Integrating Multiple-Platform Expression Data through Gene Set Features

Matěj Holec[1], Filip Železný[1], Jiří Kléma[1], and Jakub Tolar[2]

[1] Czech Technical University, Prague
[2] University of Minnesota, Minneapolis
{holecm1, zelezny, klema}@fel.cvut.cz, tolar003@umn.edu

Abstract. We demonstrate a set-level approach to the integration of multiple platform gene expression data for predictive classification and show its utility for boosting classification performance when single-platform samples are rare. We explore three ways of defining gene sets, including a novel way based on the notion of a fully coupled flux related to metabolic pathways. In two tissue classification tasks, we empirically show that the gene set based approach is useful for combining heterogeneous expression data, while surprisingly, in experiments constrained to a single platform, biologically meaningful gene sets acting as sample features are often outperformed by random gene sets with no biological relevance.

1 Introduction

The problem addressed in this paper is *set-level analysis* of gene expression data, as opposed to the more traditional gene-level analysis approaches. In the latter, one typically seeks single statistically significant genes or constructs classification models with gene expressions acting as sample features. In set-level analysis, genes are first grouped into sets apriori determined by a chosen relevant kind of background knowledge. For example, a gene set may correspond to a group of proteins acting as enzymes in a biochemical pathway or be a set of genes sharing a gene-ontology [3] term. Naturally, gene sets considered for an analysis may on one hand overlap while on the other hand their union may not exhaust the entire gene set screened in the expression data. Any gene set may then be assigned descriptive values (such as expression, fold change, significance) by statistical aggregation of the analogical values pertaining to its members. Gene sets thus may act as derived sample features replacing the original gene expressions.

The potential for set-level analysis of genomic data has been advocated recently [12, 1] on the grounds of improved interpretation power and statistical significance of analysis results. The basic idea of set-level analysis is not new. Indeed, state-of-the-art tools such as DAVID [9] have supported the established protocol of *enrichment analysis* detecting ontology terms or pathways related to a large subset of a user-supplied gene list, thus obviously following a simple form of set-level analysis. The biological utility of set-level analysis was demonstrated by the study [11] where a significantly downregulated pathway-based gene set in a class of type 2 diabetes was discovered despite no significant expression change being detected for an individual gene. In another study [18], a method based on singular value

decomposition was proposed to determine the 'level of activity' of a pathway based on the sampled expression values of its gene-members. The paper [5] reviews some common statistical pitfalls in the calculation of such statistics ascribed to gene sets. The recent work [15] suggests a more sophisticated method to estimate the activity level of a pathway, considering the pathway *structure* in addition to the expressions of the genes involved therein. Another innovative aspect of [15] is that the authors employ such pathway activities as derived features of samples and use these for sample classification by a machine learning algorithm.

The main contribution of the present work is showing that the gene set based approach naturally enables to analyze in an integrated manner gene expression data collected from heterogeneous platforms, which may even encompass different organism species. The significance of this contribution is at least twofold. First, microarray experiments are costly, often resulting in numbers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help the experimenter. Second, such integrated analysis provides the principal means to discover biological markers shared by different-genome species.

We consider three types of gene sets. The first type groups genes that share a common gene ontology [3] term. The second type groups genes acting in biological pathways formalized by the KEGG [10] database. The third gene set type represents a further novel contribution of our work and is based on the notion of a *fully coupled flux*, which is a pattern prescribing pathway partitions hypothesized by [13] to involve strongly co-expressed genes. These synergize in single gradually amplified biological functions such as enzymatic catalysis or translocation among different cellular compartments.

Research papers concerned with gene set based analysis, including the aforementioned studies, usually point out the statistical advantages of results based on gene sets in comparison with those based on single genes. We conjecture, however, that to assess the utility of the gene set approach, the relevant question that must be asked is *how data models based on biologically meaningful gene sets compare to those based on gene sets constructed randomly, with no biological relevance.* This question is important as we indeed show that even random grouping of genes into sets may lead to improved predictive accuracies. By addressing this question way we can determine whether the inclusion of background knowledge through gene sets has a positive effect on the analysis results. We are not aware of previous work considering this question[1] and it is our third contribution to address it experimentally.

