

Using *GEMS* for Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data

Alexander Statnikov
alexander.statnikov@vanderbilt.edu

Ioannis Tsamardinos
ioannis.tsamardinos@vanderbilt.edu

Constantin F. Aliferis
constantin.aliferis@vanderbilt.edu

Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Keywords : gene expression microarray analysis, decision support systems, cancer, diagnosis, biomarker discovery.

Development of cancer diagnostic models and biomarker discovery from microarray gene expression data is of great importance in bioinformatics and medicine. Currently, building of cancer diagnostic models from gene expression data has at least 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 decision support models, and its enhancement by an automated system will benefit research significantly. 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 [1].

Given a microarray gene expression dataset as input, *GEMS constructs in a supervised fashion classification models* that can be used for cancer detection and determination of correct disease subtype. During construction of these models, *GEMS allows selection of a subset of genes of minimal size that are as good as or better than the full gene set* for the diagnosis. The selection of biomarkers or genes is also useful for discovery purposes, since they suggest plausible causes and treatments of various types of cancer. Finally, *GEMS provides estimates of the models' performance in future applications* (i.e., when applied to patients not used to build the models but who come from the same patient population) and readily *allows users to apply the models to individual patients*.

We implemented in *GEMS* only best-performing methodologies according to conclusions of an extensive algorithmic evaluation involving 11 publicly available cancer microarray datasets with the total of 74 diagnostic categories and 1291 patients [2]. The algorithms currently implemented in the system are the following:

<ul style="list-style-type: none"> • Model selection & performance estimation <ul style="list-style-type: none"> ○ N-fold cross-validation ○ Leave-one-out cross-validation ○ Nested N-fold cross-validation ○ Nested leave-one-out cross-validation • Classification <u>Multicategory Support Vector Machines:</u> <ul style="list-style-type: none"> ○ One-versus-rest ○ One-versus-one ○ DAGSVM ○ Method by Crammer and Singer ○ Method by Weston and Watkins • Normalization/Rescaling <ul style="list-style-type: none"> ○ 11 methods 	<ul style="list-style-type: none"> • Gene selection <u>Univariate:</u> <ul style="list-style-type: none"> ○ Kruskal-Wallis non-parametric ANOVA ○ Signal-to-noise ratio: one-versus-rest ○ Signal-to-noise ratio: one-versus-one ○ Ratio of genes between-categories to within-categories sum of squares <u>Multivariate:</u> <ul style="list-style-type: none"> ○ HITON_PC (returns a set of parents and children genes in the causal graph) ○ HITON_MB (returns a set of Markov blanket genes in the causal graph) • Performance metrics <ul style="list-style-type: none"> ○ Accuracy ○ Relative classifier information ○ Area under ROC curve
--	---

In a preliminary evaluation of the system with 5 cancer gene expression datasets (1088 patients) not employed for the algorithmic comparison, *GEMS* completed the analysis of each dataset within 10-30 minutes (on a standard PC with Intel Pentium-IV 2.4 GHz CPU) 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 to data from a different laboratory and that nested cross-validation performance estimates well approximate the error obtained by the independent validation.

GEMS provides an intuitive wizard-like user interface abstracting the microarray data analysis process and not requiring users to be experts in data analysis. 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. Each step in the interface consists of a form with options for the specific analysis stage. The steps corresponding to construction of a classification model, one of the four tasks implemented in the system, are shown below:

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

The system implements a client-server architecture and is made of a computational engine and an interface client. The computational engine is separated from the client and incorporates functional units corresponding to different aspects of analysis. The current version of *GEMS* runs on MS Windows platforms and can be downloaded free of charge for academic use from <http://www.gems-system.org>.

References:

- [1] Statnikov A, Aliferis CF, Tsamardinos I. Methods for Multi-category Cancer Diagnosis from Gene Expression Data: A Comprehensive Evaluation to Inform Decision Support System Development. *Medinfo*. 2004;2004:813-7.
- [2] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 2005 21: 631-643.