digeR Manual

Yue Fan, Thomas Brendan Murphy and R. William G. Watson
University College Dublin, Dublin 4, Dublin, Ireland
yue.fan@ucd.ie

15th June, 2009

Background

2D Difference In-Gel Electrophoresis (2D-DIGE) or 2D gel technology has been widely used as routine method for biomarker discovery in proteomics research. The high dimensional data generated from such study requires a multivariate analysis approach to be applied. The digeR Graphical User Interface (GUI) is designed to be an easy to use tool for such tasks. digeR provides collection of tools including spots correlation analysis, plotting, classification, feature selection and power analysis. It also can be used for other data types with high dimensions such as microarray analysis.

Installation

digeR is implemented within the statistical software environment R (version >= 2.6.2) and it depends on a few other packages within R. The latest version of R software environment can be downloaded from the R project webpage at http://www.r-project.org/ [1].

Prior to the installation of digeR package, the Gtk package and the gWidgets package [2] should be installed properly in your R environment. The details of how to install these two packages can be found from John Verzani's Tutorial for gWidgets at http://cran.r-project.org/web/packages/gWidgets/vignettes/gWidgets.pdf. [3]

Quick Installation
For windows user, the Gtk package can be installed within R using command [3]:

```
> source("http://www.math.csi.cuny.edu/pmg/installPMG.R")
```

The gWidgets package can be installed using command:

```
> install.packages("gWidgetsRGtk2", dep = TRUE)
```

For Linux and Mac OS X user, please refer to John Verzani's Tutorial [3].

Other R packages required by digeR include MASS [4], pls [5], e1071 [6], randomForest [7], ROCR [8], class [9], caTools [10], adabag [11] and ellipse [12]. They can be downloaded and installed from CRAN at http://cran.r-project.org/, or they can be installed with digeR installation using command:

```
>   install.packages("digeR", dep = TRUE)
```

Load the package

Once all of the necessary packages have been installed in your R environment, the digeR package can be loaded and subsequently the digeR GUI can be activated by command:

```
> library(digeR)
> digeR()
```

It may take a few seconds because R needs to load all the dependent packages required. If other GUI toolkits are also installed, R will ask which toolkit to use for building the GUI components. gWidgetsRGtk2 toolkit is recommended for digeR and others may not be fully functional. The GUI can be selected by giving the corresponding index number, if requested. For example, if R gives a menu option such as:

```
Select a GUI toolkit

1: gWidgetsRGtk2
2: gWidgetstcltk
> 1
```

The digeR GUI will be loaded as Figure1 shown below.

Figure 1. digeR Menu

Data format

In order to upload 2D DIGE data into digeR, the input data is required to be in a

particular format. In 2D DIGE experiment, normalized spots volumes and spots coordinates usually can be exported from image analysis software (e.g. Progensis). The spots coordinates are required for spots correlation analysis. Such data can be entered into an Excel file with spots coordinates (x and y) at the first two columns and followed with by spots expression data. Each sample is stored in a column and each spot is a row. The sample names are listed in the first row with group name followed by underscore and the sample number or patient index (for example, "cancer_1"). The data should be saved as "tab separated txt". A data format example can be seen in Figure 2.

	A	B	C	D	E	F	G	H
x		y	BPH_1	BPH_2	BPH_3	BPH_4	BPH_5	BPH_6
	1728	1176	0.898024	0.032431	0.144704	-2.32764	0.086885	0.684826
	1445	1230	0.737479	0.049096	0.046183	-1.07048	0.230778	0.487572

Figure 2. An example for the data format of digeR. First two columns are the spot coordinate and followed by normalized spot volumes.

An example dataset (prostate.txt) from a prostate cancer serum 2D DIGE study (unpublished data) that includes 14 BPH and 18 Gleason 5 patients can be found at the data directory within digeR package, or can be loaded using command:

```
> data(prostate)
```

An option is provided within digeR for uploading a gel image for the correlation analysis (Jpeg, recommended resolution: 400 x 350). The correlation analysis can also be performed without this gel image. The gel image (gel.jpg) for the prostate cancer serum study described above can be found in the data directory for the digeR package.