The paper is organized as follows. In Section 2 we describe the methodological ingredients of our approach, consisting of normalization, gene set extraction, data integration and predictive classification. Section 3 describes the expression analysis case studies and the collected relevant data used for experimental validation. In Section 4 we show and discuss the experimental results. Section 5 lays out prospects for future work and concludes the paper.

[1] The suggested gene set randomization should not be confused with the standard class-permutation technique used for validation, also in the set-level analysis context [1].

2 Methods

The *input* of our workflow is a set of gene expression samples (real vectors) possibly measured by different microarray platforms. Each sample is assigned two labels. The first identifies the microarray platform from which the sample originates, the second identifies a sample class (e.g. tissue type). The *output* is a classification model, that is, a model that estimates the sample class given an expression sample and its platform label. The model is obviously applicable to any sample not present in the input ('training') data, as long as its platform label is also present in the input data. The remarkable property of the output model is that it is not a combination of separate models each pertaining to a single platform. Rather, it is a single classifier trained from the entire heterogeneous sample set and represented in terms of 'activity levels' of units that apply to all platforms, albeit the computation of these activity levels may be different across platforms. More specifically, the activity of a unit (such as a pathway) is calculated using a different gene set in each platform. We now describe the individual steps of the method in more detail.

Normalization. The first normalization step is conducted separately for each platform to consolidate same-platform samples. Quantile normalization [2] ensures that the distribution of expression values across such samples is identical. As a second step, scaling provides means to consolidate the measurements across multi-platform samples. We subtract the sample mean from all sample components, and divide them by the standard deviation within the sample. As a result, all samples independently of the platform exhibit zero mean and unit variance. We conduct these steps using the Bioconductor [4] software.

Set Construction. Here we consider three types of background knowledge in order to define apriori gene sets. Each such set will be extracted from the initial pool of all genes measured by at least one of the involved platforms.

The first type groups genes that share a common gene ontology [3] term. The second type groups genes acting in biological pathways formalized by the KEGG [10] database. A gene falls in a set corresponding to a pathway if it is mapped to a KEGG node of some organism ortholog of that pathway. The third gene set type is based on the notion of a *fully coupled flux* (FCF), motivated as follows. Many notable biological conditions are characterized by the activation of only certain parts of pathways; for example, see references [16, 19, 21]. The notion of 'pathway activation' implied by the previous gene set may thus often violate intuition and hinder interpretation. Therefore we extracted all pathway partitions which comply with the graph-theoretic notion of FCF [13]. It is known that the genes coupled by their enzymatic fluxes not only show similar expression patterns, but also share transcriptional regulators and frequently reside in the same operon in prokaryotes or similar eukaryotic multi-gene units such as the hematopoietic globin gene cluster. FCF is a special kind of network flux that corresponds to a pathway partition in which non-zero flux for one reaction implies a non-zero flux for the other reactions and vice versa. It is the strongest qualitative connectivity that can be identified in a network. The notion of an FCF is explained through an example in Fig. 1; for a detailed definition, see reference [13]. Pathway partitions forming FCF's constitute the third gene set type. Again, a gene falls in a set corresponding to a FCF if it is mapped to a KEGG node in some organism-ortholog of that FCF.

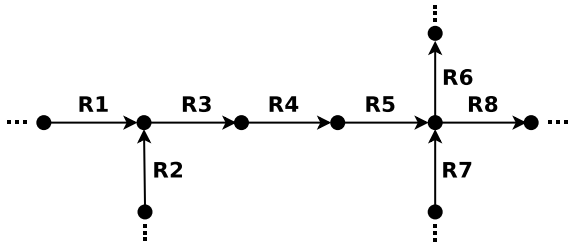

Fig. 1. Fully coupled fluxes in a simplified network with nodes representing chemical compounds and arrows as symbols for chemical reactions among them. Each arrow can be labeled by a protein. R3, R4 and R5 are fully coupled as a flux in any of these reactions implies a flux in the rest of them. Note that R1 and R3 do not constitute a FCF as a flux in R3 does not imply a flux in R1.

The extraction of fully coupled fluxes from KEGG pathways graphs was conducted in Prolog. The source code as well as the Prolog representation [8] of the pathways are available on request to the first author. The bold numbers in Table 2 display the total numbers of gene sets extracted for the respective types.

