Published in final edited form as:
Conf Proc IEEE Eng Med Biol Soc. 2009 ; 2009: 5456–5459. doi:10.1109/IEMBS.2009.5334063.

Relative Expression Analysis for Identifying Perturbed Pathways

James A. Eddy,
Department of Bioengineering, University of Illinois, Urbana, IL 61801 USA

Donald Geman, and
Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218 U S A

Nathan D. Price [Member, IEEE]
Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, IL 61801 USA (phone: 217-244-0596; fax: 217-333-5052

James A. Eddy: eddy2@illinois.edu; Donald Geman: geman@cis.jhu.edu; Nathan D. Price: ndprice@illinois.edu

Abstract

The computational identification from global data sets of stable and predictive patterns of gene and protein relative expression reversals offers a simple, yet powerful approach to target therapies for personalized medicine and to identify pathways that are disease-perturbed. We previously utilized this approach to identify a molecular classifier with near 100% accuracy for differentiating gastrointestinal stromal tumor (GIST) and leiomyosarcoma (LMS), two cancers that have very similar histopathology, but require very different treatments. Differential Rank Conservation (DIRAC) is a novel approach for studying gene ordering within pathways and is based on the relative expression ranks of participating genes. DIRAC provides quantitative measures of how pathway rankings differ both *within* and *between* phenotypes. DIRAC between pathways in a selected phenotype contrasts the scenarios where either (i) pathways are ranked similarly in all samples; or (ii) the ordering of pathway genes is highly varied. We examined gene expression inGIST and LMS tumor profiles and identified pathways that appear to be tightly regulated based on high conservation of gene ordering. The second form of DIRAC manifests as a change in ranking (i.e., shuffling) between phenotypes for a selected pathway. These variably expressed pathways serve as signatures for molecular classification, and the ability to accurately classify microarray samples provided strong validation for the pathway-level expression differences identified by DIRAC.

I. Introduction

The realization of malignant phenotypes in many diseases – notably cancer [1,2] – as intrinsically pathway-based in origin motivates the interrogation of high-throughput expression data for studying biologically meaningful pathways. Existing pathway-based expression analysis tools commonly investigate informative patterns of up- or down-regulation of grouped genes in different disease states. For example, the gene set enrichment analysis (GSEA) platform identifies pathways that are significantly enriched for over- or under-expressed genes [3,4]. Other methods employ a single statistic to represent the collective activity of a pathway (e.g., mean or median gene expression) [5,6]. Perturbed levels of pathway activity (i.e., collective up- or down-regulation) are then examined to identify those pathways most differentially expressed between phenotypes. These frameworks have been applied to diverse cancer systems and serve as a robust source of biological discovery [5,7].

Cellular regulation of a pathway can also be characterized in the context of the relative expression ranking of the participating genes (referred to herein as ordering). It is possible

that neither the individual pathway genes nor the pathway as a whole will display any notable over- or under-expression in response to environmental or disease-related stimuli. Compared to measuring only increases or decreases in expression, regulation of ordering is reflected entirely in the relative levels of expression for genes within a pathway.

The specific ordering of pathway genes is described by the corresponding ranks of expression levels (i.e., most expressed to least expressed), and is collectively referred to as a pathway ranking. We adopted a strategy for representing pathway rankings that is based on pairwise comparisons of gene expression levels (i.e., the *relative* mRNA abundance in each pair of genes). Such pairwise comparisons have been used to build two-gene predictors with simple decision rules for classification of expression profiles [8,9]. These decision rules have resulted in highly-accurate two-gene diagnostic classifiers that have proven effective for molecular identification of cancer [8–10]. As an extension of the relative expression reversal concept to pathways, we determined the pairwise ordering for each distinct pair of genes within a pathway, establishing an intuitive and computationally straightforward method for calculating pathway rankings.

