

A Knowledge-Oriented Framework for 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 Biodesign 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 ten thousand genes, but only less than a hundred of samples, many biologically irrelevant genes can gain their statistical relevance by sheer randomness. Moreover, even for genes that are biologically relevant, biologists often prefer the "trigger" to the "fire". Addressing these problems 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 Annotation database. Integrating different types of knowledge for gene selection could provide more information about the genes and samples under study. In this work, we propose a novel framework to integrate different types of knowledge for identifying biologically relevant genes. The framework converts different types of external knowledge to its internal knowledge, which is then used to rank genes. The obtained rank lists are aggregated via a probabilistic framework. Experimental results from our study on acute lymphoblastic leukemia demonstrate the novelty and efficacy of the proposed framework and shows that using different types of knowledge together can help detect genes that are biologically more relevant.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—*data mining*; I.2.6 [Artificial Intelligence]: Learning; I.5.2 [Pattern Recognition]: Design Methodology—*feature evaluation and selection*; J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms

Algorithms, Experimentation, Measurement

Keywords

Gene selection, information integration, bioinformatics

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

1. INTRODUCTION

Selecting genes that are critical to a particular biological process has been a major challenge in post-array analysis [12]. Also known as feature selection [9] 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 [14]. Given cDNA Microarray data, most existing algorithms try to identify genes that are differentially expressed over the samples. According to whether label information is used, algorithms can be either supervised [15] or unsupervised [6]. Discriminative genes help classifiers or clustering algorithms to achieve high accuracy. However, does the better accuracy necessarily indicate higher biological relevance of genes? We applied a supervised gene selection algorithm, Fisher Score [4] and an unsupervised algorithm, SPEC [32] on the expression profiling of bone marrow from 18 pediatric patients with acute lymphoblastic leukemia (ALL) [20] to select genes that may provide insight into the pathogenesis of childhood 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 using the selected genes. The result demonstrates that a gene list of higher accuracy does not necessarily contain more relevant genes. Hence, selecting genes to achieve high classification accuracy should not constitute the sole goal of biological discovery.

Table 1: Biologically relevant genes identified by gene selection algorithms, Fisher Score and SPEC

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 among others for the above proposition. First, a cDNA Microarray data usually contains more than 10000 genes but only fewer than 1000 samples. A data set of this kind usually leads to the small sample problem [21]. With so few samples, many genes, which are not biologically relevant, can easily gain their statistical relevance due to randomness [25]. Second, even genes are related, they may have different importance. For instance, to understand a specific biological process, the genes acting as the "trigger" are much more interesting than the genes acting as the "fire". And 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 KEGG pathway repository [13], Gene Ontology (GO) annotation [3] and NCI Gene-Cancer database [24], etc. Recent work has also revealed the existence of a class of small non-coding RNA species known as microRNAs (miRNAs), which are surprisingly informative for identifying cancerous tissues [17]. The availability of these various knowledge sources presents unprecedented opportunities to advance research solving previously unsolvable problems. Therefore, 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. The challenge of this project is how to address the heterogeneity in the knowledge representations.

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 [28], 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. Since most existing work is designed for specific research purpose, they can only handle one or limited types of knowledge of the same category. For instance, the model proposed in [1] can only handle different types of knowledge about genes, but not any type of knowledge about samples. To address this limitation, we propose a general framework to systematically integrate different types of knowledge to achieve knowledge-oriented gene selection. The framework is based on a probabilistic model for aggregating gene relevance ranking lists, which is obtained by using different types of knowledge. The system developed based on the framework is extensively experimented and tested. Experimental results from our acute lymphoblastic leukemia (ALL) study show that judiciously using different types knowledge can bring about significant performance improvement to assist biological discovery.

2. KOGS: A GENERAL FRAMEWORK

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

$$\mathcal{K}_i^{int} = c_i(\mathcal{K}_i^{ext}), i = 1, \dots, L \quad (1)$$

(2) **feature ranking** - assume we decide to use K sets of internal knowledge $KNOW_1, \dots, KNOW_K$ to rank genes, where $KNOW_i$ is defined as: $KNOW_i = \{\mathcal{K}_{i_1}^{int} \dots \mathcal{K}_{i_{t_i}}^{int}\}$. Let \mathcal{C}_i be a relevance criterion, $\mathbf{G} = \{g_1, \dots, g_M\}$ be a set