In what follows, gene sets act as features acquiring a real value for each sample. Formally, let π be the set of genes interrogated by a given platform, and Σ a set of gene sets of a particular type. We define a mapping

$$A_\pi : R^{|\pi|} \times \Sigma \rightarrow R$$

For an expression sample $\boldsymbol{s} = [e_1, \ldots, e_{|\pi|}] \in R^{|\pi|}$, $A_\pi(\boldsymbol{s}, \sigma)$ should collectively quantify the 'activity level' of genes in set $\sigma \in \Sigma$, in the biological situation (e.g. a tissue type) sampled by $\boldsymbol{s}$. Typically, not all members of σ will be measured by platform π, and the computation of $A_\pi(\boldsymbol{s}, \sigma)$ will be based on the expressions e_i of genes in $\sigma \cap \pi$. For transparency, in this study we define $A_\pi(\boldsymbol{s}, \sigma)$ as the average of expressions measured in $\boldsymbol{s}$ for all genes in $\sigma \cap \pi$. We only note here that more sophisticated methods have been proposed to instantiate $A_\pi(\boldsymbol{s}, \sigma)$, either linear, based e.g. on a weighted sum of expression values of the involved genes as in [18], or non-linear, based on additional structure information as in [15] but then constrained to pathway-type gene sets.

Our reasoning above assumes the aggregation of gene expression measurements. Precisely speaking, genes themselves aggregate one or more measurements since multiple probesets can represent the same gene. Here, the expression of a gene is simply defined as the average of the corresponding normalized probeset measurements, despite certain caveats of this approach.[2]

Data Integration. The goal of this methodological step is to integrate heterogeneous expression samples into a single-tabular representation (that is, into a set of samples sharing a common feature set) that predictive classification algorithms

[2] For example, Affymetrix chips contain probesets representing the same gene that cannot be consolidated into unique measures of transcription due to alternative splicing, use of alternative poly(A) signals, or incorrect annotations [17].

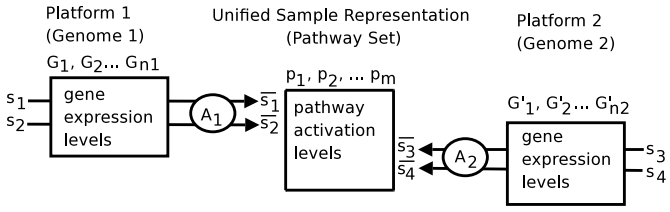

Fig. 2. Integrating expression data collected from heterogeneous platforms into a unified tabular representation of pathway activations. If these platforms pertain to different organisms, we assume that (an ortholog of) each pathway p_i exists in each of the organisms.

can process. Formally, we have a set of expression samples $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots\}$ in which for all i

$$\boldsymbol{s}_i \in \cup_j R^{|\pi_j|},\ \pi_j \in \Pi$$

where Π is the set of the considered platforms. We wish to obtain a new representation $\bar{S} = \{\bar{\boldsymbol{s}}_1, \bar{\boldsymbol{s}}_2, \ldots\}$ where each $\bar{\boldsymbol{s}}_i \in R^n$, $n \in N$.

This aim is achieved using the above introduced 'gene set activation' concepts. Formally, using gene set type $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$, for each sample $\boldsymbol{s}_i$ labeled with platform π we stipulate

$$\bar{\boldsymbol{s}}_i = [A_\pi(\boldsymbol{s}_i, \sigma_1), \ldots, A_\pi(\boldsymbol{s}_i, \sigma_m)]$$

Naturally, sample $\bar{\boldsymbol{s}}_i$ then inherits the class label from $\boldsymbol{s}_i$. The integration principle is exemplified in Fig. 2 with pathways p_i playing the role of gene sets σ_i. The described representation conversion is part of the functionality of the aforementioned Prolog code.

Classification and Validation. The final step of the workflow is to employ machine learning algorithms to induce predictive classification models of the integrated samples. As the achieved unified representation $\bar{S}$ can be processed by virtually any machine learning algorithm, the choice appears rather arbitrary. Since one of the usual arguments in favor of gene set based analysis is the ease of interpretation, we decided to test decision-tree classifiers enabling direct human inspection. Specifically, we experimented with the J48 decision tree learner included the machine learning environment Weka [20].

The design of the experiments and the validation protocol is dictated by the following questions we wish to address empirically.

