Identifying Biologically Relevant Genes via Multiple Heterogeneous Data Sources

Zheng Zhao† Jiangxin Wang‡ Huan Liu† Jieping Ye† 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, huan.liu, jieping.ye, yung.chang}@asu.edu

ABSTRACT

Selection of genes that are differentially expressed and critical to a particular biological process has been a major challenge in post-array analysis. Recent development in bioinformatics has made various data sources available such as mRNA and miRNA expression profiles, biological pathway and gene annotation, etc. Efficient and effective integration of multiple data sources helps enrich our knowledge about the involved samples and genes for selecting genes bearing significant biological relevance. In this work, we studied a novel problem of multi-source gene selection: given multiple heterogeneous data sources (or data sets), select genes from expression profiles by integrating information from various data sources. We investigated how to effectively employ information contained in multiple data sources to extract an intrinsic global geometric pattern and use it in covariance analysis for gene selection. We designed and conducted experiments to systematically compare the proposed approach with representative methods in terms of statistical and biological significance, and showed the efficacy and potential of the proposed approach with promising findings.

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

1. INTRODUCTION

Decades of research have proven that cancer is a heterogenous cellular disorder caused by the deregulation of many interacting cellular pathways that generate tumor formation and growth. Functional genomics and proteomics techniques, such as global-expression profiling and molecular interaction screen, are being used to explore how these genes work together to trigger (or maintain) growth, survival, progression, and invasiveness of malignant cells. Microarrays allow biological researchers to examine the messenger RNA (mRNA) expression levels of tens of thousands of genes at once, and possibly to analyze the expression profiles of an entire genome. Despite these advances of microarray technologies, bottlenecks still exist, primarily in identification and selection of genes critical to the biological phenotype and processes of particular interests. Although steady progress has been made recently in statistical data analysis [16, 21], biomedical researchers using microarrays still face daunting challenges in understanding and determining the biological significance of "wide data" with tens of thousands of genes but relatively small numbers of samples. For example, fold differences identified via statistical analysis offers limited or inaccurate selection of biological features, leading to inaccurate groups of genes among different types of cancer [19, 25]. In order to elucidate mechanisms underlying oncogenesis or other biological processes, much remains to be explored on how to organize and interpret microarrays based on both biological relevance and statistical significance. Since the number of samples cannot be increased considerably in the near future, additional data sources can be exploited to enhance our understanding of the data in hand.

In this work, we studied a novel problem of gene selection using multiple heterogeneous data sources. It is a problem arising from a study of bioinformatical cancer research in which cancerous samples are collected with both mRNA and microRNA (miRNA) (i.e., two different data sources) with gene properties and relationships being provided by gene function annotation [14], gene ontology annotation [4] and biology pathway (i.e., another three data sources). Our goal is to find pertinent genes from the mRNA expression profiles (henceforth, the target) with the help of four additional data sources. The additional data sources fall into two categories: (1) the data sources about genes - gene function annotation, gene ontology annotation and biology pathway; and (2) the data sources about relationships among samples - miRNA and miRNAs expression profiles sharing the same set of samples. Recent studies suggest the role of miRNA as both oncogenes and anti-oncogenes [11, 28]. For instance, *miR-15* and *miR-16* have been shown to induce apoptosis by targeting *BCL2* [6], acting as anti-oncogenes, whereas *miR-21* targets the tumor suppressor gene *TPM1* and blocks apopto-

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

sis, thereby functioning as oncogenes and promoting tumor growth [23]. Studies also reveal that miRNA profiles are surprisingly informative for depicting sample relationships in terms of separating tissues of cancer and noncancer, as well as different types of cancers [19, 13]. It is, however, not trivial to integrate data sources describing the relationships among samples as well as data source describing the relationships among genes. Since disparate data sources contain information of different perspectives, effectively integrating them for gene selection remains a challenge. To address the problem, we proposed a framework for multi-source gene selection. We focus on unlabeled data in this work. Thus the proposed framework is unsupervised by definition. We later showed that if class label is available, the framework can be adapted to handle labeled data in a straightforward way.

Gene selection is also known as feature selection [17] in machine learning. Although various approaches for unsupervised feature selection have been proposed [12, 32, 8], the potential of using multiple heterogeneous data sources for unsupervised feature selection has not been explored to help biological investigation. We notice that gene prioritization using multiple data sources is previously studied in [3]. However, it is essentially a classification problem in which genes are treated as instances and only data sources describing genes relationships are used. Therefore despite its significance, gene selection using multiple heterogeneous data sources is a problem that has yet been addressed. Before presenting the framework for solving the problem, we first provide notations used in this work.

2. NOTATIONS

In the paper we assume the given mRNA expression profile $\mathcal{D}$ containing d genes and n samples. We use $F_1, \ldots, F_d$ to denote the d genes and $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d$ the corresponding vectors in the profile. We use $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ to denote the n samples, and for each sample we have, $\mathbf{x}_i \in R^d$. In feature selection, $F_1, \ldots, F_d$ are also called features, $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d$ are the corresponding feature vectors. Besides the mRNA expression profile, we also have h^G additional data sources, $\mathcal{D}_1^G, \ldots, \mathcal{D}_{h^G}^G$ depicting the relationships among genes; and h^S additional data sources, $\mathcal{D}_1^S, \ldots, \mathcal{D}_{h^S}^S$ depicting the relationships among samples. The representation for the relationships among genes and samples can be heterogenous. In the work, we use spectral graph theory [5] as a tool to explore the geometric structure of a data. Let $\mathbb{G}$ denote a graph, its *similarity matrix* is denoted by $W(i,j) = w_{ij}$. Let $\mathbf{d}$ denote the vector: $\mathbf{d} = \{d_1, \ d_2, \ldots, \ d_n\}$, where $d_i = \sum_{k=1}^{n} w_{ik}$, the *degree matrix* D of the graph $\mathbb{G}$ is defined by: $D(i,j) = d_i$ if $i = j$, and 0 otherwise. Given D and W, the *Laplacian matrix* L and the *normalized Laplacian matrix* $\mathcal{L}$ are defined as: $L = D - W$; $\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. It is well known that the leading eigenvectors of L and $\mathcal{L}$ form soft cluster indicators of the data [29]. We use $\mathbf{S}_+$ to denote the space of symmetric positive semidefinite matrices; K, a kernel matrix; $\mathbf{C}$, the covariance matrix; $\mathbf{1}$, the vector with all elements equal to 1; and I, the identity matrix.

3. MULTI-SOURCE GENE SELECTION

Given multiple data sources carrying information about gene and sample relationships with heterogeneous representations, we exploit both types of information for effective gene selection. The key to the problem necessitates a common way to represent knowledge that meets the following requirements: (1) information can be easily extracted from both types of information; (2) information can be combined for integration; and (3) information can be effectively used for gene selection. Hence, we propose to use a geometric pattern among samples as the common representation. We show: (1) given relationships among genes and the mRNA expression profiles, we can use the relationships to obtain a geometric pattern of samples; (2) upon obtaining local geometric patterns from various data sources, we can combine them to form a global pattern; and (3) using the obtained global pattern we can effectively select genes that bear both biological and statistical relevance. As shown in Figure 1, given multiple heterogeneous data sources, we first extract local sample geometric patterns from each data source, then combine them to form a global pattern, based on which discriminative genes can be identified. We first show how to select genes with a given geometric pattern of samples. We propose *Geometry-Dependent Covariance* and employ it for gene selection. In the following, we adopt the notations from feature selection, and use *features* and *genes* interchangeably.