Rank conservation for a pathway describes the extent to which the ordering of genes is maintained over a population, or the manner in which a pathway ranking is maintained (i.e., the specific ordering observed). We have developed a new method, Differential Rank Conservation (DIRAC), to evaluate how patterns of rank conservation for pathways change in different phenotypes. Specifically, differential rank conservation occurs in two forms. The first is differential rank conservation between pathways in a phenotype, where either (i) pathways are ranked similarly in all samples (high rank conservation); or (ii) pathways for which gene ordering is highly varied (low rank conservation). In the second case, differential rank conservation can manifest as a change in ranking (i.e., shuffling) between two phenotypes for a selected pathway.

We applied DIRAC to analyze gene expression profiles obtained from primary intestinal tumors in patients with two related sarcomas: gastrointestinal stromal tumor (GIST) or leiomyosarcoma (LMS).

II. Materials and Methods

A. Microarray Data

The gene expression profiles from 68 sarcoma patients were previously analyzed to identify a two-gene relative expression classifier that accurately differentiates GIST and LMS tumors[10]. Given the list $\{g_1, \ldots, g_G\}$ of G genes on a microarray, we let $\mathbf{X} = (X_1, \ldots, X_G)$ denote the corresponding expression profile, where X_i is the expression of gene g_i. Our data then consists of a G x N matrix; the n'th column represents the expression profile $\mathbf{x}_n$ of the n'th sample, $n = 1, \ldots, N$. In addition, each sample is labeled by a class (e.g., phenotype) $Y \in \{1, 2,\ldots,K\}$; $K = 2$ for binary classification. The labeled training set is $F = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. Expression profiles $\mathbf{X}$ and phenotype labels Y are regarded as random variables, and the elements of F represent independent and identically distributed samples from some underlying probability distribution of ($\mathbf{X}$, Y).

B. Rank Template Matching for Pathways

Knowing the ordering of the gene expressions within each profile is equivalent to knowing all of the pairwise orderings, i.e., whether $X_i < X_j$ or $X_i > X_j$ for each distinct pair of genes $1 \leq i, j \leq G$. For example, various rankings for the GS pathway in GIST patients are shown in Fig. 1. In order to define a template representing the expected ranking of pathway genes within a class, we consider the probabilities $\Pr(X_i < X_j \mid Y = k)$ for each pair of genes $g_i < g_j$ and for each class k. These probabilities are estimated from the training set by computing the

fraction of samples in each phenotype for which gene g_i is expressed less than gene g_j. The class k rank template for a fixed pathway m is the binary vector $T^{(m,k)}$ of length $G(G-1)/2$ where the i,jth component is 1 if $\Pr(X_i < X_j \,|Y = k) > 0.5$ and 0 if $\Pr(X_i < X_j \,|Y = k) \leq 0.5$. The rank template for the GS pathway in GIST patients is highlighted in Fig. 1.

Given an expression profile **x**, there is then a natural measure for how well the sample matches the template $T^{(k)}$. The matching score of sample **x** is denoted by $R^{(k)}(\mathbf{x})$ and is defined to be the fraction of the $G(G-1)/2$ pairs for which the observed ordering within **x** matches the template – those expected for class k. Rank matching scores corresponding to each unique ranking of the GS pathway in GIST patients are shown in Fig. 1.

C. Rank Conservation Indices

Averaging the rank matches over all the samples in a class k yields an rank conservation index denoted by $\mu_R^{(k)} = E(R^{(k)}|Y = k)$. It is estimated in practice by averaging the scores $R^{(k)}(\mathbf{x})$ over all the samples $(\mathbf{x}, y)$ in the training set for which $y = k$. This index can be seen as a measure of stability in rankings among genes in the class. Two extreme cases correspond to (i) pure random shuffling of the expression values in the class from sample to sample, in which case $\mu_R^{(k)} \approx 5$; and (ii) all samples displaying exactly the same ordering, in which case $\mu_R^{(k)} \approx 1$. In general, however, there are many gene pairs g_i and g_j which are expressed on different scales, and hence $x_i < x_j$ across nearly all samples and phenotypes. As a result, one generally finds $\mu_R^{(k)} \ll 5$. This index is similar to entropy in the sense that values of $\mu_R^{(k)} \ll 1$ indicate a highly disorganized state in which there is a great deal of variation among the rankings in class k from sample to sample and values of $\mu_R^{(k)} \ll 1$ indicate a highly ordered state in which samples have very similar, and hence predictable, orderings among the genes.