- (Q1) How do classifiers based on original *single gene* expressions compare in terms of predictive accuracy to those based on activations of biologically meaningful *gene sets*?
- (Q2) How do classifiers based on *biologically meaningful gene sets* compare in terms of predictive accuracy to those based on *gene sets constructed randomly*, with no biological relevance?
- (Q3) How do classifiers learned from *single-platform* data compare in terms of predictive accuracy to those learned from data integrated from *heterogeneous platforms*?

In the case of (Q2), we constructed three families of random gene sets corresponding to the three respective kinds of genuine gene sets, for each of the involved platforms. The correspondence is in that a particular type of random gene sets contains exactly the same number of set-elements and exactly the same set-cardinality distribution as its genuine counterpart. For each platform, the members of each random gene set were drawn randomly without replacement from a uniform probability distribution cast on the genes measured by the platform.

We are interested in the insights Q1-Q3 for both the 'data-rich' and 'data-poor' situation, i.e. for both small and large sets of expression samples. Therefore the preferred means of assessment is through *learning curves* which are diagrams plotting an unbiased estimate of the classifier's predictive accuracy against the proportion p of the available data set used for its training. The accuracy estimate for each measured p was obtained by inducing a classifier 20 times with a randomly chosen subset (of proportional size p) of the entire data set and testing its accuracy on the remaining data not used for training. In each such step, the 20 empirical accuracy results were averaged into the reported value. We let p range from 0.2 to 0.8 to prevent statistical artifacts arising from overly small sets used for training or testing, respectively.

3 Classification Tasks and Data

Here we validate our methodology in biological classification tasks. In order to avoid domain bias, we chose not to tackle overly special classification cases such as those addressing particular diseases. We therefore address two general tasks of tissue type classification. The first experiment focuses on distinct features of blood-forming (hematopoietic; '*heme*' in figure legends) and supportive (stromal; '*stroma*') cellular compartments in the bone marrow. The second assesses differences in brain, liver and muscle tissues. Both experiments are of biological significance as they tackle novel challenges in understanding of cellular behavior: the former in the complex functional unit termed hematopoietic stem cell niche, where inter-dependent hematopoietic and stromal cell functions synergize in the blood-forming function of the bone marrow; the latter in comparison of cell fate determined by the tissue origin from the separate layers of the embryo: ectoderm (brain), endoderm (liver) and mesoderm (muscle). While of general character, the chosen classification tasks are not just random biological exercises as these studies may illuminate cellular functions determined by gene expression signatures in complex cell system seeded by cell-type-heterogeneous undifferentiated populations (hematopoietic and stromal stem cells in the cell niche), and in the cell-type-homogeneous differentiated tissues (brain, liver and muscle), respectively.

For both the first (2-class) and the second (3-class) classification problems, samples were downloaded from the Gene Expression Omnibus database [14]. We only downloaded control (non-treated, non-pathological) samples of each tissue in question. For ease of gene functional annotation, we only downloaded samples measured with platforms provided by Affymetrix. Table 1 provides the statistics on sample distribution among classes and platforms. Table 2 then shows statistics derived from the application of apriori constructed gene sets onto the collected expression samples.

Platform	1261	339	341	570	81	91	96	97
Organism	mmu	mmu	rno	hsa	mmu	hsa	hsa	hsa
Heme	46	7		4	19	6	18	18
Stroma	19		8	47			26	33

Platform	1261	91	96
Organism	mmu	hsa	hsa
Brain	6	15	20
Liver	11	2	6
Muscle	11	22	41

Table 1. Sample size statistics. Platforms are identified by NCBI's GPL keys. Organism keys stand for *mus musculus* (mmu), *homo sapiens* (hsa) and *rattus norvegicus* (rno)

.

Set type	**Total**	**Probesets contained**			
		Min	Max	Avg	Median
FCF	**901**	0	83	5.47	2
Pathway	**251**	0	457	52.09	33
GO term	**5164**	1	7605	25.75	3
Gene	12808*	1	49	1.58	1

*average across platforms

Table 2. Gene sets statistics. Numbers in bold are independent of the specific platforms measuring the expression data, being only determined by the respective types of background knowledge. The 'Probesets contained' columns capture statistics over all involved platforms. The first three rows correspond to the apriori defined sets. For accuracy, we list their sizes in terms of probesets, rather than genes. The statistical relation between genes and probes are in turn shown in the last row.