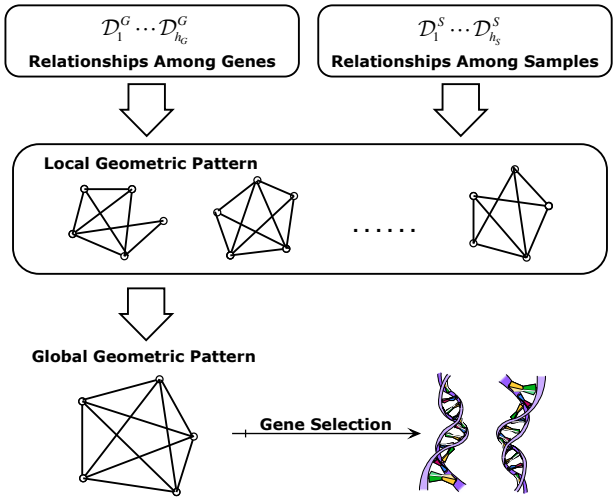

Figure 1: Integration of multiple heterogeneous data sources for gene selection: Local geometric patterns are extracted from data sources and are combined to form a global pattern for identifying relevant genes.

3.1 Gene Selection Using Geometric Patterns

3.1.1 Geometry-Dependent Covariance

Let $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be a data set containing n samples. The sample covariance matrix [27] of X is given by:

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

Here $\bar{\mathbf{x}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$ and $\mathbf{C} \in \mathbf{S}_+$ is an unbiased estimator of the covariance matrix. $\mathbf{C}$ captures the correlation between all possible feature pairs. Let $\mathbf{C}_{i,j}$ be the ij-th element of $\mathbf{C}$. It measures the covariance between features F_i and F_j, and is given by:$\mathbf{C}_{i,j} = \frac{1}{n-1} \left(\boldsymbol{f}_i - \bar{f}_i \cdot \mathbf{1}\right)^T \left(\boldsymbol{f}_j - \bar{f}_j \cdot \mathbf{1}\right)$. Here, $\bar{f}_i = \sum_{k=1}^{n} f_{ik}$. Assuming data has been centralized:

$\bar{\mathbf{x}} = 0$, it can be verified that $\forall i, \bar{f}_i = 0$, and we have:

$$\mathbf{C}_{i,j} = \frac{1}{n-1} \boldsymbol{f}_i^T \boldsymbol{f}_j. \quad (2)$$

To simplify the notation, in the following, we assume X has been centralized. In this work we propose to adjust the covariance measures between features according to a given geometric pattern of samples. Given W, the similarity matrix that captures the geometric pattern of samples, and L, the corresponding Laplacian matrix. We give the concept of the geometry-dependent sample covariance.

DEFINITION 1. *Geometry-Dependent Sample Covariance with L. Given $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, the feature vectors of features F_i and F_j, and L, a Laplacian matrix, the Geometry-Dependent Sample Covariance between F_i and F_j is given by:*

$$\widehat{\mathbf{C}}_{i,j} = \frac{1}{n-1} \boldsymbol{f}_i{}^T \Gamma(L) \boldsymbol{f}_j, \quad (3)$$

where $\Gamma : \mathbf{S}_+ \to \mathbf{S}_+$ is a prescribed spectral matrix function which is induced from an non-increasing real function of positive input, $\gamma(\cdot) : (0, \infty) \to (0, \infty)$.

Let $A \in \mathbf{S}_+$, $A = U \Sigma U^T$ be the singular value decomposition (SVD) [10] of A, so $U^T U = I$ and $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Here $\text{diag}(a_1, \ldots, a_n)$ denotes the diagonal matrix with a_i as its i-th diagonal element. We can calculate $\Gamma(A)$ by $\Gamma(A) = U \hat{\Sigma} U^T, \hat{\Sigma} = \text{diag}(\gamma(\lambda_1), ..., \gamma(\lambda_n))$. We require $\gamma(\cdot)$ to be non-increasing and an example is to define $\gamma(\lambda) = \frac{1}{\lambda + \epsilon}$, where $\epsilon > 0$. This gives $\Gamma(L) = (L + \epsilon I)^{-1}$.

In Equation (3), $\boldsymbol{f}$ and L are used to calculate the covariance measure, however the scale of the two factors can affect the measure arbitrarily. An effective way to handle this issue is to apply normalization on $\boldsymbol{f}$ and L. Let the normalized feature vector[1] of $\boldsymbol{f}$ be defined as:

$$\widetilde{\boldsymbol{f}} = \frac{D^{\frac{1}{2}} \boldsymbol{f}}{\left\| D^{\frac{1}{2}} \boldsymbol{f} \right\|} \quad (4)$$

We can define a geometry-dependent covariance measure using normalized Laplacian matrix $\mathcal{L}$:

DEFINITION 2. *Geometry-Dependent Sample Covariance with $\mathcal{L}$. Given $\widetilde{\boldsymbol{f}}_i$ and $\widetilde{\boldsymbol{f}}_j$, the normalized feature vectors of F_i and F_j, and $\mathcal{L}$, the normalized Laplacian matrix, Geometry-Dependent Sample Covariance between F_i and F_j is:*

$$\widetilde{\mathbf{C}}_{i,j} = \frac{1}{n-1} \widetilde{\boldsymbol{f}}_i{}^T \Gamma(\mathcal{L}) \widetilde{\boldsymbol{f}}_j, \quad (5)$$

In the two definitions, we assume data has been centralized. We have the following theorem for $\widehat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$:

Theorem 1. *$\widehat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$ have following properties:*
1. *Assume $\gamma(0) = 0$ and all features have unit norm[2], let $W = \frac{1}{n} \cdot \mathbf{1}\mathbf{1}^T$, we have $\widehat{\mathbf{C}} = \widetilde{\mathbf{C}} = \gamma(1) \cdot \mathbf{C}$, where $\mathbf{C}$ is the standard covariance matrix defined in Equation (1).*
2. *$\widehat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$ are symmetric positive semidefinite.*

[1]The form of normalization is commonly used in spectral dimension reduction and clustering, e.g. Laplacian Eigenmap [1]. Basically, the feature vectors are adjusted according to their nearby data density before normalization.

[2]It can be verified that if $\bar{\mathbf{x}} \neq 0$, the proposition still can hold when $\gamma(0) = 0$. Also if features are not normalized, the proposition still holds for $\widehat{\mathbf{C}}$ but not $\widetilde{\mathbf{C}}$.

