M@CBETH — Tutorial

a MicroArray Classification BEnchmarking Tool on a Host server

Nathalie L.M.M. Pochet, Frizo A.L. Janssens, Frank De Smet,
Kathleen Marchal, Johan A.K. Suykens and Bart L.R. De Moor

K.U.Leuven,
ESAT-SCD,
Kasteelpark Arenberg 10,
B-3001 Leuven (Heverlee),
Belgium

Email: {nathalie.pochet,frizo.janssens,frank.desmet,kathleen.marchal,
johan.suykens,bart.demoor}@esat.kuleuven.be

When reporting results obtained by M@CBETH, one should refer to

Nathalie L. M. M. Pochet, Frizo A. L. Janssens, Frank De Smet,
Kathleen Marchal, Johan A. K. Suykens, and Bart L. R. De Moor
M@CBETH: a microarray classification benchmarking tool
Bioinformatics, Jul 2005; 21: 3185 - 3186.

Contents

1 Introduction to the M@CBETH web service **3**

2 The M@CBETH services **4**

2.1 Benchmarking service 5

2.1.1 Dataset and class label files 5

2.1.2 Classification methods 8

2.1.3 Randomizations 8

2.1.4 Normalization 9

2.2 Prediction service 10

2.2.1 Dataset and class label files 10

2.2.2 Optimal prediction models 10

3 Examples **11**

3.1 Example 1: Benchmarking analysis 12

3.2 Example 2: Benchmarking analysis with immediate evaluation of prospective data (calculation of prospective accuracy) 14

3.3 Example 3: Benchmarking analysis with immediate evaluation of class labels for prospective data (prediction of prospective samples) 16

3.4 Example 4: Prediction analysis for later evaluation of prospective data (calculation of prospective accuracy) 18

3.5 Example 5: Prediction analysis for later evaluation of prospective data (prediction of class labels for prospective samples) 20

1 Introduction to the M@CBETH web service

Microarray classification can be useful to support clinical management decisions for individual patients in for example oncology. However, comparing classifiers and selecting the best for each microarray dataset can be a tedious and nonstraightforward task.

The M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) web service offers the microarray community a simple tool for making optimal two-class predictions. M@CBETH aims at finding the best prediction among different classification methods by using randomizations of the benchmarking dataset.

The M@CBETH web service intends to introduce an optimal use of clinical microarray data classification.

The M@CBETH web service is available at http://www.esat.kuleuven.be/MACBETH/.

2 The M@CBETH services

The M@CBETH website offers two services: *benchmarking* and *prediction*.

After registration and logging on to the web service, users can request benchmarking or prediction analyses. Users are notified by email about the status of their analyses running on the host server.

Users can also check the status and the results of their analyses on the *analysis results* page. This page gives an overview of all analyses and contains links to corresponding *results* pages.

To guarantee an efficient use of this web service, a maximum of 5 benchmarking analyses is allowed to run simultaneously on the host server, the rest will be scheduled. Since prediction analyses are very fast compared to benchmarking analyses, they are always allowed to start immediately.

2.1 Benchmarking service

Benchmarking, the main service on the M@CBETH website, involves selection and training of an optimal model based on the submitted benchmarking dataset and corresponding class labels. This model is then stored for immediate or later use on prospective data.

Benchmarking results in a table showing summary statistics (leave-one-out cross-validation (LOO-CV), training set accuracy (ACC) and area under the Receiver Operating Characteristic curve performance (AUC), test set ACC and AUC) for all selected classification methods, highlighting the best method in red.

Prospective data can be submitted and evaluated immediately during the same benchmarking analysis. If the corresponding prospective labels are submitted, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples. This latter application is useful for classifying new unseen patients in clinical practice.

2.1.1 Dataset and class label files

The M@CBETH web service is intended for classification of patient samples, supposing microarray data is represented by an expression matrix characterized by high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient.

Two kinds of data formats are accepted. First, spreadsheet-like tab-delimited (comma or space-delimited is also possible) text files (see Figures 1 and 2) with extension '.txt' are allowed. Further, also matrix-like matlab files (see Figures 3 and 4) with extension '.mat' are accepted.

More specific information on the data format:

- Datasets are not allowed to contain missing values.
- Class labels are restricted to '+1' (or just '1') and '-1'.
- All data must be numeric data (numbers may contain points, but no commas). All gene expression and sample descriptors must therefore be removed.
- The number of gene expression levels in a particular prediction dataset must be the same as the number of gene expression levels in the corresponding benchmarking datasets.
- The number of samples in a particular dataset must be the same as the number of corresponding class labels.
- The number of gene expression levels is supposed to be much larger than the number of samples. This allows that rows as well as columns are allowed to represent gene expression levels as well as samples.

- The filesize is limited to 40000000 bytes for safety reasons.