4 Results

Here we show the empirical results obtained by processing the data described in Section 3 by the method explained in Section 2 and comment on their relevance to questions Q1-Q3 formulated in the latter section. Results are of two types: single-platform (experiments conducted on a single type of microarray) and cross-platform (experiments on the integrated heterogeneous expression data). Single-platform experiments are shown in both classification tasks for the sample-richest platform pertaining to the homo sapiens organism (GPL97 and GPL96 respectively).

The principal trends observed are as follows. Q1 is addressed by the top two panels of Fig. 3. While they do not provide a conclusive performance ranking of the four types of sample representation, they clearly demonstrate that predictive accuracy is not sacrificed by converting the representation from genes to gene sets. On the contrary, the gene set representation based on GO terms quite systematically outperforms the original gene based representation. The lower two panels of Fig. 3 compare the three gene set based approaches in the cross-platform experiments where the gene based representation is not applicable. In the Heme-Stroma task, a clear ranking is observable with fully coupled fluxes performing best, followed by GO terms and lastly pathways. Ranking induced by the Brain-Liver-Muscle task is much less crisp.

Figures 4 and 5 relate to Q2. Fig. 4 provides the surprising finding that none of the three genuine gene set representations strictly outperforms its randomized counterpart in both tasks performed in the single-platform setting; with the pathway based gene set representation being strikingly outperformed in the Brain-Liver-Muscle task. To make sure that these results were not a statistical artifact we regenerated all the randomized gene sets and arrived at principally same results. Combining these results with the top row of Fig 3, we deduce another observation that the random gene set approach often improves classification accuracy upon the basic classification based on gene expressions. This latter observation can however be explained rather naturally by viewing the random gene set approach as a form of stochastic feature extraction [7] reducing the dimensionality of the data and thus suppressing the variance component [6] of the classification error. The trends are significantly different in the cross-platform setting (Fig. 5) where all genuine gene set types strictly outperform their random counterparts in both tasks. Here the value of biologically meaningful gene sets manifests itself clearly in that the sets act as links connecting diverse genes distributed across platforms. Such a link is obviously broken when the gene sets are randomized.

Finally, to answer Q3 we compare the upper panels of Fig. 3 against its lower panels. With large training data sizes, accuracy differences between single-platform (upper panels) and cross-platform (lower panels) learning are insignificant, letting us conclude that the assembling of multiple-platform data did not have a detrimental effect on classification performance. More importantly still, in the cross-platform setting, high accuracies are achieved much earlier along the x axis than in the single-platform setting. While the reason is obvious (the same sample set proportion corresponds to a higher absolute number of samples in the cross-platform case), this observation is reassuring. An experimenter possessing a sample set too small for reliable model induction may benefit from employing the gene set based approach to include further relevant public expression samples, however coming from diverse microarray platforms.

5 Conclusions and Future Work

We have demonstrated a set-level approach to the integration of multiple-platform gene expression data for predictive classification and argued its utility for boosting classification performance when single-platform samples are rare. We explored three ways of defining gene sets, including a novel way based on the notion of a fully coupled flux related to metabolic pathways. In two tissue classification tasks, we showed that the gene set based representation is unquestionably useful for combining heterogeneous expression data. This may be for sakes of assembling a larger sample set or to obtain general biological insights not limited to a particular organism. On the other hand, in experiments constrained to a single platform, biologically meaningful gene sets were often outperformed by random gene sets with no biological relevance. Further studies are obviously needed to conclusively compare the performance of biologically relevant gene sets with their randomized counterparts; such studies would especially be interesting in problems where the genuine gene set approach was shown successful, such as in [18, 11]. Another natu-

