can also help us select genes that perform the function but are still unknown.

4 Ranking Genes with Knowledge

Having the various types of internal knowledge ready, we study how to use them to rank genes as well as how to combine various ranking lists to obtain a final list.

4.1 Ranking Using Internal Knowledge The internal knowledge can be used to rank genes in various ways. Selecting genes using $\mathcal{K}_{CAT}^{int}$, corresponding to traditional supervised gene selection algorithms, has been well studied. Below we show how to rank genes using the other three types of internal knowledge.

4.1.1 Geometric Consistency Checking with $\mathcal{K}_{GEO}^{int}$ Given $\mathcal{K}_{GEO}^{int}$ carrying the distribution information of samples, one way to estimate gene relevance is to measure its consistency with the given distribution, called geometric consistency, which leads to distance based algorithms [44] for gene selection. The intuition is that the distribution of samples reflects the structure of the underlying model. For instance, samples that are near to each other usually belongs to the same category. Therefore selecting genes whose expressions are consistent with the distribution corresponds to select genes whose expression is influenced by (or influence) the underlying model. Here the consistency means that a gene expresses similarly on samples that are near to each. The geometric consistency can be measured by applying spectral analysis. Given $W_s \in \mathbb{R}^{N \times N}$ of N samples, the similarity samples matrix derived from $\mathcal{K}_{GEO}^{int}$, the laplacian matrix $L_s = D_s - W_s$ forms a consistency (or smoothness) estimator [37], where D_s is a diagonal matrix with d_{ii}^s = $\sum_k w_{ik}^s$. Let $\mathbf{g}$ be a vector carrying the expression levels of a gene over the N samples, the geometric consistency of $\mathbf{g}$ can be evaluated by:

$$(4.8) \qquad \mathbf{g}^T L_s \mathbf{g} = \sum_{i,j} w_{i,j}^s (g_i - g_j),$$

and the smaller the value, the more consistent the vector $\mathbf{g}$. This measurement is improved in [44] with:

$$(4.9) \qquad \varphi(\boldsymbol{g}_i) = \frac{\widehat{\boldsymbol{g}}_i^T \gamma(\mathcal{L}_s) \widehat{\boldsymbol{g}}_i}{1 - \left(\widehat{\boldsymbol{g}}_i^T \xi_0 \right)^2}.$$

In the equation, $\widehat{\boldsymbol{g}}_i = (D_s^{\frac{1}{2}} \boldsymbol{g}_i) \cdot \|(D_s^{\frac{1}{2}} \boldsymbol{g}_i)\|^{-1}$ is the normalized feature vector; $\mathcal{L} = D_s^{-\frac{1}{2}} L_s D_s^{-\frac{1}{2}}$ is the normalized laplacian

matrix; and $\gamma(\cdot)$ is a spectral matrix function [8], induced from an increasing real function, which is used to rescale the eigenvalues of $\mathcal{L}_s$ for reducing noise. As shown in [44], compared to Equation (4.8), Equation (4.9) is more robust to noise and has better performance. In this work, we use Equation (4.9) to measure the geometric consistency.

4.1.2 Relevance Propagating with $\mathcal{K}_{CON}^{int}$
Given $\mathcal{K}_{CON}^{int}$, the knowledge of gene connections, we can derive a graph $\mathbb{G}$ to depict the knowledge. Given a set of genes $\mathcal{G} = \{g_1, \ldots, g_t\}$, which are known to be relevant, we can propagate their relevance on the graph to nearby nodes. Assuming $\mathcal{K}_{CON}^{int}$ is built from $\mathcal{K}_{SIM}^{ext}$, the knowledge of gene similarity, relevance propagation corresponds to the hypothesis that if a gene is relevant, the genes, which are similar to it, may also be relevant. We can formulate the idea using the concept from random walk theory. Assume W_g is the affinity matrix of $\mathbb{G}$, which is derived from $\mathcal{K}_{CON}^{int}$ of genes, the transition probability matrix is defined as:

$$
\begin{aligned}
P_g &= D_g^{-1} W_g, \\
D_g &= \mathrm{diag}\left(d_1^g, \ldots, d_M^g\right), \\
d_i^g &= \sum_k w_{ik}^g.
\end{aligned}
$$

Assuming $\mathbf{r}$ is the vector containing the initial relevance of genes, then the final relevance of genes is given by:

$$
\begin{aligned}
\mathbf{r}^* &= \mathbf{r} + \ldots + (\lambda P_g)^k \mathbf{r} + \ldots + (\lambda P_g)^\infty \mathbf{r} \\
(4.10) \quad &= (I - \lambda P_g)^{-1} \mathbf{r}.
\end{aligned}
$$

In the above equation, $(\lambda P_g)^k \mathbf{r}$ corresponds to the relevance gained by genes after k steps of propagation, and $0 < \lambda < 1$ is the decay parameter which is used to reduce the magnitude of the relevance when it is propagated from one node to another node. After obtained $\mathbf{r}^*$, genes can be ranked according to their corresponding value in $\mathbf{r}^*$.

