Learning pathway-based decision rules to classify microarray cancer samples

*Enrico Glaab, Jonathan M. Garibaldi and Natalio Krasnogor
School of Computer Science, University of Nottingham, United Kingdom

enrico.glaab@cs.nott.ac.uk

Abstract:

Despite recent advances in DNA chip technology current microarray gene expression studies are still affected by high noise levels, small sample sizes and large numbers of uninformative genes. Combining microarray data with cellular pathway data by using new integrative analysis methods could help to alleviate some of these problems and provide new biological insights.

We present a method for learning simple decision rules for class prediction from pairwise comparisons of cellular pathways in terms of gene set expression levels representing the up- and down- regulation of pathway members. The procedure generates compact and comprehensible sets of rules, describing changes in the relative ranks of gene expression levels in pairs of pathways across different biological conditions. Results for two large-scale microarray studies, containing samples from prostate cancer and B-cell lymphoma patients, show that the method provides robust and accurate rule sets and new insights on differentially regulated pathway pairs. However, the main benefit of these predictive models in comparison to other classification methods like support vector machines lies not in the attained accuracy levels but in the ease of interpretation and the insights they provide on the relative regulation of cellular pathways in the biological conditions under consideration.

1 Introduction

Classification of microarray gene expression samples often suffers from several limitations resulting from the high dimensionality of the data, a typically small number of available samples, and from various sources of technical and biological noise. In recent years, several methods have extended or replaced classical machine learning methods to provide more compact, robust and easily interpretable classification models. These approaches reduce the prediction model complexity and increase its robustness by using regularization and shrinkage techniques [AMD+05, GHT07], by generating more human-interpretable machine learning models, which are based on simple decision rules [A+06, BK08], or by using more robust data representations and model formulations, e.g. computing rank scores [WEB05] or only considering relative expression values by comparing pairs of genes [G+04a, TNX+05].

In this paper, we address the problem of low model robustness due to noise by combining ideas from the techniques mentioned above with an approach to analyse the data at the

level of pathways instead of at the single-gene level. Briefly, we map the genes in a microarray study onto cellular pathways and processes from public databases and learn simple decision rules for sample classification by comparing gene expression levels in pairs of pathways. Rules describing single pathway-pairs are then weighted and combined into a unified classification model by applying a boosting algorithm. The approach can be understood as a methodological extension of the "top-scoring pairs" (TSP) algorithm [G$^+$04a, TNX$^+$05], which identifies discriminative pairs of genes in microarray data, and has therefore been named "top-scoring pathway pairs" (TSPP) algorithm. Moreover, we draw inspiration from other pathway-based microarray analysis approaches, which use summarized expression values for genes in cellular pathways and processes for enrichment analysis (e.g. the methods GSEA [S$^+$05], MaxMean [ET07] and the global test [G$^+$04b]) or as features for sample classification [G$^+$05].

In contrast to previous methods comparing single gene expression values or summarized expression values for single pathways against fitted threshold values, TSPP provides increased robustness by at the same time combining expression levels of multiple genes into "pathway expression fingerprints" and making pairwise, relative comparisons between pathways. In summary, the TSPP approach is not designed to compete with existing microarray sample classification and data mining methods, but to complement them with the following added benefits:

- New biological insights can be gained from easily interpretable decision rules on the relative up- and down-regulation of cellular pathways.
- The prediction models are applicable to data from other microarray platforms without requiring that all platforms contain the same genetic probes and that cross-study normalization is applied (the integration takes place at the level of pathways, and the gene expression values are replaced by rank scores).
- By summarizing the expression values of multiple genes belonging to the same pathway, the dimensionality of the data is reduced (from about 50.000 genes to a few hundred pathways) and the summarized "pathway expression fingerprints" have a higher robustness than single gene expression vectors (however, at the expense of losing detail; therefore single-gene based methods should be applied additionally).

2 Methods

The TSPP algorithm identifies, scores and combines decision rules based on pathway-pairs according to the following five-step procedure:

1. **Rank score transformation**:

   A gene expression matrix X with dimension $n \times p$ (n: number of samples, p: number of genes) and class labels y for the samples is read as input and transformed into a "rank matrix" R by sorting the expression values for each gene across the n samples

and replacing them with their position index in the sorted vector (ties are handled by replacing equal values by the mean of the corresponding position indices).