D. Rank Difference Scores

We consider two phenotypes $Y = 1, 2$ and a fixed pathway m. If m is tightly regulated in one phenotype, the samples from that class, say $Y = 1$, will have high $R^{(m,1)}$ values on average. But if $\mu_R^{(k)}$ is large for both $k = 1$ and $k = 2$, and if the two rank templates $T^{(m,1)}$ and $T^{(m,2)}$ are significantly different, then the samples from class $Y = 1$ will generally have low values for the statistic $R^{(m,2)}$ as well as high values for the statistic $R^{(m,1)}$, and vice-versa for the samples from class $Y = 2$. We want to capture this phenomenon, namely low variance of pathway ranking within classes, but variance between classes, with a single statistic or metric. The natural measure is the difference $\Delta(m,\mathbf{x}) = R^{(m,1)}(\mathbf{x}) - R^{(m,2)}(\mathbf{x})$. Clearly, $-1 \leq \Delta(m,\mathbf{x}) \leq 1$ with positive (respectively, negative) values providing evidence that the phenotype of sample **x** is $Y = 1$ (resp., $Y = 2$). The characteristics captured by the rank difference score are illustrated in Fig. 2 for the EDG1 pathway. The difference score provides a classifier for phenotype identification based on the degree of regulation of the genes in pathway m. A new sample **x** is predicted to belong to class $Y = 1$ if $\Delta(m,\mathbf{x}) > 0$ and to class $Y = 2$ if $\Delta(m,\mathbf{x}) \leq 0$. The classification rate for pathway m is then: $\eta(m) = \Pr(\Delta(m,\mathbf{X}) > 0|Y = 1)*\Pr(\mathrm{Y} = 1) + \Pr(\Delta(m,\mathbf{X}) \leq 0|Y = 2)*\Pr(Y = 2)$.

For example, if $Y = 1$ denotes GIST and $Y = 2$ denotes LMS, and if we assume that the two phenotypes are *a priori* equally likely, then $\eta(m)$ is simply the average of sensitivity and specificity relative to identifying GIST. In order to determine the most differentially expressed pathways between two given classes, we calculate rank templates for each class, evaluate the differential metric for each sample in the training set and choose the pathways with the largest estimated classification rate.

III. Results and Discussion

A. Tightly Regulated Pathways in GIST and LMS

The 20 most tightly regulated pathways in GIST and LMS, as measured by rank conservation indices, are shown in Table I. Large rank conservation index values indicate that gene orderings in these pathways are very similar among all samples of each phenotype.

One example of a tightly regulated pathway in both GIST and LMS is the GS pathway (illustrated in Fig. 1). The GS pathway comprises major signaling proteins downstream of G-protein coupled receptors, including guanine nucleotide binding proteins alpha (*GNAS*), beta (*GNB1*), and gamma (*GNGT1*); adenylate cyclase 1 (*ADCY1*); and both the catalytic (*PRKACA*) and regulatory (*PRKAR1A*) subunits of the cAMP-dependent protein kinase C (PKC). Determining the relative expression level for each distinct pair among the six pathway genes resulted in an overall ranking defined by 15 pairwise orderings. We found that one pathway ranking was shared by 27 out of 37 GIST samples (73%) and 23 out of 31 LMS samples (74%); as the probability for each pairwise ordering is much greater than 50%, it follows that the rank templates are identical for the two phenotypes. Furthermore, six other samples in GIST and LMS (12 total) displayed only a single mismatch. PKC family members phosphorylate a wide variety of protein targets and are known to be involved in diverse cellular signaling pathways, such as those associated with cell adhesion, cell transformation, cell cycle checkpoint, and cell volume control.