-214	-139	-76	-135	-106	-138	-72	-413	5	-88
-153	-73	-49	-114	-125	-85	-144	-260	-127	-105
-58	-1	-307	265	-76	215	238	7	106	42
88	283	309	12	168	71	55	-2	268	219
-295	-264	-376	-419	-230	-272	-399	-541	-210	-178
-558	-400	-650	-585	-284	-558	-551	-790	-535	-246
199	-330	33	158	4	67	131	-275	0	328
-176	-168	-367	-253	-122	-186	-179	-463	-174	-148
252	101	206	49	70	87	126	70	24	177
206	74	-215	31	252	193	-20	-169	506	183

Figure 1: Extract of spreadsheet-like tab-delimited text dataset file.

-1	-1	-1	-1	-1	-1	1	1	1	1

Figure 2: Extract of spreadsheet-like tab-delimited text class labels file.

8589.4	9164.3	3825.7	6246.4	3230.3	2510.3	7126.6	4028.7	9330.7	5271.5
5468.2	6719.5	6970.4	7823.5	3694.4	1960.7	3779.1	3156.2	7017.2	4740.8
4263.4	4883.4	5370	5955.8	3400.7	1566.3	3705.6	2870.3	4723.8	3318.5
4064.9	3718.2	4705.6	3975.6	3463.6	3072.8	6594.5	4417.6	9491.5	6792.3
1997.9	2015.2	1166.6	2002.6	2181.4	1810.2	2460.9	1854.1	5346.5	2632.9
5282.3	5569.9	1572.2	2130.5	2922.8	1673.6	3775.7	2828.3	1557.1	5449.2
2169.7	3849.1	1325.4	1531.1	2069.2	1290.4	2621.4	1427.5	1969.1	4623.2
2773.4	2793.4	1472.3	1714.6	2948.6	2465.8	2047.3	3390.7	2295.4	3277.4
7526.4	7017.7	3297	3869.8	3303.4	1675.5	6411.3	4373	6880.3	4488.1
4607.7	4802.3	2786.6	4989.4	3109.4	1312.8	3857.1	3080.5	6162.9	3343.8

Figure 3: Extract of matrix-like matlab dataset file.

-1	1	-1	1	-1	1	-1	1	-1	1

Figure 4: Extract of matrix-like matlab class labels file.

Several publicly available microarray datasets (preprocessing and missing value estimation as proposed by the original authors and discussed in (Pochet et al., 2004)) are present on the *example data* page (Figure 5) in correct data format as an example. Remark that users are responsible for preprocessing and missing value estimation of their own data before submitting it to the M@CBETH web service.



Figure 5: Download page containing several publicly available microarray datasets in correct data format as an example.

2.1.2 Classification methods

Different classification methods - based on Least Squares SVM (LS-SVM) (Suykens et al., 2002) (based on linear and Radial Basis Function (RBF) kernels), Fisher Discriminant Analysis (FDA), Principal Component Analysis (PCA) and kernel PCA (Suykens et al., 2002) (based on linear and RBF kernels) - are considered.

These are:

1. LS-SVM with linear kernel
2. LS-SVM with RBF kernel
3. Fisher Discriminant Analysis (FDA)
4. PCA (unsupervised PC selection) +FDA
5. PCA (supervised PC selection) +FDA
6. Kernel PCA with linear kernel (unsupervised PC selection) +FDA
7. Kernel PCA with linear kernel (supervised PC selection) +FDA
8. Kernel PCA with RBF kernel (unsupervised PC selection) +FDA
9. Kernel PCA with RBF kernel (supervised PC selection) +FDA

Users can select the classification methods that will be compared. The default selection is set to the best overall and most efficient methods from the benchmarking study, namely 1, 2, 6 and 7.

Remark that PCA and Kernel PCA are based on centered expression and kernel matrices respectively.

More detailed information on these methods can be found in (Pochet et al., 2004).

2.1.3 Randomizations

It is possible to change the number of randomizations. The default value is 20. Users should keep in mind that results are more reliable when the number of randomizations is large (preferably at least 20). The maximum number of randomizations allowed is 100 because of time complexity reasons.

To get an idea how long a specific analysis will take on your dataset, it is recommended to first perform the analysis you prefer with zero randomizations (this means that only a first dataset split is used). The time needed for this relatively short analysis can then be multiplied by the number of requested randomizations in order to have an idea of the dureation of the complete analysis. Of course, due to other users using the web service, this is only a rough estimation.

2.1.4 Normalization

Users can switch off normalization, although performing normalization is better from a statistical viewpoint.

Normalization is done by standardizing each gene expression of the data to have zero mean and unit standard deviation. Normalization of training sets as well as test sets is done by using the mean and standard deviation of each gene expression profile of the training sets.

2.2 Prediction service

Via the *prediction* service, the M@CBETH website offers a way for later evaluation of prospective data by reusing an existing optimal prediction model (built in a previous benchmarking analysis by the same user). If the corresponding prospective labels are submitted, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples. This latter application is useful for classifying new unseen patients in clinical practice.