Proof: The second property can be easily verified, here we provide a proof for the first one. Given the definition of W, it can be verified that $D = I$, therefore $L = \mathcal{L}$. Since all features have unit norm, we have $\widetilde{\boldsymbol{f}}_i = \boldsymbol{f}_i$ which means $\widehat{\mathbf{C}} = \widetilde{\mathbf{C}}$. Let $L = U \Sigma U^T$ be the SVD of L, it can be verified $\Sigma = diag(1, \ldots, 1, 0)$, and 1 is the $n-1$ repeat eigenvalues of L. Thus we have $\Gamma(L) = \gamma(1) \cdot L$, also $\widehat{\mathbf{C}} = \widetilde{\mathbf{C}} = \frac{1}{n-1} \gamma(1) \cdot XLX$. Since $L = I - \frac{1}{n} \cdot \mathbf{1}\mathbf{1}^T$, we have $L = L^2$. Substituting in L^2 and by noticing $XL = \left(X - \bar{\mathbf{x}}\mathbf{1}^T\right)$, we have:

$$\begin{aligned} \widehat{\mathbf{C}} = \widetilde{\mathbf{C}} &= \frac{1}{n-1} \gamma(1) \cdot \left(X - \bar{\mathbf{x}}\mathbf{1}^T\right) \left(X - \bar{\mathbf{x}}\mathbf{1}^T\right)^T \\ &= \gamma(1) \cdot \mathbf{C} \end{aligned}$$ ■

The theorem says that (1) by setting W in a special form, the geometry-dependent covariance matrix is equivalent to a scaled standard covariance matrix; and (2) since any symmetric positive semidefinite matrix is a valid covariance matrix, we see $\widehat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$ are valid as covariance matrices.

Given two feature vectors, $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, geometry-dependent covariance measures their covariance by considering geometric pattern captured by the Laplacian matrix L. To see this, let $L = U \Sigma U^T$, $U = (\mathbf{u}_1, \ldots, \mathbf{u_n})$, we have

$$\widehat{\mathbf{C}}_{i,j} = \left(\Gamma(\Sigma)^{\frac{1}{2}} U^T \boldsymbol{f}_i\right)^T \left(\Gamma(\Sigma)^{\frac{1}{2}} U^T \boldsymbol{f}_j\right). \quad (6)$$

To compare $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, geometry-dependent covariance first projects the two vectors into a space spanned by $(\mathbf{u}_1, \ldots, \mathbf{u_n})$, then weight the transformed vectors by $\gamma(\lambda_1)^{\frac{1}{2}}, \ldots, \gamma(\lambda_d)^{\frac{1}{2}}$. Let $a_i = \mathbf{u}_i^T \boldsymbol{f}$, where a_i measures how well $\boldsymbol{f}$ and $\mathbf{u}_i$ align to each other. We can write weighted transformed vector as:

$$\boldsymbol{f}_w = \left(a_1 \gamma(\lambda_1)^{\frac{1}{2}}, \ldots, a_n \gamma(\lambda_n)^{\frac{1}{2}}\right)^T \quad (7)$$

Each eigenvalue of L measures the consistency between the corresponding eigenvector and the given geometric pattern captured by L [29]. The bigger the value, the more inconsistent the eigenvector. A non-increasing $\gamma(\cdot)$ ensures to assign small weights to eigenvectors which are inconsistent to the given geometric pattern. If a feature vector $\boldsymbol{f}$ only aligns well to inconsistent eigenvectors, the weighting mechanism ensures that all elements in $\boldsymbol{f}_w$ will be small. Thus, in covariance calculation, the inner product between $\boldsymbol{f}_w$ and another weighted transformed vector will be small, even if the two vectors align well. This explains why the proposed measure is *geometry-dependent*. Similar analysis holds for $\widetilde{\mathbf{C}}$.

3.1.2 Gene Selection

The global geometry pattern obtained from multiple data sources reflects some intrinsic relationship among instances. Therefore, genes supporting this relationship need to be consistent with the obtained pattern [32]. With the global geometry pattern, one can build a geometry-dependent covariance matrix, based on which features can be selected using two methods: (1) sorting the diagonal of the covariance matrix and choosing the features with the biggest variances; and (2) applying sparse principle component analysis (SPCA) [7] to select a set of features that can maximally retain the total variance, assuming that all features have similar scales. According to Equation (7), the diagonal elements of $\widehat{\mathbf{C}}$ is given by: $\widehat{\mathbf{C}}_{i,i} = \sum_{i=1}^{n} a_k^2 \gamma(\lambda_k)$. As we discussed, for $\widehat{\mathbf{C}}_{i,i}$ to achieve big value, $\boldsymbol{f}_i$ must align well

to the eigenvectors that are consistent with the given geometric pattern depicted by L. In other words, a bigger $\widehat{\mathbf{C}}_{i,i}$ indicates that $\boldsymbol{f}_i$ is more consistent with the given geometric pattern. Therefore, selecting features according to the first method is equivalent to selecting features which are consistent with the given geometric pattern. Similar analysis holds for $\widetilde{\mathbf{C}}$, where all features are automatically normalized to have similar scale. Since the first method measures feature relevance individually, it may select a feature set with redundancy. The second method, applying SPCA, considers the interaction among features, and is able to select a feature set containing less redundancy. Gene selection aims to identify all relevant genes as candidates for examination. Hence, the first method is more proper becuase the process of removing statistical redundancy may cause the removal of potentially biologically relevant genes.

3.1.3 Efficient Calculation

The calculation of geometry-dependent covariance involves a spectral matrix function induced from a non-increasing real function, which may require a full eigen decomposition on L or $\mathcal{L}$, and has a time complexity of $O(n^3)$, where n is the number of samples. This does not cause problems for gene selection, since the involved data are usually of small sample size. However, if the sample size is large, calculating $\Gamma(L)$ or $\Gamma(\mathcal{L})$ will be expensive. With the following theorem, we show that for the geometry-dependant covariance defined with $\mathcal{L}$, given $W \in \mathbf{S}_+$ [3], we are able to obtain $\Gamma(\mathcal{L})$ in $O(n^2)$ with a special $\gamma(\cdot)$. The theorem can be verified via the fact that the i-th eigenvalue and eigenvector of $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ is given by $1-\lambda_i$ and u_i of $\mathcal{L}$, and $u_1 = \frac{D^{\frac{1}{2}}1}{\left\|D^{\frac{1}{2}}1\right\|}$.

Theorem 2. *Given $\gamma(\cdot)$ defined as:*

$$\gamma(\lambda) = \begin{cases} 0, & \lambda = 0 \\ 1-\lambda, & \lambda > 0 \end{cases} \tag{8}$$

We have

$$\Gamma(\mathcal{L}) = D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - \frac{D^{\frac{1}{2}}11^TD^{\frac{1}{2}}}{1^TD1}$$

$$\widetilde{\mathbf{C}} = \Pi X\left(W - \frac{W\mathbf{1}\mathbf{1}^TW}{\mathbf{1}^TW\mathbf{1}}\right)X^T\Pi, \tag{9}$$

where Π is the diagonal matrix with $\Pi_{i,i} = \left\|D^{\frac{1}{2}}\boldsymbol{f}_i\right\|^{-1}$.