B. Differentially Expressed Pathways in GIST and LMS

A total of 165 pathways were identified that significantly differentiated between expression profiles of GIST and LMS (*P*-value less than 0.05), the top 20 of which are listed in Table II.

The EDG1 pathway was identified as one of the most differentially expressed pathways in GIST and LMS, achieving a classification rate of 97.3% when used to separate expression profiles in the training data. The principal features governing the formulation of the rank difference metric, and also an example of how it is applied for molecular classification are illustrated for the EDG1 pathway in Fig. 2. Here, R denotes the rank matching score for a profile, and superscripts indicate the phenotype of the rank template (e.g., R^{GIST} represents the rank matching score for a sample when compared to the ordering defined in the GIST template). The rank difference values calculated for the EDG1 pathway are also shown in Fig. 2, along with the corresponding class predictions (i.e., GIST where positive, LMS if negative).

C. Classification with DIRAC

We used leave-one-out cross validation to estimate how accurately the top pathways – selected as those achieving the highest apparent classification rate for predicting sample classes based on rank difference scores – were able to predict the class of future samples (Fig. 3).

As a means for comparison, we used the top scoring pair (TSP) algorithm and support vector machines (SVM) to classify samples in each of the datasets. We found that our method performed well in a number of the datasets, including estimated accuracies between 90–98% in gastrointestinal sarcoma, leukemia, and prostate cancer (Fig. 3). In cases with poor accuracies such as breast cancer, lung cancer, and melanoma, we saw that the other methods used also failed to accurately classify samples. We thus suspect that the poor performance in these cases is a factor of unclear differences in phenotypes, rather than a shortcoming of our method. The foremost goal of our method is to aid in biological discovery and hypothesis

generation, and the excellent classification accuracy overall affirms the robustness of the pathway rank regulation measure.

IV. Conclusions

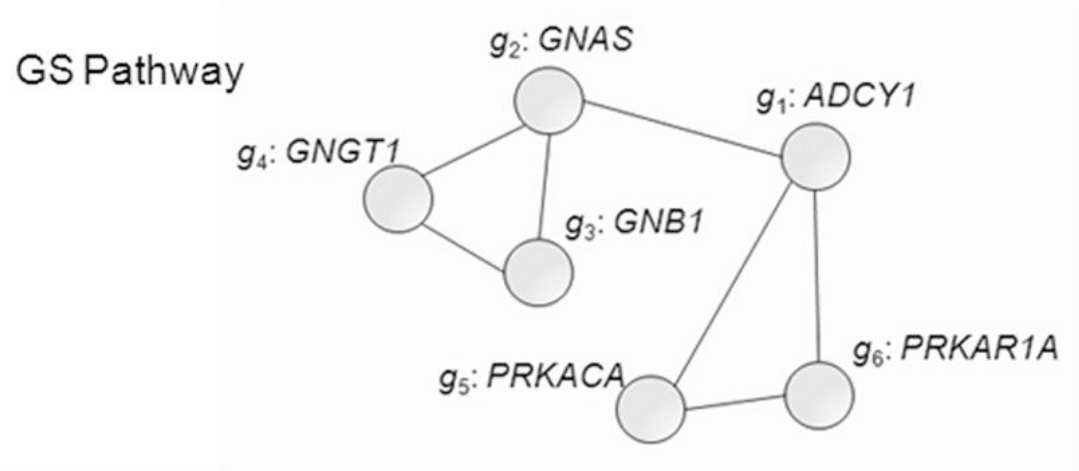
In this study we demonstrate a novel method to identify highly discriminative biological pathways based on differing patterns of gene expression ranking within pathways. Importantly, this method not only identifies perturbed pathways, but does so in such a way that it can be used for classification of samples. Thus, predictive accuracy becomes a strong measure for the validity of the perturbed pathway being a reproducible hallmark of the disease phenotype. Studying rank regulation of biologically relevant gene sets is thus a promising tool for measuring pathway behavior within and across different populations.