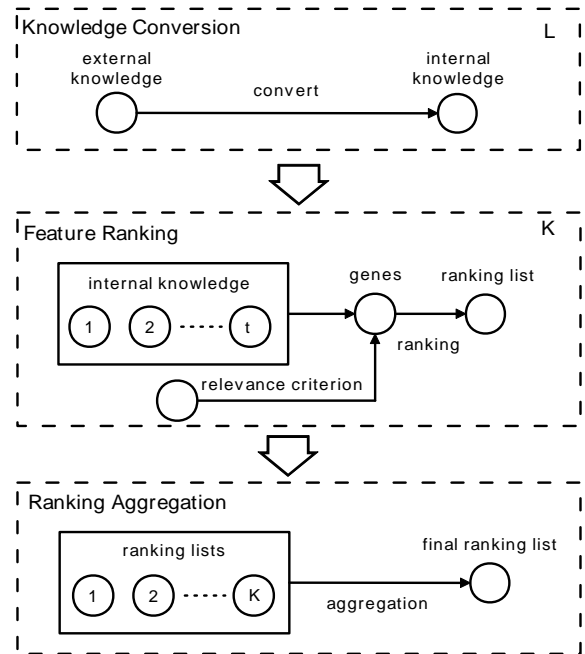


Figure 1: A framework for integrating different types of domain knowledge to assist gene selection

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:

$$R_i^{rank} = \mathcal{R}(KNOW_i, \mathcal{C}_i, \mathbf{G}) \quad (2)$$

(3) **ranking aggregation** - after obtaining 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:

$$R_F^{rank} = \mathcal{A}(R_1^{rank}, \dots, R_K^{rank}, \mathcal{C}) \quad (3)$$

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 [3], 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 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 pathways, 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 [27], 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 on 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

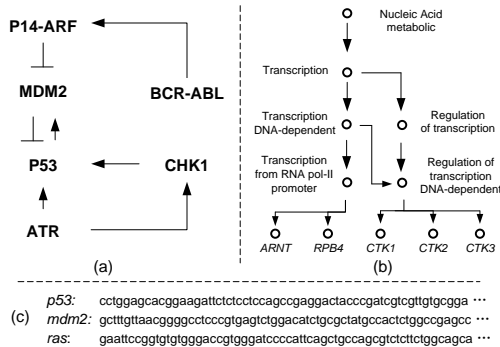


Figure 2: An example of three different types of knowledge about genes, (a) Metabolic Pathway, (b) Gene Ontology Annotation, and (c) Gene Sequence.

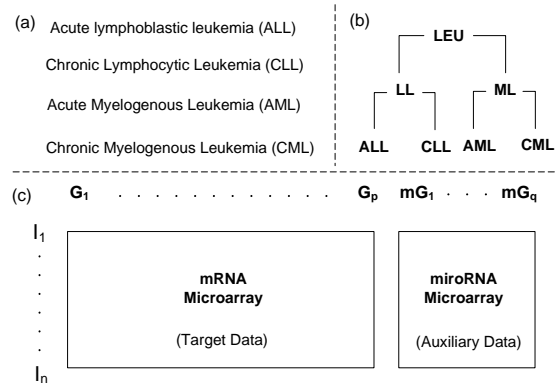


Figure 3: Different types of knowledge about samples, (a) the class label information, (b) sample hierarchy, and (c) an example of the auxiliary data.

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 different types of cancers [11]. 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 points of view. 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}_{CAT}^{int} , (2) sample geometric pattern, \mathcal{K}_{GEO}^{int} ; and knowledge about genes: (3) gene connection, \mathcal{K}_{CON}^{int} , and (4) gene function, \mathcal{K}_{FUN}^{int} . 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

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

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

the network derived from \mathcal{K}_{CON}^{int} . KOGS is not restricted to the four types of internal knowledge defined above. As far as new knowledge can be used to rank genes, it can be treated as a type of internal knowledge. This ensures the extensibility 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
$\mathcal{K}_{GEO}^{ext}, \mathcal{K}_{FUN}^{ext}, \mathcal{K}_{SIM}^{ext}$	\mathcal{K}_{GEO}^{int}
$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{INT}^{ext}$	\mathcal{K}_{CON}^{int}
\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{FUN}^{int}
\mathcal{K}_{CAT}^{ext}	\mathcal{K}_{CAT}^{int}