2. **Pathway mapping**:

   Gene sets representing cellular pathways and processes are extracted from a public database (e.g. KEGG, Gene Ontology, BioCarta or Reactome). Pathway assignments are computed for the p genes in the microarray input data by testing whether they occur in these gene sets. For genes which cannot be assigned to a pathway the corresponding rows are removed from matrix R.

3. **Scoring of pathway pairs**:

   To score a pair of pathways as being useful for discriminating between two sample class labels 1 and 2, e.g. "tumour (1) vs. normal (2)" or "drug treatment (1) vs. no treatment (2)"), the pathway-submatrices R_1 and R_2, corresponding to these two samples classes, are extracted from matrix R based on the mappings from step 2. The matrices R_1 and R_2 are then reduced to vectors r_1 and r_2 by replacing each column of expression level ranks by its median value. For a two-class problem, the score for a pathway-pair is then obtained by comparing the median ranks in pathway 1 to those in pathway 2 and computing the maximum of two relative frequencies: The relative frequency of samples which are up-regulated for class 1 and down-regulated for class 2, and vice versa, the relative frequency of cases which are down-regulated for class 1 and up-regulated for class 2 (i.e. there are two possibilities for the relation of sample ranks in two pathways to differ across the sample classes). Given the sets of column indices for two sample classes S_1 and S_2, the final score can thus be computed as follows:

$$partial\_score_1 = \sum_{i \in S1} I(r_{1i} >= r_{2i}) + \sum_{i \in S2} I(r_{1i} < r_{2i}) \tag{1}$$

$$partial\_score_2 = \sum_{i \in S1} I(r_{1i} < r_{2i}) + \sum_{i \in S2} I(r_{1i} >= r_{2i}) \tag{2}$$

$$score = \frac{max(partial\_score_1, partial\_score_2)}{|S1| + |S2|} \tag{3}$$

   where I is the indicator function. For a multi-class problem, a similar score can be obtained by computing the mean of the scores obtained for all pairs of sample classes.

4. **Identification of top-scoring pairs**:

   By default top-scoring pathway pairs (TSPPs) are identified by performing an exhaustive search across all pairs of pathways. This should be feasible in most practical applications, because the number of pathways is typically much smaller than the number of genes, and the scoring method is kept simple. Moreover, the method does not assume that all genes in a pathway are either up- or down-regulated, but

searches for pairs of pathways for which many genes occurring in the first pathway change their relation of expression level ranks across the sample classes to genes in the second pathway. Nevertheless, it might be beneficial to investigate whether alterations in the pathway definitions can provide improved results. Therefore, the user can alternatively let the algorithm introduce "mutations" into the pathway gene sets, by randomly adding or deleting genes up to a small user-defined maximum number of mutations, and replacing the exhaustive search by a previously published evolutionary search algorithm [JUA05]. Only one modification is applied to this algorithm: A genome contains two bit-vectors representing two pathways and mutations are only applied to one of these bit-vectors, selected randomly. The scoring function in the evolutionary algorithm is the same as for the exhaustive search.

5. **Classification model generation**:

   Each TSPP provides a simple decision rule for classifying microarray samples depending on the relative median expression value ranks of their genes in a pair of pathways. To combine multiple TSPPs into a unified classification model, we use the TSPP decision rules as "base classifiers" in the Adaboost.M1 algorithm [FS96], adding one decision rule at a time to the boosting model based on the order of the TSPP-scores computed in step 3. This boosting scheme assigns weights to each decision rule in the combined ensemble model, accounting for a rule's prediction accuracy and capacity to correctly classify samples that were misclassified by decision rules added in previous iterations of the algorithm. Previous experiments with boosting and ensemble techniques applied to microarray data [GGK09, HPG+] have shown that improvements can be obtained both in terms of robustness and accuracy.