4.1.3 Functional Relevance Voting with $\mathcal{K}_{FUN}^{int}$
The functions of genes, for example, the ones provided in the Gene Ontology (GO) database [4], are usually depicted by a controlled vocabulary. In this cases, the terms can be regarded as the hyper features of genes. Let g_i be the ith gene in the gene list, whose function is obtained from $\mathcal{K}_{FUN}^{int}$ and is described by a vector $\mathbf{f}_i = (f_{i,1}, \ldots, f_{i,T})$, where T is the total number of functions, and $f_{i,j} = 1$, if and only if gene i is related to function j, otherwise $f_{i,j} = 0$. Assume we know the relevance of all the functions, which is described by a vector $\mathbf{r}^{\mathrm{fun}} = \left(r_1^{\mathrm{fun}}, \ldots, r_T^{\mathrm{fun}}\right)$, the relevance of gene i can be obtained by the following equation:

$$
(4.11) \qquad r_i = \sum_{l=1}^T f_{i,l} r_l^{\mathrm{fun}}.
$$

The equation sums the relevance of all the functions related to the gene as its relevance sore. $\mathbf{r}_{\mathrm{fun}}$ can be either assigned by researchers according to their research purpose or learnt automatically. In the experimental part we will show how to learnt the relevance of the GO gene function annotation terms by using the gene-cancer association information.

4.2 Aggregating Gene Ranking Lists
Using different types of knowledge, we can obtain multiple lists that rank genes in different ways. Aggregating these rankings into a joint one has been studied as rank aggregation in both machine learning and information retrieval [33]. In this work we propose a probabilistic model for rank aggregation. While existing rank aggregation algorithms, such as the methods presented in [33], treat different ranking lists equally in the combination process, our proposed method is able to automatically learn a set of combination coefficients according to the importance of different ranking lists. And this is achieved by maximizing the likelihood of gene relevance in a given gene set. When the gene set only contains genes which are known to be relevant, the model achieves rank aggregation in a supervised way. When the gene set contains all genes, it combines ranking lists in an unsupervised way. Let g_i denote gene i, $1 \leq i \leq M$, and its rank in ranking list l be $r_{l,i}$, we define the probability of g_i to be relevant according to its rank in the ranking list l to be:

$$
\begin{aligned}
P(r_{l,i}) &= \frac{1}{B} \exp\left(\frac{1}{r_{l,i}}\right), \\
B &= \sum_{j=1}^M \exp\left(\frac{1}{j}\right).
\end{aligned}
$$

In the equation, B is the normalization factor for the distribution. For defining the probability, the exponential function $\exp(\cdot)$ is adopted to emphasize the top ranked genes. Given L ranking lists $R_1, \ldots, R_L$, let the prior probability of picking the lth ranking list, R_l, to rank genes as π_l with $\pi_1 + \ldots + \pi_L = 1$. π_l reflects the reliability of R_l. To construct a mixture model [3], for each gene g_i, we introduce an L dimensional latent variable $\mathbf{z}_i = \{z_{i,1}, \ldots, z_{i,L}\}$ indicating using which ranking list we rank g_i, that is if g_i's rank is taken from its rank in R_l, then $z_{i,l} = 1$ and all other elements in $\mathbf{z}_i$ are set to 0. Based on these definitions, we can formulate the joint likelihood of the relevance of a gene set $\mathbf{G} = \{g_1, \ldots, g_K\}$ as below:

$$
\begin{aligned}
& p\left(g_1, \ldots, g_K, Z | R_1, \cdots, R_L, \Theta\right) \\
(4.12) \qquad & = \prod_{i=1}^K \prod_{l=1}^L \pi_l^{z_{i,l}} P\left(r_{l,i}\right)^{z_{i,l}}.
\end{aligned}
$$

In Equation (4.12), Z is the set of latent variables $Z = (z_{i,l})_{K \times L} = (\mathbf{z}_1, \ldots, \mathbf{z}_K)$. And the prior probabilities, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_L\}$, can be obtained by maximizing the joint likelihood specified in Equation (4.12) with an EM algorithm.