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 $\mathcal{K}_{GEO}^{ext} \rightarrow \mathcal{K}_{GEO}^{int}$, $\mathcal{K}_{CAT}^{ext} \rightarrow \mathcal{K}_{CAT}^{int}$, $\mathcal{K}_{SIM}^{ext} \rightarrow \mathcal{K}_{CON}^{int}$, $\mathcal{K}_{FUN}^{ext} \rightarrow \mathcal{K}_{FUN}^{int}$, and $\mathcal{K}_{INT}^{ext} \rightarrow \mathcal{K}_{CON}^{int}$ are straightforward. For example, \mathcal{K}_{SIM}^{ext} , the similarity among genes, and \mathcal{K}_{INT}^{ext} , the interaction among genes, can be directly used to construct gene connection graphs, corresponding to \mathcal{K}_{CON}^{int} . Below, we show how to perform conversions: $\mathcal{K}_{SIM}^{ext} \rightarrow \mathcal{K}_{GEO}^{int}$ and $\mathcal{K}_{FUN}^{ext} \rightarrow \mathcal{K}_{GEO}^{int}$. 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 [23]. The pairwise distance can also be conveniently used in well studied distance based gene selection algorithms. Figure 4 shows how to convert \mathcal{K}_{SIM}^{ext} and \mathcal{K}_{FUN}^{ext} to \mathcal{K}_{GEO}^{int} . The basic idea is to involve the two types of knowledge in the calculation of the pairwise similarity among samples.

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 [18], which is defined as:

$$\|\mathbf{x} - \mathbf{y}\|_M^2 = (\mathbf{x} - \mathbf{y})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{y}). \quad (4)$$

In the equation, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$ are two samples with M genes g_1, \dots, 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} [10]. In real applications, \mathbf{C} is usually estimated from the data

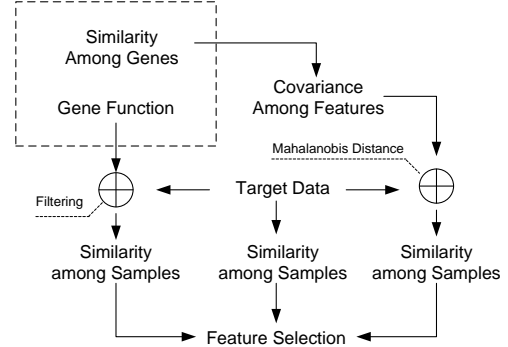


Figure 4: Obtaining the knowledge of sample geometry, using different types of knowledge of genes.

by the following equation:

$$\mathbf{C} = \frac{1}{N-1} \sum_{k=1}^N (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T, \quad (5)$$

where $\mathbf{x}_1, \dots, \mathbf{x}_N$ are the N samples of the data, with $\bar{\mathbf{x}}$ being their mean. Although Equation (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 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 \mathbf{C} can be calculated as:

$$\mathbf{C} = K \left(I - \frac{1}{l} U \mathbf{1} \mathbf{1}^T U^T \right) K. \quad (6)$$

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 [7] 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 some research purpose. Given \mathcal{K}_{FUN}^{ext} , the knowledge of gene functions, and \mathcal{F} , a set of biological functions of interests, we can use the genes associated with the functions to filter the data,

$$X_{\mathcal{F}} = \Pi_{G_{\mathcal{F}}}(X), \quad (7)$$

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 biological 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 types of knowledge ready, we study how to use different types internal knowledge 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 various ways for ranking genes using the other three types of internal knowledge.

4.1.1 $\mathcal{K}_{GEO}^{int} + \text{Geometric Consistency}$

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 [32] 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 , the similarity matrix of samples, the laplacian matrix $L = D - W$ forms a consistency (or smoothness) estimator [26], where D is a diagonal matrix with $d_{ii} = \sum_k w_{ik}$. With L , the geometric consistency of a vector \mathbf{g} can be measured by the following way:

$$\mathbf{g}^T L \mathbf{g} = \sum_{i,j} w_{i,j} (g_i - g_j), \quad (8)$$

and the smaller the value, the more consistent the vector \mathbf{g} . This measurement is improved in [32] with:

$$\varphi(\mathbf{g}_i) = \frac{\hat{\mathbf{g}}_i^T \gamma(\mathcal{L}) \hat{\mathbf{g}}_i}{1 - (\hat{\mathbf{g}}_i^T \xi_0)^2}. \quad (9)$$

In the equation, $\hat{\mathbf{g}}_i = (D^{\frac{1}{2}} \mathbf{g}_i) \cdot \|(D^{\frac{1}{2}} \mathbf{g}_i)\|^{-1}$ is the normalized feature vector; $\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ is the normalized laplacian matrix; and $\gamma(\cdot)$ is a spectral matrix function [7], induced from an increasing real function, which is used to rescale the eigenvalues of \mathcal{L} for reducing noise. As shown in [32], compared to Equation (8), Equation (9) is more robust to noise and has better performance. In this work, we use Equation (9) to measure the geometric consistency.

