Appendix D. *GEMS* user's manual

The graphics user interface of the *GEMS* system consists of a single form with a menu bar on top. Below we briefly describe the meaning of each labeled section in the form (**Figure D1**).

Section A. This section is used to specify input data files:

- Dataset (gene expression dataset in a tab/space separated ASCII format, with columns corresponding to genes/variables and rows to observations, the first column is the target variable which is encoded with integers starting from 0);
- Gene names (ASCII file with the list of gene names – one line per gene, first line is not used, line indices correspond to columns in the dataset);
- Gene accession numbers (ASCII file with the list of gene accession numbers – one line per gene, first line is not used, line indices correspond to columns in the dataset);

The user has to specify a dataset. Gene names and accession numbers (optional fields) will be used only for generation of experimental report in HTML format.

Section B. This section is used to select experimental design: either (1) N-fold cross-validation or (2) leave-one-out cross-validation (LOOCV). In case N-fold cross-validation is used, it is necessary to input the number of folds.

Section C. This section is used only when the experimental task is to estimate performance of the best model (see section M) and LOOCV design is employed. In that case, the user has to input the number of cross-validation folds in the inner loop of LOOCV design.

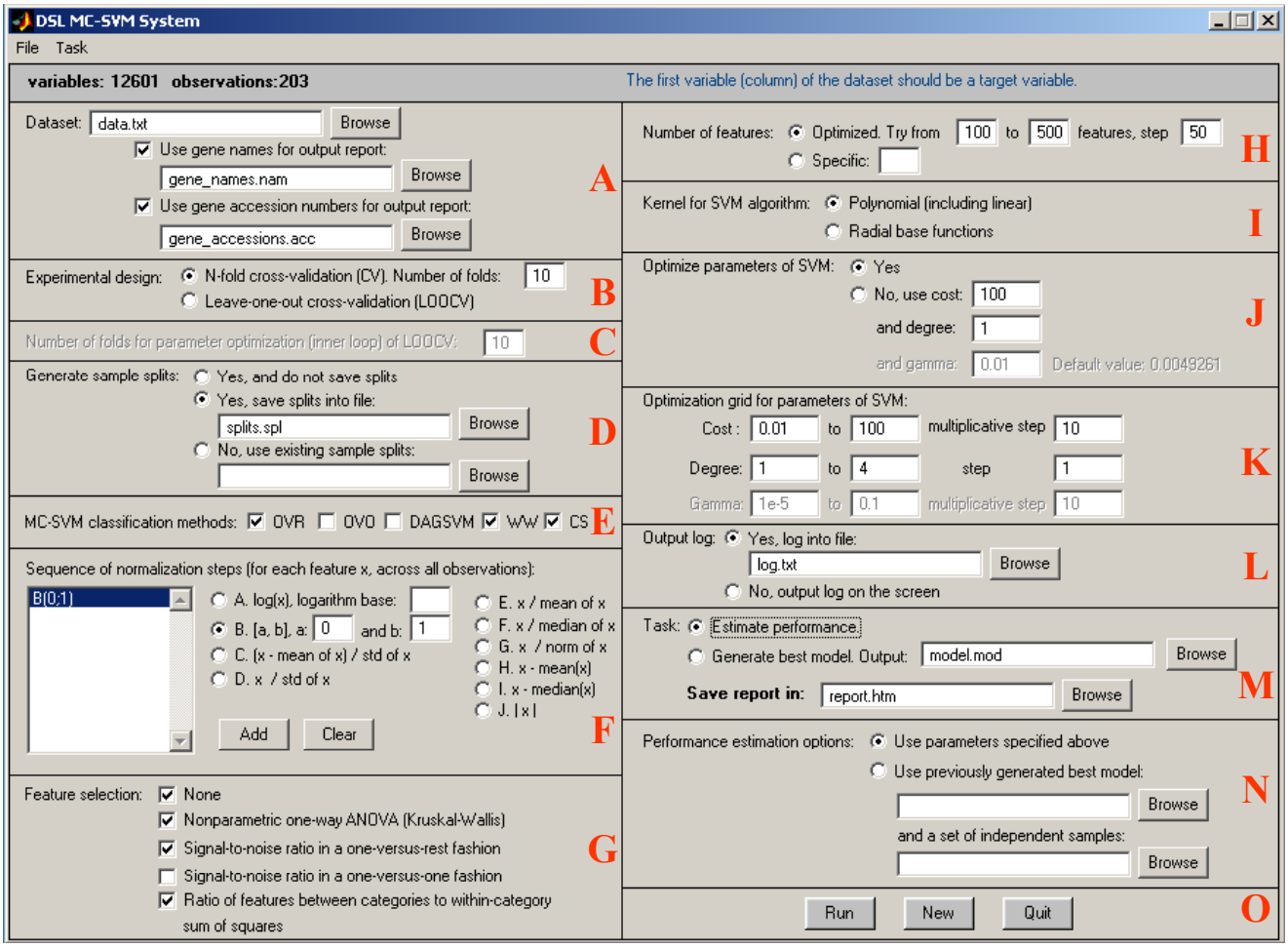

Figure D1. *Screenshot of the GEMS system. Many fields are automatically filled out with default values. Most experiments in this study can be replicated using the system with a few clicks of the mouse.*

Section D. Using this section one can either (1) generate randomly stratified sample splits for N-fold cross-validation and either discard them after experiments or save them into a file; or (2) load already generated sample splits. Sample splits are stored in an ASCII file, where each line contains indices of samples participating in a single fold (sample indices are delimited by spaces).

Section E. This section is used to select MC-SVM classification methods. If the user selects multiple classification algorithms, the system will perform optimization of the algorithms by cross-validation and derive a single algorithm yielding the largest cross-validation accuracy.

Section F. This section allows users to specify a sequence of data normalization steps for each gene x in the dataset. Normalization is always performed based solely on training dataset, so that the final results are not overfitted. It is suggested to use normalization "B" ($x \to [a, b]$ with $a = 0$ and $b = 1$) to speed up training of MC-SVM algorithms. Notice that one may need to apply normalization $|x|$ before applying $log(x)$ to ensure that the dataset does not contain negative values.