4.2.1 An EM Algorithm for Computing π EM is a standard iterative approach for finding the maximum likelihood estimates of parameters in a probabilistic model [3]. The probabilistic model specified in Equation 4.12 can be solved by the EM approach in the following way:

E Step. Assume π is known, we can show that the posterior distribution of Z takes the following form:

$$P\left(Z|R_1, \cdots, R_L, \mathbf{G}\right) \propto P\left(Z\right) P\left(\mathbf{G}|K_1, \cdots, K_L, Z\right)$$

$$= \prod_{i=1}^{N} \prod_{l=1}^{L} \pi_l^{z_{i,l}} \prod_{i=1}^{N} \prod_{l=1}^{L} \mathcal{N}\left(t_i | m_{i,c}^{(l)}, \left(\sigma_{i,c}^{(l)}\right)^2\right)^{z_{i,l}}$$

$$= \prod_{i=1}^{K} \prod_{l=1}^{L} \{\pi_l P\left(r_{l,i}\right)\}^{z_{i,l}}.$$

Using standard techniques, we can show that the responsibility of L_l for g_i is given by the following equation:

$$(4.13) \qquad \gamma_{i,l} = E\left(z_{i,l}\right) = \frac{\pi_l P\left(r_{l,i}\right)}{\sum\limits_{j=1}^{L} \pi_j P\left(r_{l,i}\right)}.$$

The responsibilities can be used to determine the expectation of the complete log likelihood, which defines the Q function [3] specified as below:

$$Q\left(\Theta, \Theta^{\text{old}}\right) = E_z\left(\ln P\left(\mathbf{G}, Z|\Theta\right)\right)$$

$$= \sum_{i=1}^{K} \sum_{l=1}^{L} \gamma_{i,l} \{\ln \pi_l + \ln P\left(r_{l,i}\right)\}.$$

M Step. Assume Z is known, we can find the Θ by maximizing the Q function under the constraint of $\pi_1 + \ldots + \pi_L = 1$, which leads to the following updating:

$$(4.14) \qquad \pi_l^{new} = \frac{1}{K} \sum_{i=1}^{K} \gamma_{i,l}.$$

The algorithm is guaranteed to converge. After obtained π, the probability of g_i to be relevant can be calculated by marginalizing the joint probability $P\left(g_i, R_l\right)$.

$$P(g_i) = \sum_{l=1}^{L} P\left(g_i, R_l\right) = \sum_{l=1}^{L} P\left(g_i | R_l\right) P(R_l)$$

$$(4.15) \qquad = \sum_{l=1}^{L} P\left(r_{l,i}\right) P(R_l) = \sum_{l=1}^{L} P\left(r_{l,i}\right) \pi_l.$$

The final gene ranking list can be obtained by ranking the obtained relevance probability of genes.

5 Experimental Results

We empirically evaluate the effect of using knowledge to assist gene selection. Different types of knowledge about both samples and genes can be combined differently, which leads to different gene ranking methods. Genes selected by different ranking methods are compared on their statistical as well as biological relevance. Algorithms are implemented in Matlab and will be made publicly available.

5.1 Data and Knowledge Sources

5.1.1 Pediatric ALL Data The data is obtained from the Gene Expression Omnibus (GEO)[1]. The data contains the expression profiling of **4,670** genes in bone marrow from **18** pediatric patients with acute lymphoblastic leukemia (ALL): **10** B-cell ALL, **5** T-cell ALL, and **3** B-cell ALL with the MLL/AF4 chromosomal rearrangement. Each bone marrow is measured twice, resulting in totally **36** samples in the data. The data provides insight into the pathogenesis of childhood ALL. We choose this data since our biologist collaborators' research background is closely related to leukemia study.

5.1.2 Knowledge Sources Five different knowledge sources are used in the experiments: (1) **Sample Category**, patients are assigned to one of the three classes, B-ALL, T-ALL, or MLL/AF4. The sample category information forms one type of $\mathcal{K}_{CAT}^{ext}$. (2) **Gene Expression**, the expression profiles of genes are used to obtain sample pairwise similarity with Mahalanobis distance, forming one type of $\mathcal{K}_{GEO}^{ext}$. (3) **Metabolic Pathway**, the 208 Homo sapiens metabolic pathways are obtained from the KEGG pathway repository [15]. 6 ALL-related pathways, including B-CELL RECEPTOR pathway and T-CELL RECEPTOR pathway are selected by the biologist. These pathways form one type of the $\mathcal{K}_{Fun}^{ext}$ (gene function), and the genes involved in these pathways are used to filter data for calculating $\mathcal{K}_{GEO}^{int}$. (4) **Cancer-Gene Annotation**, the cancer gene annotation data are obtained from three knowledge sources: IPA gene annotation[2], NIC Gene-Cancer database [35] and Cancer Gene Census project[3]. The cancer gene annotation data form one type of $\mathcal{K}_{Fun}^{ext}$, which is used to construct both $\mathcal{K}_{GEO}^{int}$ and $\mathcal{K}_{Fun}^{int}$. (5) **Gene Ontology (GO)**, we obtain the GO annotations for genes from the Gene Ontology Database [4]. The information forms one type of $\mathcal{K}_{Fun}^{ext}$ and one type of $\mathcal{K}_{SIM}^{ext}$ (gene similarity). $\mathcal{K}_{SIM}^{ext}$ is extracted from GO annotation using an information content based measure proposed in [27]. The obtained $\mathcal{K}_{SIM}^{ext}$ is used to construct $\mathcal{K}_{GEO}^{int}$ with Mahalanobis distance and $\mathcal{K}_{CON}^{int}$ for relevance propagation.