3.2 Geometric Pattern Extraction

Given relationships among samples, it is straightforward to extract geometric patterns. We now show how to extract a geometric pattern of samples using gene relationships and study how to construct a global pattern via linearly combining obtained local patterns.

3.2.1 Exploiting Gene Relationships

Given relationships among genes, we propose to use the relationships to construct a covariance matrix $\mathbf{C}_g$ for genes, which can be used to calculate the Mahalanobis distance [20] among samples to reveal sample geometric patterns:

$$d(\mathbf{x},\mathbf{y})^2 = (\mathbf{x}-\mathbf{y})^T\mathbf{C}_g^{-1}(\mathbf{x}-\mathbf{y}). \tag{10}$$

[3]For example, W is provided by a kernel matrix. This condition ensures the eigenvalues of $\mathcal{L}$ is bounded by 1 from above, so that $\gamma(\cdot)$ is valid.

In Equation (10), $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ are gene expression profile of two samples. By incorporating $\mathbf{C}_g$ for calculating sample distance, we adjust the sample geometric pattern by considering the given relationships among genes. In real applications, the relationships among genes are usually specified by: graphs (or kernels), e.g. biological pathway and protein-protein interaction. Or can be derived from gene descriptions, e.g. gene annotation [4]. In the second case, each gene is described by several terms from a fixed vocabulary. If we treat genes as features, $\mathbf{C}_g$ can be obtained directly via Equation (1). However in the first case, the explicit expression of genes is unavailable, therefore in the case we can not obtain $\mathbf{C}_g$ directly. To handle the issue, we propose to calculate an embedding [2] from a given affinity matrix W or kernel matrix K and use the obtained embedding to construct $\mathbf{C}_g$ with Equation (1).

Specifically, given an affinity matrix W, we propose to construct commute time embedding [18], which is shown to be effective in real applications [29]. Let $L = D - W$ and $L = U\Sigma U^T$ be the SVD of L, the embedding of $F_1,\ldots,F_d$ are given by $Y = \left(\Sigma^+\right)^{\frac{1}{2}}U^T$, where each column of Y corresponding to an $\boldsymbol{f}_i$. By transposing Y we obtain the explicit expression of features: $X^W_{EM} = U\left(\Sigma^+\right)^{\frac{1}{2}}$. By substituting X^W_{EM} in Equation (1), we can obtained a $\mathbf{C}_g$.

Theorem 3. *Given an affinity matrix W depicting relationships among genes, using commute time embedding, $\mathbf{C}_g$ is given by:*

$$\mathbf{C}_g = U\left(\Sigma^+ - \frac{1}{l}\left(\Sigma^+\right)^{\frac{1}{2}}\mathbf{1}\mathbf{1}^T\left(\Sigma^+\right)^{\frac{1}{2}}\right)U^T \tag{11}$$

Similarly, we can show that given a kernel matrix K, using kernel PCA [22] for constructing embedding, we have $\mathbf{C}_g$ is given by: $K\left(I - \frac{1}{l}U\mathbf{1}\mathbf{1}^TU^T\right)K$, where $K = U\Sigma U^T$.

3.2.2 Combining Local Geometric Patterns

Given multiple local geometric patterns, the global pattern can be obtained by linearly combining local patterns [33].

$$W_{global} = \sum_{i=1}^{h^G}\alpha_i^GW_i^G + \sum_{j=1}^{h^S}\alpha_j^SW_j^S \tag{12}$$

In the equation, W_i^Gs are the geometric patterns extracted from relationships among genes; W_i^Ss are the ones extracted from relationships among samples, and α_i^G and α_j^S are the combination coefficients, which can be assigned by domain experts according to their domain knowledge[4] [33], or, if the label information is available, learned automatically via convex optimization based on a set of kernel matrices carrying the information about the local geometric patterns. We refer readers to literature for comprehensive study on the research issues of kernel combination [15, 31].

4. MSGS - THE FRAMEWORK

The above technical discussion has paved way for us to propose a framework for *M*ulti-*S*ource *G*ene *S*election: *MSGS*. The detail of the framework can be found in Algorithm 1. It consists of three major steps: (1) obtaining local geometry pattern from each data source (Lines 3-5); (2) combining the local patterns to construct a global geometry pattern (Line

[4]This provides a way to incorporate domain knowledge

6); and (3) using the global pattern to form a geometry-dependent covariance matrix and selecting genes (Lines 7,8). Below we give analysis on time complexity for *MSGS*.

Algorithm 1 *MSGS*: *Multi-Source Gene Selection*

1: INPUT: $\mathcal{D}_1^G \dots \mathcal{D}_{h^G}^G$, $\mathcal{D}_1^S \dots \mathcal{D}_{h^S}^S$ and X
2: OUTPUT: $List_{gene}$ - the selected gene list
3: **for** each $\mathcal{D}_i \in \left(\mathcal{D}_1^G \dots \mathcal{D}_{h^G}^G, \mathcal{D}_1^S \dots \mathcal{D}_{h^S}^S\right)$ **do**
4: construct W_i^Gs and W_j^Ss, the local geometric patterns;
5: **end for**
6: obtain global pattern W from W_i^Gs and W_j^Ss;
7: form the geometry-dependent covariance matrix $\mathbf{C}$;
8: select genes according to $\mathbf{C}$ and form $List_{gene}$;
9: **return** $List_{gene}$

Since the representation of the $h^G + h^S$ data sources are heterogenous, the time complexity of extracting local patterns from data sources can vary greatly. Assuming for data sources $\mathcal{D}_1^G \dots \mathcal{D}_{h^G}^G$, each data sources provides us an affinity matrix depicting gene relationships and involving l genes, then constructing $\mathbf{C}_g$ using Equation (11) and calculating its (pseudo) inverse requires $O(l^3)$ operations. Computing Mahalanobis distance among n^2 pairs of instances and forming a RBF kernel require $O(l^2n^2)$ operations. Crossing the h^G data source, the total cost is $O\left((l^2n^2 + l^3)h^G\right)$. Assuming for $\mathcal{D}_1^S \dots \mathcal{D}_{h^S}^S$, each data source has d features depicting the same set of n samples, and we use RBF kernel to represent the local sample geometric pattern. The time complexity of forming a RBF kernel on each data source is $O(dn^2)$. And crossing h^S data source, the cost is $O\left(h^s n^2 d\right)$. Therefore, the total cost of the first step is $O\left((l^2n^2 + l^3)h^G + n^2dh^S\right)$. Assuming we linearly combine W_is with a set of prescribed combination coefficients, the cost is $O\left((h^G + h^S)n^2\right)$. Using the method specified in Theorem 2 to form $\widetilde{\mathbf{C}}$, the cost is $O(dn^2 + d^2)$. Selecting genes based on $\widetilde{\mathbf{C}}$ requires $O(d \log d)$ operations, if we use gene variance; or $O(d^3)$, if we use the SPCA approach proposed in [7]. Hence the overall time complexity of *MSGS* is $O\left((l^2n^2 + l^3)h^G + n^2dh^S + d^2\right)$, using gene variance; otherwise, $O\left((l^2n^2 + l^3)h^G + n^2dh^S + d^3\right)$.