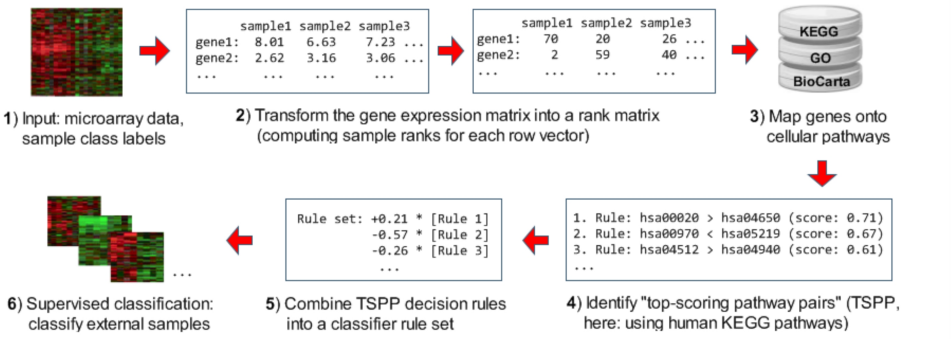

Figure 1: An overview of the workflow in the TSPP algorithm (example data is derived from a human prostate cancer microarray dataset [S+02b])

3 Results

The TSPP algorithm was applied to the gene expression matrices from two public microarray studies covering different types of cancer: B-cell lymphoma [S$^+$02a] (7129 genes and 77 samples) and prostate cancer [S$^+$02b] (12600 genes and 102 samples). Both datasets contain samples from two biological classes: In the B-cell lymphoma dataset 58 samples were obtained from patients suffering from diffuse large B-cell lymphoma (class D), while the remaining samples derive from a related follicular B-cell lymphoma (class F). The prostate cancer expression measurements were obtained from 50 healthy control tissues (class C) and 52 tumour tissues (class T) (for details on the normalization and preprocessing of the datasets, see the Data Sets section).

To evaluate the predictive accuracy for TSPP-models generated for these datasets, we applied an external leave-one-out cross-validation (LOOCV) procedure using different numbers of top-scoring pairs k (for k = 1, 3, 5, 10 and 15) and including all modelling steps in the cross-validation procedure. The parameter k can be regarded as a bias/variance trade-off, enabling the user to control the complexity of the generated classifiers. The cross-validation results, computed both for mappings of genes to KEGG pathways and to Gene Ontology (GO) terms, include the average accuracy, sensitivity and specificity for each LOOCV run and are shown in Tables 1 and 2.

Table 1: **Leave-one-out cross-validation results (TSPP on KEGG database)**

Dataset	No. of top-scoring pairs	Sensitivity (%)	Specificity (%)	Avg. Accuracy (%)
Prostate cancer	1	83.7	71.7	77.5
	3	87.8	73.6	80.4
	5	85.7	77.4	81.4
	10	77.6	73.6	75.5
	15	79.6	64.2	71.6
Lymphoma	1	64.9	85.0	70.1
	3	68.4	90.0	74.0
	5	78.9	90.0	81.8
	10	77.2	90.0	80.5
	15	75.4	90.0	79.2

In summary, average classification accuracies above 70% were obtained in all cases, and for both datasets the best accuracies (prostate cancer: 81.4%, DLBCL: 81.8%) were achieved when using 5 top-scoring pairs, suggesting that k = 5 represents a reasonable bias/variance trade-off. The sensitivity and specificity scores were in a roughly similar percentage range.

Apart from using the decision rules for class prediction, their simplicity also makes them suitable for direct human interpretation. The ten top-scoring pathway pairs for each dataset are shown in Tables 4 and 5. Interestingly, the top-ranked rule for the prostate cancer dataset contains the KEGG-pathways "Prostate cancer" and "Insulin signaling", which are both known to be de-regulated in the disease [SK03, H$^+$01]. However, the results

Table 2: **Leave-one-out cross-validation results (TSPP on GO database)**

Dataset	No. of top-scoring pairs	Sensitivity (%)	Specificity (%)	Avg. Accuracy (%)
Prostate cancer	1	83.7	67.9	75.5
	3	89.8	67.9	78.4
	5	89.8	69.8	79.4
	10	91.8	66.0	78.4
	15	85.7	67.9	76.5
Lymphoma	1	68.4	80.0	71.4
	3	57.9	90.0	66.2
	5	71.9	90.0	76.6
	10	52.6	90.0	62.3
	15	71.9	85.0	75.3