4.1.2 $\mathcal{K}_{CON}^{int} + \text{Relevance Propagation}$

Given \mathcal{K}_{CON}^{int} , the knowledge of gene connections, a graph \mathbb{G} can be derived. Given a set of genes $\mathcal{G} = \{g_1, \dots, g_t\}$,

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}_{STM}^{ext} , the knowledge of gene similarity, relevance propagation corresponds to the hypothesis that if a gene is relevant, the genes, which is similar to it, may also be relevant. We can formulate the idea using the concept from random walk theory. Assuming the affinity matrix of \mathbb{G} is W , the transition probability matrix is defined as:

$$P = D^{-1}W, \quad D = \text{diag}(d_1, \dots, d_M), \quad d_i = \sum_k w_{ik}. \quad (10)$$

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

$$\begin{aligned} \mathbf{r}^* &= \mathbf{r} + \dots + (\lambda P)^k \mathbf{r} + \dots + (\lambda P)^\infty \mathbf{r} \\ &= (I - \lambda P)^{-1} \mathbf{r}. \end{aligned} \quad (11)$$

In the above equation, $(\lambda P)^k \mathbf{r}$ 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 obtaining \mathbf{r}^* , genes can be ranked according to their corresponding value in \mathbf{r}^* .

4.1.3 $\mathcal{K}_{FUN}^{int} + \text{Functional Relevance Voting}$

The knowledge of gene function, \mathcal{K}_{FUN}^{int} , can also be used to rank genes. For instance, in the Gene Ontology (GO) annotation database [3], genes' functions are provided by a controlled vocabulary. In this cases, the terms can be regarded as the hyper features of genes. Let g_i be the i th gene in the gene list, whose function is described by a vector $\mathbf{f}_i = (f_{i,1}, \dots, f_{i,T})$, where T is the total number of functions, and $f_{i,j} = 1$, if and only if gene i is related to function j . Assume we know the relevance of all the functions, which is described by a vector \mathbf{r}_{fun} , the relevance of gene i can be obtained by the following equation:

$$r_i = \mathbf{f}_i^T \mathbf{r}_{\text{fun}}. \quad (12)$$

The equation sums the relevance of all the functions related to the gene as its relevance score. \mathbf{r}_{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 ranking aggregation in both machine learning and information retrieval [22]. In this work we propose a probabilistic model for ranking aggregation. While existing rank aggregation algorithms, such as the methods presented in [22], treat different ranking lists equally in the combination process, the proposed method is able to automatically learn a set of combination coefficients to weight the importance of different ranking lists. And this is achieved by maximizing the likelihood of the relevance of genes in a given gene set. When the gene set contains genes which are known to be relevant, the model achieves ranking aggregation in a supervised way. When the gene set contains all the genes, the model maximizes the joint relevance likelihood of all the genes, which corresponds to combine ranks 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:

$$P(r_{l,i}) = \frac{1}{B} \exp\left(\frac{1}{r_{l,i}}\right), \quad B = \sum_{j=1}^M \exp\left(\frac{1}{j}\right). \quad (13)$$

In the equation, B is the normalization factor for the distribution. Given L ranking lists R_1, \dots, R_L , let the prior probability of picking the l th ranking list, R_l , to rank genes as π_l with $\pi_1 + \dots + \pi_L = 1$. π_l reflects the reliability of R_l . To construct a mixture model [2], for each gene g_i , we introduce an L dimensional latent variable \mathbf{z}_i 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. By using these definitions, we can formulate the joint likelihood of the relevance of a gene set $\mathbf{G} = \{g_1, \dots, g_K\}$ as below:

$$\begin{aligned} p(g_1, \dots, g_K, Z | R_1, R_2, \dots, R_L, \Theta) \\ = \prod_{i=1}^K \prod_{l=1}^L \pi_l^{z_{i,l}} P(r_{l,i})^{z_{i,l}}. \end{aligned} \quad (14)$$

In Equation (14), Z is the set of latent variables $Z = (z_{i,l})_{K \times L} = (\mathbf{z}_1, \dots, \mathbf{z}_K)$. And the prior probabilities, $\Pi = \{\pi_1, \dots, \pi_L\}$, can be obtained by maximizing the joint likelihood specified in Equation (14) with the EM algorithm proposed below.