Acknowledgments

This work was supported in part by the NIH-NCI Howard Temin Pathway to Independence Award in Cancer Research (NDP).

References

1. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science Sep 26;2008 321(5897):1801–6. [PubMed: 18772397]
2. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature Oct 23;2008 455(7216):1061–8. [PubMed: 18772890]
3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–15550. [PubMed: 16199517]
4. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics 2007;23(23):3251. [PubMed: 17644558]
5. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007;3(140)
6. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring Pathway Activity toward Precise Disease Classification. PLoS Comput Biol 2008;4(11)
7. Auffray C. Protein subnetwork markers improve prediction of cancer outcome. Mol Syst Biol 2007;3(141)
8. Geman D. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. Stat Appl Genet Mol Biol 2004;3
9. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 2005;21(20):3896–3904. [PubMed: 16105897]
10. Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, Hood L, Shmulevich I, Zhang W. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proc Natl Acad Sci U S A 2007;104(9):3414. [PubMed: 17360660]



Pairwise orderings	Rank template	Pathway rankings among samples						
g1 < g2	1	1	1	1	1	1	1	1
g1 < g3	1	1	1	1	1	0	0	1
g1 < g4	0	0	0	0	1	0	1	1
g1 < g5	1	1	0	1	1	0	0	1
g1 < g6	1	1	1	1	1	1	1	1
g2 < g3	0	0	0	0	0	0	0	0
g2 < g4	0	0	0	0	0	0	0	1
g2 < g5	0	0	0	0	0	0	0	0
g2 < g6	0	0	0	0	0	0	0	0
g3 < g4	0	0	0	0	0	1	1	1
g3 < g5	0	0	0	1	0	0	0	1
g3 < g6	1	1	1	1	1	1	1	1
g4 < g5	1	1	1	1	1	0	0	0
g4 < g6	1	1	1	1	1	1	1	0
g5 < g6	1	1	1	1	1	1	1	0
Mismatches (of 15):		0	1	1	1	4	5	7
Rank matching score:		1.000	0.933	0.933	0.933	0.733	0.667	0.533
Observed rankings (of 37):		27	1	1	4	1	1	2

Fig. 1.

Example of tightly regulated pathway in GIST. A simplified diagram of the GS pathway, comprising six signaling proteins downstream of G-protein couple receptors, is shown above. The majority of GIST samples match the pairwise orderings in the GIST rank template exactly.