5. EXPERIMENTAL STUDY

We empirically evaluate the performance of *MSGS* for multi-source gene selection in terms of statistical significance and biological relevance. Correspondingly, we choose **accuracy** and **hit ratio**[5] as performance measures. *MSGS* is an unsupervised gene selection algorithm using multi-source data. Hence, label information is only used after selecting genes to calculate accuracy during the testing phase to measure how good selected genes are. In addition, we also measure *robustness* of selected genes by varying the number of classes, if class information is used in gene selection. The results presented in this section are based on $\widetilde{\mathbf{C}}$, the geometry-dependent sample covariance matrix defined with $\mathcal{L}$. This is because in the experiment we found $\widetilde{\mathbf{C}}$ provides more robust performance than $\widehat{\mathbf{C}}$, i.e., the normalization step does lead to better performance. We next introduce the data sets used in this work.

[5]For accuracy, we first filter the data with selected genes and build a classifier on the filtered data, then obtain its accuracy as performance measure. For hit ratio, given a list of genes, we check how many of the genes are cancer related by using the gene function annotation information provided by Ingenuity Pathways Analysis (IPA) system [14].

5.1 Data Sets

Human Cancer Data. It consists of five heterogeneous data sources. Two sets of gene expression profiles from a mixture of 88 normal and cancerous tissue samples[6]: a miRNA expression profile for 151 human miRNAs and a mRNA expression profile for 16,063 human mRNAs [19, 13]. miRNA profile provides relationships among samples and it is observed in [19] that comparing with mRNA, miRNA expression profiles is of more power in terms of discriminating cancer from noncancer tissues as well as cancer of different types of tissues. Three gene information profiles are provided: *Gene Function Annotation*, *Biological Pathway* and *Gene Ontology Annotation*. The three profiles are generated as follows. (1) *Gene Function Annotation*: we used the name of involved tissues (e.g., colon, lung, ...) as keywords to search in IPA system [14] for cancer related processes, for each matching process, we add all genes involved in the process to the graph and connected them with edges. This resulted in a graph containing 535 genes. (2) *Biological Pathway*: with the mRNA expression profile, we used the IPA system to infer the relevant pathways, which results in about 40 networks of molecules, connected these networks and obtained a graph involving 571 genes. (3) *Gene Ontology Annotation*: 68 Gene Ontology annotation [4] terms, related to cell cycle, cell growth, differentiation, apotoposis, cancer development and so on, were provided by domain experts. According to whether a term is assigned to a gene, we obtained a matrix with a size of 16063×68, which can be used to compute a covariance matrix among genes. The three profiles provide relationships information among genes. Using the five profiles, we obtained two sets of data: 2C-DATA and 4C-DATA. 2C-DATA contains all 88 tissue samples of the original data and class label is assigned to a sample according to whether the sample is a normal or a tumor tissue. 4C-DATA contains 33 tissue samples from 4 types cancerous tissues, *Mesothelioma, Uterus, Colon and Pancreas*, each has at least 7 samples. A summary of the two data sets is given in Table 1.

Type	Data Sources	Genes	Samples
	2C DATA		
$\mathcal{D}$	mRNA Expression Profile	16063	88
$\mathcal{D}_1^S$	miRNA Expression Profile	151	88
$\mathcal{D}_1^G$	Gene Function Annotation	535	-
$\mathcal{D}_2^G$	Biological Pathway	571	-
$\mathcal{D}_3^G$	Gene Ontology Annotation	4385	-
	4C DATA		
$\mathcal{D}$	mRNA Expression Profile	16063	33
$\mathcal{D}_1^S$	miRNA Expression Profile	151	33
$\mathcal{D}_1^G$	Gene Function Annotation	535	-
$\mathcal{D}_2^G$	Biological Pathway	571	-
$\mathcal{D}_3^G$	Gene Ontology Annotation	4385	-

Table 1: A summary of the Human Cancer Data.

5.2 Experiment Setup

In the experiment, we compare gene selection using single (target) data source with using multiple data sources. We

[6]Tissues involved:colon, pancreas, kidney, bladder, prostate, ovary, uterus, lung, mesothelioma, melanoma and breast.

chose 3 feature selection algorithms for gene selection using single data sources. They are two unsupervised algorithms, Laplacian Score [12] and PathSPCA [7], and one supervised algorithms, ReliefF [24]. Experiments are performed in the Matlab environment. We use *RBF* kernel to represent the local and global geometric patterns. A domain expert determined two sets of combination coefficients for linear combination of the local geometric patterns: COMB-1, using the combination coefficients (0.0, 0.5, 0.5, 0.0, 0.0) for 2C-DATA and (0, 0.3, 0.4, 0.3, 0) for 4C-DATA; and COMB-2, using (0.1, 0.3, 0.3, 0.2, 0.1) for 2C-DATA and (0.1, 0.2, 0.4, 0.2, 0.1) for 4C-DATA. The usefulness of each data source can also be determined by checking the quality of the genes selected by *MSGS* using the geometric pattern extracted from the data source. We also tried to use class label to learn combination coefficients via a kernel learning approach proposed in [31]. We apply 1NN classifier to data with selected genes, and use its accuracy to measure the quality of the gene set. Reported results are based on averaging the results from 10 trials of experiments.

5.3 Empirical Findings

Experimental results are organized in terms of two sets of multi-source data with different numbers of classes. When label information is available, we study how to incorporate it in *MSGS* and its effect. We examine the biological relevance of the genes selected by *MSGS*. In the following, *MSGS*-VAR stands for selecting genes using gene variance, and *MSGS*-SPCA for using the sparse PCA method.

5.3.1 Results on 2C-DATA

Figure 2-(a) compares unsupervised baseline gene selection algorithms using mRNA profiles with *MSGS*-VAR using five individual profiles as well as the combinations of them. The results show that by using miRNA profiles, COMB-1 and COMB-2, *MSGS*-VAR selects genes that provide the best accuracy which are significantly better than those achieved by genes selected using baseline algorithms. Table 2 shows the detailed accuracy results as well as hit ratio. According to hit ratio, averagely, using COMB-2, the gene list provided by *MSGS*-VAR containing the most known cancer relevant genes (7), which is following by using COMB-1 (6.6), miRNA (6.6) and Function Annotation (5). According to hit ratio, Biology Pathway is also helpful (3.6). The averaged number of selected known cancer related genes is 5.8 for SPCA and 3.2 for Laplacian score. We notice that the first two genes selected by *MSGS*-VAR with mRNA profile, Function Annotation, COMB-1 and COMB-2 are all known to be cancer relevant. We will provide biological relevance analysis for selected genes later. The observations suggest that (1) the geometric pattern obtained from miRNA profile indeed possesses better discriminative power, which is consistent with the findings in [19], and (2) given a geometric pattern with higher quality, *MSGS*-VAR can select better genes. This support the use of multiple data sources in gene selection, and show that *MSGS* is effective. (3) By combining multiple heterogeneous data sources we are able to achieve better performance than using any individual data source. This is consistent with the observations in [15]. Figure 2-(c) plots the performance of *MSGS*-VAR when different combination coefficients are used to combine local geometry patterns obtained from miRNA profile and Function Annotation. We observe that the highest accuracy is achieved by using the two data sources together. This indicates the existence of complimentary information, which helps to improve the estimation of the geometric pattern of the underlying model. Similar trends can be observed in the experimental results of *MSGS*-SPCA.