4.2.1 An EM Algorithm for Computing Π

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

$$\begin{aligned} P(Z | R_1, \dots, R_L, \mathbf{G}) &\propto P(Z) P(\mathbf{G} | K_1, \dots, K_L, Z) \\ &= \prod_{i=1}^K \prod_{l=1}^L \{\pi_l P(r_{l,i})\}^{z_{i,l}}. \end{aligned} \quad (15)$$

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

$$\gamma_{i,l} = E(z_{i,l}) = \frac{\pi_l P(r_{l,i})}{\sum_{j=1}^L \pi_j P(r_{j,i})}. \quad (16)$$

The responsibilities can be used to determine the expectation of the complete log likelihood, which defines the Q function [2] specified as below:

$$\begin{aligned} Q(\Theta, \Theta^{\text{old}}) &= E_z(\ln P(\mathbf{G}, Z | \Theta)) \\ &= \sum_{i=1}^K \sum_{l=1}^L \gamma_{i,l} \{\ln \pi_l + \ln P(r_{l,i})\}. \end{aligned}$$

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

$$\pi_l^{\text{new}} = \frac{1}{K} \sum_{i=1}^K \gamma_{i,l}. \quad (17)$$

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 $P(g_i, R_l)$.

$$\begin{aligned} P(g_i) &= \sum_{l=1}^L P(g_i, R_l) = \sum_{l=1}^L P(g_i | R_l) P(R_l) \\ &= \sum_{l=1}^L P(r_{l,i}) P(R_l) = \sum_{l=1}^L P(r_{l,i}) \pi_l. \end{aligned} \quad (18)$$

The final gene ranking list can be obtained by ranking the obtained relevance probability of genes. Below we conduct experiments to evaluate the efficacy of the proposed framework for knowledge-oriented gene selection.

5. EXPERIMENTAL RESULTS

We empirically evaluate the effect of using knowledge to assist gene selection. Different types of knowledge about both samples and genes are used in the experiments, 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 Set and Knowledge Sources

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 pediatric **18** 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. 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. Below we introduce the knowledge sources being used in the experiments.

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 $\mathcal{K}_{CAT}^{\text{ext}}$. (2) **Gene Expression**, the expression profiles of genes are used to obtain sample pairwise similarity with Mahalanobis distance, forming one type of $\mathcal{K}_{GEO}^{\text{ext}}$. (3) **Metabolic Pathway**, the 208 Homo sapiens metabolic pathways are obtained from the KEGG pathway repository [13]. 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 $\mathcal{K}_{Fun}^{\text{ext}}$ (gene function), and the genes involved in these pathways are used to filter data for calculating $\mathcal{K}_{GEO}^{\text{int}}$. (4) **Cancer-Gene Annotation**, the cancer gene annotation data are obtained from three knowledge sources: IPA gene annotation², NIC Gene-Cancer database [24] and Cancer Gene Census project³. The cancer gene annotation data form one type of $\mathcal{K}_{Fun}^{\text{ext}}$, which is used to construct both $\mathcal{K}_{GEO}^{\text{int}}$ and $\mathcal{K}_{Fun}^{\text{int}}$. (5) **Gene Ontology (GO) Annotation** information, we obtain the GO annotations for genes from the Gene Ontology Database [3]. The information forms one type of $\mathcal{K}_{Fun}^{\text{ext}}$ and one type of $\mathcal{K}_{SIM}^{\text{ext}}$ (gene similarity). $\mathcal{K}_{SIM}^{\text{ext}}$ is extracted from GO annotation using an information content based measure proposed in [19]. The obtained $\mathcal{K}_{SIM}^{\text{ext}}$ is used to construct $\mathcal{K}_{GEO}^{\text{int}}$ with Mahalanobis distance and $\mathcal{K}_{CON}^{\text{int}}$ 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: Description of the gene ranking methods obtained by using different types of knowledge