Data upload

"File ->Open" can be used to select the tab separated txt file to be upload. The gel image can be uploaded for correlation analysis through "File-> Upload_gel_image" (Optional).

If the data loading is successful, the path for the data, the corresponding sample names and the group names will appear in the R console as below.

```
[1] "C:\\Program Files\\R\\R-2.9.0\\library\\digeR\\data\\prostate.txt"

[1] "BPH_1"  "BPH_2"  "BPH_3"  "BPH_4"  "BPH_5"  "BPH_6"  "BPH_7"  "BPH_8"

[9] "BPH_9"  "BPH_10" "BPH_11" "BPH_12" "BPH_13" "BPH_14" "G5_1"   "G5_2"

[17] "G5_3"   "G5_4"   "G5_5"   "G5_6"   "G5_7"   "G5_8"   "G5_9"   "G5_10"

[25] "G5_11"  "G5_12"  "G5_13"  "G5_14"  "G5_15"  "G5_16"  "G5_17"  "G5_18"

[1] "BPH" "G5"
```

Spots correlation analysis

One of the advantages of the 2D DIGE or 2D gel technique is the ability to separate one particular protein into its different isoforms resulting from Post-Translational Modification (PTM). The expression level of those isoforms should be highly correlated as certain ratios of those isoforms are expected to be maintained within a relative stable range in a health sample. The correlation between spots can provide additional information for the biological question being asked. The changing of such correlation may reflect that certain isoforms of the protein may have been modified through disrupted PTM pathway involving in the disease process.

The correlation between the spots can be visualized on a graphical representation of 2D gel. The gel image can be attached to the GUI for a better comparison. The x and y axis of the plot represent PI and MASS, respectively. On the simulated 2D gel plot, the diameter of the spot is determined by the absolute value of correlation coefficient, positive and negative correlations are represented as red and black circles, respectively. The correlation differences between one group to another can be seen by switching the Dataset option from one group to another. The correlation threshold can be changed using the Correlation Coefficient slider on the bottom left hand side of the digeR-Spot Correlation window.
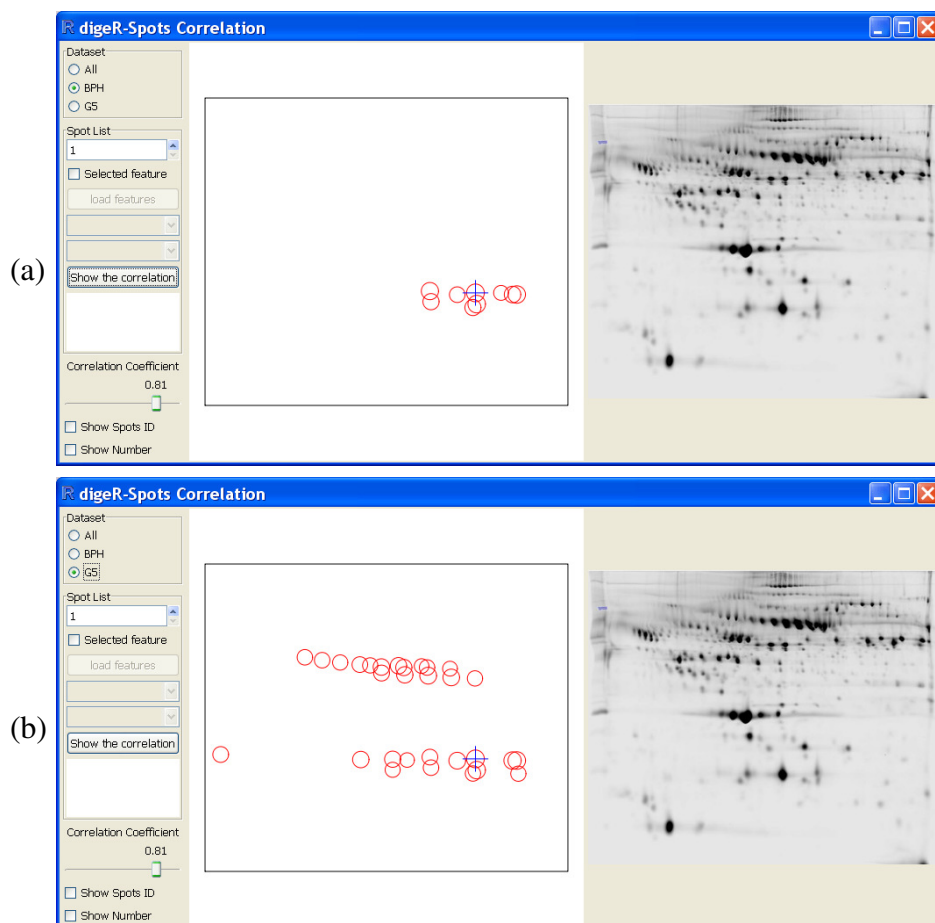
(a)