5.3.2 Results on 4C-DATA

Figure 2-(d, e) and Table 2 contain the results on 4C-DATA. We obtained largely similar observations as those from using 2C-DATA. Since the number of samples becomes smaller but the number of classes becomes larger, we observe that the unsupervised baseline algorithms perform worse comparing with on 2C-DATA. However, the performance of *MSGS* using miRNA profile, Function Annotation, Biological Pathway, COMB-1 and COMB-2 are consistently good in terms of accuracy and hit ratio. This indicates that *MSGS* can effectively select relevant genes given high quality geometric patterns.

5.3.3 Incorporating Label Information

When label information is available, we can incorporate it in *MSGS* in two ways: (1) using a supervised metric learning algorithm to learn a geometric pattern on the data and input the obtained pattern into *MSGS*. In this work we implement *MSGS*-OLDA: we use OLDA [30] to learn a geometric pattern and feed it into *MSGS*-VAR. (2) Learning the combination coefficient automatically with a supervised kernel learning algorithm and input the combined pattern into *MSGS*. In this work we implement *MSGS*-kerCB: we use the approach proposed in [31] to learn the combination coefficients[7]. For comparison we also include a supervised selection algorithm ReliefF [24] in the experiment. The upper part of Tabel 3 shows the performance of supervised gene selection algorithms on 2C-DATA. For accuracy, *MSGS*-OLDA performs the best (0.92). For average hit ratio, *MSGS*-KerCB performs the best (5.6), while we observed that the averaged hit ratio of *MSGS*-VAR with COMB-2 is 7.

Furthermore, we experimented how robust supervised gene selection is by applying the genes selected on 2C-DATA by supervised algorithms and *MSGS*-VAR to 4C-DATA to check their discriminative power. The results are shown in the lower part of Table 3 and Figure 2-(f). We observe that the genes selected by *MSGS*-VAR using miRNA profile, COMB-1 and COMB-2 can discriminate cancer from different types of tissues, however, those selected by supervised algorithms cannot. This finding suggests that (1) results of supervised feature selection hinge upon the target concept and the change of class information can results in the selection of different genes; and (2) the geometric pattern corresponding to miRNA profiles as well as COMB-1 and COMB-2 is relatively stable - genes selected by *MSGS* can discriminate data of different numbers of classes, indicating that it may be consistent with intrinsic structure of the underlying model and more robust. *MSGS* is an unsupervised algorithm and obtains intrinsic patterns that do not vary with class definitions from multiple data sources. This is not so for supervised algorithms: with different class definitions, different sets of genes will be selected, which is clearly shown in Table 3.

[7]The learnt coefficients on 2C-DATA is: (0,0.11,0.89,0,0)

Algorithms	2	5	10	20	30	Ave	2	5	10	20	30	Ave
	UNSUPERVISED, 2C-DATA						UNSUPERVISED, 4C-DATA					
SPCA	0.31	0.31	0.38	0.66	0.60	0.45	0.25	0.25	0.21	0.21	0.27	0.24
	1	**4**	5	8	11	5.8	**2**	2	2	3	6	3
Laplacian Score	0.62	0.66	0.70	0.65	0.73	0.67	0.22	0.22	0.22	0.22	0.22	0.22
	0	0	2	6	8	3.2	1	2	3	8	10	4.8
	MSGS-VAR, 2C-DATA						MSGS-VAR, 4C-DATA					
MSGS	0.68	0.78	0.72	0.74	0.78	0.74	0.15	0.43	0.43	0.37	0.24	0.32
mRNA	0	0	2	3	6	2.2	0	0	0	1	2	0.6
MSGS	0.69	0.74	0.88	0.91	0.88	0.82	0.60	0.51	0.63	0.82	0.78	0.67
miRNA	**2**	**4**	6	8	**13**	6.6	1	2	5	10	13	6.2
MSGS	0.69	**0.79**	0.86	0.85	**0.89**	0.82	0.42	0.61	0.76	0.85	0.79	0.69
Function	**2**	**4**	5	6	8	5	**2**	3	7	**13**	16	8.2
MSGS	0.61	0.64	0.65	0.75	0.72	0.67	0.61	**0.81**	0.88	0.91	0.88	0.82
Pathway	0	1	3	6	8	3.6	**2**	3	5	12	**17**	7.8
MSGS	0.69	0.76	0.84	0.73	0.65	0.73	0.33	0.33	0.40	0.50	0.38	0.39
GO Terms	0	0	2	2	6	2	1	1	3	4	6	3
MSGS	**0.82**	0.74	0.88	**0.93**	**0.89**	**0.85**	**0.67**	0.79	0.88	**0.94**	0.91	**0.84**
COMB-1	**2**	**4**	6	8	**13**	6.6	**2**	**5**	**8**	11	**17**	**8.6**
MSGS	0.69	0.74	**0.90**	0.90	**0.89**	0.82	0.33	**0.81**	**0.94**	**0.94**	**0.94**	0.79
COMB-2	**2**	**4**	**7**	**9**	**13**	**7**	**2**	**5**	6	11	15	7.8
	MSGS-SPCA, 2C-DATA						MSGS-SPCA, 4C-DATA					
MSGS	0.68	0.78	0.72	0.75	0.77	0.74	0.23	0.21	0.30	0.24	0.25	0.25
mRNA	0	0	2	4	6	2.4	0	0	1	2	2	1
MSGS	0.69	0.64	0.72	**0.89**	0.85	0.76	0.30	0.30	0.57	0.70	0.69	0.51
miRNA	**2**	**3**	**5**	**9**	**14**	**6.6**	1	3	**7**	**10**	**14**	**7**
MSGS	0.69	**0.81**	0.83	0.85	**0.90**	**0.82**	0.51	0.54	0.54	0.75	0.78	0.62
Function	2	3	3	7	10	5	**2**	**5**	**7**	10	11	**7**
MSGS	0.61	0.54	0.63	0.73	0.69	0.64	0.40	**0.58**	**0.86**	0.67	0.88	0.67
Pathway	0	1	2	5	8	3.2	1	2	5	7	11	5.2
MSGS	0.69	0.75	**0.84**	0.75	0.65	0.74	0.19	0.46	0.48	0.39	0.29	0.36
GO Terms	0	1	2	2	6	2.2	1	1	2	4	6	2.8
MSGS	**0.82**	0.75	0.75	0.82	0.82	0.79	**0.61**	0.52	0.76	**0.88**	**0.91**	**0.74**
COMB-1	2	3	4	6	6	4.2	2	3	4	7	11	5.4
MSGS	0.69	0.63	0.72	0.88	0.84	0.75	0.33	0.42	0.51	0.76	0.63	0.53
COMB-2	**2**	**3**	**5**	**9**	**14**	**6.6**	**2**	3	6	**10**	12	6.6

Table 2: Accuracy and hit ratio when using 2, 5, 10, 20, 30 genes selected by algorithms. In the table, numbers with bold typeface indicate the highest accuracy or hit ratio achieved with each number of selected genes. *MSGS*-VAR stands for selecting genes according to gene variance. *MSGS*-SPCA stands for selecting genes using sparse PCA method.