METHODS	DESCRIPTION
SPEC	Expression of genes are used to construct \mathcal{K}_{GEO}^{ext} with Mahalanobis distance.
Fisher Score	\mathcal{K}_{CAT}^{ext} (Label information), is used with Fisher Score to select genes.
Pathway-FILT	Genes in selected pathways are used to filter the data, \mathcal{K}_{GEO}^{ext} is obtained on the filtered data.
GO-REL-VOTE	GO terms are weighed according to their relevance, then used to rank genes. See Section 4.1.3
GO-MAH	GO based gene similarity is used to construct Mahalanobis distance to extract \mathcal{K}_{GEO}^{ext} .
GO-CAN-MAH	Similar to GO-MAH, but only cancer related GO terms are used to calculate gene similarity.
GO-REL-PROP	Relevance is propagated on the graph constructed from the GO based gene similarity. See Section 4.1.2.
Leukemia-FILT	Use genes with ALL-related functions to filter the data, and \mathcal{K}_{GEO}^{ext} is obtained on the filtered data.

Table 5: The building components in different gene ranking methods. ext. knw. cat. stands for categories of external knowledge and int. knw. cat. stands for categories of internal knowledge.

METHODS	KNOWLEDGE SOURCES	EXT. KNW. CAT.	INT. KNW. CAT.	RANKING CRITERION
SPEC	Expression	\mathcal{K}_{GEO}^{ext}	\mathcal{K}_{GEO}^{int}	Geometric Consistency
Fisher Score	Category	\mathcal{K}_{CAT}^{ext}	\mathcal{K}_{CAT}^{int}	Supervised Gene Selection
Pathway-FILT	Pathway	\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{GEO}^{int}	Geometric Consistency
GO-REL-VOTE	GO Anno	\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{FUN}^{int}	Functional Relevance Voting
GO-MAH	GO Anno	\mathcal{K}_{SIM}^{ext}	\mathcal{K}_{GEO}^{int}	Geometric Consistency
GO-CAN-MAH	GO Anno, CAN Anno	$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{FUN}^{ext}$	\mathcal{K}_{GEO}^{int}	Geometric Consistency
GO-REL-PROP	GO Anno, CAN Anno	$\mathcal{K}_{SIM}^{ext}, \mathcal{K}_{FUN}^{ext}$	$\mathcal{K}_{CON}^{int}, \mathcal{K}_{FUN}^{int}$	Relevance Propagation
Leukemia-FILT	CAN Anno	\mathcal{K}_{FUN}^{ext}	\mathcal{K}_{GEO}^{int}	Geometric Consistency

5.2 Experiment Setup

Using different types of knowledge results different gene ranking algorithms. In the experiment, we tried **8 different methods**, using various types of knowledge to rank genes. The description and the building components of the algorithms can be found in Table 4 and 5. 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 ranking lists obtained from the 8 methods is aggregated in three ways: $KOGS_{Borda}$, $KOGS_{Prob}$ and $KOGS_{Prob-SUP}$, which correspond to use Borda count [5] and the probabilistic model proposed in Section 4.2 using all genes and only acute lymphoblastic leukemia (ALL) related genes as **G** respectively. Borda count is a representative rank aggregation algorithm based on majority voting, which is used in the experiment as a baseline rank aggregation algorithm. To evaluate the performance of different methods, we use four different types of evaluation criteria. They are: (1) **Accuracy**: accuracy of INN 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 **HIT_{leu}**, the hit counts of known cancer related genes and acute lymphoblastic leukemia related genes in the top ranked genes provided by methods; and (4) **REL_{pos}**, the number of possible biologically relevant genes found by biologists in the top 50 genes provided by each algorithm. The possible relevant gene lists are obtained from our biologist collaborators, who use their domain knowledge to remove those irrelevant genes from the original gene list. The genes in the possible relevant gene lists are either known to be leukemia-related according to literature or are unknown to be related, but their relevance cannot be ruled out according to their biological

functions or roles in biological process. In the experiment, given \mathcal{K}_{GEO}^{int} , the geometric consistency of genes is evaluated by SPEC [32], which is also introduced in Section 4.1.1. And when \mathcal{K}_{CAT}^{int} is given, Fisher Score is used to rank genes.

5.3 Empirical Findings

Table 6 contains the experimental results obtained from methods ranking genes using different types of knowledge. Based on the results we report the following observations.

