# A Web Interface for Automatic DNA-Microarray Analysis

The University of Nottingham

School of Computer Science, Jubilee Campus, Nottingham, NG8 1BB

Enrico Glaab, Jonathan M. Garibaldi, Natalio Krasnogor
{egg, jmg, nxk}@cs.nott.ac.uk

www.arraymining.net

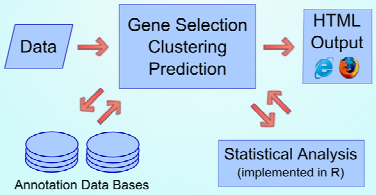
## 1 Introduction

DNA microarray experiments provide a means to understand cancer and genetic diseases on a molecular level, improve diagnosis and identify new drug targets. However, choosing appropriate data processing methods and parameters is a difficult and time-consuming task, particularly for researchers without prior experience in this field.

We present **ArrayMining.net**, a free web-service for automatic microarray analysis to address these issues. ArrayMining.net covers three major areas in statistical microarray analysis - **Feature Selection, Clustering and Prediction** - providing access to several algorithms for each of these tasks based on a single, easy-to-use interface.

## 2 Workflow

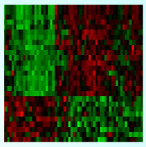
The ArrayMining server consists of **three PHP-modules** linked to the R statistical programming environment [1] and to several online **annotation data bases** (e.g. ENSEMBL [2] and DAVID [3]). Users can upload their own data or use pre-normalized public data sets as input. **Automatic parameter selection** is carried out and all results are combined into a single HTML report.



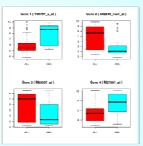

## 3 Gene Selection Module

Finding genes which are functionally related to changes in biological conditions can promote the understanding of many diseases. Thus, our server provides a diverse choice of gene selection algorithms including **filters, wrappers and ensemble feature selection**.

The resulting web-reports list all selected genes, provide **heat maps** and **box plots** for their expression values and links to **functional annotation** data bases. Selected genes can also be passed over to an external web-tool for functional annotation clustering.



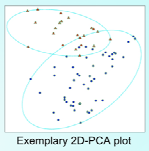
Exemplary heatmap



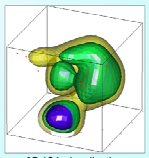
Box plots for selected genes

## 4 Clustering Module

To analyze gene expression data with unknown sample labels our web-service features both **partition-based clustering** methods (e.g. SOM, PAM, k-Means) and various **hierarchical approaches** (e.g. DIANA, AGNES). For all algorithms the number of clusters is determined automatically based on multiple cluster validity indices. Various **visual aids** are available to interpret the results, e.g. a 2D-principe components plot, a Silhouette plot and a 3D visualisation of an Independent Component Analysis (ICA).



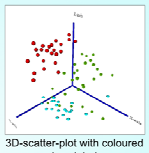
Exemplary 2D-PCA plot



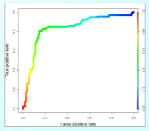
3D-ICA visualization

## 5 Prediction Module

When experimenters want to classify the biological state of new samples based on training data, our prediction module provides access to statistical learners like **SVMs**, **RF**, **kNN and PAM** [4] in combination with various feature selection methods. Cross-validated accuracies can be obtained using the widely accepted **two-level external cross-validation** methodology [5] and further evaluation statistics and analysis plots assist the user in comparing the performance for different combinations of selection and classification methods.



3D-scatter-plot with coloured class labels



Predictor evaluation based on ROC-curves

## 6 Conclusion

**ArrayMining.net is a free web-service for microarray analysis providing**

- **easy-to-interpret visual & tabular outputs**
- **automatic parameter selection**
- **integration with annotation data bases**

## References

[1] R. Ihaka, R. Gentleman, *R: A Language for Data Analysis and Graphics*, Journal of Computational and Graphical Statistics 1996, 5, 299-314
[2] T.J.P. Hubbard et al., *Ensembl 2009*, Nucleic Acids Research 2009, 37, D690-D697
[3] G. Dennis Jr. et al., DAVID: *Database for Annotation, Visualization, and Integrated Discovery*, Genome Biology 2003, 4, 2003-2004
[4] R. Tibshirani et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression*, Proceedings of the National Academy of Sciences 2002, 99, 6567-6572
[5] I.A. Wood, P. M. Visscher, K. L. Mengersen, *Classification based upon gene expression data: bias and precision of error rates*, Bioinformatics, 23, 1363-1370

We acknowledge support by Marie-Curie grant MEST-CT-2004-007597, by EPSRC grant EP/E017215/1 and BBSRC grant BB/F01855X/1

EPSRC Engineering and Physical Sciences Research Council

BBSRC bioscience for the future

MARIE CURIE ACTIONS