5.3.4 A further Study of Gene Biological Relevance

In the experiments above, we used hit ratio to measure biological relevance of selected genes. In order to closely examine biological relevance of selected genes, we performed a further study in which our biologist collaborators examined the top 20 genes selected by *MSGS*-VAR with COMB-2 on 2C-DATA.[8] It turned out that 17 of the top 20 genes were experimentally confirmed to be cancer related, except for *SMNT*, *LAD1* and *LMOD1*. Detailed information of the selected genes can be found in Table 4. Among them, nine were already annotated to be related to cancer or tumor in the IPA system. The other genes are found to be differentially expressed in unique or several cancer cell lines and are supported by literature. Enzymes involved in metabolism, such as *FABP1* and *GPX2* (well-known oxidoreductase responsive to oxidative stress) were selected. *FABP1* plays an important role in lipid metabolism and investigations have demonstrated altered systemic lipid metabolism in cancer patients [9]. *GPX2* is one of the major cellular antioxidants as biomakers for reactive oxygen species (ROS) producing in animal, and plants. Hydrogen peroxides, one of ROS, has been shown to act as tumor promoters [26]. Several tumor suppressor genes were also identified, including *FHL1*, *SPARCL1* and *SYNPO2*. Interestingly, most of these genes encode transmembrane proteins, implicating their role in signal transduction during cancer development. Thus, our analyses show that multi-source gene selection is able to select genes bearing both statistical significance and biological relevance.

The obtained results on Human Cancer Data confirm the efficacy of multi-source gene selection algorithm *MSGS* as well as the potential of multi-source gene selection. The *MSGS* software will be made available for research purposes.

[8]The reason is that this list results in high accuracy on both 2C-DATA and 4C-DATA and its hit ratio is also high.

6. CONCLUSIONS

Multi-source gene selection presents a new way of dealing with wide data. It leverages information from disparate data sources aiming to form a global pattern describing intrinsic structures embedded in multi-source data. We proposed an unsupervised gene selection algorithm for multi-source data, *MSGS*. Five data sources are used in this study which demonstrated that *MSGS* is effective in multi-source gene selection in terms of statistical significance and biological relevance. In addition, we discussed the nuanced roles of supervised and unsupervised gene selection. Furthermore, the identified genes through our multi-data analysis bear important biological relevance, highlighting a new way of gene selection, and offering guidance for experimental biologists.

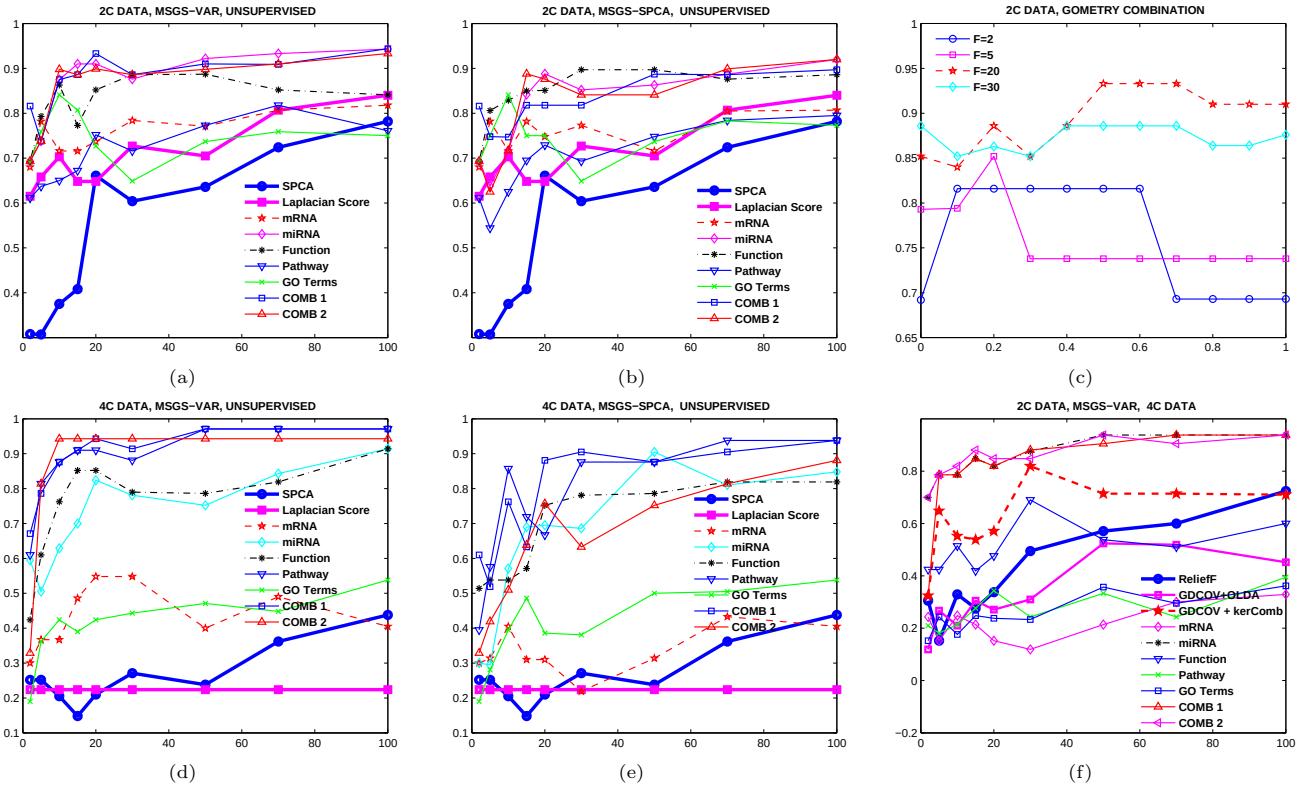

Figure 2: Charts (a) (b) (d), (e): Accuracy (Y axis) vs. different numbers of selected genes (X axis). In the charts, SPCA and Laplacian Score are two baseline unsupervised algorithm for comparison. *MSGS*-VAR stands for selecting genes according to gene variance. *MSGS*-SPCA stands for selecting genes using sparse PCA method. Chart (c): Accuracy (Y axis) vs. different combination coefficient (X axis) on 2C data with local geometry patterns from miRNA and gene function annotation as the inputs. Chart (f): Accuracy (Y axis) vs. different numbers of selected genes (X axis). The accuracy is obtained by first selecting genes with each algorithm on 2C data, then using the selected genes to build a classifier on 4C data and obtaining its accuracy. The chart shows genes selected by *MSGS* with global geometric patterns and several local patterns are discriminative on both 2C and 4C data, but those selected by supervised algorithms are not.