Comparing on accuracy, among the 8 methods using only one or two types of knowledge. Only Fisher Score achieved accuracy higher than 0.9. High accuracy indicates that the genes in the list are statistically relevant, since they can separate samples from different categories. According to the results, we can see that among the 8 methods, only the genes selected by Fisher Score bear high statistical relevance. We also notice that comparing with SPEC, GO-MAH achieved a 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 Section 3.3.1 is effective. For the methods derived from KOGS, we observed that all three methods result in accuracy higher than 0.9, indicating that the genes selected by the ranking methods derived from the KOGS framework have high statistical relevance.

Comparing on the four biological relevance measurements: Sim_{anno} , HIT_{canc} , HIT_{leu} and REL_{pos} , the two methods using \mathcal{K}_{FUN}^{int} (GO-REL-VOTE and GO-REL-PROP) achieve good performance, while Fisher Score does not perform well. This is reasonable, since in Sim_{anno} , HIT_{canc} , HIT_{leu} and REL_{pos} , we actually use \mathcal{K}_{FUN}^{int} to evaluate the relevance of 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; and this is analogous to the case when \mathcal{K}_{CAT}^{int} is provided, Fisher Score can achieve better perfor-

Table 6: Performance comparison for gene ranking methods using one type of knowledge with using multiple types of knowledge. ACC-10, ACC-30 and ACC-50 correspond to the accuracy achieved on the top 10, 30 and 50 genes provided by different algorithms. ACC-AVE is the averaged accuracy achieved using the top 10, 15, . . . , 50 genes provided by algorithms. Results show that methods generated from the KOGS framework can select genes bearing both statistical and biological relevance.

METHODS	ACC-10	ACC-30	ACC-50	ACC Ave	Sim _{anno}	HIT _{canc}	HIT _{leu}	REL _{pos}
SPEC	0.64	0.66	0.83	0.65	797	2	0	21
Fisher Score	0.97	0.97	0.97	0.97	823	8	2	14
Pathway-FILT	0.61	0.81	0.89	0.81	807	4	0	19
GO-REL-VOTE	0.56	0.69	0.83	0.64	7686	26	8	25
GO-MAH	0.69	0.80	0.86	0.82	759	3	0	14
GO-CAN-MAH	0.62	0.83	0.86	0.80	2996	5	1	17
GO-REL-PROP	0.70	0.78	0.86	0.74	7688	22	15	33
Leukemia-FILT	0.55	0.62	0.64	0.62	687	4	1	20
KOGS _{Borda}	0.91	0.97	0.97	0.96	1723	6	2	16
KOGS _{Prob}	0.97	0.94	0.94	0.95	6954	21	12	20
KOGS _{Prob-SUP}	0.94	0.91	0.91	0.93	7766	24	17	29

Table 7: The biologically relevant genes in the top 50 gene list provided by KOGS_{Prob-SUP}.

LMO1, CBFA2T3, TYROBP, STAT5B, IGFBP3, JUN, USP33, GSN, BTG1, TFRC, PTK2, PDE7A, TIMP1, AKT1, FLT1, CEBPD, TIMP2, TIMP4, TYK2, CDK4, SERPINF2, PRKACA, NCOR1, SIVA1, BRD8, CAPN7, SPATA2, PRKAR1A, PPARA

mance than other methods in terms of accuracy. 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 four biological relevance measurements. For the methods derived from KOGS, we observed that the two methods use the probabilistic model proposed in Section 4.2 achieve good performance. Although KOGS_{Borda} does not perform as good as KOGS_{Prob} and KOGS_{Prob-SUP}, it still achieves better performance comparing to the average level achieved by the 8 methods without aggregating ranks. We also noticed that compared to KOGS_{Prob}, KOGS_{Prob-SUP} selects more biologically relevant genes. This clearly suggests that the supervision information used in KOGS_{Prob-SUP} helps.

5.4 Further Study on Biological Relevance

In order to closely examine the biological relevance of the selected genes, we performed a further study, in which our biologist collaborators examined the top 50 genes selected by KOGS_{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_{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 sam-

ples but not in T-ALL samples [1]. 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 [30]. TIMP1 and TIMP2, members of Tissue Inhibitor of MetalloProteinases, were found related to the infiltration of ALL leukemia cells into extramedullary organs [29]. 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[31]. 12 genes are found to be possibly leukemia or cancer related due to the following reasons: (1) their functions on tumorigenesis and cell cycle control (e.g., PPARA, TIMP4 and CDK4); (2) their cAMP-dependence (PRKACA and PRKAR1A); (3) transcription factors (BRD8 and NCOR1), whose expressions 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 antiapoptotic activity [8]. The study of these genes may help identify new biomarkers crucial to leukemia tumorigenesis.

