Using the *GEMS* System for Supervised Analysis of Cancer Microarray Gene Expression Data

Alexander Statnikov, M.Sc., Ioannis Tsamardinos, Ph.D., Constantin F. Aliferis, M.D., Ph.D.
Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

The authors will demonstrate a system called **GEMS** (Gene Expression Model Selector) for the automated development and evaluation of high-quality cancer diagnostic and outcome prediction models and bio-marker discovery from microarray gene expression data [1]. Manual approaches to building such models (a) require specialized training in statistics, bioinformatics, or machine learning; (b) take several weeks to months to accomplish in typical academic settings; and (c) may suffer from pitfalls introduced by human analysts such as *overfitting* the data (i.e., building models that are very good for the training set but perform poorly on future independent patient samples). **GEMS** *performs these tasks quickly, automatically, without overfitting, and without requiring the user to have expertise in data analysis*.

We will guide the users through a sequence of examples, when given a microarray gene expression dataset as input, **GEMS** *constructs in a supervised fashion classification models* that can be used for cancer detection and/or 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 or another outcome of interest. 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 population) and *allows users to apply the models to individual patients*.

We implemented in **GEMS** only the best-performing methodologies according to conclusions of an extensive algorithmic evaluation involving 11 publicly available cancer microarray datasets with a total of 74 diagnostic categories and 1291 patients.

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 the error obtained by the independent validation. Many of the aforementioned analyses will be rerun live during this demonstration.

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 (see Figure 1). The steps corresponding to construction of a classification model, (one of the four tasks implemented in the system) are the following:

- 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.
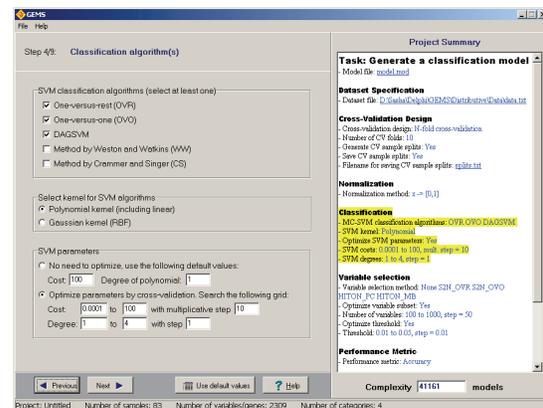
Figure 1: Example of a screen-shot of **GEMS**. The left part of the screen contains options for the current analysis step (classification algorithm). The summary of the entire project is shown in the right part of the screen.

Reference:

[1] Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF. *GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data.* Int J Med Inform, 2005.