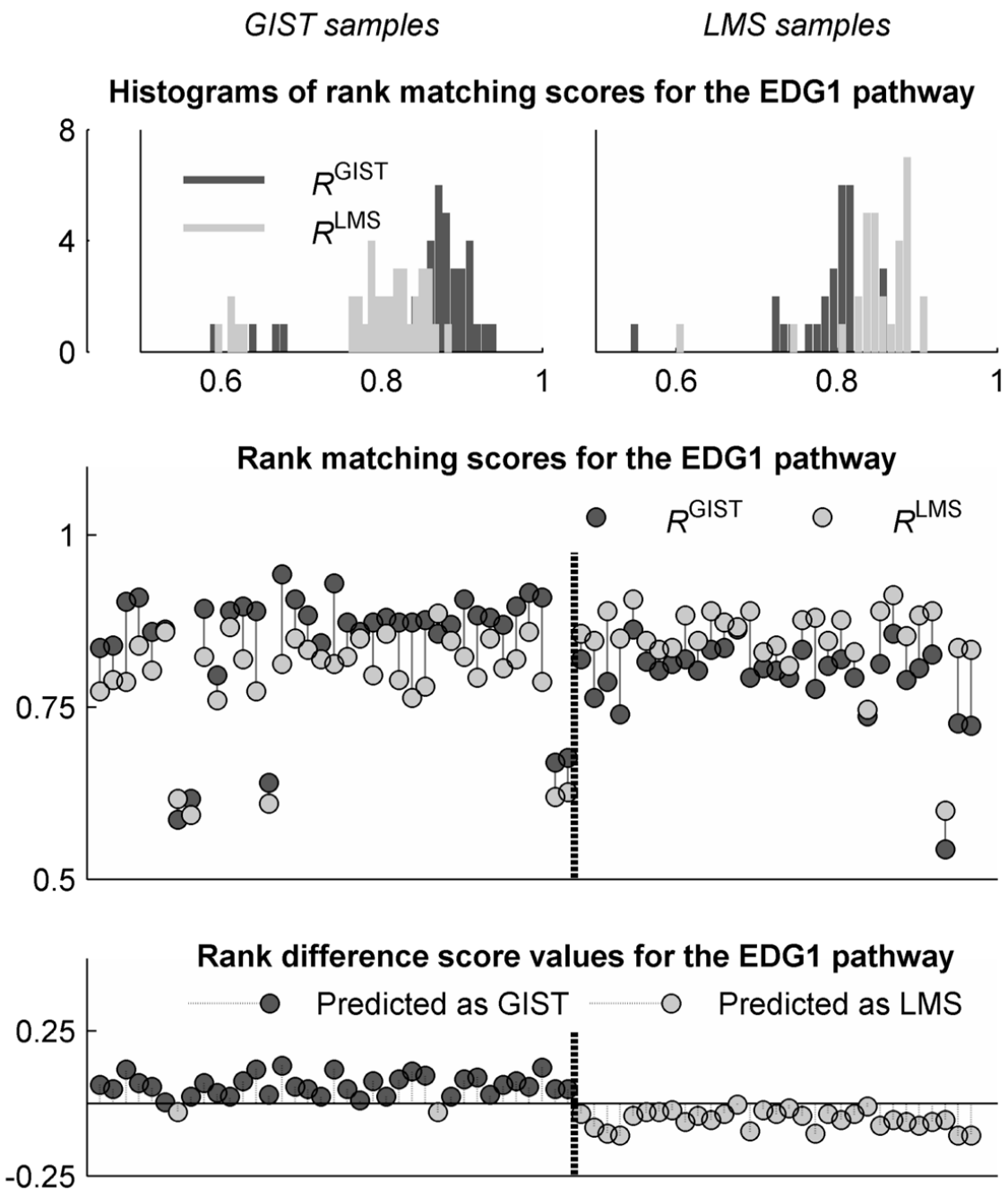


Fig. 2.

Differential rank conservation of the EDG1 pathway in GIST and LMS. The GIST template matching scores (R^{GIST}) are higher on average in GIST samples than LMS template matching scores (R^{LMS}). In LMS samples, R^{LMS} scores are higher on average than R^{GIST} scores. Comparing the two rank matching scores in each sample, GIST samples match the GIST template more than the LMS template in all but two cases; LMS samples match the LMS template more than the GIST template in all cases. Samples are classified as GIST if the difference score is positive, and as LMS if the difference is negative.

