International Journal of Medical Informatics (2005) **74**, 491—503

ELSEVIER

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data

Alexander Statnikov, Ioannis Tsamardinos, Yerbolat Dosbayev, Constantin F. Aliferis*

Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, 2209 Garland Avenue, Nashville, TN 37232, USA

Received 30 October 2004; accepted 2 May 2005

KEYWORDS
Gene expression microarray analysis;
Decision support systems;
Neoplasms;
Diagnosis;
Computer-assisted;
Artificial intelligence

Summary The success of treatment of patients with cancer depends on establishing an accurate diagnosis. To this end, we have built a system called GEMS (gene expression model selector) for the automated development and evaluation of high-quality cancer diagnostic models and biomarker discovery from microarray gene expression data. In order to determine and equip the system with the best performing diagnostic methodologies in this domain, we first conducted a comprehensive evaluation of classification algorithms using 11 cancer microarray datasets. In this paper we present a preliminary evaluation of the system with five new datasets. The performance of the models produced automatically by GEMS is comparable or better than the results obtained by human analysts. Additionally, we performed a cross-dataset evaluation of the system. This involved using a dataset to build a diagnostic model and to estimate its future performance, then applying this model and evaluating its performance on a different dataset. We found that models produced by GEMS indeed perform well in independent samples and, furthermore, the cross-validation performance estimates output by the system approximate well the error obtained by the independent validation. GEMS is freely available for download for non-commercial use from http://www.gems-system.org.

1. Introduction

Development of cancer diagnostic models and discovery from DNA microarray data is of great interest in bioinformatics and medicine. Diagnostic models from gene expression data go beyond traditional histopathology and provide accurate, resource-efficient, and replicable diagnosis [1]. Furthermore, biomarker discovery in high-dimensional microarray data facilitates learning about the biology of cancer [2]. Currently, building of cancer diagnostic models from microarray gene expression

* Corresponding author.
E-mail address: constantin.aliferis@vanderbilt.edu (C.F. Aliferis).

1386-5056/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2005.05.002

data has three challenging components: collection of samples, assaying, and statistical analysis. A typical statistical analysis process takes from a few weeks to several months and involves many specialists: clinical researchers, statisticians, bioinformaticians, and programmers. As a result, statistical analysis is a serious bottleneck in the development of cancer diagnostic models, and its enhancement by an automated or semi-automated system will benefit research significantly. Our goal is thus to build a system that takes microarray data as input and outputs a high-quality cancer diagnostic model, produces a reliable performance estimate, allows application of this model to unseen patients, and enables biomarker discovery. In order for the system to be clinically successful, it should implement the best-known methodologies applicable to this domain and use sound techniques for model selection and performance estimation in an automated fashion. An ideal system should achieve the same or better quality than human analysts and complete the entire process within minutes or a few hours requiring minimal human effort.

The paper is organized as follows: in Section 2, we describe existing software systems for cancer diagnosis from microarray gene expression data as well as prior research in this field. Section 3 summarizes results of the comprehensive algorithmic evaluation conducted in [3,4] and describes the system GEMS (gene expression model selector). In Section 4, we conduct an evaluation of GEMS by applying the system to cancer microarray datasets not used in algorithmic evaluation and by assessing performance of developed diagnostic models using microarray datasets from different laboratories. The results of this section constitute the primarily novel contribution of the presented work. Section 5 provides directions for future research and current limitations of GEMS. The paper concludes with Section 6.

2. Related work

2.1. Existing software systems

Currently, there exist many dozens of software systems designed for microarray gene expression data analysis [5,6]. Since prior research has demonstrated superiority of supervised classification methods for cancer diagnosis over unsupervised techniques [7], we focused only on systems implementing supervised classification algorithms. Using this criterion, we identified 16 software systems (Table 1): 6 are commercial (names are shown with boldface in Table 1) and 10 can be used free of charge for non-profit research. All systems have several of the following limitations. First, the performance quality of the learning algorithms selected for inclusion into the systems is unknown. Typically, the algorithmic palette reflects the authors' preferences and their prior publication history; there is often limited evidence that these algorithms are indeed appropriate for this domain and equally important, that they are among the best performing ones. Second, many classification algorithms implemented in the software systems are not able to handle multicategory diagnosis, despite that most diagnostic tasks involve several diseases and that powerful multicategory classification methods do exist in machine learning. Third, none of the systems automatically optimizes the parameters and the choice of both classification and gene selection algorithms (also known as model selection) while simultaneously avoiding overfitting.[1] The user of these systems is left with two choices: either to avoid rigorous model selection and possibly discover a suboptimal model, or to experiment with many different parameters and algorithms and select the model with the highest cross-validation performance. The latter is subject to overfitting primarily due to multiple-testing, since parameters and algorithms are selected after all the testing sets in cross-validation have been seen by the algorithms [3]. All these problems are addressed in GEMS.

2.2. Prior methodological studies

Previous studies in cancer diagnosis model creation from gene expression data provide limited evidence for selecting the best performing learning techniques. We identified 193 primary gene expression-based cancer diagnosis studies using the ONCOMINE Cancer Microarray Database [8], the UPITT Cancer Gene Expression Data Set Link Database [9], and the Stanford Microarray Database [10]. A review of these studies and publications that reanalyzed publicly available datasets (identified by querying ISI Web of Science Cited Reference Search and PubMed Central Citation Search for citations of primary microarray studies) revealed the following:

[1] Only one commercial software system, Partek Predict by Partek Inc., attempts to automatically conduct a rigorous optimization of the parameters and the choice of algorithms while providing unbiased performance estimates. Unfortunately, the current version 6.0 of Partek Predict does not completely implement this methodology, since it does not allow optimization of the choice of gene selection algorithms.