(b)

Figure 3. Spots correlation changes from control BPH (a) to Gleason 5 (b).

Score Plot

Instead of looking at each spot one at a time, Principal Component Analysis (PCA) and Partial Least Square Regression (PLSR) allow us to reduce the data dimension to a 2 or 3 dimensional space, which can be used for plotting. It provides a way of summarizing the data and looking for potential outlier samples. PCA is an unsupervised method, and PLSR is a supervised method with group information used in forming the model. A detailed description of PCA and PLSR is beyond the scope of this manual, and the introduction of these methods can be found through online tutorials.

By selecting the Plot Type (PCA or PLSR) from the drop list in score plot window, corresponding PCA or PLSR score plot will be drawn. Options are provided to look at the "Top N Component" or "pairwise" component in PCA or PLSR score plot. The color of each group can be specified in the color drop list. Different scaling methods are also provided. An example of top 5 PCA component plotting is shown in Figure 4 and 5.

The label of each sample can be added to the pairwise component plot by ticking the "With label" checkbox. The positions for the labels can be adjusted using the slider on the bottom left of the digeR-Score Plot window. The examples for PLSR score plot of the first 2 components with label are shown in Figure 6 and 7.
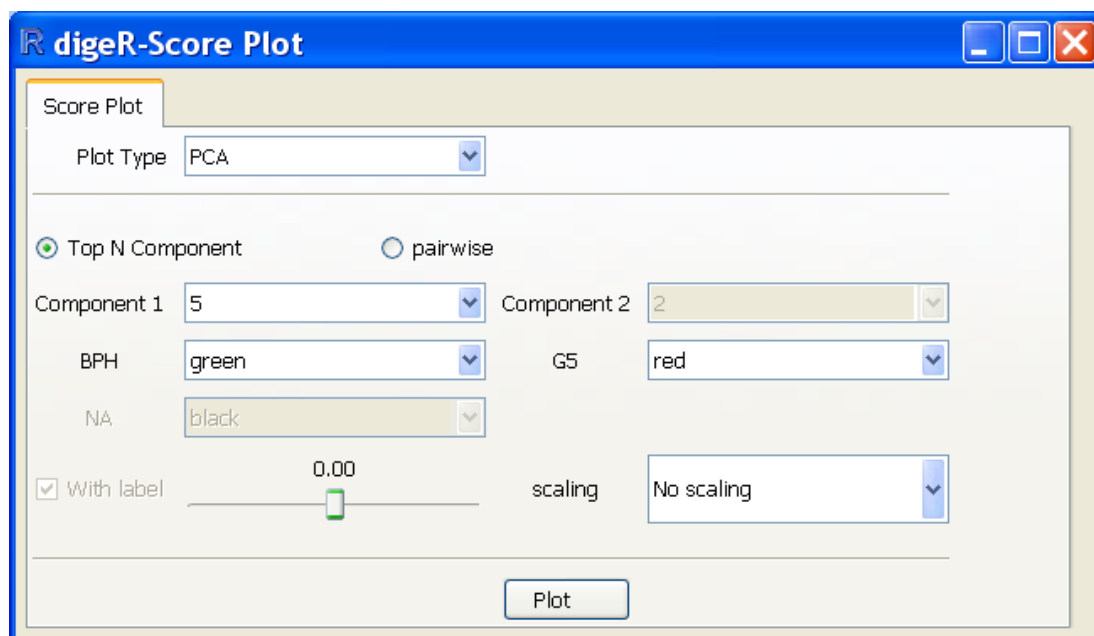