Table 3: **Leave-one-out cross-validation results (Gene-based: eBayes & SVM)**

Dataset	No. of features (genes)	Sensitivity	Specificity	Avg. Accuracy (%)
Prostate cancer	2	88.0	84.6	86.3
	6	96.0	88.5	92.2
	10	96.0	86.5	91.2
	20	90.0	88.5	89.2
	30	90.0	90.4	90.2
Lymphoma	2	91.4	68.4	85.7
	6	93.1	78.9	89.6
	10	94.8	94.7	94.8
	20	96.6	84.2	93.5
	30	98.3	100.0	98.7

also point to relative de-regulations in other pathways with less obvious associations to the cancer disease, e.g. "Pyrimidine metabolism" and "Glycerolipid metabolism", with a score close to the best-ranked pair. Similarly, for the B-cell dataset the top-ranked pathway pairs contain pathways known to be associated with B-cell neoplasia, e.g. the "Wnt signaling pathway" [QERR03, LB03], whereas for other pathways no direct and specific associations with the disease are known. In spite of the class-imbalance in this dataset, the prediction models did not display a preference to assign samples to the majority class; however, similar to other statistical methods for microarray data analysis, problems with robustness can occur when the sample size per condition is very small. Thus, when planning a microarray study, the experimenter might first want to study the literature on sample size estimation [LHC10], microarray study design [Chu02] and sampling techniques to alleviate these problems [VHKNW09].

It is also important to note that in a top-scoring pathway pair (TSPP) not necessarily both pathways are differentially regulated across the sample classes, but one pathway might have a constant expression, while the other pathway is highly de-regulated in one of the sample classes. The main benefit of comparing pairs of pathways lies in the possibility to avoid comparing single pathways against fitted thresholds, which would more likely

be affected by experimental bias and thus provide prediction models with higher generalization error. However, if a user's main goal is not to obtain a prediction model from the TSPP-algorithm, but to identify pathway associations, then TSPPs in which one of the pathways is not differentially regulated across the sample classes can easily be identified and filtered out by computing the variance for the corresponding gene expression vectors and removing TSPPs containing a pathway with low variance.

When using the evolutionary search methodology and allowing the algorithm to introduce small numbers of random gene deletions and insertions into the pathways (up to five genes), in spite of the higher flexibility of this method, in all experiments the prediction accuracies are either similar or lower than those obtained for the original pathways using an exhaustive search (data not shown). The weaker performance might result from an entrapment in local minima due to the expansion of the search space, but could also suggest that the original pathways and processes are already well defined and therefore hard to optimize based on an evolutionary search procedure.

Overall, the results from the cross-validation analysis and the lists of top-scoring pathways show that the method can generate compact predictive models with both high interpretability and high accuracy in comparison with a random model predictor (when measuring this using the "proportional chance criterion" by Huberty [Hub94], we obtain p-values < 0.01 in all cases). To put these results into relation with existing machine learning methods based on single genes as predictors, we applied a C-SVM from the e1071 R software package [DHL$^+$05], a wrapper for the well-known LibSVM library [CL01], with different kernel functions, including the radial basis function and polynomial kernels with a degree up to 3 (the results for the best kernel, a linear SVM, are reported in Table 3). The gene-based SVM-models achieve higher average accuracies than pathway-based models, with the best models reaching more than 90% accuracy on both datasets; however, these models only contain information on the relevance of single genes for the prediction and do not enable an interpretation of the data on the level of cellular pathways and processes. Although the simple decision rules generated by the TSPP algorithm do not reach the highest accuracies obtained by the support vector machine on single genes, their high interpretability and significant predictive information content allow the user to quickly identify cases, in which the relative gene expression in pathway pairs is differentially regulated across different biological conditions.

To investigate the utility of top-scored pathway pairs (TSPPs) in more detail, we have mapped the genes in these pathways onto their corresponding proteins in a large-scale protein-protein interaction network, consisting of 38857 interactions between 9392 proteins assembled from direct binary interactions in a previous study [GBKV10]. Figure 2 a) shows the largest connected component of an example mapping for the TSPP with the highest score on the Prostate cancer dataset, "hsa05215 Prostate cancer" vs. "hsa04910 Insulin signaling pathway" (see also Figure 1), revealing a strong network of interactions between these pathways, which also share a significantly large set of overlapping genes/proteins (q-value = 5.1E-17, when testing the hsa04910 pathway against all other KEGG pathways using the one-sided Fisher exact test and adjusting for multiple testing with the Benjamini-Hochberg method [BH95]). However, the TSPP-method also points the user to differentially regulated pathway pairs which would not be detected as signif-