Table 1 Software systems for gene expression-based cancer diagnosis (supervised classification only)

Name	Version	Developer	Supervised classification	Cross-validation for performance estimation	Automatic model selection for classifier and gene selection methods	URL
ArrayMiner ClassMarker	5.2	Optimal Design, Belgium	•K-Nearest Neighbors •Voting	**Yes**	**No**	http://www.optimaldesign.com/ArrayMiner
Avadis Prophetic	3.3	Strand Genomics, USA	•Decision Trees •Neural Networks •Support Vector Machines	**Yes**	**No**	http://avadis.strandgenomics.com/
BRBArrayTools	3.2 Beta	National Cancer Institute, USA	•Compound Covariate Predictor •Diagonal Linear Discriminant Analysis •Nearest Centroid •K-Nearest Neighbors •Support Vector Machines	**Yes**	**No**	http://linus.nci.nih.gov/BRB-ArrayTools.html
caGEDA	(accessed 10/2004)	University of Pittsburgh and University of Pittsburgh Medical Center, USA	•Nearest Neighbors Methods •Naïve Bayes Classifier	**Yes**	**No**	http://bioinformatics.upmc.edu/GE2/GEDA.html
Cleaver	1.0 (accessed 10/2004)	Stanford University, USA	•Linear Discriminant Analysis	**Yes**	**No**	http://classify.stanford.edu/
GeneCluster2	2.1.7	Broad Institute, Massachusetts Institute of Technology, USA	•Weighted Voting •K-Nearest Neighbors	**Yes**	**No**	http://www.broad.mit.edu/cancer/software
GeneLinker Platinum	4.5	Predictive Patterns Software, Canada	•Neural Networks •Support Vector Machines •Linear Discriminant Analysis •Quadratic Discriminant Analysis •Uniform/Gaussian Discriminant Analysis	**Yes**	**No**	http://www.predictivepatterns.com/
GeneMaths XT	1.02	Applied Maths, Belgium	•Neural Networks •K-Nearest Neighbors •Support Vector Machines	**Yes**	**No**	http://www.applied-maths.com/genemaths/genemaths.htm
GenePattern	1.2.1	Broad Institute, Massachusetts Institute of Technology, USA	•Weighted Voting •K-Nearest Neighbors •Support Vector Machines	**Yes**	**No**	http://www.broad.mit.edu/cancer/software
Genesis	1.5.0	Graz University of Technology, Austria	•Support Vector Machines	**No**	**No**	http://genome.tugraz.at/Software/Genesis/Genesis.html
GeneSpring	7	Silicon Genetics, USA	•K-Nearest Neighbors •Support Vector Machines	**Yes**	**No**	http://www.silicongenetics.com/
GEPAS	1.1 (accessed 10/2004)	National Center for Cancer Research (CNIO), Spain	•K-Nearest Neighbors •Support Vector Machines •Diagonal Linear Discriminant Analysis	**Yes**	**Limited** (for number of genes)	http://gepas.bioinfo.cnio.es/tools.html

Table 1 (*Continued*)

Name	Version	Developer	Supervised classification	Cross-validation for performance estimation	Automatic model selection for classifier and gene selection methods	URL
MultiExperiment Viewer	3.0.3	The Institute for Genomic Research, USA	•K-Nearest Neighbors •Support Vector Machines	**Yes**	**No**	http://www.tigr.org/software/tm4/mev.html
PAM	1.21a	Stanford University, USA	•Nearest Shrunken Centroids	**Yes**	**Limited** (for a single parameter of the classifier)	http://www-stat.stanford.edu/~tibs/PAM/
Partek Predict	6.0	Partek, USA	•K-Nearest Neighbors •Nearest Centroid Classifier •Discriminant Analysis	**Yes**	**Limited** (does not allow optimization of the choice of gene selection algorithms)	http://www.partek.com/
Weka Explorer	3.4.3	University of Waikato, New Zeland	•K-Nearest Neighbors •Decision Trees •Rule Sets •Bayesian Classifiers •Support Vector Machines •Multi-Layer Perceptron •Linear Regression •Logistic Regression •Meta-Learning Techniques (Boosting, Bagging)	**Yes**	**No**	http://www.cs.waikato.ac.nz/ml/weka/

- a typical study applies only a few (usually, 2—3) classification algorithms to a single cancer microarray dataset;
- the majority of diagnostic tasks pursued by the studies are binary (i.e. with two possible outcomes), whereas real-life diagnostic problems are generally multicategory;
- researchers often apply parametric classifiers without rigorous optimization of their parameters;
- different computational experimental designs employed by the studies (e.g., *N*-fold cross-validation, leave-one-out cross-validation, hold-out cross-validation, bootstrapping, etc.) make the findings incomparable.

We also located two meta-analyses covering the scope of our research [11,12]. According to [11], only 26% of studies in this domain attempted independent validation or cross-validation of their findings. This questions whether published results will generalize well to unseen patients. Unfortunately, neither of these two meta-analyses is aimed at the identification of the best performing methodologies, nor can be used to do so. The meta-analysis by Ntzani and Ioannidis [11] examined the predictive performance of DNA microarrays for cancer diagnosis and prognosis in general, without resorting to specific algorithms. The meta-analysis by Rhodes et al. [12] is geared toward biomarker assessment across 40 studies and uses a single simplistic biomarker discovery method and an equally simplistic and a non-standard classifier.

In addition, two recent bioinformatics studies [13,14] performed comparative analyses of multicategory classification algorithms in the cancer gene expression domain. However, the results of these evaluations cannot serve as a basis for the development of cancer diagnostic decision support system for the following two reasons: first, both evaluations are limited only to two microarray datasets and second, neither study optimized parameters of the classifiers in all datasets, which is likely to result in suboptimal application of diagnostic methods.

For the above reasons, and given the plethora of classification algorithms applicable to gene expression-based cancer diagnostic problems, it was unclear what constitutes a small subset of methods that perform optimally across many datasets and cancer types. Therefore, we decided to conduct such an evaluation de novo in order to base our system on the currently best techniques for the chosen task and domain. The methods and results of this evaluation are discussed in two com-

panion publications [3,4] and summarized in the next section.

3. GEMS methods and development

3.1. Nested cross-validation procedure

At the core of our algorithmic evaluation is a *nested cross-validation* procedure [15,16]. This experimentation protocol allows the simultaneous optimal selection of parameters (tuning) of a classifier and the unbiased (non-overfitted) estimation of the performance of the final diagnostic model. Cross-validation is a method for providing an estimate of the performance of a diagnostic model produced by a learning procedure A on available data D. First, one partitions the data D into N non-overlapping and balanced subsets of cases. Then, the following is repeated N times: A is trained on the N-1 subsets (training set) and tested on the hold-out subset (testing set). Finally, the average performance ρ of A over the N testing sets is reported. This methodology produces an unbiased performance estimate ρ of the model produced by A by training on all the available data D (i.e., all N subsets comprise the training set). The pseudo-code of the procedure referred to as *cross-validation for performance estimation* is shown in Fig. 1, where for simplicity A is a classifier with a fixed parameter α.