Gene name	Gene Description	Functions and Biological Process	Disease
LGALS4	lectin, galactoside-binding soluble, 4	sugar binding, cell adhesion	colon cancer
CNN1	calponin 1	calmodulin binding,actin filament binding	bone cancer
MYH11	myosin heavy chain 11	calmodulin binding,actin binding	lung cancer
FUT6	fucosyltransferase 6	integral to membrane,L-fucose catabolism	colon cancer
LTF	lactotransferrin	ubiquitin ligase complex	prostate cancer
OLFM4	oflactomedin 4	latroxin receptor activity	gastric Cancer
FABP1	fatty acid binding protein 1	fatty acid metabolism,GABA-A receptor	prostate cancer
SYNPO2	synaptopodin 2	actin binding	prostate cancer
MYLK2	myosin, light chain kinase	protein serine/threonine kinase activity	colorectol cancer
GPX2	glutathione peroxidase 2	oxidoreductase,response to oxidative stress	breast cancer
SPARCL1	SPARC-like 1	calcium ion binding	brain cancer
TM4SF3	tetraspanin 8	signal transducer activity	colon, prostate cancer
TACSTD1	tumor-assoc calcium signal tran 1	plasma membrane	ovarian cancer
KRT15	keratin 15	structural constituent of cytoskeleton	breast cancer
FHL1	Four and a half LIM domains 1	extracellular space,complement activation	brain cancer
LMOD1	leiomodin 1	tropomyosin binding	-
NFIB	nuclear factor I/B	transcription factor activity	breast,brain cancer
LAD1	ladinin 1	anchoring filament	-
CDH17	cadherin 17	transporter activity, calcium ion binding	colon cancer
SMTN	smoothelin	actin binding,muscle development	-

Table 4: The top 20 genes selected by *MSGS*-VAR with COMB-2 on 2C-DATA. Genes with boldface names are the ones directly detected by the IPA system as cancer related. In the table genes are ordered according to their relevance scores assigned by *MSGS*-VAR (from highest to lowest).

Algorithms	2	5	10	20	30	Ave
ON 2C-DATA						
RELIEFF	0.83	**0.90**	**0.92**	**0.95**	**0.96**	0.91
	0	1	3	9	12	5
MSGS	**0.93**	**0.90**	**0.92**	0.92	0.93	**0.92**
OLDA	0	1	3	5	8	3.4
MSGS	0.82	0.81	0.79	0.81	0.84	0.81
KerCB	**2**	3	5	8	10	5.6
MSGS	0.69	0.74	0.90	0.90	0.89	0.82
COMB-2	**2**	**4**	**7**	**9**	**13**	**7**
ON 4C-DATA						
RELIEFF	0.31	0.15	0.33	0.34	0.50	0.32
MSGS-OLDA	0.12	0.27	0.21	0.27	0.31	0.24
MSGS-kerCB	0.32	0.65	0.55	0.57	0.82	0.58
LAPLACIAN	0.21	0.39	0.40	0.34	0.39	0.34
SPCA	0.22	0.25	0.22	0.34	0.33	0.27
mRNA	0.24	0.15	0.25	0.15	0.12	0.18
miRNA	**0.70**	**0.79**	0.79	0.82	**0.88**	0.79
Function	0.42	0.42	0.51	0.48	0.69	0.51
Pathway	0.21	0.18	0.21	0.34	0.24	0.24
GO Terms	0.15	0.24	0.18	0.24	0.23	0.21
COMB-1	0.32	**0.79**	0.79	0.82	**0.88**	0.72
COMB-2	**0.70**	**0.79**	**0.82**	**0.85**	0.85	**0.80**

Table 3: Upper: accuracy and hit ratio by supervised algorithms. *MSGS*-VAR with COMB-2 is unsupervised and listed for comparison. Lower: accuracy on 4C-DATA with genes selected by from 2C-DATA. It shows that genes selected by *MSGS*-VAR are discriminative on both 2C and 4C DATA, but not so for those selected by supervised algorithms.

7. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *NIPS*, 15, 2003.

[2] Y. Bengio, O. Delalleau, N. L. Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel pca. *Neural Comput.*, 16(10):2197–2219, 2004.

[3] T. D. Bie, L. C. Tranchevent, L. Oeffelen, and Y. Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23:i125–i132, 2007.

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

[5] F. Chung. *Spectral graph theory*. AMS, 1997.

[6] A. Cimmino, etal. mir-15 and mir-16 induce apoptosis by targeting bcl2. *PNAS*, 102:13944–13949, 2005.

[7] A. d'Aspremont, F. Bach, and L. E. Ghaoui. Optimal solutions for sparse principal component analysis. Technical report, Princeton University, 2007.

[8] J. Dy. Unsupervised feature selection. In H. Liu and H. Motoda, editors, *Computational Methods of Feature Selection*. Chapman and Hall/CRC Press, 2007.

[9] C. Gercel-Taylor, D. L. Doering, F. B. Kraemer, and D. D. Taylor. Aberrations in normal systemic lipid metabolism in ovarian cancer patients. *Gynecologic Oncology*, 60:35–41, 1996.

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

[11] J. Hagan and C. Croce. Micrornas in carcinogenesis. *Cytogenet Genome Res*, 118:252–259, 2007.

[12] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems 18*. MIT Press, 2005.

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

[14] Ingenuity-Systems. Ingenuity pathways analysis. http://www.ingenuity.com.

[15] G. R. G. Lanckriet, etal. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.

[16] 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.

[17] H. Liu and H. Motoda, editors. *Computational Methods of Feature Selection*. Chapman and Hall/CRC Press, 2008.

[18] L. Lovasz. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2:353–397, 1993.

[19] J. Lu, etal. Microrna expression profiles classify human cancers. *Nature*, 435:834–838, 2005.

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

[21] H. Pingzhao, G. Bader, D. Wigle, and A. Emili. computational prediction of cancer-gene function. *Nature Reviews Cancer*, 7:23–34, 2007.

[22] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. Technical report, Max Planck Institut, 1996.

[23] M.-L. Si, etal. mir-21-mediated tumor growth. *Oncogene*, 26:2799–2803, 2007.

[24] M. R. Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, 53:23–69, 2003.

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

[26] T. J. Slaga, etal. Skin tumor-promoting activity of benzoyl peroxide, a widely used free radical-generating compound. *Science*, 213:1023–1025, 1981.

[27] M. R. Spiegel. *Theory and Problems of Probability and Statistics*. New York: McGraw-Hill, 2nd edition, 1992.

[28] H. Tazawa, etal. Tumor-suppressive mir-34a induces senescence-like growth arrest through modulation of the e2f pathway in human colon cancer cells. *PNAS*, 104:15472–15477, 2007.

[29] U. von Luxburg. A tutorial on spectral clustering. Technical report, Max Planck Institute for Biological Cybernetics, 2007.

[30] J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *J. Mach. Learn. Res.*, 6:483–502, 2005.

[31] J. Ye, J. Chen, and S. Ji. Discriminant kernel and regularization parameter learning via semidefinite programming. In *ICML.*, 2007.

[32] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML.*, 2007.

[33] D. Zhou and C. Burges. Spectral clustering and transductive learning with multiple views. In *Intl. Conf. on Mach. Learn.*, 2007.