Table 4: **Top-ranked pathway pairs (Prostate cancer data)**

Rank	Pathway 1	Pathway 2	Direction	Score
1	hsa05215 Prostate cancer	hsa04910 Insulin signaling pathway	down	0.81
2	hsa00240 Pyrimidine metabolism	hsa00561 Glycerolipid metabolism	up	0.80
3	hsa04540 Gap junction	hsa05210 Colorectal cancer	up	0.78
4	hsa04115 p53 signaling pathway	hsa00230 Purine metabolism	down	0.75
5	hsa04510 Focal adhesion	hsa00071 Fatty acid metabolism	down	0.75
6	hsa04514 Cell adhesion molecules (CAMs)	hsa04610 Complement and coagulation cascades	up	0.72
7	hsa03050 Proteasome	hsa01430 Cell Communication	up	0.69
8	hsa04920 Adipocytokine signaling pathway	hsa04730 Long-term depression	up	0.69
9	hsa04810 Regulation of actin cytoskeleton	hsa04530 Tight junction	down	0.65
10	hsa04512 ECM-receptor interaction	hsa04110 Cell cycle	down	0.63

The 10 top-ranked pathways for the prostate cancer dataset based on the TSPP-score (Direction "down" means that in the healthy control samples, pathway 1 is down-regulated in relation to pathway 2, whereas in the prostate cancer samples, pathway 1 is up-regulated in relation to pathway 2, and respectively, "up" means the pathways have opposite relations in the two sample classes).

icantly associated based on an overlap-based significance test, e.g. Figure 2 b) shows the largest connected component for the TSPP "hsa04115 p53 signaling pathway" vs. "hsa00230 Purine metabolism", with only two overlapping proteins, but a multitude of direct binary protein-protein interactions between the two pathways. Further experimental evidence for an association between these pathways is provided by a study showing that the inhibition of de novo purine synthesis by the drug "AG2034", which also inhibits prostate cancer cell growth, increases the expression levels of p53 [OKM09]. Thus, although the up- and down-regulation of top-scoring pathway pairs does not necessarily result from a regulatory relationship between the pathways, the analysis of the TSPPs can help to point the user to associations between pathways, which would remain unnoticed by other methods, such as an overlap-based Fisher test.

.

3.1 Data sets

3.1.1 Diffuse large B-cell lymphoma (DLBCL)

The DLBCL data set [S$^+$02a] contains expression values for 7,129 genes and 77 microarray samples, 58 of which were obtained from patients suffering from diffuse large B-cell lymphoma (D), while the remaining samples derive from a related B-cell lymphoma, called follicular lymphoma (F). The experiments in this microarray study had been carried out on an Affymetrix HU6800 oligonucleotide platform [Aff01].

To pre-process the raw data, we applied the "Variance stabilizing normalization" [HvHS$^+$02]

Table 5: **Top-ranked pathway pairs (B-Cell lymphoma data)**

Rank	Pathway 1	Pathway 2	Direction	Score
1	hsa00020 Citrate cycle (TCA cycle)	hsa04310 Wnt signaling pathway	down	0.88
2	hsa00052 Galactose metabolism	hsa04664 Fc epsilon RI signaling pathway	down	0.87
3	hsa04670 Leukocyte transendothelial migration	hsa03050 Proteasome	up	0.87
4	hsa04514 Cell adhesion molecules (CAMs)	hsa00030 Pentose phosphate pathway	up	0.86
5	hsa04730 Long-term depression	hsa00240 Pyrimidine metabolism	up	0.85
6	hsa00562 Inositol phosphate metabolism	hsa00051 Fructose an mannose metabolism	up	0.84
7	hsa00220 Urea cycle and metabolism of amino groups	hsa00980 Metabolism of xenobiotics by cytochrome P450	down	0.84
8	hsa04540 Gap junction	hsa00330 Arginine and proline metabolism	up	0.84
9	hsa00252 Alanine and aspartate metabolism	hsa04630 Jak-STAT signaling pathway	down	0.84
10	hsa00970 Aminoacyl-tRNA biosynthesis	hsa04912 GnRH signaling pathway	down	0.81

The 10 top-ranked pathways for the B-Cell lymphoma dataset based on the TSPP-score (Direction "down" means that in the DLBCL samples, pathway 1 is down-regulated in relation to pathway 2, whereas in the follicular B-cell lymphoma samples, pathway 1 is up-regulated in relation to pathway 2, and respectively, "up" means the pathways have opposite relations in the two sample classes).