Typically, however, a classifier used for learning is parametric and the optimal value of parameters should be estimated and used to produce the final model. Let us assume that the classifier can be applied with parameter α taking m values = $\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{m-1}, \alpha_m\}$, where α_i is a vector with the following parameters:

- choice of classification algorithms (e.g., K-Nearest Neighbors, Support Vector Machines (SVMs));
- parameters of the specific classification algorithms (e.g., number of neighbors K for K-Nearest Neighbors, penalty parameter C for Support Vector Machines);
- choice of algorithms applied prior to classification, such as gene selection, normalization, imputation, and others (e.g., gene selection by signal-to-noise ratio, gene selection by ANOVA);
- parameters of algorithms applied prior to classification (e.g., number of genes to be used for classification).

To estimate the optimal value of the parameter α, cross-validation is used again. The performance $P(i)$ of learner A trained with parameter α_i is estimated for $i = 1, \ldots, m$ by cross-validation. The final model is built by training A on all available data D using the parameter α_j, where $j = \text{argmax}\, P(i)$ for $i = 1, \ldots, m$ (Fig. 2). Notice that in Fig. 2 cross-validation is used only for model selection and it does not provide an unbiased performance estimate for the final model and so we call this procedure *cross-validation for model selection*.

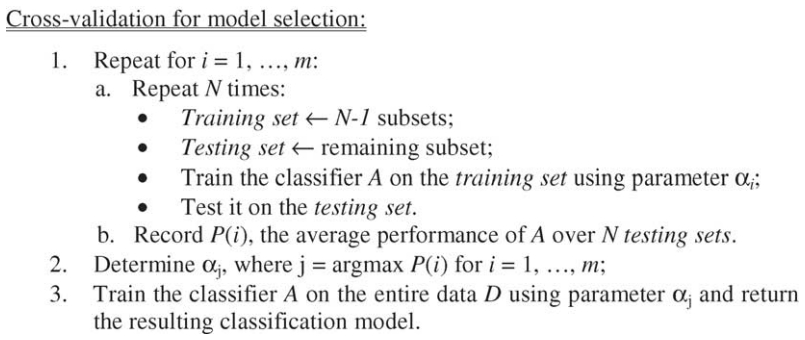
In order to combine optimal model selection and unbiased performance estimation, the *cross-validation for model selection* is ''nested'' inside the *cross-validation for performance estimation* to obtain the *nested cross-validation* procedure (Fig. 3). The dashed box in Fig. 3 corresponds to *cross-validation for model selection* (steps 1.1, 1.2, and 1.3) ''nested'' into the steps 1 and 2 belonging to *cross-validation for performance estimation*. Since the optimized classifier is each time evaluated on a testing set not used for learning, the resulting performance estimate ρ is unbiased.

The algorithm in Fig. 3 avoids the following common pitfall in estimating the performance of a diagnostic model produced by a parametric classifier: quite often, the procedure in Fig. 2 is used to identify the best parameter values and to build the final model; however, the best cross-validation performance $P(j)$, where $j = \text{argmax}\, P(i)$ for $i = 1, \ldots, m$ is often reported as an estimate of performance of the final model, instead of applying a second cross-validation loop over the whole model selection procedure as in Fig. 3. For a sufficiently large number of attempted parameter values, one is likely to be found that by chance alone provides a high estimate of cross-validation performance. The less the available sample is, and the more complex models the

Cross-validation for performance estimation:

1. Repeat N times:
   - *Training set* ← *N-1* subsets;
   - *Testing set* ← remaining subset;
   - Train the classifier A on the *training set* using parameter α;
   - Test it on the *testing set*.
2. Return ρ, the average performance of A over N *testing sets*.

Fig. 1 Cross-validation for performance estimation.


Cross-validation for model selection:

1. Repeat for $i = 1, \ldots, m$:
   a. Repeat N times:
      - *Training set* ← *N-1* subsets;
      - *Testing set* ← remaining subset;
      - Train the classifier A on the *training set* using parameter α_i;
      - Test it on the *testing set*.
   b. Record $P(i)$, the average performance of A over N *testing sets*.
2. Determine α_j, where j = argmax $P(i)$ for $i = 1, \ldots, m$;
3. Train the classifier A on the entire data D using parameter α_j and return the resulting classification model.

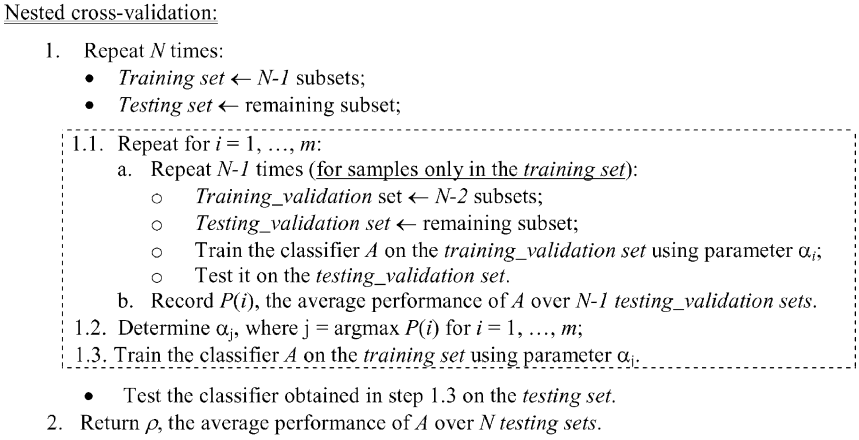
Fig. 2 Cross-validation for model selection.

classifier can build, the more acute becomes the problem. In contrast, the described *nested cross-validation* protocol will be able to identify whether the model selection procedure is selecting values that by accident produce models that perform well on the test sets, or indeed they generalize well to unseen cases.