---

[1] http://www.ncbi.nlm.nih.gov/geo. Access ID: GSE2604
[2] http://www.ingenuity.com/
[3] http://www.sanger.ac.uk/genetics/CGP/Census/

Table 4: The details of how ranking lists are generated. SPEC and Fisher score correspond to traditional unsupervised and supervised gene selection algorithms based microarray data, respectively.

KNOLWEDGE SOURCES	EXTERNAL KNW.	INTERNAL KNW.	RANKING CRITERION	RANKING METHOD
cDNA Expression	$\mathcal{K}_{GEO}^{ext}$	$\mathcal{K}_{GEO}^{int}$	Geometric Consistency	SPEC
Sample Category	$\mathcal{K}_{CAT}^{ext}$	$\mathcal{K}_{CAT}^{int}$	Supervised Gene Selection	Fisher Score
Metabolic Pathway	$\mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{GEO}^{int}$	Geometric Consistency	Pathway-FILT
Gene Ontology	$\mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{FUN}^{int}$	Functional Relevance Voting	GO-REL-VOTE
Gene Ontology	$\mathcal{K}_{SIM}^{ext}$	$\mathcal{K}_{GEO}^{int}$	Geometric Consistency	GO-MAH
Gene Ontology, Cancer-Gene	$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{GEO}^{int}$	Geometric Consistency	GO-CAN-MAH
Gene Ontology, Cancer-Gene	$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{CON}^{int}, \mathcal{K}_{FUN}^{int}$	Relevance Propagation	GO-REL-PROP
Cancer-Gene	$\mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{GEO}^{int}$	Geometric Consistency	Leukemia-FILT

Table 5: The conversion of different types of external knowledge to internal knowledge.

RANKING METHOD	KNOWLEDGE CONVERSION
SPEC	The whole gene expression data are used to construct $\mathcal{K}_{GEO}^{ext}$ with Mahalanobis distance.
Fisher Score	$\mathcal{K}_{CAT}^{ext}$, the label information, is used as $\mathcal{K}_{CAT}^{int}$ in supervised gene selection.
Pathway-FILT	Genes in the selected pathways ($\mathcal{K}_{FUN}^{ext}$) are used to filter the whole data, $\mathcal{K}_{GEO}^{int}$ is obtained on the filtered data.
GO-REL-VOTE	GO terms ($\mathcal{K}_{FUN}^{ext}$) are directly used as $\mathcal{K}_{FUN}^{INT}$, and are weighed according to their relevance for ranking genes.
GO-MAH	GO based gene similarity ($\mathcal{K}_{SIM}^{ext}$) is used to construct Mahalanobis distance to extract $\mathcal{K}_{GEO}^{int}$. See Section 3.3.1
GO-CAN-MAH	Similar to GO-MAH, but only cancer related GO terms ($\mathcal{K}_{FUN}^{ext}$) are used to calculate gene similarity ($\mathcal{K}_{SIM}^{ext}$).
GO-REL-PROP	Relevance ($\mathcal{K}_{FUN}^{int}$) is propagated on the graph ($\mathcal{K}_{CON}^{ext}$) constructed from the GO based gene similarity ($\mathcal{K}_{SIM}^{ext}$).
Leukemia-FILT	Use genes with ALL-related functions ($\mathcal{K}_{FUN}^{ext}$) to filter the data, and $\mathcal{K}_{GEO}^{ext}$ is obtained on the filtered data.

5.2 Experiment Setup Using different types of knowledge and their different combinations results in 8 representative ranking methods that generate 8 different ranking lists. The detail information of how these lists are obtained can be found in Tables 4 and 5. Among the 8 lists, SPEC and Fisher score correspond to using the traditional unsupervised and supervised gene selection on Microarray data to select genes, respectively. The other 6 ranking lists correspond to using one or two types of external knowledge to select genes, which are analogous to the existing methods for gene selection with certain types of knowledge. The 8 lists are used as baselines in the experiment for comparison. In the experiment, for GO-REL-VOTE and GO-CAN-MAH, the relevance of a GO term is determined by M_{can}/M_{all}, where M_{all} denotes the number of the genes associated to the term and M_{can} denotes the number of the cancer related genes associated to the term. The 8 ranking lists are aggregated in three ways: KOGS$_{Borda}$, KOGS$_{Prob}$ and KOGS$_{Prob-SUP}$, which correspond to using Borda count [6] and the probabilistic model proposed in Section 4.2 using all genes and only acute lymphoblastic leukemia (ALL) related genes respectively. Borda count is a representative rank aggregation algorithm based on majority voting, which is also used as a baseline for comparison in the experiment.

5.3 Performance Evaluation To evaluate the performance of different methods, we use four evaluation criteria: (1) **Accuracy**: accuracy of 1NN achieved on the top ranked genes provided by different algorithms; (2) **Sim**$_{anno}$: the similarity between selected genes and the known ALL related genes according to GO annotation; (3) **Hit**$_{canc}$ and (4) **HIT**$_{leu}$, the counts of known cancer related genes and ALL related genes in the top ranked genes provided by methods.