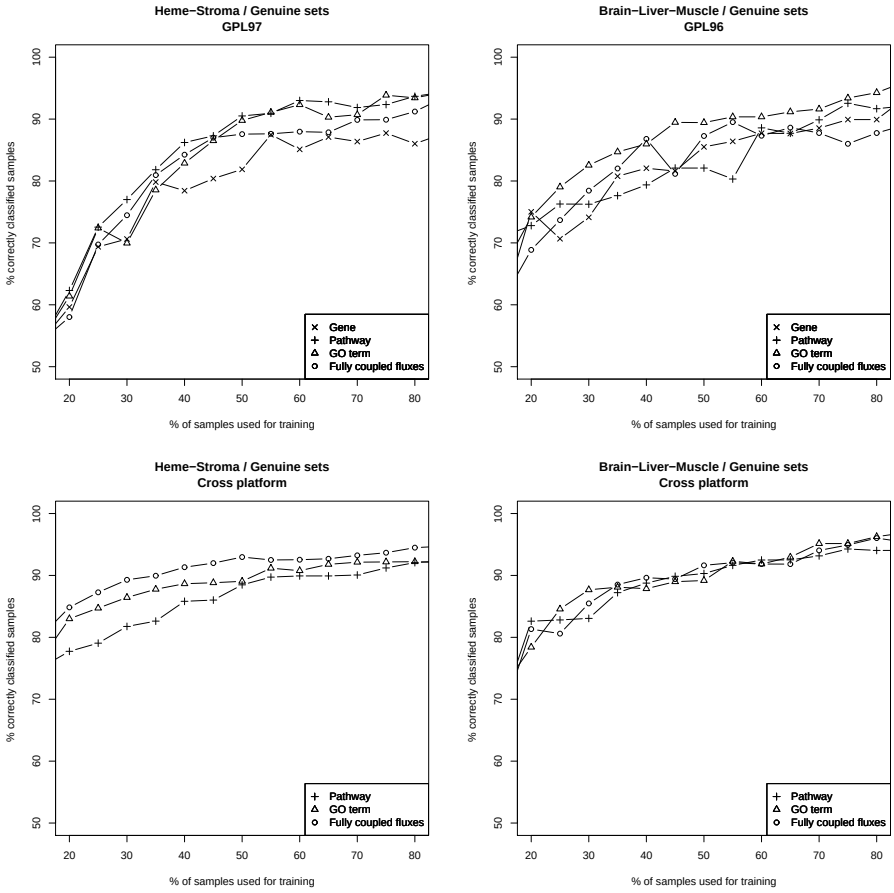

Fig. 3. Overall comparison of predictive classification performance using genes (only single-platform) and genuine gene sets. **Top:** single-platform, **Bottom:** cross-platform

ral extension of this work would be in the adoption of a less elementary approach to determine the pathway activation levels, e.g. along the lines of the study [15].

Acknowledgements. The authors are supported by the Czech Grant Agency through project 201/09/1665 (MH), the Czech Ministry of Education through projects ME910 (FZ) and MSM6840770012 (JK), and by the Children's Cancer Research Fund of the University of Minnesota (JT).

References

1. Andrea Bild and Phillip George Febbo. Application of a priori established gene sets to discover biologically important differential expression in microarray data. *PNAS*, 102(43):15278–15279, 2005.
2. B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003.