3.2. Algorithmic evaluation

To determine which algorithms should comprise the basis of our system, we conducted a comprehensive algorithmic evaluation [3,4] (Table 2). Performance estimation and model selection were performed by two nested cross-validation designs: one based on 10-fold cross-validation and another based on leave-one-out cross-validation. We used 11 classification algorithms that are: (I) relatively insensitive to low sample-to-variable ratio, (II) considered robust in bioinformatics and machine learning research, (III) allow multicategory classification, and (IV) represent major families of learning machines. Since the employed classifiers are different in a sense that they give preference to different models, the final classification performance may be improved via use of algorithms that combine outputs of individual classifiers, so-called *ensembles of classifiers*. Therefore, we explored seven types of ensemble classification algorithms. In the case of ensemble classification, the input dataset consisted of attributes corresponding to the outputs of classifiers and the original class labels. We also applied four gene selection techniques to investigate the improvement of classification performance by reduction of the number of predictive variables. Two performance metrics were used to assess classifications: accuracy and the entropy-based relative classifier information metric. Whereas the former metric allows comparison with the prior studies, the latter one is insensitive to unbalanced distribution of diagnostic categories. Finally, a randomized permutation procedure was devised and employed to determine whether the classifiers produce different predictions to a statistically significant degree. Details on the application of the methods are provided in [3,4] and references therein.

The previous algorithms were applied to 11 microarray cancer gene expression datasets (see Table 3 and [3,4] for references to the primary studies). Most of the datasets were obtained using


Nested cross-validation:

1. Repeat N times:
   - *Training set* ← *N-1* subsets;
   - *Testing set* ← remaining subset;

   1.1. Repeat for $i = 1, \ldots, m$:
      a. Repeat *N-1* times (for samples only in the *training set*):
         - *Training_validation* set ← *N-2* subsets;
         - *Testing_validation set* ← remaining subset;
         - Train the classifier A on the *training_validation set* using parameter α_i;
         - Test it on the *testing_validation set*.
      b. Record $P(i)$, the average performance of A over *N-1 testing_validation sets*.
   1.2. Determine α_j, where j = argmax $P(i)$ for $i = 1, \ldots, m$;
   1.3. Train the classifier A on the *training set* using parameter α_j.

   - Test the classifier obtained in step 1.3 on the *testing set*.
2. Return ρ, the average performance of A over N *testing sets*.

Fig. 3 Nested cross-validation for performance estimation in the outer loop and model selection in the inner loop (dashed box).

Table 2 Methods used for comprehensive evaluation of algorithms for cancer diagnosis from microarray gene expression data

Classification algorithms
- K-Nearest Neighbors
- Backpropagation Neural Networks
- Probabilistic Neural Networks
- Multi-Class SVM: one-versus-rest
- Multi-Class SVM: one-versus-one
- Multi-Class SVM: DAGSVM
- Multi-Class SVM by Weston and Watkins
- Multi-Class SVM by Crammer and Singer
- Weighted Voting: one-versus-rest
- Weighted Voting: one-versus-one
- Decision Trees: CART

Ensemble classification algorithms

Based on outputs of Multi-Class SVM methods
- Majority voting
- Decision Trees: CART
- Multi-Class SVM: DAGSVM
- Multi-Class SVM: one-versus-rest
- Multi-Class SVM: one-versus-one

Based on outputs of all classifiers
- Majority voting
- Decision Trees: CART

Computational experimental design
- Leave-one-out cross-validation for performance estimation (outer loop) and 10-fold cross-validation for model selection (inner loop)
- 10-fold cross-validation for performance estimation (outer loop) and 9-fold cross-validation for model selection (inner loop)

Gene selection methods
- Signal-to-noise ratio in one-versus-rest fashion
- Signal-to-noise ratio in one-versus-one fashion
- Kruskal—Wallis nonparametric one-way ANOVA
- Ratio of genes between-categories to within-category sum of squares

Performance metrics
- Accuracy
- Relative classifier information (entropy-based performance metric)

Statistical comparison among classifiers
- Custom randomized permutation procedure

oligonucleotide technology. These datasets were chosen because they contain a relatively large sample and are often used in bioinformatics as benchmarks. Overall, the 11 datasets had 2—26 distinct diagnostic categories, 50—308 samples (patients), and 2308—15,009 variables (genes) after removing genes with absent calls [3].

In order to make our experiments computationally feasible, we followed a staged experimental design. In the first stage, the two nested cross-validation designs (10-fold and leave-one-out cross-validation) were executed with all classifiers and all datasets without gene selection. In the second stage, we applied four gene selection techniques only to the datasets where the resulting accuracy was not near perfect (operationally defined as <90%). The staged factorial design allowed us to complete experiments within 4 single CPU months and led to development of 2.6 million diagnostic models.

The results of this evaluation are described in detail in [3,4] and the online supplement to [3] available for download from http://www.gems-system.org. In summary, we concluded the following:

- for multicategory classification of cancer from microarray gene expression data, Support Vector Machines are the best performing family among the tested algorithms outperforming K-Nearest Neighbors, Backpropagation Neural Networks, Probabilistic Neural Networks, Decision Trees, and Weighted Voting classifiers to a statistically significant degree;
- among multicategory Support Vector Machines, the best performing techniques are: one-versus-rest, the method by Weston and Watkins, and the method by Crammer and Singer;
- the diagnostic performance can be moderately improved for SVMs and significantly improved for the non-SVM methods by gene selection;
- ensemble classification does not improve performance of the best non-ensemble diagnostic models;
- the obtained results favorably compare with the primary literature on the same datasets.

3.3. System functionality and implementation

GEMS incorporates the results of the algorithmic evaluation by providing to the user an implementation of all and only the best performing learning algorithms in this domain. Given a microarray dataset on input, the system can automatically perform one the following tasks:

I. Generate a classification model optimizing the parameters of classification and gene selection algorithms as well as the choice of the classifier and gene selection methods using *cross-validation for model selection* (Fig. 2).