Figure 4. Score plot window (Options for top 5 components PCA plot)
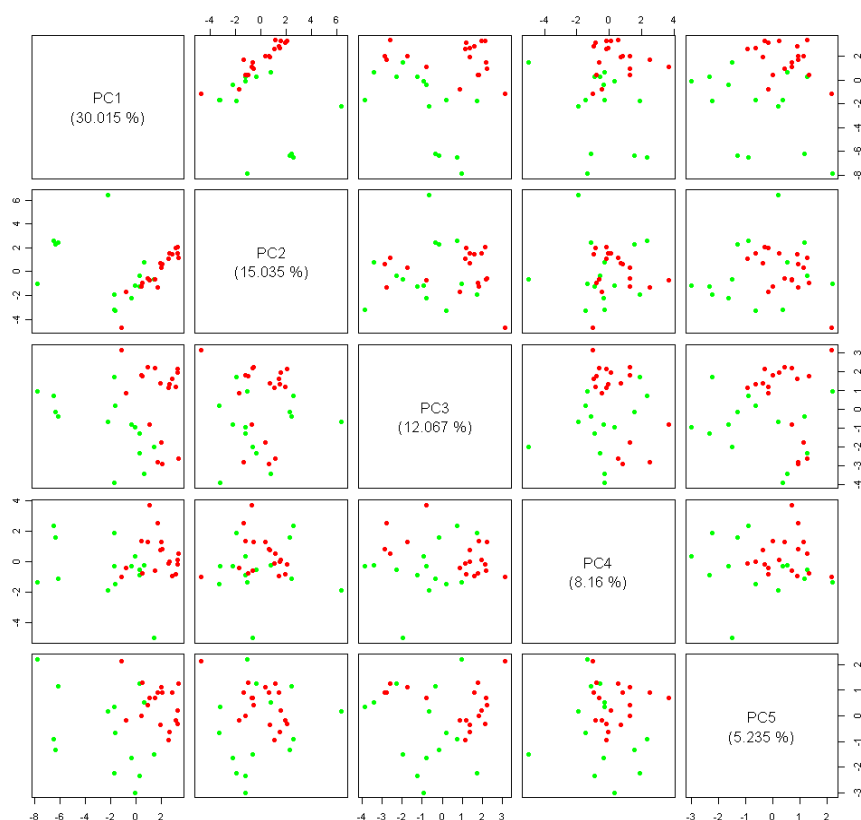
Figure 5. Top 5 components of PCA score plot. Green: BPH, Red: Gleason 5. The proportion of variation explained by each component is shown in diagonal plot.
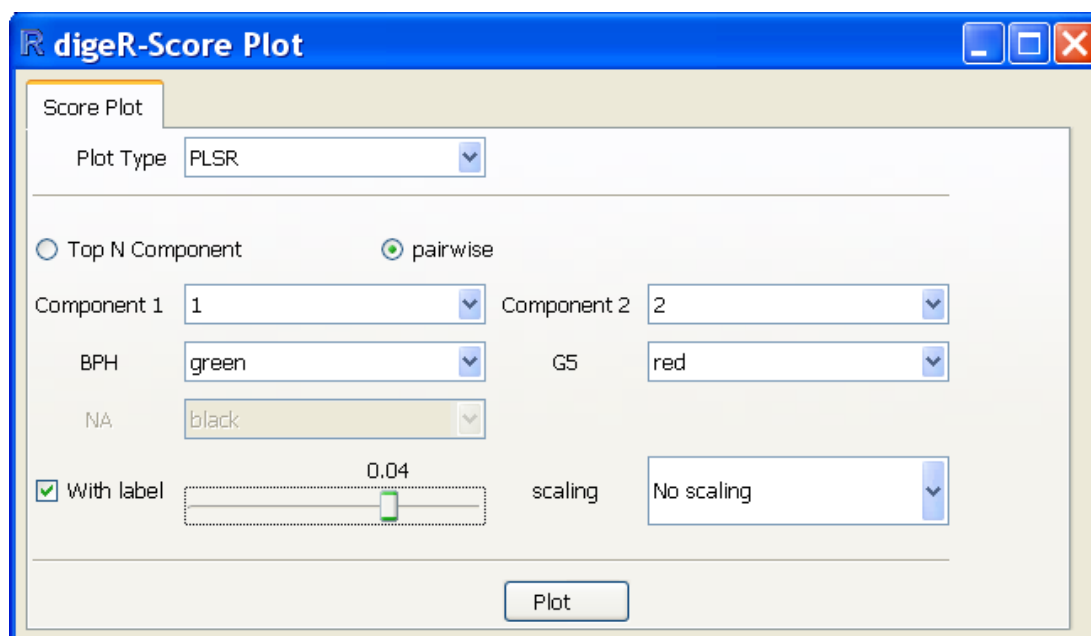