Among the four, **Accuracy** is the standard criterion for evaluating the statistical relevance of the selected genes. For genes that are related to the biological process inducing different phenotypes, their expression pattern should be different on samples of different phenotypes. Therefore using these genes in classification or clustering should result in high accuracy. However, due to the small sample problem in cDNA microarray analysis, genes that result in high accuracy may not be biologically relevant. The three criteria: **Sim**$_{anno}$, **Hit**$_{canc}$ and **HIT**$_{leu}$ are designed to provide evidence on how many selected genes are biologically relevant according to literature. The hypothesis is that if a gene list results in high accuracy and contains many genes that are biologically relevant according to literature, it indicates that (1) the corresponding algorithm can select biologically relevant genes; and (2) others genes in the list may also be biologically relevant. Achieving high value on the three evidence criteria, with low accuracy indicates that genes do not have discriminative expression patterns on the samples of different phenotypes for the current study. Therefore, it requires both high accuracy and strong supports from evidence criteria to confirm the biological relevance of a gene list. In the following we compare ranking lists obtained by using tradi-

Table 6: Performance comparison for gene ranking lists generated from different methods. ACC-10, ACC-30, and ACC-50 correspond to the accuracy achieved on the top 10, 30 and 50 genes provided by different algorithms, respectively. ACC-AVE is the averaged accuracy achieved by genes using the top 10, 30 and 50 genes provided by the algorithms. Sim_{anno} is the functional similarity between selected genes and known ALL related genes according to GO annotation. Hit_{canc} and HIT_{leu} are the hit ratios of known cancer and leukemia related genes, respectively. To confirm the biological relevance of a gene list requires both high accuracy and strong supports from evidence criteria.

RANKING METHODS	ACC-10	ACC-30	ACC-50	ACC Ave	Sim_{anno}	HIT_{canc}	HIT_{leu}
SPEC	0.64	0.66	0.83	0.65	797	2	0
Fisher Score	**0.97**	**0.97**	**0.97**	**0.97**	823	8	2
Pathway-FILT	0.61	0.81	0.89	0.81	807	4	0
GO-REL-VOTE	0.56	0.69	0.83	0.64	**7686**	**26**	8
GO-MAH	0.69	0.80	0.86	0.82	759	3	0
GO-CAN-MAH	0.62	0.83	0.86	0.80	2996	5	1
GO-REL-PROP	0.70	0.78	0.86	0.74	**7688**	**22**	**15**
Leukemia-FILT	0.55	0.62	0.64	0.62	687	4	1
KOGS$_{Borda}$	**0.91**	**0.97**	**0.97**	**0.96**	1723	6	2
KOGS$_{Prob}$	**0.97**	**0.94**	**0.94**	**0.95**	**6954**	**21**	**12**
KOGS$_{Prob-SUP}$	**0.94**	**0.91**	**0.91**	**0.93**	**7766**	**25**	**17**

tional gene selection algorithms, using one or two types of knowledge, and using multiple types of knowledge.

5.4 Empirical Findings Table 6 contains the experimental results obtained from methods using different types of knowledge. We report the following observations.

First, comparing on accuracy, the gene lists obtained from Fisher socre, KOGS$_{Borda}$, KOGS$_{Prob}$, and KOGS$_{Prob-SUP}$ achieve good performance. High accuracy indicates that the genes in these lists are statistically relevant, since they can separate samples from different phenotypes. We also notice that comparing with SPEC, GO-MAH achieved higher accuracy. Both SPEC and GO-MAH use Mahalanobis distance, but GO-MAH uses the gene covariance learnt from GO based gene similarity. This suggests that the strategy proposed in Figure 4 is effective.

Second, comparing on the three evidence criteria (Sim_{anno}, HIT_{canc} and HIT_{leu}), the two methods using $\mathcal{K}_{FUN}^{int}$ (GO-REL-VOTE and GO-REL-PROP) and the two methods generated from KOGS (KOGS$_{Prob}$ and KOGS$_{Prob-SUP}$) achieve good performance. While Fisher score and other ranking methods do not perform well. This is reasonable, since in Sim_{anno}, HIT_{canc}, and HIT_{leu} we actually use $\mathcal{K}_{FUN}^{int}$ to evaluate genes. As GO-REL-VOTE and GO-REL-PROP are provided with $\mathcal{K}_{FUN}^{int}$, it is understandable that they can achieve better performance. We noticed that by using only the terms related to cancer for learning gene similarity, GO-CAN-MAH achieves a better performance than GO-MAH according to the three evidence criteria. For the methods derived from KOGS, the two methods use the probabilistic model proposed in Section 4.2