Table 3 Cancer-related human gene expression datasets used for evaluation

Dataset name	Diagnostic task	Number of				Max. prior (%)
		Samples	Variables (genes)	Categories	Variables/ samples	
11_Tumors	11 various human tumor types	174	12533	11	72	15.5
14_Tumors	14 various human tumor types and 12 normal tissue types	308	15009	26	49	9.7
9_Tumors	9 various human tumor types	60	5726	9	95	15.0
Brain_Tumor1	5 human brain tumor types	90	5920	5	66	66.7
Brain_Tumor2	4 malignant glioma types	50	10367	4	207	30.0
Leukemia1	Acute myelogenous leukemia (AML), acute lympboblastic leukemia (ALL) B-cell, and ALL T-cell	72	5327	3	74	52.8
Leukemia2	AML, ALL, and mixed-lineage leukemia (MLL)	72	11225	3	156	38.9
Lung_Cancer	4 lung cancer types and normal tissues	203	12600	5	62	68.5
SRBCT	Small, round blue cell tumors (SRBCT) of childhood	83	2308	4	28	34.9
Prostate _Tumor	Prostate tumor and normal tissues	102	10509	2	103	51.0
DLBCL	Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas	77	5469	2	71	75.3

The column ''Max. prior'' indicates the prior probability of the dominant diagnostic category.

II. Estimate classification performance of the optimized model by nested *cross-validation* (Fig. 3).
III. Perform tasks I and II, i.e. generate a classification model and estimate its performance.
IV. Apply an existing model to a new set of patients.

In order to execute the tasks mentioned above, the user may select the type of the experimental design (*N*-fold cross-validation or leave-one-out cross-validation), the algorithm(s) to be used for classification, gene selection, and normalization, and the ranges of parameters over which optimization should take place. Table 4 summarizes all implemented algorithms. As the system evolved and based on discussions with our biomedical colleagues, we added new functionality to the system, namely, several simple gene expression normalization methods, area under ROC curve performance metric (for binary diagnostic problems), and two state of the art local causal discovery algorithms [17,18] shown with boldface in Table 4. To guide the user's choices according to the available computational power and time, the system outputs the number of models to be generated while the user is selecting analysis options. GEMS provides an intuitive wizard-like user interface abstracting the microarray data analysis process and not requiring users to be experts in data analysis.[2] Each step in the interface consists a form with options for a specific stage of analysis (Fig. 4):

- overall task selection
- dataset specification
- cross-validation design
- normalization
- classification
- gene selection
- performance estimation
- logging
- report generation
- execution of analysis

Since the system can perform one out of four tasks outlined above, each task corresponds to a different sequence of steps. The overall software architecture of GEMS is shown in Fig. 5. The system implements a client-server architecture consisting of a computational engine and an interface client. The computational engine is separated from the client and consists of intercommunicating functional units corresponding to different aspects of analysis. Upon completion of analysis, a detailed report is generated in HTML format with links to system input and output files as well as links to NCBI website with information on selected genes. The GEMS graphics user interface is implemented using Borland Delphi 6.0 and the computational engine is programmed in Mathworks Matlab 6.5.1 and Microsoft Visual C++ 6.0.

[2] Unlike the version 2.0 of GEMS discussed in this paper, the previous version 1.0 of the system introduced in [3,4] possesses a power-user interface with all analysis options placed on a single form.

Table 4 Algorithms implemented in GEMS

Classification algorithms
- Multi-Class SVM: one-versus-rest
- Multi-Class SVM: one-versus-one
- Multi-Class SVM: DAGSVM
- Multi-Class SVM by Weston and Watkins
- Multi-Class SVM by Crammer and Singer

Gene selection methods
- Signal-to-noise ratio in one-versus-rest fashion
- Signal-to-noise ratio in one-versus-one fashion
- Kruskal–Wallis nonparametric one-way ANOVA
- Ratio of genes between-categories to within-category sum of squares
- **HITON_PC**
- **HITON_MB**

Normalization techniques
- For every gene $x \rightarrow$ [a, b]
- **For every gene $x \rightarrow [x - \text{mean}(x)]/\text{std.}(x)$**
- **For every gene $x \rightarrow x/\text{std.}(x)$**
- **For every gene $x \rightarrow x/\text{mean}(x)$**
- **For every gene $x \rightarrow x/\text{median}(x)$**
- **For every gene $x \rightarrow x/||x||$**
- **For every gene $x \rightarrow x - \text{mean}(x)$**
- **For every gene $x \rightarrow x - \text{median}(x)$**
- **For every gene $x \rightarrow |x|$**
- **For every gene $x \rightarrow x + |x|$**
- **For every gene $x \rightarrow \log(x)$**

Computational experimental design
- Leave-one-out cross-validation for performance estimation (outer loop) and N-fold cross-validation for model selection (inner loop)
- N-fold cross-validation for performance estimation (outer loop) and $(N-1)$-fold cross-validation for model selection (inner loop)
- Leave-one-out cross-validation for model selection
- N-fold cross-validation for model selection

Performance metrics
- Accuracy
- Relative classifier information (entropy-based performance metric)
- **Area under ROC curve (AUC)**

4. Preliminary evaluation of the system

In order to debug the interface and the algorithmic engine of the system, we repeated most of experiments performed in [3,4] using GEMS, and we found that published results completely matched the system outputs. Next, we performed two studies to evaluate the system. First, we applied GEMS to several microarray datasets, not included in our previous study, and compared the resulting system performance with the published models. Second, we performed a cross-dataset evaluation of the system. This involved using the system to build a classifier from a gene-expression dataset, estimating its cross-validation performance on the same dataset, and then applying that classifier to a different dataset (using the same genes and diagnostic target) produced by an independent research group.

4.1. Application of GEMS to new datasets

We selected five human cancer microarray gene expression datasets to test our system: *6_Tumors* [19], *Leukemia3* [20], *Lung_Cancer2* [21], *DLBCL2* [22], and *Lung_Cancer3*[3] [23] (see Table 5 for description of datasets and diagnostic tasks). All five datasets were produced using Affymetrix oligonucleotide technology and processed as in [3]. None of these datasets was included in our previous studies [3,4].

The results of application of GEMS to these datasets are presented in Table 6.[4] The analyses completed within 10–30 min per dataset and yielded performance results comparable or better than ones obtained by human analysts and previously published in literature.

4.2. Cross-dataset evaluation of the system

Many researchers believe that even though small cross-validation error is an important finding, it still requires further validation on an independent data [7]. There are two reasons for doing this: (1) cross-validation performance estimates in very small samples may have large variance, and (2) the dataset may not be representative of the general population. Therefore, we used GEMS to conduct two analyses with construction and evaluation of the classifier on one dataset and consecutive independent validation on another data. The results are summarized in Table 7, and the text below describes our experiments and findings in detail.