Figure 6. Score plot window (Options for Pair wise (first two components) PLSR score plot with sample labels)
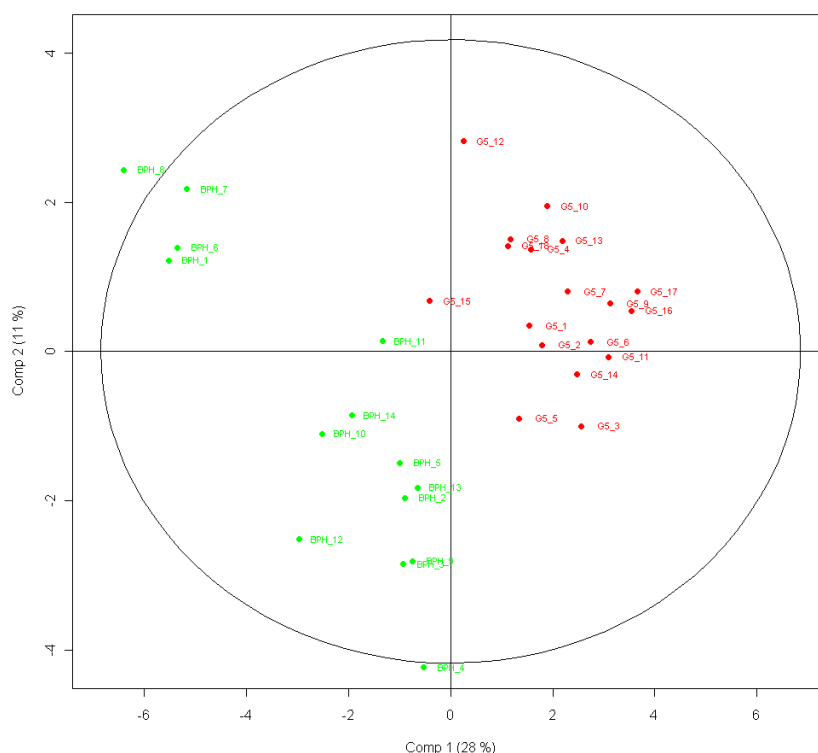
Figure 7. PLSR score of first 2 components with sample labels. Green: BPH, Red: Gleason 5.

Classification

A classification model can be built for predicting the correct group label (disease vs. control) for new unlabeled samples. The classifier is estimated with the training data set and the prediction is made on the testing samples. Leave one out cross validation, N fold cross validation or bootstrapping can be used to obtain an estimate of the prediction accuracy of the classifier. The classification in digeR can be performed with Linear Discriminant Analysis (LDA), Principal Component Regression (PCR), Partial Least Square Regression (PLSR), Logistic Regression and Support Vector Machine (SVM). Parameters for each method can be set using the Argument tabs, if required.

In order to perform a classification analysis, the classification method and the parameters for the method can be selected or specified. Before running the classification, you can select one item in the "Prediction results" table on the right hand side of the classification window, where your classification result will be temporarily saved. By pressing the Run Classification, the classification will be performed. Figure 8 shows an example of the parameters setting for a PLSR analysis with leave one out cross validation in the prostate cancer study.

The classification can also be performed on the selected spot features produced using the feature selection windows. By ticking the "Use selected features", the option of "loading features" becomes enabled, and then the R workspace (.RData) containing selected feature can be uploaded. Then the feature list from a particular feature

selection method can be selected from the drop list.