2.2.1 Dataset and class label files

This is already discussed in Section 2.1.1.

2.2.2 Optimal prediction models

An existing optimal prediction model (built in a previous benchmarking analysis) needs to be selected for evaluation of prospective data. Remark that by viewing the *results* of an existing model, this model will be automatically selected for *prediction*.

3 Examples

This section shows examples of 5 different possible applications of the M@CBETH web service:

1. Performance of a benchmarking analysis

2. Performance of a benchmarking analysis with immediate evaluation of prospective data (calculation of prospective accuracy)

3. Performance of a benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples)

4. Performance of a prediction analysis for later evaluation of prospective data (calculation of prospective accuracy)

5. Performance of a prediction analysis for later evaluation of prospective data (prediction of prospective samples)

3.1 Example 1: Benchmarking analysis

A benchmarking dataset with corresponding class labels is submitted to the *benchmarking* service. All classification methods are selected. The number of randomizations is 20 and normalization is switched off. Figure 6 shows the completed *benchmarking* page based on the prostate cancer data of (Singh et al., 2002).

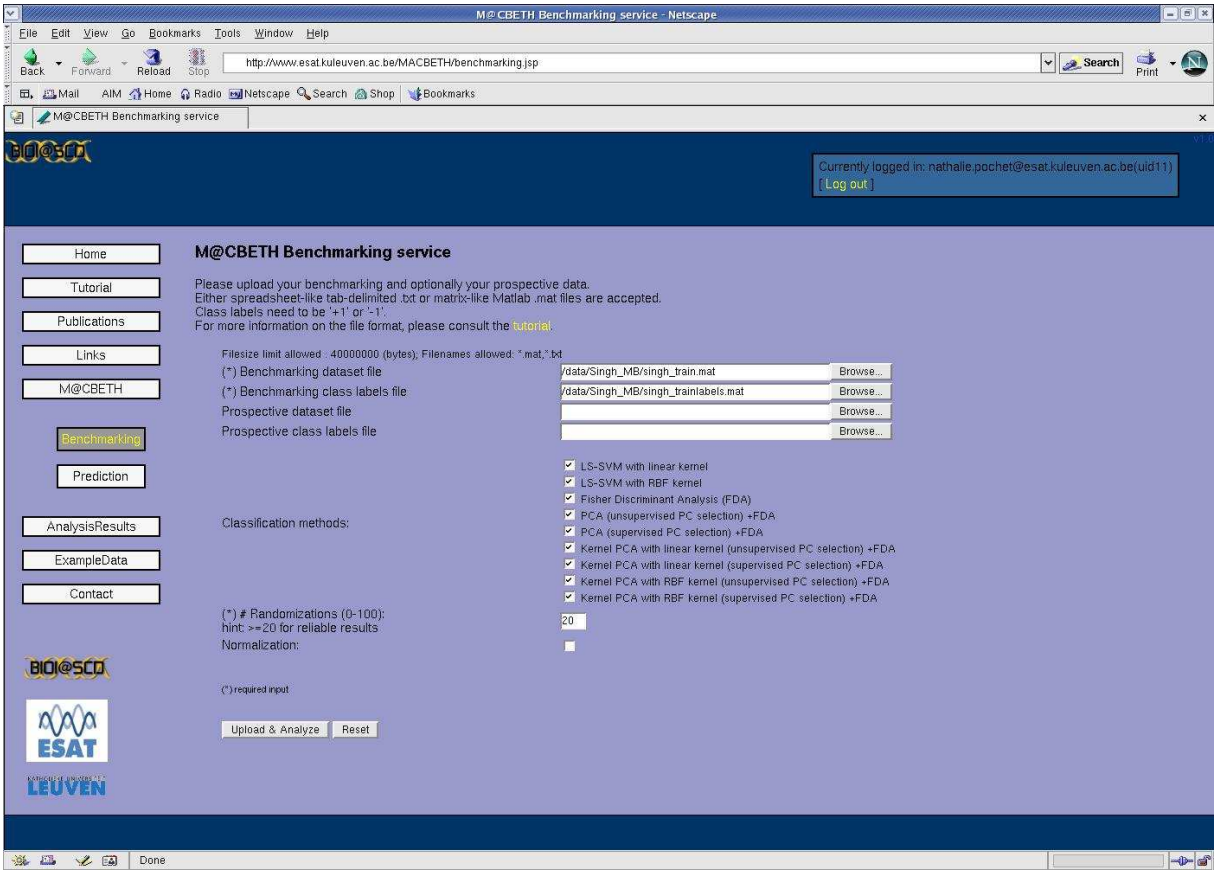

Figure 6: Completed benchmarking page for performance of a benchmarking analysis.

The *results* page of this benchmarking analysis is presented in Figure 7. This consists of a table comparing all selected classification methods and highlighting the best in red.