[3] This dataset contains the same adenocarcinoma samples as in previously analysed *Lung_Cancer* data [24]. However, *Lung_Cancer3* dataset contains additional mesothelioma samples and is now used to solve a different diagnostic problem (adenocarcinoma versus mesothelioma) compared to diagnosis developed using *Lung_Cancer* data (adenocarcinoma versus squamous versus small-cell lung cancer versus pulmonary carcinoids versus normal tissues).

[4] Notice that for *DLCBL2* dataset Savage et al. reported 88.7% accuracy in their paper [22], however, in the supplement the authors clarified that their classification procedure might be biased, since they optimized their classifier based on testing sets. When experiments were repeated using nested cross-validation, the authors obtained 83.9% accuracy (see supplement to [22]).

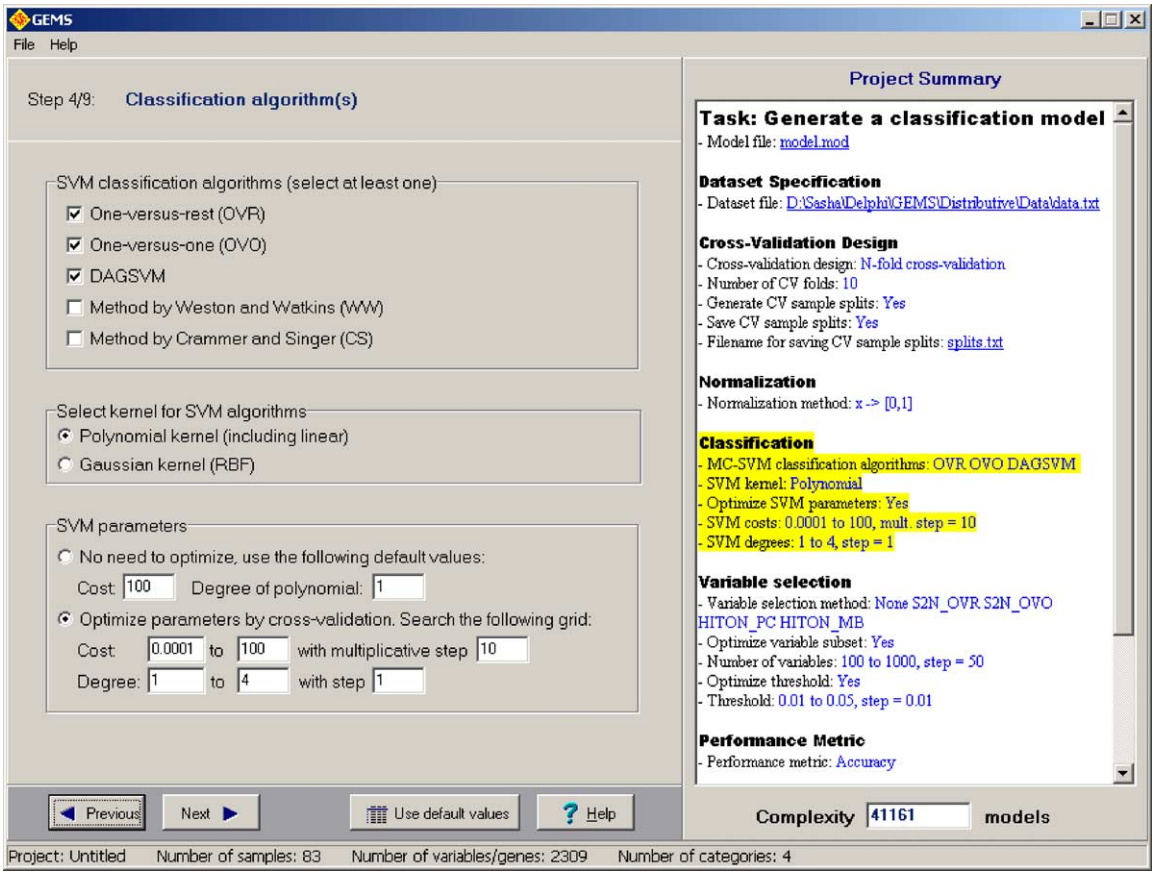


Fig. 4 An example screen-shot of GEMS. The left part of the screen contains options for the current analysis step (classification algorithm). The summary of the entire project is shown in the right part of the screen.

First, we used the *Lung_Cancer* [24] and *Lung_Cancer2* [21] datasets with the diagnostic task being to differentiate between cancerous and normal tissues. The *Lung_Cancer* dataset contains 186 tumor and 17 normal samples, and *Lung_Cancer2* dataset contains 86 tumor and 10 normal samples. The datasets were produced using different microarray technologies: *Lung_Cancer* dataset was obtained using Affymetrix Human Genome U95A chips with 12,600 oligonucleotide probes, while *Lung_Cancer2* dataset was obtained using Affymetrix HuGeneFL chips with 7129 oligonucleotide probes. The mapping of 6623 probes from HuGeneFL to 7094 probes from Human Genome U95A was derived using Affymetrix array comparison spreadsheets [25]. Next, we used GEMS

Table 5 Cancer-related human gene expression datasets used for preliminary evaluation of GEMS system

Dataset name	Diagnostic task	Number of				Max. prior (%)
		Samples	Variables (genes)	Categories	Variables/samples	
6_Tumors	Six various human tumor types	353	7069	6	20	32.0
Leukemia3	Six types of leukemia	248	12135	6	49	31.9
Lung_Cancer2	Lung cancer and normal tissues	96	7129	2	74	89.6
Lung_Cancer3	Mesothelioma and adenocarcinoma	181	12533	2	69	82.9
DLBCL2	Diffuse large B-cell lymphomas (DLBCL) and mediastinal large B-cell lymphomas (MLBCL)	210	32404	2	154	83.8

The column ''Max. prior'' indicates the prior probability of the dominant diagnostic category.