achieve good performance. Compared with KOGS$_{Prob}$, KOGS$_{Prob-SUP}$ achieves better performance on the evidence criteria. This clearly suggests that the supervision information used in KOGS$_{Prob-SUP}$ helps. Both KOGS$_{Prob}$ and GO-REL-PROP generate gene lists that have strong supports from evidence criteria. However in accuracy, GO-REL-PROP's performance is about 20% lower than that of KOGS$_{Prob}$. To intuitively observe the expression pattern of genes in each list, we applied cluster analysis on the genes selected by the two algorithms. The obtained heatmaps are presented in Figure 5. Results show that, although many genes selected by the GO-REL-PROP are reported to be leukemia related in other studies, most of these genes do not show discriminative expression patterns on the current data. When doing cluster using these genes, samples of different phenotypes are mixed up. The fact suggests that these genes may not be related to the current study. As compared to GO-REL-PROP, we observed that the genes selected by KOGS$_{Prob}$ show discriminative expression patterns and lead to good clustering performance.

Last, considering both accuracy and evidence criteria, experiment results in Table 6 show that the traditional gene selection algorithms and the algorithms using just one or two types of knowledge in gene selection can only achieve either high statistical relevance, or strong supports from evidence criteria, but not both. Comparing with these algorithms, the algorithms derived from KOGS can achieve high performance on both types of criteria. The results clearly demonstrated the efficacy of the proposed integrative approach on identifying biologically relevant genes.

Table 7: The biologically relevant genes in the top 50 gene list provided by KOGS$_{\text{Prob-SUP}}$. The upper part contains 17 genes which are known to be leukemia related according to literature. And lower part contains 12 genes whose biological relevance can not be ruled out according to their biological functions or roles in pediatric ALL.

Rank	Gen Symbol	Gene Name	Related Cancers
Genes Are Known to Be Leukemia Related (17)			
1	LMO1	LIM domain only 1 (rhombotin 1)	leukemia
2	CBFA2T3	core-binding factor, runt domain, alpha subunit 2; translocated to, 3	leukemia, breast cancer, +2 more
4	TYROBP	TYRO protein tyrosine kinase binding protein	leukemia
5	STAT5B	signal transducer and activator of transcription 5B	leukemia, breast cancer, +2 more
6	IGFBP3	insulin-like growth factor binding protein 3	leukemia, breast cancer, +4 more
7	JUN	jun oncogene	leukemia, breast cancer, +4 more
8	USP33	ubiquitin specific peptidase 33	leukemia
9	GSN	gelsolin (amyloidosis, Finnish type)	leukemia, bladder tumours
10	BTG1	B-cell translocation gene 1, anti-proliferative	leukemia, ovarian carcinomas
11	TFRC	transferrin receptor (p90, CD71)	leukemia, breast cancer, +2 more
13	PTK2	PTK2 protein tyrosine kinase 2	leukemia, lung cancer, +2 more
15	PDE7A	phosphodiesterase 7A	leukemia
16	TIMP1	TIMP metallopeptidase inhibitor 1	leukemia, bladder cancer, +11 more
17	AKT1	v-akt murine thymoma viral oncogene homolog 1	leukemia, prostate cancer, +4 more
19	FLT1	fms-related tyrosine kinase 1	leukemia, breast cancer, +4 more
47	CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	leukemia
48	TIMP2	TIMP metallopeptidase inhibitor 2	leukemia, bladder cancer, +6 more
Potential Leukemia Related Genes (12)			
18	TIMP4	TIMP metallopeptidase inhibitor 4	breast cancer, glioma
23	TYK2	tyrosine kinase 2	fibrosarcoma
25	CDK4	cyclin-dependent kinase 4	retinoblastoma, melanoma, glioma
31	SERPINF2	serpin peptidase inhibitor, clade F, member 2	
32	PRKACA	protein kinase, cAMP-dependent, catalytic, alpha	pituitary tumor
34	NCOR1	nuclear receptor co-repressor 1	prostate cancer, breast cancer
36	SIVA1	SIVA1, apoptosis-inducing factor	
38	BRD8	bromodomain containing 8	pancreatic cancer
40	CAPN7	calpain 7	
43	SPATA2	spermatogenesis associated 2	
49	PRKAR1A	protein kinase, cAMP-dependent, regulatory, type I, alpha	adrenocortical cancer, myxoma,
50	PPARA	peroxisome proliferator-activated receptor alpha	colorectal cancer, bladder cancer