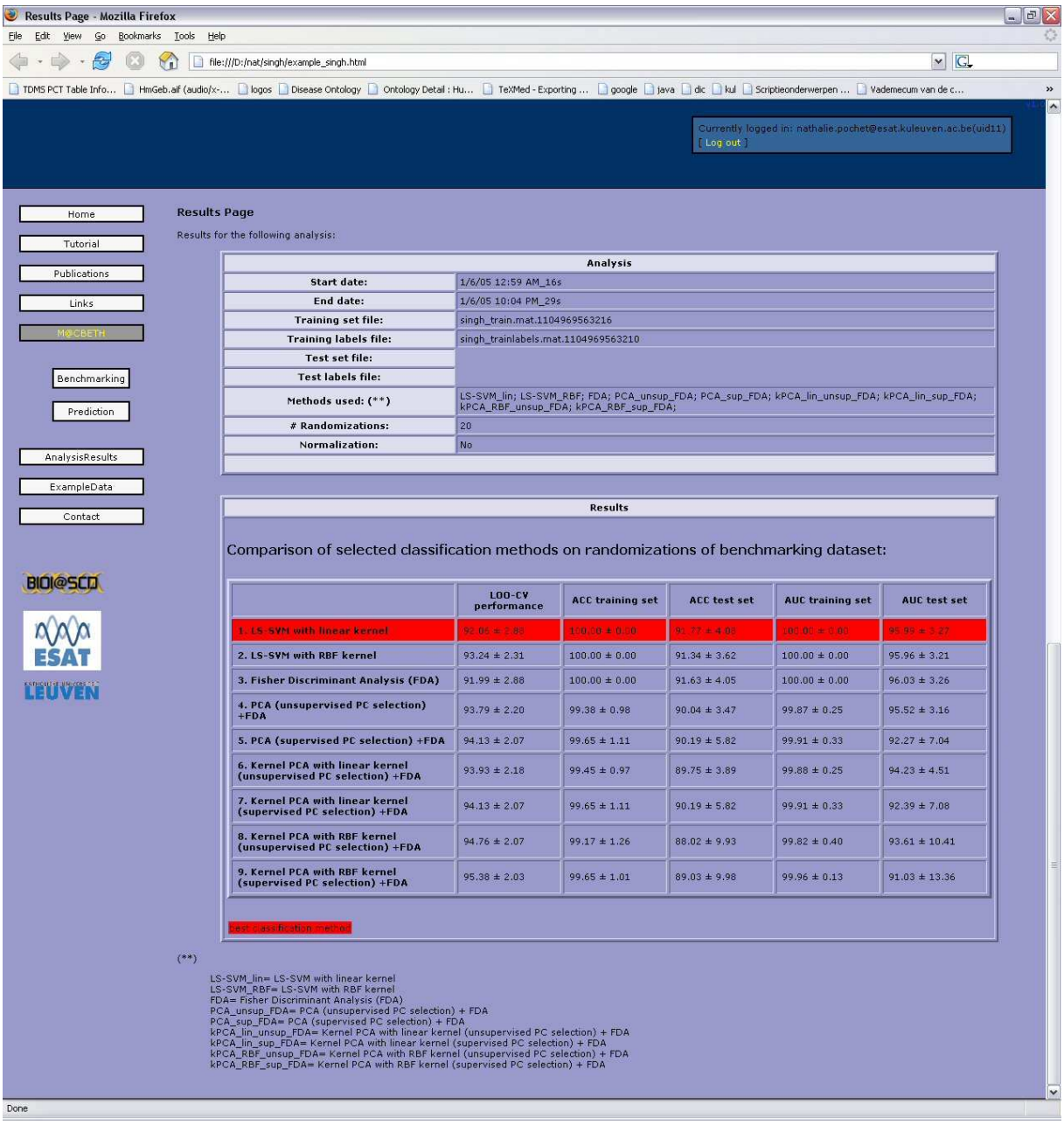

Figure 7: Results page for performance of a benchmarking analysis.

3.2 Example 2: Benchmarking analysis with immediate evaluation of prospective data (calculation of prospective accuracy)

A benchmarking dataset with corresponding class labels is submitted to the *benchmarking* service, as well as a prospective dataset with corresponding class labels. The default selection of classification methods is preserved. The number of randomizations is 30 and normalization is selected. Figure 8 shows the completed *benchmarking* page based on the colon cancer data of (Alon et al., 1999).