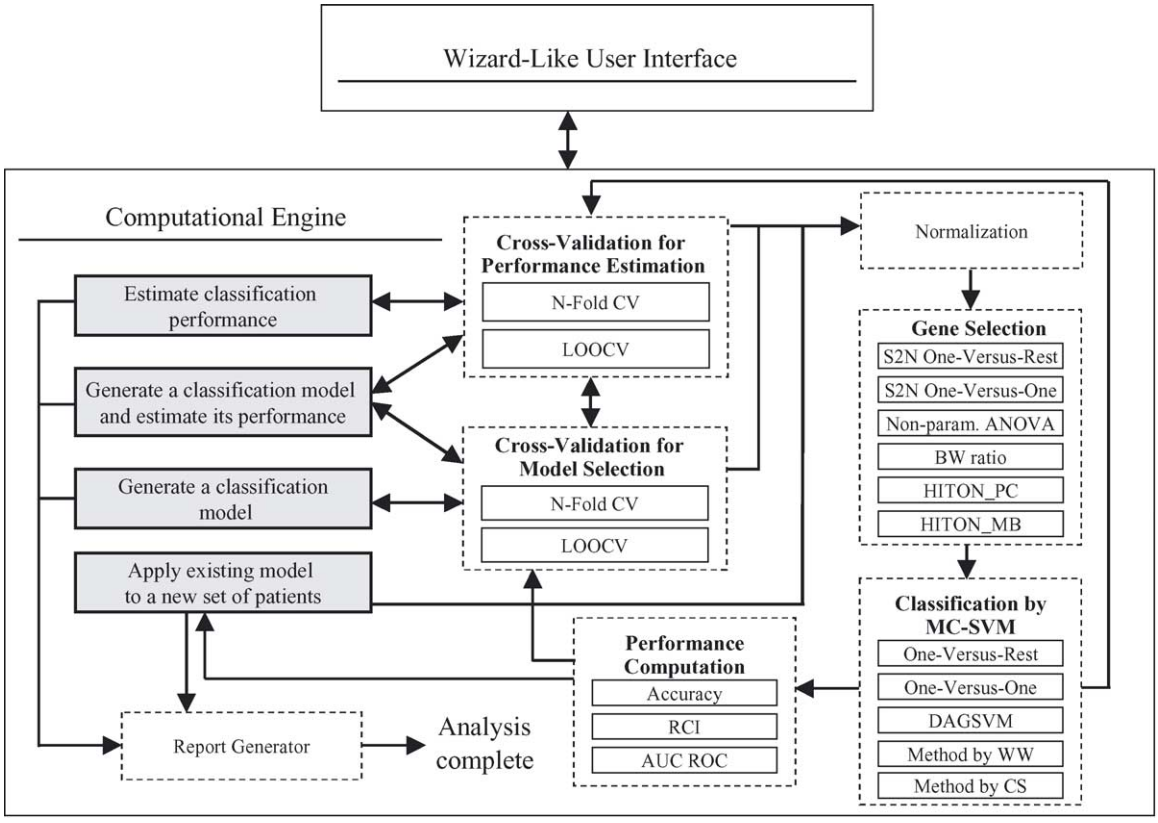


Fig. 5 Software architecture of GEMS.

to generate a classification model and estimate its performance in a nested cross-validation fashion using *Lung_Cancer* dataset. We decided to use area under ROC curve (AUC) as a performance metric since both datasets are not balanced in terms of distribution of cancerous and normal samples. GEMS created a classification model and estimated its cross-validation performance to be 100% AUC. When this model was applied to *Lung_Cancer2* data, the actual performance was again 100% AUC. We emphasize that, *Lung Cancer2* was never seen by the model neither during training, nor during the performance estimation phase.

Table 6 Results of application of GEMS to five microarray datasets not employed for algorithmic evaluation

Dataset name	GEMS classification accuracy (%)	Published classification accuracy (%)
6_Tumors	97.2	96.0
Leukemia3	98.4	98.4
Lung_Cancer2	100.0	100.0
Lung_Cancer3	99.4	99.3
DLBCL2	87.1	83.9

Similarly, we used *Leukemia1* [1] and *Leukemia2* [26] datasets with the goal to build a classifier to predict whether a patient has acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML). The *Leukemia1* dataset contains 47 ALL and 25 AML samples, and *Leukemia2* dataset contains 24 ALL and 28 AML samples. Again, the datasets were produced using different microarray technologies: *Leukemia2* dataset was obtained using Affymetrix Human Genome U95A chips, while *Leukemia1* dataset was obtained using Affymetrix HuGeneFL chips. We used a similar approach as described above to map probes between datasets. Next, we fed *Leukemia2* dataset to GEMS to create a classification model and estimate its performance in a nested cross-validation fashion (AUC = 100%). When this model was applied to *Leukemia1* data, the final classification performance was 99.15% AUC.

In summary, the performance of the models as estimated by the system on one dataset is approximately equal to its performance on the independent dataset. This provides further confidence on the use of nested-cross-validation design both for generating the models and for estimating their performance.

Table 7 Results of cross-dataset experiments: first, we used a dataset to build a diagnostic model and to estimate its future performance by cross-validation, and then we applied this model and computed its performance on a different dataset

Dataset used for construction of a classification model		Performance estimate of the model[a] (AUC, %)	Dataset used for independent validation of the classification model		Performance on the independent dataset (AUC, %)
Name	Distribution of samples		Name	Distribution of samples	
Lung_Cancer	186 tumors, 17 normals	100.00	*Lung_Cancer2*	86 tumors, 10 normals	100.00
Leukemia2	24 ALL, 28 AML	100.00	*Leukemia1*	47 ALL, 25 AML	99.15

More details on datasets used for these experiments are provided in Tables 3 and 5.
[a] This performance estimate was obtained by nested cross-validation on the dataset used for construction of the model.

5. Limitations and future research

Although GEMS is a highly robust system for cancer diagnosis and discovery, it can be improved in several ways. Since many biomedical researchers and practitioners are interested in causal discovery, we are planning to extend and perform an evaluation of the computational causal discovery algorithms implemented in the system. The system evaluation presented in this paper was laboratory based with authors functioning as users of the system. In the future, we plan to conduct a fielded evaluation of the system, ideally, with various types of users from different institutions and organizations. We also believe that gene selection capabilities of the system can be extended by SVM-based gene selection, such as the RFE algorithm [27], and additional Markov-blanket based techniques [17,18]. The current version of GEMS communicates with SVM classifiers by a file input/output interface. A dynamic linked library or similar interface can provide significant speed-up of GEMS by eliminating necessity to write and read multi-megabyte microarray data files. Finally, the output report produced by the system provides minimal links to existing knowledge about genes. In particular, it will be useful to link the report on selected genes to Go terms and known pathways and interactions.