5.5 Discussion on Biological Relevance In order to closely examine the biological relevance of the selected genes, we performed some further study, in which our biologist collaborators examined the top 50 genes selected by KOGS$_{\text{Prob-SUP}}$. The information of relevant genes is summarized in Table 7. The upper part of the table contains the genes whose relevance to leukemia has been confirmed by the literature. And the lower part of the table contains the genes, whose relevance is unknown but cannot be ruled out. Analyses of these genes may yield finding of new leukemia-related genes. 17 leukemia relevant genes are selected by KOGS$_{\text{Prob-SUP}}$. This list involves several crucial genes, such as the USP33, LMO1, TIMP1, TIMP2 and STAT5B, which play important roles in the leukemia related tumorigenesis and may lead to different subtype of acute lymphoblastic leukemia (ALL). For instance, USP33 is reported to be consistently over-expressed in B-ALL samples but not in T-ALL samples [29]. LMO1 is mapped to an area of consistent chromosomal translocation in chromosome 11, disrupting it in T-cell ALL. The LMO1 gene family was also defined as a class of T-cell oncogenes [42]. TIMP1 and TIMP2, members of Tissue Inhibitor of MetalloProteinases, were found related to the infiltration of ALL leukemia cells into extramedullary organs [41]. STAT5B is a member of the Signal Transducers and Activator of Transcription (STAT), the dysregulation of the signaling pathways mediated by this protein may be the cause of the ALL and other human cancers[43]. 12 genes are found to be possibly leukemia or cancer related due to the following reasons: (1) their functions on tumorigensis and cell cycle control (e.g., PPARA, TIMP4 and CDK4); (2) their cAMP-dependence (PRKACA and PRKAR1A); (3) transcription factors (BRD8 and NCOR1), whose expres-
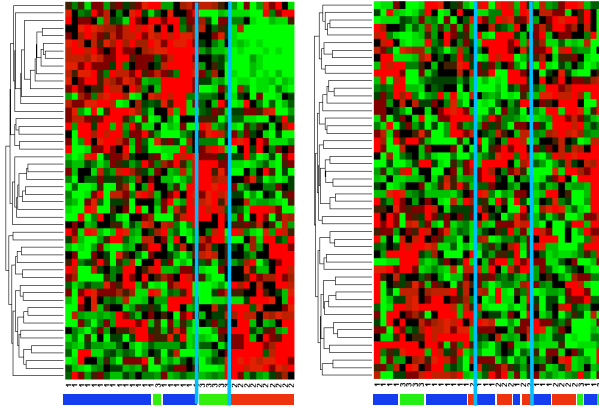
Figure 5: Cluster analysis on the genes selected by KOGS$_{Prob}$ (left) and GO-REL-PROP (right), respectively. The color lines on the bottom of the figure correspond to the samples from patients of B-cell ALL (blue), T-cell ALL (red), and B-cell ALL with the MLL/AF4 chromosomal rearrangement (green), respectively.

sions were closely related to other known ALL genes mentioned above; (4) their known highly expression in leukemia (e.g. SIVA). Recent research results revealed a role of SIVA inactivation in leukemia related tumorigenesis, presumably through enhancing NF-kappaB-mediated anti-apoptotic activity [9]. The study of these genes may help identify new biomarkers crucial to leukemia tumorigenesis.

6 Conclusion

In this work, we proposed KOGS, an integrative approach for using multiple types of knowledge in gene selection. The approach is able to convert different types of external knowledge to its internal knowledge for genes ranking. Given multiple gene ranking lists, KOGS can aggregate them to form a final gene ranking list by considering various definitions of gene relevance. For KOGS, the knowledge conversion step effectively ensures the extendability of the approach. And the rank aggregation step provides an efficient way for knowledge integration and improves the flexibility. Experimental results demonstrated the methods derived from KOGS is able to provide superior performance and select biologically relevant genes.

In [45], the authors studied the problem of gene selection using multiple data sources and proposed a gene selection approach named MSGS. The two systems are different in that (1) KOGS explicitly defines the concepts of external and internal knowledge, and organizes different types of knowledge into well defined categories, while no knowledge related concept is proposed in MSGS; (2) In the current work, the coefficient combination can be automatically learned, while this problem is not addressed in MSGS; and (3) KOGS is based on combining ranking lists, while the one in MSGS replies on combining sample similarity, which restricts the model flexibility. We noticed that supervised rank aggregation is also studied in [23], but it requires to provide

the supervision information via partial orders among entry pairs, which is not intuitive in our application.

The developed KOGS approach forms our preliminary work for integrative gene selection. Our ongoing work includes: (1) understanding the roles of different types of knowledge in gene selection, and (2) developing a user friendly toolbox for integrative gene selection to assist biologists' study.

Acknowledgments

References

[1] Stein Aerts, and *et al.* Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537–545, 2006.

[2] Ramon Aragues, Chris Sander, and Baldo Oliva. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics*, 9:172, 2008.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Evelyn Camon, *et al.* The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32:262–266, 2004.

[5] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2 edition, 2001.

[6] C. Dwork, R. Kumar, M. Naor, and D. R. Sivakumar. Aggregation methods for the web. In *In Proceedings of the 10th International World Wide Web Conference*, 2001.

[7] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, 5:845–889, 2004.