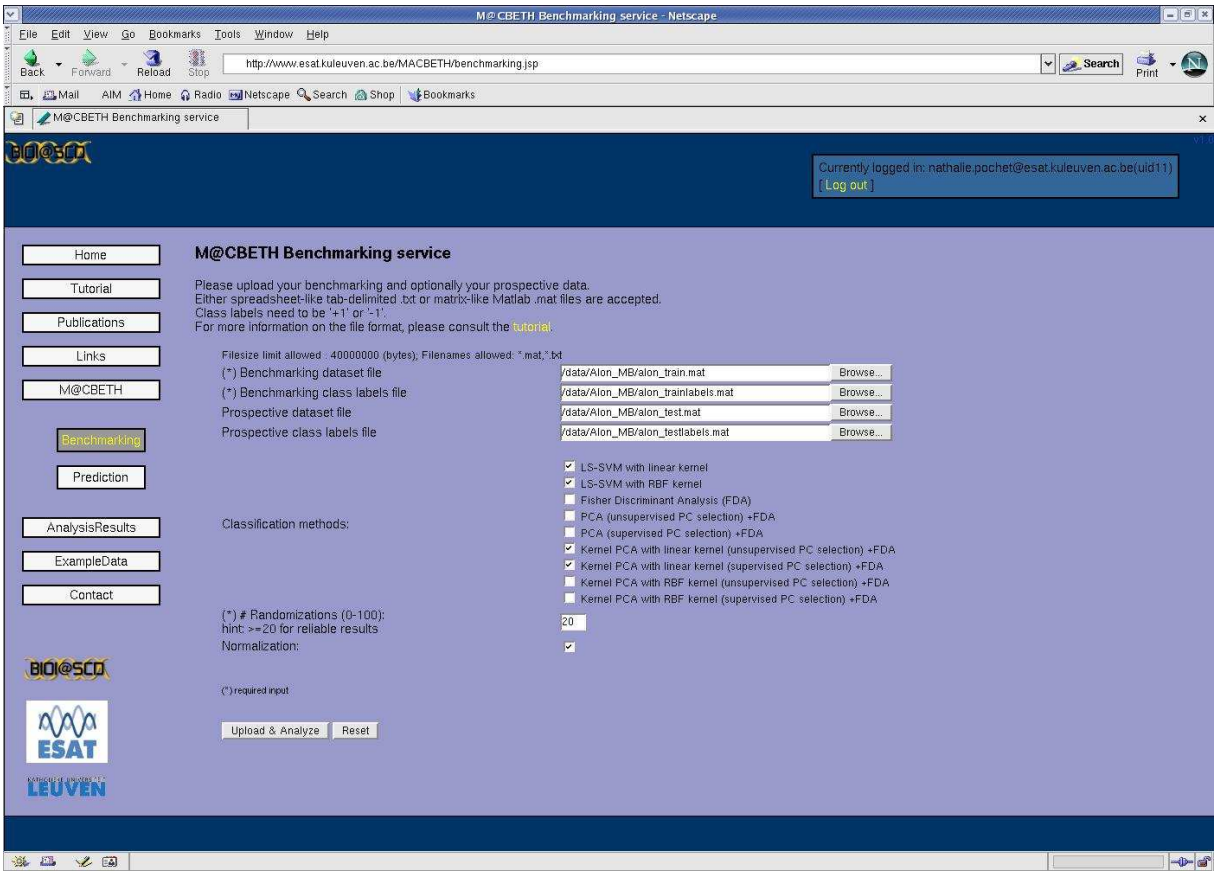

Figure 8: Completed benchmarking page for performance of a benchmarking analysis with immediate evaluation of prospective data (calculation of prospective accuracy).

The *results* page of this benchmarking analysis is presented in Figure 9. This starts with a table comparing all selected classification methods and highlighting the best in red. This is followed by the evaluation of the best model onto the prospective data, given by the prospective accuracy and a description of all misclassified samples.



Figure 9: Results page for performance of a benchmarking analysis with immediate evaluation of prospective data (calculation of prospective accuracy).

3.3 Example 3: Benchmarking analysis with immediate evaluation of class labels for prospective data (prediction of prospective samples)

A benchmarking dataset with corresponding class labels is submitted to the *benchmarking* service, as well as a prospective dataset (without corresponding class labels). The default selection of classification methods is preserved. The number of randomizations is 20 and normalization is switched off. Figure 10 shows the completed *benchmarking* page based on the high-grade glioma data of (Nutt et al., 2003).