Section G. This section is used to specify gene selection algorithms. If the user selects multiple gene selection algorithms, the system will perform optimization of the algorithms by cross-validation and derive a single algorithm yielding the largest cross-validation accuracy.

Section H. If gene selection techniques are employed, this section allows selection of cardinalities of the gene subsets. One can either (1) use a gene subset of the fixed size, or (2) consider multiple gene subsets and the system will derive a single gene subset yielding the largest cross-validation accuracy.

Section I. This section is used to select a class of kernel functions for SVM algorithm – either polynomial or radial basis functions.

Section J. This section indicates if it is necessary to optimize SVM parameters by cross-validation. If optimization is not desired, the user needs to input values of cost and degree or gamma parameters. Otherwise, one has to input ranges for optimization in section K.

Section K. This section contains ranges for optimization of SVM parameters by cross-validation. The system will select a single instantiation of parameters cost and degree or gamma yielding the largest cross-validation accuracy.

Section L. This section is used to specify whether log is displayed on the screen or saved in a file.

Section M. This section is used to select an experimental task and specify the output report file. The user has two options: either (1) estimate performance, or (2) generate the best model and save it. This section also contains a field for the output report HTML file.

Section N. In case the user wants to estimate performance, this section allows either (1) to run the entire experiment with model selection (i.e. estimate performance) using Design I or II, or (2) to use already generated best model and apply it to new samples. In the latter case, one has to specify a model and a testing dataset. The testing dataset should have the same format (i.e. should contain the same variables in the same order) as the dataset in section A. If the user does not want the system to compute final accuracies, the true values of the first (target) column in the testing dataset should be substituted with an arbitrary integer number.

Section O. This is the control section of the user interface. It contains three buttons:

- Run (estimate complexity of the experiment and execute it);
- New (reset the form to default values);
- Quit.

In addition to sections described above, one can use menu bar to open and save project files which can be also created or edited with a simple text editor.