The prediction result can be assessed using Receiver operating characteristic (ROC) curves. The Area Under the Curve (AUC) value can be used as an estimate of prediction accuracy. Prediction results from different methods can be compared by simply selecting the stored prediction results (click + shift for multiple results) and followed by the ROC button. The legend shows each set of prediction results with different color and also reports an AUC value. The ROC curves for the prostate cancer data using LDA, PLSR (ncomp=3) and SVM with leave one out cross validation are shown in Figure 9.
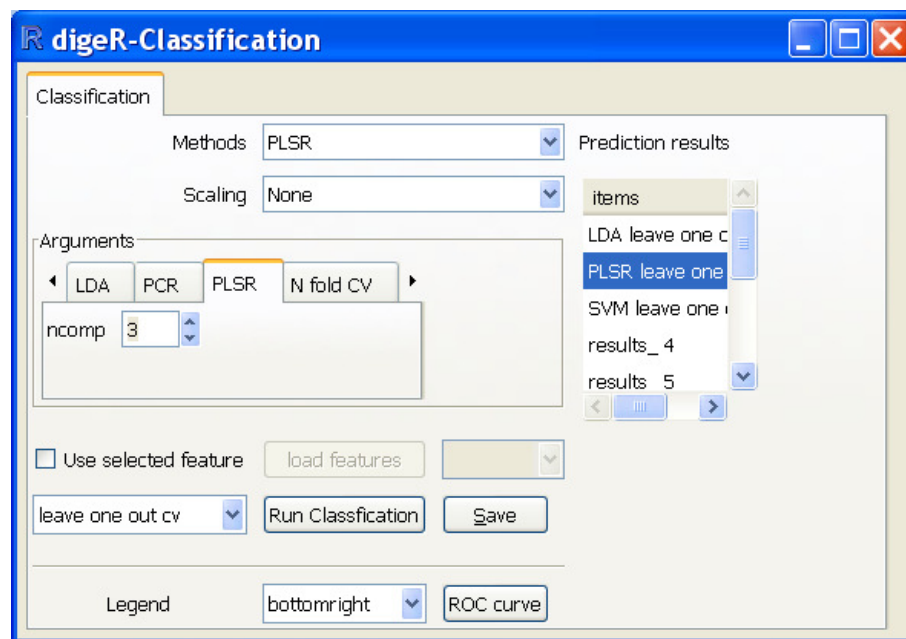

Figure 8. Parameter Setting in Classification windows. Three methods were used: LDA, PLSR and SVM.
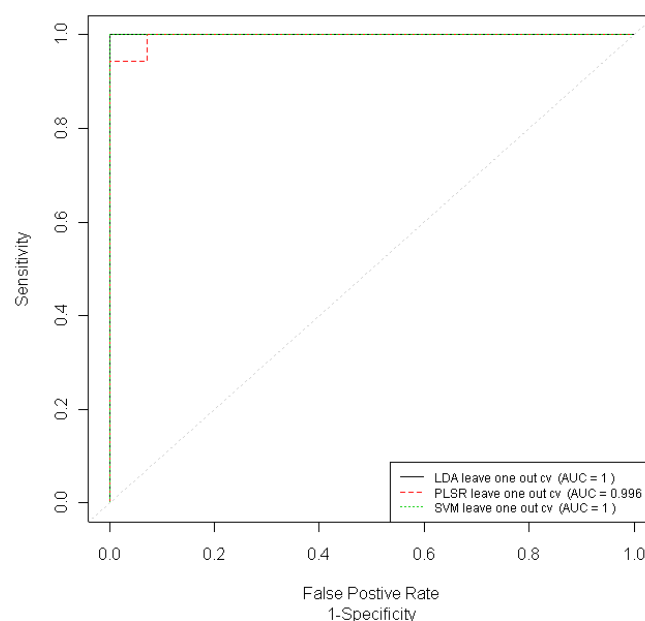

Figure 10. ROC curve for BPH vs. Gleason 5 using 3 different classification methods:

LDA, PLSR and SVM.

The prediction results can also be saved as an **R** workspace (.RData) by pressing the "Save" button beside the "Run classification" button.