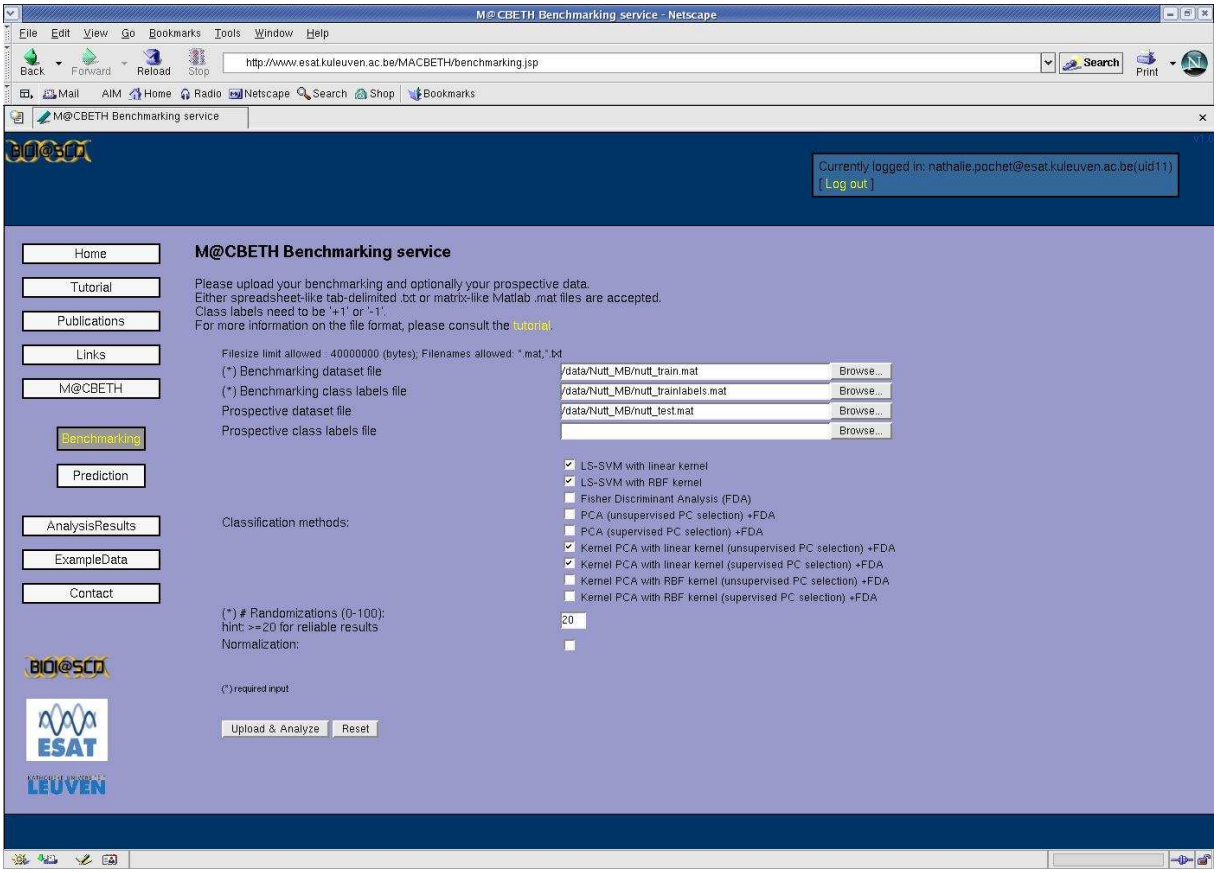

Figure 10: Completed benchmarking page for performance of a benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples).

The *results* page of this benchmarking analysis is presented in Figure 11. This starts with a table comparing all selected classification methods and highlighting the best in red. This is followed by the evaluation of the best model onto the prospective data, given by the predicted class labels for all prospective samples.