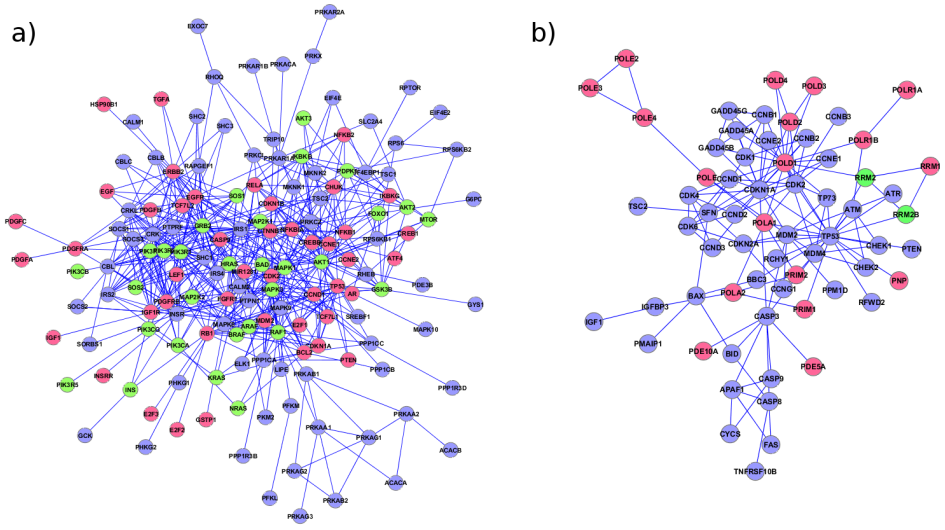

Figure 2: Analysing TSPPs in a protein-protein interaction network: a) Largest connected component for KEGG pathways: "Prostate cancer" and "Insulin signaling" (blue: Prostate cancer, red: Insulin signaling, green: members in both pathways); b) Largest connected component for KEGG pathways "P53 signaling" and "Purine metabolism" (blue: P53 signaling, red: Purine metabolism, green: members in both pathways)

to filter out intensity-dependent variance (this was done using the vsn-library and the expresso-package in the R statistical learning environment [Tea10]). Moreover, we applied thresholding based on the suggestions in the supplementary material of the original publication [S$^+$02a] and a "fold change"-filter to remove all genes with less than a 3-fold change between the maximum and minimum expression value.

3.1.2 Prostate cancer

The prostate cancer data set [S$^+$02b] consists of expression measurements for 12,600 genetic probes across 50 healthy control tissues (C) and 52 prostate cancer tissues (C). All experiments have been carried out on Affymetrix Hum95Av2 arrays [Aff01]. Due to the large number of samples and memory limitations of the expresso-package (used to normalize the other two data sets), we applied the fast GeneChip RMA (GCRMA) normalization algorithm [WI05]. Moreover, we employed thresholding based on the suggestions in the original publication of the dataset [S$^+$02b] and a fold change filter to remove all probes with less than a 2-fold change between the maximum and minimum expression value.

Table 6: **Data sets used in this paper**

Data set	Platform	No. of genes	No. of samples class 1; class 2	references
B-cell lymphoma	Affymetrix	7,129	58 (D) ; 19 (F)	[S$^+$02a]
Prostate cancer	Affymetrix	12,600	52 (T) ; 50 (C)	[S$^+$02b]

4 Conclusion

We present a new method for extracting pathway-based decision rules from combined gene expression data and gene sets representing cellular pathways and processes. When applying prediction models derived from these decision rules for sample classification on two public microarray cancer datasets, we obtain compact and easily interpretable models with significant predictive information content. The generated decision rules are robust against monotonic transformations of the data, and the algorithm is easy to implement and has a comparatively short run-time due to the reduction of the data dimensionality when considering summarized pathway expression values instead of gene expression values. Moreover, these models also enable a different interpretation of microarray data by analysing the data at the level of pathways. Specifically, the top-scoring pathway pairs can point the user to regulatory relationships or other functional associations between the corresponding pathways. In summary, the TSPP algorithm provides both a novel method to generate compact and accurate classification models and a new exploratory tool to analyse microarray data at the level of pairwise pathway-relations.