Feature selection

Feature selection can be used to look for the spots with good discrimination power. Four methods are available for such analysis: LDA, PLSR, RandomForest and Adaboost. The feature selection results can be stored in the "Selected features" table on the right hand side of the feature selection window. The data can be scaled through the "Scaling" option. The parameters of the feature selection method can be set in the Arguments tabs. The number of features reported can be set from "Top". The feature selection results can be saved in the right hand panel "Selected features" in the same way as in the classification window. The feature selection results can also be output as an R workspace by pressing "Save features" at the bottom right of the window. The R workspace can be uploaded later on for correlation analysis and classification. An example of using RandomForest method in the feature selection window can be seen in Figure 11.

The selected top N features will be ranked and plotted as green bars after running each feature selection method. With RandomForest method, the plot of error rate estimation from Out Of Bag (OBB) appears first and then you can click it to proceed to the variable importance plot. An example of variable importance plot for RandomForest can be seen in Figure 12.
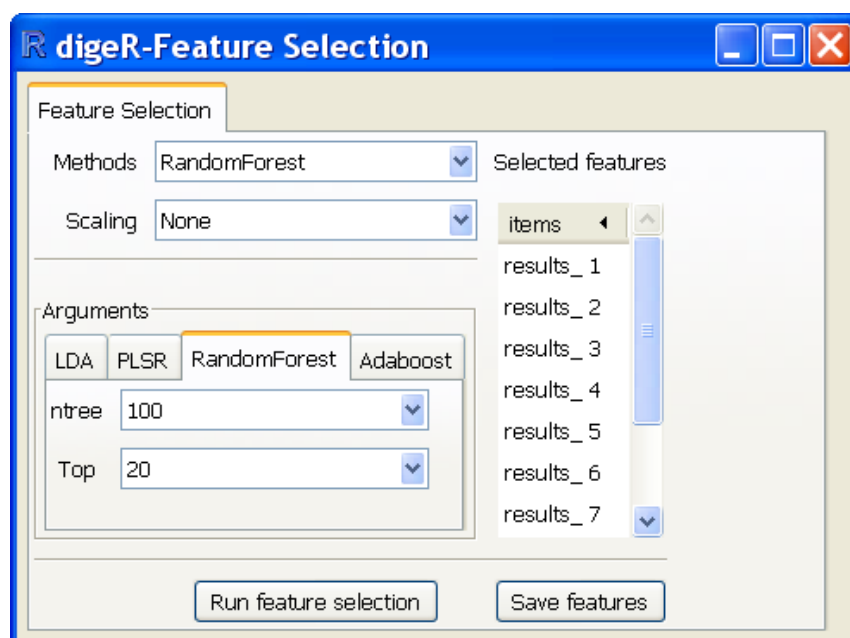

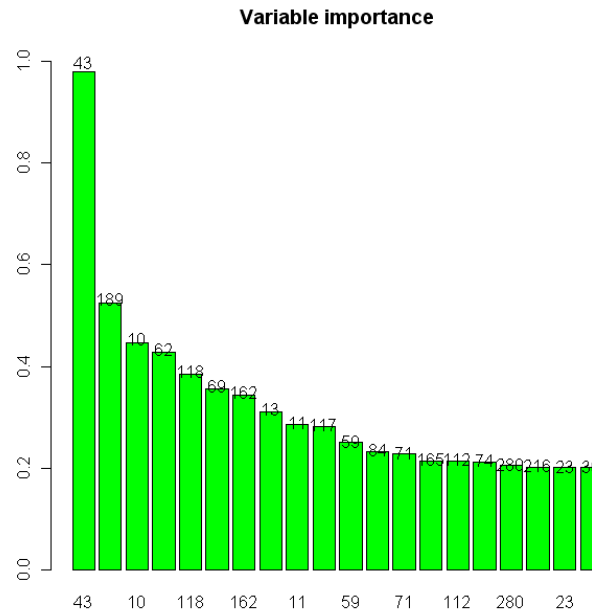Figure 11. Feature selection window (RandomForest top 20 Features).

Figure 12. The variable importance plot with top 20 variables from Randomforest method.

Power analysis

Power analysis can be performed for single spots and also for the entire gel experiment. The total number of gels needed for achieving certain power and significance level for particular spot can be calculated by putting the parameters on the Arguments section. By entering two of three parameters in the Arguments section, the remaining one is calculated. The total number of gels required for the entire experiment to achieve certain power and significance is calculated using the method proposed by Hwang *et al* [13]. An example of power analysis for spot no.1 can be found in Figure 13.