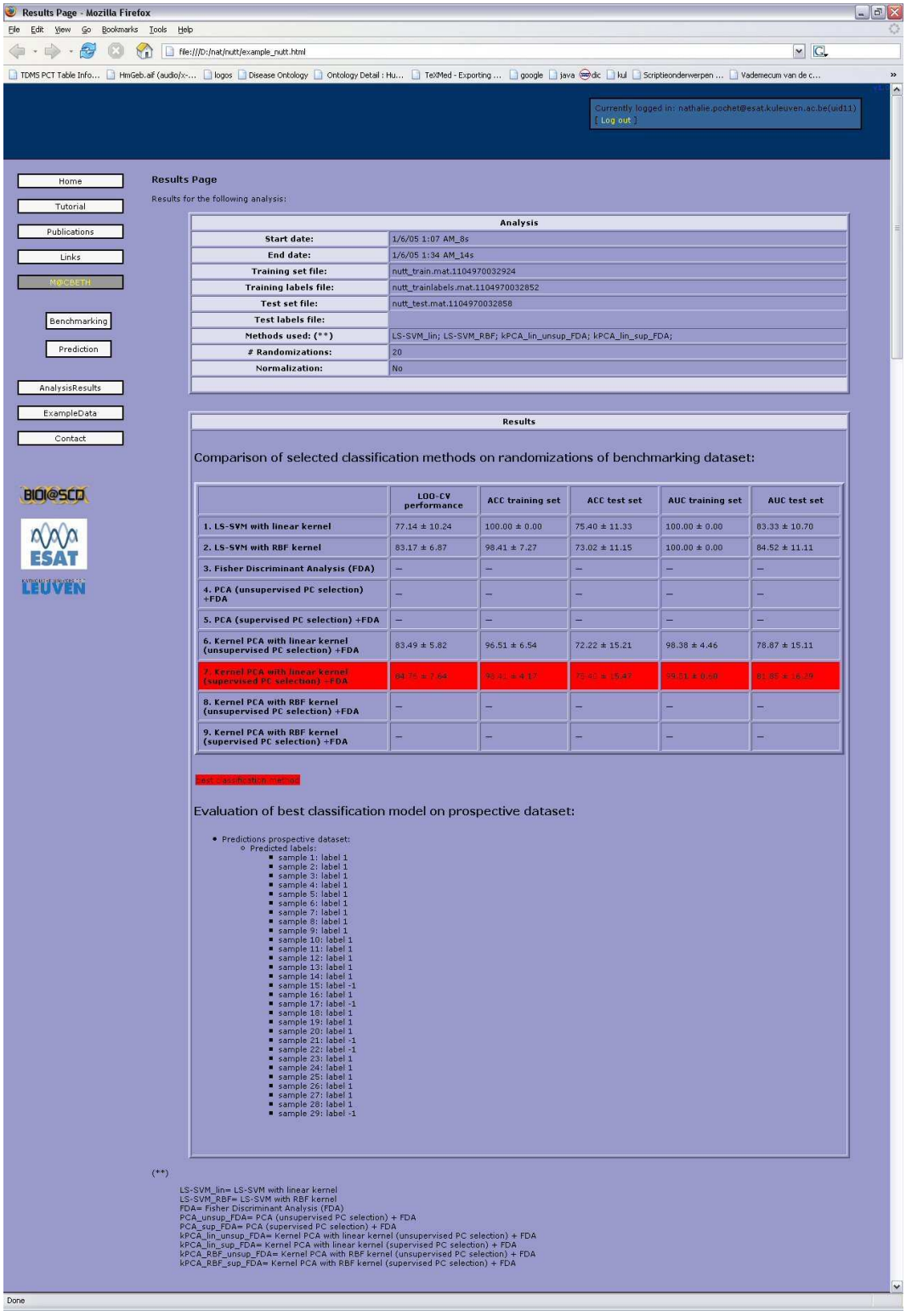

Figure 11: Results page for performance of a benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples).

3.4 Example 4: Prediction analysis for later evaluation of prospective data (calculation of prospective accuracy)

A prospective dataset with corresponding class labels is submitted to the *prediction* service. An existing optimal model is selected. Figure 12 shows the completed *prediction* page based on the high-grade glioma data of (Nutt et al., 2003).