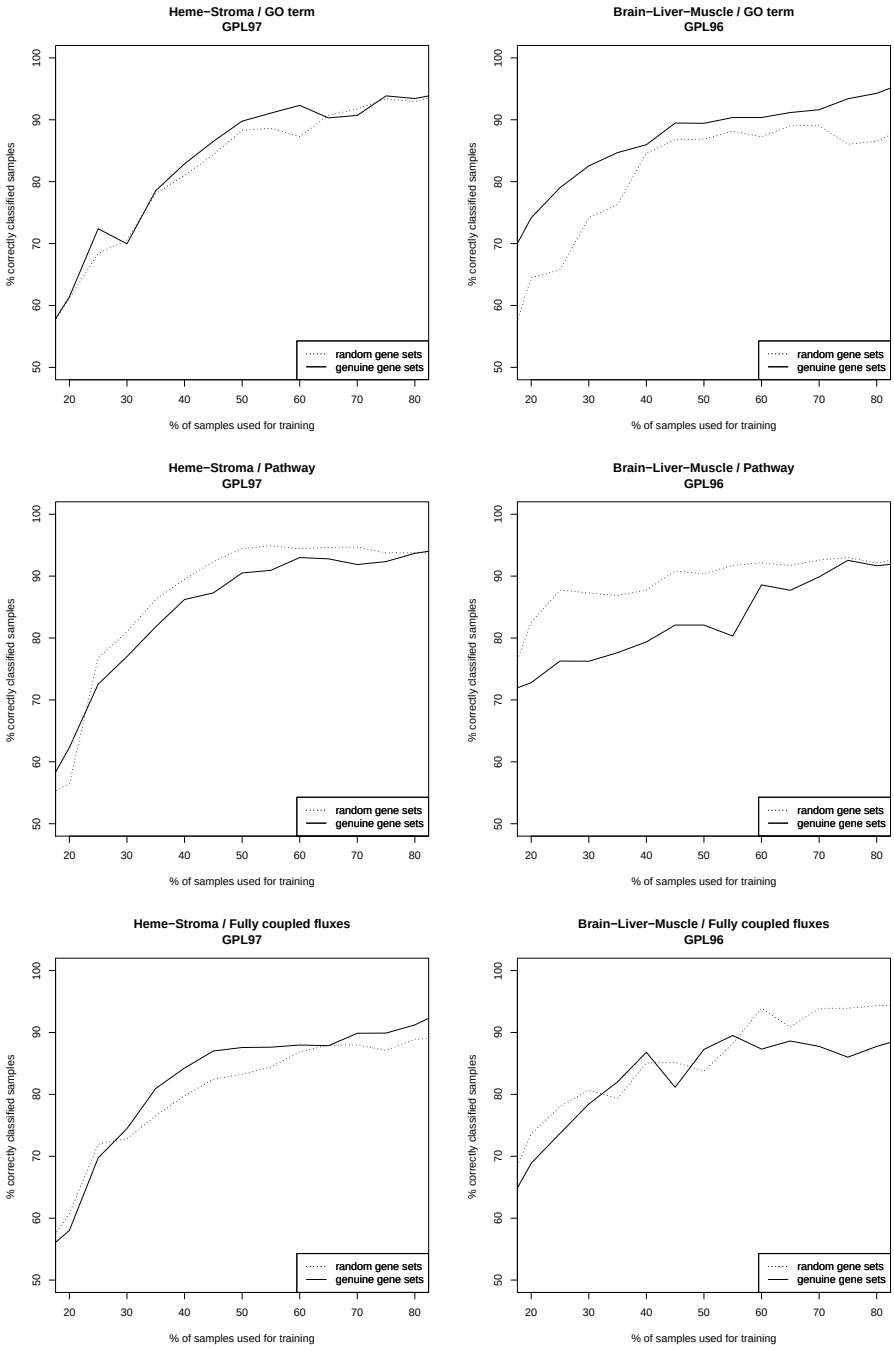

Fig. 4. Single-platform experiments comparing performance of predictive classification using genuine gene sets with that using random gene sets as sample features. Rows correspond to different gene set types, columns to different classification tasks.