Figure 13. Power analysis windows. Calculating the sample size per group for spot no.1 to achieve 80% power and 95% of significant level.

Reference:

1. R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

2. John Verzani. Based on the iwidgets code of Simon Urbanek, suggestions by Simon Urbanek, Philippe Grosjean and Michael Lawrence (2009). gWidgets: gWidgets API for building toolkit-independent, interactive GUIs. R package version 0.0-35. http://CRAN.R-project.org/package=gWidgets.

3. John Verzani, "Examples for gWidgets", Tech. Rep., 2009. [Online]. URL: http://cran.r-project.org/web/packages/gWidgets/vignettes/gWidgets.pdf.

4. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

5. Ron Wehrens and Bjørn-Helge Mevik (2007). pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR). R package version 2.1-0. http://mevik.net/work/software/pls.html

6. Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer and Andreas Weingessel (2009). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-19.

7. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

8. Tobias Sing, Oliver Sander, Niko Beerenwinkel and Thomas Lengauer (2007). ROCR: Visualizing the performance of scoring classifiers. R package version 1.0-2. http://rocr.bioinf.mpi-sb.mpg.de/

9. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

10. Jarek Tuszynski (2008). caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.. R package version 1.9.

11. Esteban Alfaro Cortes, Matias Gamez Martinez and Noelia Garcia Rubio. adabag: Applies Adaboost.M1 and Bagging. R package version 1.1.

12. Duncan Murdoch and E. D. Chow (porting to R by Jesus M. Frias Celayeta) (2007). ellipse: Functions for drawing ellipses and ellipse-like confidence regions. R package version 0.3-5.

13. Daehee Hwang, William A. Schmitt, George Stephanopoulos, and Gregory Stephanopoulos (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics 18: 1184-1193.

Appendix

File

Open	upload the txt file
Upload_gel_image	upload the JPG image as a reference for spots correlation analysis
Quit	dispose the GUI

Correlation

Dataset	select the group to look at
Spot List	select the spot to look at
Selected feature	upload the feature list from feature selection
Load features	upload the feature list from a saved R workspace
Show the correlation	plot the spots with required correlation
Correlation Coefficiency	change the coefficiency threshold
Show spot ID	plot spots with ID
Show number	Show ID for those spots with required correlation

Score Plot

Plot Type	select either PCA or PLSR score plot
Top N component	plot score plot with top N components
Pairwise	plot selected 2 components
Component 1 and 2	two components in the pairwise plot
Group	set the color for the two groups
With label	plot the sample ID
Scaling	scale the data before plotting
Plot	plot the score plot

Classification

Methods	select the method for the classification
Scaling	scale the data before classification
Arguments	
Method	way of estimating the covariance matrix.
"moment"	standard estimators of the mean and variance
"mle"	MLEs,
"mve"	to use cov.mve
"t"	robust estimates based on a t distribution
nComp	number of component for fitting PCR or PLSR
N fold CV	number of fold in the cross validation
nboot	number of bootstrap in the classification
Selected feature	upload the feature list from feature selection
Load features	upload the feature list from a saved R workspace
leave one out cv	classification with leave one out cross validation
N fold cv	classification with N fold cross validation

Bootstrap	classification with bootstrap
Run classification	press button to do the classification
Save	save the prediction results into an R workspace
Legened	where the legend will be put
ROC curve	generate ROC plot
Prediction result	store the classification results in the selected items

Feature Selection

Method	select feature selection method
Scaling	scale the data before feature selection
Argument	
Method	same as Method in Classification
Ncomp	same as ncomp in Classification
Top	select the top n variables from the feature selection
Ntree	Number of trees to grow in randomForest
Mfinal	the number of iterations for which boosting is run or the number of trees to use
Run feature selection	Press to start feature selection
Select featuers	store the selected features in the selected items
Save features	save the features into an R workspace

Power

Single Spots	univariate power analysis
Gel	multivariate power analysis for experiment design
Significant level	set the significant level
Power	set the power level to be achieved
Sample size per group	the sample size for achieving certain significant level and power in each group
Spot Number	set the spots to calculated
Calculate	calculate the one being left blank (either power, sample size or significant level)