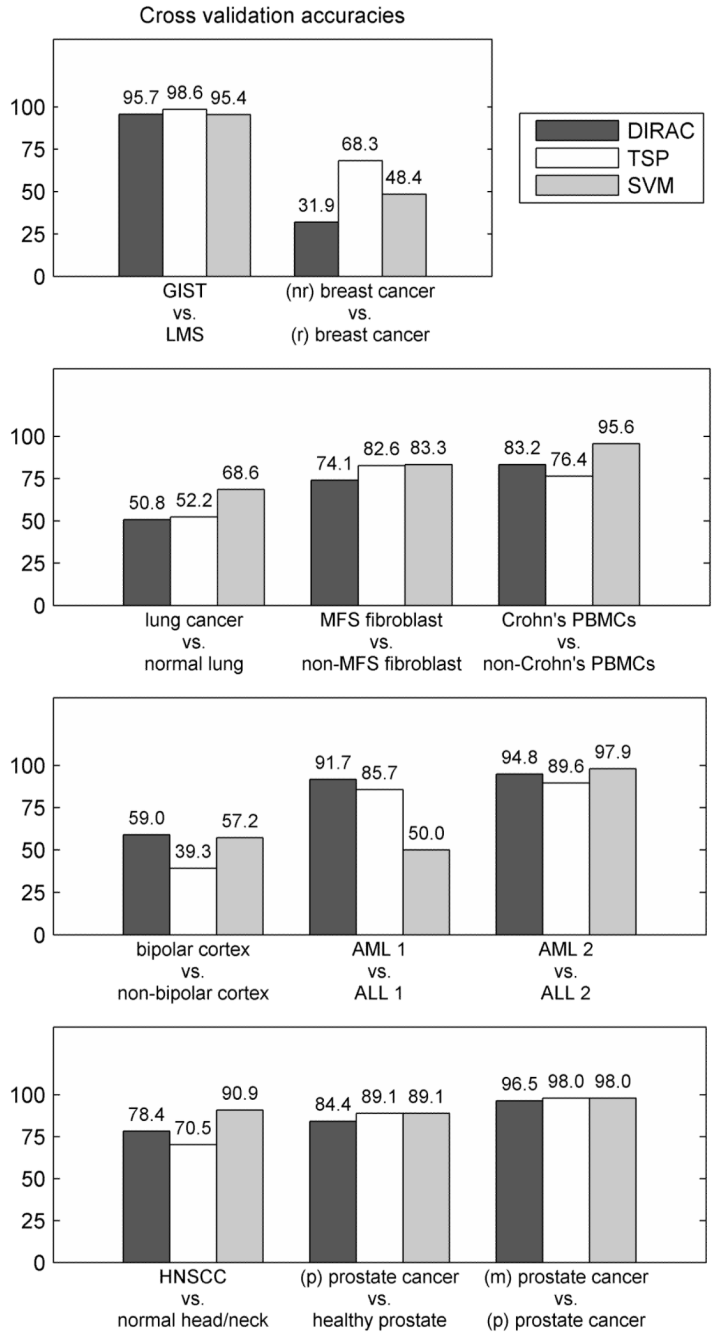


Fig. 3.
Comparison of classification with DIRAC to other methods.

TABLE I

Tightly Regulated Pathways in GIST and LMS

Pathway	Number of		Rank
	Genes	Gene Pairs	Conservation
GIST			
GS	6	15	0.948
BETAOXIDATION	6	15	0.941
IFNG	6	15	0.930
ETC	10	45	0.915
CELL2CELL	13	78	0.906
LMS			
RAN	5	10	0.968
GS	6	15	0.966
FEEDER	9	36	0.943
CDC42RAC	14	91	0.939
ETC	10	45	0.938

TABLE II

Differentially Expressed Pathways in GIST and LMS

Pathway	Number of		Apparent Accuracy	*P*-value
	Genes	Gene Pairs		
GH	6	15	0.973	< 1.0E-07
EDG1	6	15	0.973	< 1.0E-07
EIF4	6	15	0.970	< 1.0E-07
ATM	10	45	0.959	< 1.0E-07
CREB	13	78	0.950	< 1.0E-07
KERATINOCYTE	5	10	0.932	1.6E-06
P53HYPOXIA	6	15	0.931	1.6E-06
FEEDER	9	36	0.930	1.6E-06
CDC42RAC	14	91	0.927	2.4E-06
ETC	10	45	0.925	2.4E-06