6. Conclusion

In this work we described GEMS, a system for automated development and evaluation of cancer diagnostic models and biomarker discovery from microarray gene expression data. Unlike past efforts, this system is informed by a comparative evaluation of many classification and related algorithms (e.g., cross-validation, gene selection, etc.) applicable for this task and domain. In a preliminary evaluation of the system with five cancer gene expression datasets not employed for the algorithmic comparison, GEMS completed the analysis of each dataset within 10—30 min and the output model performed as well as or better than previously published models obtained by human analysts. Also, we used this system to perform cross-dataset analysis of cancer diagnostic models using two pairs of different datasets corresponding to two different diagnostic tasks. We found that the diagnostic models obtained by GEMS in one dataset generalize well in data from a different laboratory and that nested cross-validation performance estimates well approximate the error obtained by the independent validation. The system is available for download from http://www.gems-system.org free of charge for non-commercial use.

Acknowledgements

This research was supported by NIH grants RO1 LM007948-01 and P20 LM 007613-01. We also acknowledge all developers of the systems listed in Table 1 for access to their software. In particular, we would like to acknowledge Partek Inc. for arranging a web-conference with demonstration of their software.

References

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531—537.
[2] A. Balmain, J. Gray, B. Ponder, The genetics and genomics of cancer, Nat Genet. 33 (Suppl.) (2003) 238—244 (Review).
[3] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics 21 (5) (2005) 631—643.

[4] A. Statnikov, C.F. Aliferis, I. Tsamardinos, Methods for multicategory cancer diagnosis from gene expression data: a comprehensive evaluation to inform decision support system development, Medinfo 2004 (2004) 813—817.
[5] G. Parmigiani, E.S. Garrett, R. Irizarry, S.L. Zeger (Eds.), The Analysis of Gene Expression Data: Methods and Software, Springer, New York, 2003.
[6] H.C. Causton, J. Quackenbush, A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide, Blackwell Publishing, 2003.
[7] R. Simon, M.D. Radmacher, K. Dobbin, L.M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, J. Natl. Cancer Inst. 95 (1) (2003) 14—18.
[8] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, A.M. Chinnaiyan, ONCOMINE: a cancer microarray database and data-mining platform, Neoplasia 6 (1) (2004) 1—6.
[9] University of Pittsburgh, Benedum Oncology Informatics Center, UPITT Cancer Gene Expression Data Set Link Database, http://bioinformatics.upmc.edu/Help/UPITTGED.html (accessed 10/2004).
[10] J. Gollub, C.A. Ball, G. Binkley, J. Demeter, D.B. Finkelstein, J.M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J.C. Matese, M. Schroeder, P.O. Brown, D. Botstein, G. Sherlock, The Stanford microarray database: data access and quality assessment tools, Nucleic Acids Res. 31 (1) (2003) 94—96.
[11] E.E. Ntzani, J.P. Ioannidis, Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment, Lancet 362 (9394) (2003) 1439—1444.
[12] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, A.M. Chinnaiyan, Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, Proc. Natl. Acad. Sci. U.S.A. 101 (25) (2004) 9309—9314.
[13] D.P. Berrar, C.S. Downes, W. Dubitzky, Multiclass cancer classification using gene expression profiling and probabilistic neural networks, Pac. Symp. Biocomput. (2003) 5—16.
[14] C. Romualdi, S. Campanaro, D. Campagna, B. Celegato, N. Cannata, S. Toppo, G. Valle, G. Lanfranchi, Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification, Hum. Mol. Genet. 12 (8) (2003) 823—836.
[15] S. Dudoit, M.J. van der Laan, Asymptotics of cross-validated risk estimation in model selection and performance assessment, U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 126, February 5, 2003.
[16] T. Scheffer, Error estimation and model selection, Ph.D. thesis, Technischen Universität Berlin, School of Computer Science, 1999.
[17] C.F. Aliferis, I. Tsamardinos, A. Statnikov, HITON: a novel Markov Blanket algorithm for optimal variable selection, AMIA Annu. Symp. Proc. (2003) 21—25.
[18] I. Tsamardinos, C.F. Aliferis, A. Statnikov, Time and sample efficient discovery of markov blankets and direct causal relations, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.
[19] K.A. Shedden, J.M. Taylor, T.J. Giordano, R. Kuick, D.E. Misek, G. Rennert, D.R. Schwartz, S.B. Gruber, C. Logsdon, D. Simeone, S.L. Kardia, J.K. Greenson, K.R. Cho, D.G. Beer, E.R. Fearon, S. Hanash, Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework, Am. J. Pathol. 163 (5) (2003) 1985—1995.
[20] E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve, L. Wong, J.R. Downing, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer Cell. 1 (2) (2002) 133—143.
[21] D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M. Taylor, M.D. Iannettoni, M.B. Orringer, S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (8) (2002) 816—824.
[22] K.J. Savage, S. Monti, J.L. Kutok, G. Cattoretti, D. Neuberg, L. De Leval, P. Kurtin, P. Dal Cin, C. Ladd, F. Feuerhake, R.C. Aguiar, S. Li, G. Salles, F. Berger, W. Jing, G.S. Pinkus, T. Habermann, R. Dalla-Favera, N.L. Harris, J.C. Aster, T.R. Golub, M.A. Shipp, The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma, Blood 102 (12) (2003) 3871—3879.
[23] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, W.G. Richards, M.T. Jaklitsch, D.J. Sugarbaker, R. Bueno, Using gene expression ratios to predict outcome among patients with mesothelioma, J. Natl. Cancer Inst. 95 (8) (2003) 598—605.
[24] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. U.S.A. 98 (24) (2001) 13790—13795.
[25] H. Jiang, Y. Deng, H.S. Chen, L. Tao, Q. Sha, J. Chen, C.J. Tsai, S. Zhang, Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes, BMC Bioinformatics 5 (1) (2004) 81.
[26] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, S.J. Korsmeyer, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nat. Genet. 30 (1) (2002) 41—47.
[27] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389—422.