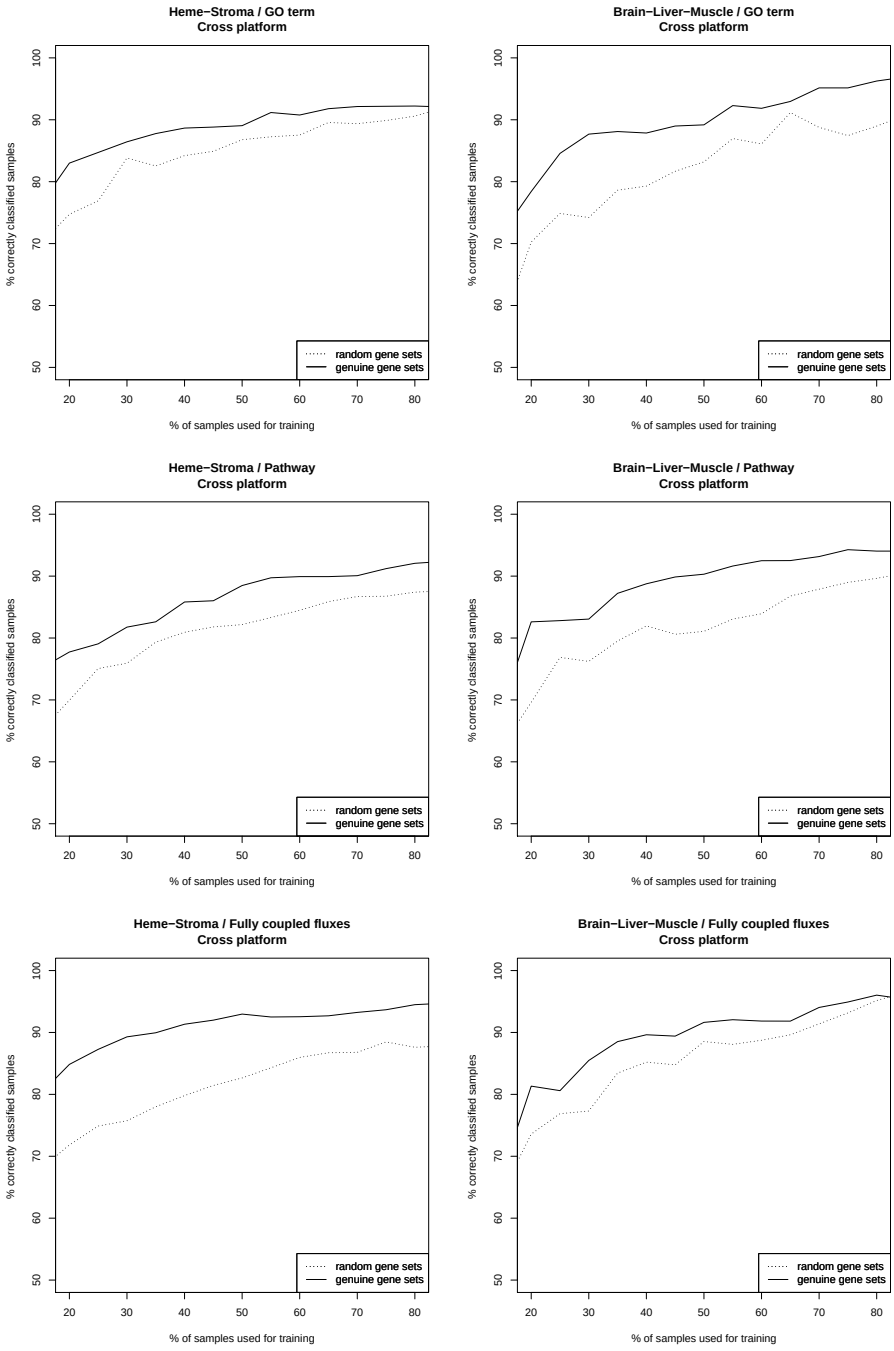

Fig. 5. Cross-platform experiments comparing performance of predictive classification using genuine gene sets with that using random gene sets as sample features. Rows correspond to different gene set types, columns to different classification tasks.

3. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25, 2000.
4. Robert C Gentleman, Vincent J. Carey, and Douglas M. Bates et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.
5. Jelle Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.
6. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.
7. Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.
8. M. Holec, F. Zelezny, and J. Klema et al. Using bio-pathways in relational learning. In *Late Breaking Papers, 18th International Conference on Inductive Logic Programming (ILP'08)*, 2008.
9. Da Wei Huang, Brad T Sherman, and Richard A Lempick. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4:44 – 57, 2009.
10. Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32:277–280, 2004.
11. VK Mootha, C Lindgren, and S Laureta et al. Pgc-1-alpha-responsive genes involved in oxidative phosphorylation are coorinately down regulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.
12. Dan L. Nicolae, Omar De la Cruz, William Wen, Baoguan Ke, and Minsun Song. Invited keynote talk: Set-level analyses for genome-wide association data. In *ISBRA'08: 4th International Symposium on Bioinformatics Research and Applications*. Springer, 2008.
13. Richard A. Notebaart, Bas Teusink, Roland J. Siezen, and Balzs Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLOS Computational Biology*, 4(1), 2008.
14. R Edgar R, M Domrachev M, and AE Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–10, 2002.
15. Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007.
16. AS Shaw and EL Filbert. Scaffold proteins and immune-cell signalling. *Nat Rev Immunol.*, 9(1):47–56, 2009.
17. Maria A. Stalteri and Andrew P. Harrison. Interpretation of multiple probe sets mapping to the same gene in affymetrix genechips. *BMC Bioinformatics*, 8:13+, 2007.
18. John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6, 2005.
19. T Weichhart and MD Semann. The PI3K/Akt/mTOR pathway in innate immune cells: emerging therapeutic applications. *Ann Rheum Dis.*, Suppl 3:iii:70–4, 2008.
20. Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.
21. Y Sun Y and J. Chen. mTOR signaling: PLD takes center stage. *Cell Cycle*, 7(20):3118–23, 2008.