The results obtained from the experiment demonstrated that the methods derived from the proposed knowledge oriented gene selection framework, KOGS, is able to select genes which bear both statistical significance and biological relevance. Therefore the proposed framework is effective.

6. CONCLUSION

In this work, we proposed KOGS, a general framework for knowledge oriented gene selection to convert different types of external knowledge to internal knowledge for ranking genes. Given multiple gene ranking lists, KOGS can aggregate them to form a final list considering various gene relevance. Experimental results demonstrated the methods derived from KOGS can select genes bearing both statistical significance and biological relevance. In [33], we studied the problem of gene selection using multiple data sources. 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 [33]; (2) In the current work, the coefficient combination can be automatically learned, while this problem is not addressed in [33]; and (3) KOGS is based on combining ranking lists, while our earlier replies on combining sample similarity, which restricts the model flexibility. We noticed that supervised ranking aggregation is also studied in [16], but it requires to provide the supervision information via partial orders among entry pairs, which is not intuitive in our application. The developed KOGS framework forms our preliminary work for knowledge oriented gene selection. Our ongoing work includes: (1) understanding the roles of different types of knowledge in gene selection, (2) including more types of knowledge in KOGS, and (3) developing a user friendly toolbox for knowledge oriented gene selection to assist biologists' research.

7. REFERENCES

- [1] S. Aerts, et al. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537–545, 2006.
- [2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [3] E. Camon, et al. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32:262–266, 2004.
- [4] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2 edition, 2001.
- [5] C. Dwork, R. Kumar, M. Naor, and D. R. Sivakumar. Aggregation methods for the web. In *Proceedings of the 10th Int. World Wide Web Conference*, 2001.
- [6] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, 5:845–889, 2004.
- [7] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.
- [8] R. Gudi, 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.
- [9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- [11] J. C. Huang, et al. Using expression profiling data to identify human microRNA targets. *NATURE METHODS*, 4:1045–1049, 2007.
- [12] N. Jones and P. Pevzner. *An Introduction to Bioinformatics Algorithms*. The MIT Press, 2004.
- [13] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucl Acids Res*, 28:27–30, 2000.
- [14] 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.
- [15] J. Liao and K.-V. Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *BIOINFORMATICS*, 23:1945–1951, 2007.
- [16] Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li. Supervised rank aggregation. In *Proceedings of the 16th int. conference on World Wide Web*, 2007.
- [17] J. Lu, et al. MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838, 2005.
- [18] P. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 12:49–55, 1936.
- [19] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcao, and F. M. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9:S4, 2008.
- [20] C. D. Pitta, L. Tombolan, M. C. Dell’Orto, and B. Accordi. A leukemia-enriched cdna microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia. *Haematologica*, 90:890–898, 2005.
- [21] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13:252–264, 1991.
- [22] F. Schalekamp and A. van Zuylen. Rank aggregation: Together we’re strong. In *Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2009.
- [23] B. Scholköpfung and A. J. Smola. *Learning with Kernels*. The MIT Press, 2002.
- [24] C. M. Schueller, et al. Towards a comprehensive catalog of gene-disease and gene-drug relationships in cancer. Tech. report, National Cancer Institute, 2005.
- [25] C. Sima and E. R. Dougherty. What should be expected from feature selection in small-sample settings. *Bioinformatics*, 22:2430–2436, 2006.
- [26] A. Smola and I. Kondor. Kernels and regularization on graphs. In *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, 2003.
- [27] C. Stark, et al. Biogrid: A general repository for interaction datasets. *Nuc Acids Res*, 34:535–539, 2006.
- [28] A. Subramanian, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102:15545–15550, 2005.
- [29] 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.
- [30] T. Boehm, et al. 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.
- [31] H. Yu and R. Jove. The stats of cancer – new molecular targets come of age. *Nature Reviews Cancer*, 4:97–105, 2004.
- [32] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *International Conference on Machine Learning (ICML)*, 2007.
- [33] Z. Zhao, J. Wang, H. Liu, J. Ye, and Y. Chang. Identifying biologically relevant genes via multiple heterogeneous data sources. In *SIGKDD*, 2008.