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

[9] R. Gudi, and *et al.* Siva-1 negatively regulates nf-kappab activity: effect on t-cell receptor-mediated activation-induced cell death (aicd). *Oncogene*, 8:3458–62, 2006.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[12] D. Herold, and *et al.* Comparison of unsupervised and supervised gene selection methods. *Conf Proc IEEE Eng Med Biol Soc*, 1:5212–5215, 2008.

[13] J. C. Huang, and *et al.* Using expression profiling data to identify human microrna targets. *NATURE METHODS*, 4:1045–1049, 2007.

[14] N.C. Jones and P.A. Pevzner. *An Introduction to Bioinformatics Algorithms*. The MIT Press, 2004.

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

[16] Fumiaki Katagiri and Jane Glazebrook. Overview of mrna expression profiling using dna microarrays. *Current Protocols in Molecular Biology*, 22.4:s85, 2009.

[17] Y. B. Kim and J. Gao. Unsupervised gene selection for high dimensional data. In *Proc. Sixth IEEE Symposium on BioInformatics and BioEngineering BIBE 2006*, pages 227–234, 16–18 Oct. 2006.

[18] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 2008.

[19] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *BIOINFORMATICS*, 20:2429–2437, 2004.

[20] J.G. Liao and Khew-Voon Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *BIOINFORMATICS*, 23:1945–1951, 2007.

[21] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers, 1998.

[22] H. Liu and H. Motoda, editors. *Computational Methods of Feature Selection*. Chapman and Hall/CRC Press, 2007.

[23] Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. Supervised rank aggregation. In *Proceedings of the 16th international conference on World Wide Web*, 2007.

[24] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub. Microrna expression profiles classify human cancers. *Nature*, 435:834–838, 2005.

[25] P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 12:49–55, 1936.

[26] Carl Murie, Owen Woody, Anna Lee, and Robert Nadon. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, 10(1):45, Feb 2009.

[27] Catia Pesquita, Daniel Faria, Hugo Bastos, Antonio EN Ferreira, Andre O Falcao, and Francisco M Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9:S4, 2008.

[28] John H. Phan, Qiqin Yi Goen, Andrew N. Young, and May D. Wang. Improving the efficiency of biomarker identification using biological knowledge. In *Pacific Symposium on Biocomputing*, pages 427–38, 2009.

[29] C. D. Pitta and et. al. A leukemia-enriched cdna microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia. *Haematologica*, 90:890–898, 2005.

[30] Jianlong Qi and Jian Tang. Gene ontology driven feature selection from microarray gene expression data. In *Computational Intelligence and Bioinformatics and Computational Biology*, 2006.

[31] Sarunas J. Raudys and Anil K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13:252–264, 1991.

[32] Yvan Saeys, Iki Inza, and Pedro Larraga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, Oct 2007.

[33] Frans Schalekamp and Anke van Zuylen. Rank aggregation: Together we're strong. In *Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2009.

[34] B. Scholköpf and A. J. Smola. *Learning with Kernels*. The MIT Press, 2002.

[35] Christine M.E. Schueller, Andreas Fritz, Eduardo Torres Schumann, Karsten Wenger, Kaj Albermann, George A. Komatsoulis, Peter A. Covitz, Lawrence W. Wright, and Frank Hartel. Towards a comprehensive catalog of gene-disease and gene-drug relationships in cancer. Technical report, National Cancer Institute, 2005.

[36] C. Sima and E. R. Dougherty. What should be expected from feature selection in small-sample settings. *Bioinformatics*, 22:2430–2436, 2006.

[37] A.J. Smola and I.R. Kondor. Kernels and regularization on graphs. In *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, 2003.

[38] Shireesh Srivastava, Linxia Zhang, Rong Jin, and Christina Chan. A novel method incorporating gene ontology information for unsupervised clustering and feature selection. *PLoS ONE*, 3(12):e3860, 2008.

[39] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: A general repository for interaction datasets. *Nucleic Acids Res*, 34:535–539, 2006.

[40] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102:15545–15550, 2005.

[41] A. Suminoe, A. Matsuzaki, H. Hattori, Y. Koga, E. Ishii, and T. Hara. Expression of matrix metalloproteinase (mmp) and tissue inhibitor of mmp (timp) genes in blasts of infant acute lymphoblastic leukemia with organ involvement. *Leuk Res*, 10:1437–40, 2007.

[42] et al. T. Boehm. The rhombotin family of cysteine-rich lim-domain oncogenes: distinct members are involved in t-cell translocations to human chromosomes 11p15 and 11p13. *Proc Natl Acad Sci*, 88:4367–71, 1991.

[43] Hua Yu and Richard Jove. The stats of cancer – new molecular targets come of age. *Nature Reviews Cancer*, 4:97–105, 2004.

[44] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *International Conference on Machine Learning (ICML)*, 2007.

[45] Z. Zhao, J. Wang, H. Liu, J. Ye, and Y. Chang. Identifying biologically relevant genes via multiple heterogeneous data sources. In *The Fourteenth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (SIGKDD 2008)*, 2008.