References

[A+06] G. Alexe et al. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Res*, 8(4):R41, 2006.

[Aff01] Affymetrix. *Affymetrix Microarray Suite User Guide, Version 5*, 2001.

[AMD+05] N. Ancona, R. Maglietta, A. D'Addabbo, S. Liuni, and G. Pesole. Regularized least squares cancer classifiers from DNA microarray data. *BMC Bioinformatics*, 6(Suppl 4):S2, 2005.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B (Methodological)*, 57:289–300, 1995.

[BK08] J. Bacardit and N. Krasnogor. Fast rule representation for continuous attributes in genetics-based machine learning. In *Genet Evol Comput Conf*, pages 1421–1422. ACM, 2008.

[Chu02] G.A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nat Genet*, 32:490–495, 2002.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[DHL+05] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, and M.F. Leisch. Misc functions of the department of statistics (e1071), TU Wien, 2005. R-Package e1071 version 1.5-19.

[ET07] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Ann Appl Stat*, 1(1):107–129, 2007.

[FS96] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proc Int Conf Mach Learn*, pages 148–156. ACM, 1996.

[G+04a] D. Geman et al. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*, 3(19), 2004.

[G+04b] J.J. Goeman et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[G+05] Z. Guo et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6(1):58, 2005.

[GBKV10] E. Glaab, A. Baudot, N. Krasnogor, and A. Valencia. TopoGSA: network topological gene set analysis. *Bioinformatics*, 26(9):1271–1272, 2010.

[GGK09] E. Glaab, J.M. Garibaldi, and N. Krasnogor. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinformatics*, 10(1):358, 2009.

[GHT07] Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

[H+01] A.W. Hsing et al. Prostate cancer risk and serum levels of insulin and leptin: a population-based study. *J Natl Cancer Inst*, 93(10):783–789, 2001.

[HPG+] H. O. Habashy, D. G. Powe, E. Glaab, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, G. Ball, A. R. Green, C. Caldas, and I. O. Ellis. RERG (Ras-related and oestrogen-regulated growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype. *Breast Cancer Research and Treatment*. (online first).

[Hub94] C. J. Huberty. *Applied Discriminant Analysis*. John Wiley, New York, 1994.

[HvHS+02] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(1):96–104, 2002.

[JUA05] T. Jirapech-Umpai and S. Aitken. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6(1):148, 2005.

[LB03] B. Lustig and J. Behrens. The Wnt signaling pathway and its role in tumor development. *J Cancer Res Clin Oncol*, 129(4):199–221, 2003.

[LHC10] W.J. Lin, H.M. Hsueh, and J.J. Chen. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, 11(1):48, 2010.

[OKM09] O. Obajimi, J.C. Keen, and P.W. Melera. Inhibition of de novo purine synthesis in human prostate cells results in ATP depletion, AMPK activation and induces senescence. *The Prostate*, 69(11):1206–1221, 2009.

[QERR03] Y.W. Qiang, Y. Endo, J.S. Rubin, and S. Rudikoff. Wnt signaling in B-cell neoplasia. *Oncogene*, 22(10):1536–1545, 2003.

[S+02a] M.A. Shipp et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*, 8(1):68–74, 2002.

[S+02b] D. Singh et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[S+05] A. Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*, 102(43):15545–15550, 2005.

[SK03] P. Stattin and R. Kaaks. Prostate cancer, insulin, and androgen deprivation therapy. *Br J Cancer*, 89(9):1814–1815, 2003.

[Tea10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

[TNX+05] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, and D. Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.

[VHKNW09] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, and R. Wald. Feature Selection with High-Dimensional Imbalanced Data. In *2009 IEEE International Conference on Data Mining Workshops*, pages 507–514. IEEE, 2009.

[WEB05] P. Warnat, R. Eils, and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6(1):265, 2005.

[WI05] Z. Wu and R.A. Irizarry. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. *J Comput Biol*, 12(6):882–893, 2005.