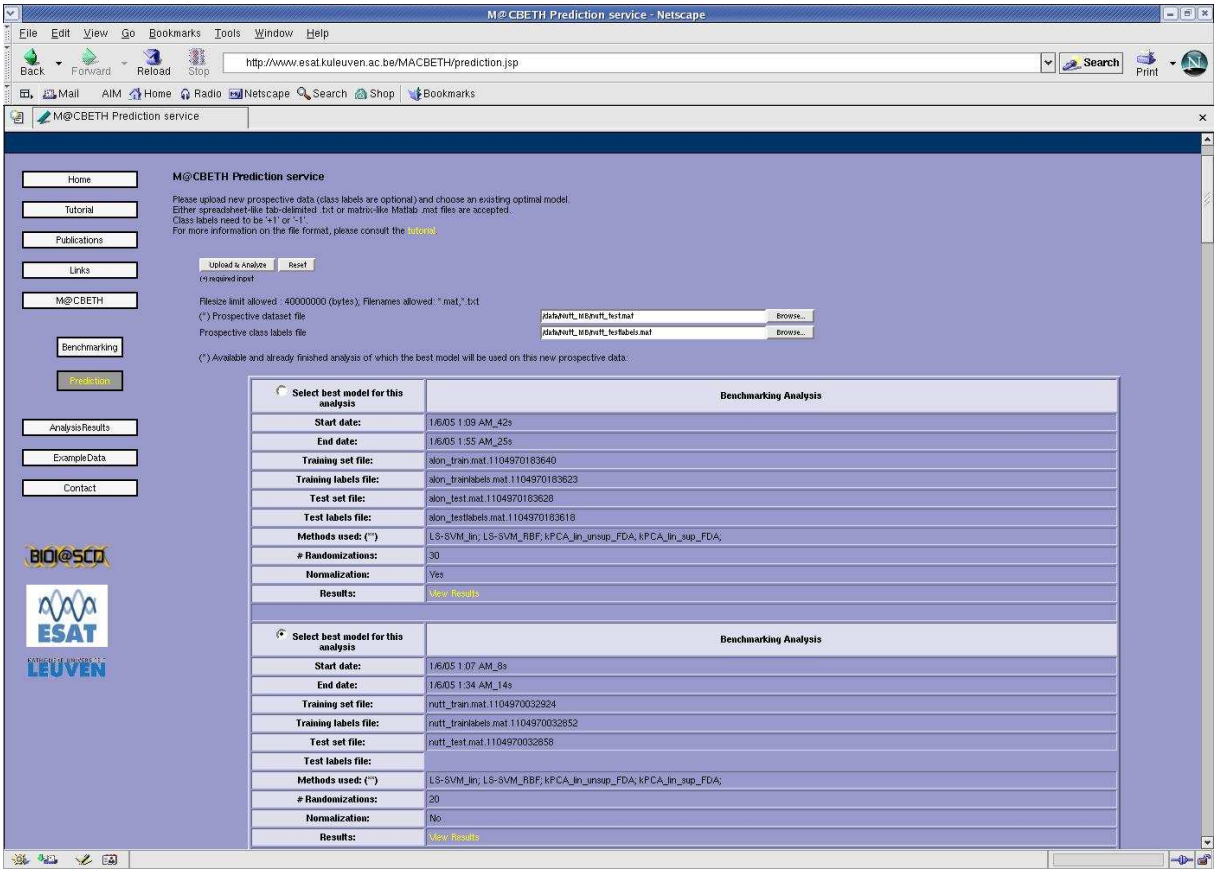

Figure 12: Completed benchmarking page for performance of a prediction analysis for later evaluation of prospective data (calculation of prospective accuracy).

The *results* page of this prediction analysis is presented in Figure 13. This consists of the evaluation of the selected optimal model onto the prospective data, given by the prospective accuracy and a description of all misclassified samples.

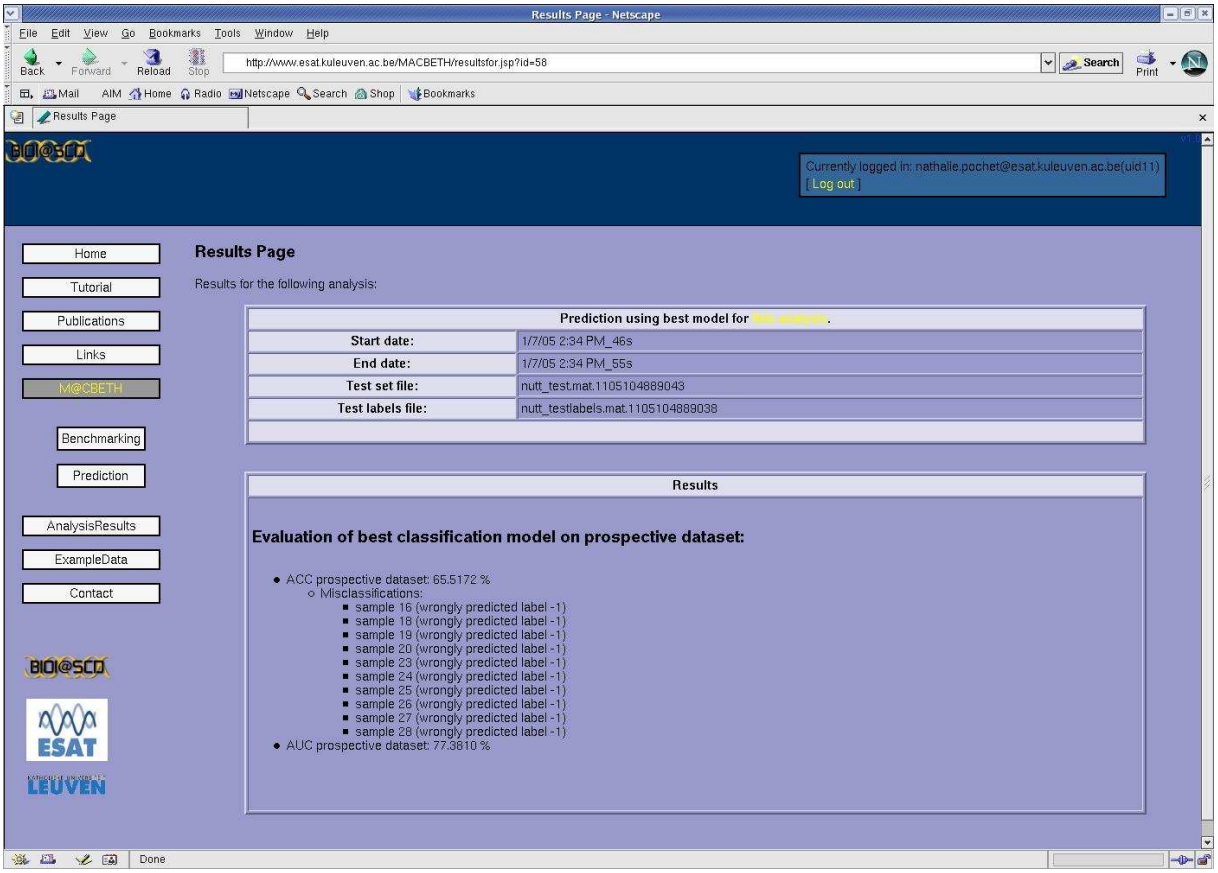

Figure 13: Results page for performance of a prediction analysis for later evaluation of prospective data (calculation of prospective accuracy).

3.5 Example 5: Prediction analysis for later evaluation of prospective data (prediction of class labels for prospective samples)

A prospective dataset (without corresponding class labels) is submitted to the *prediction* service. An existing optimal model is selected. Figure 14 shows the completed *prediction* page based on the colon cancer data of (Alon et al., 1999).

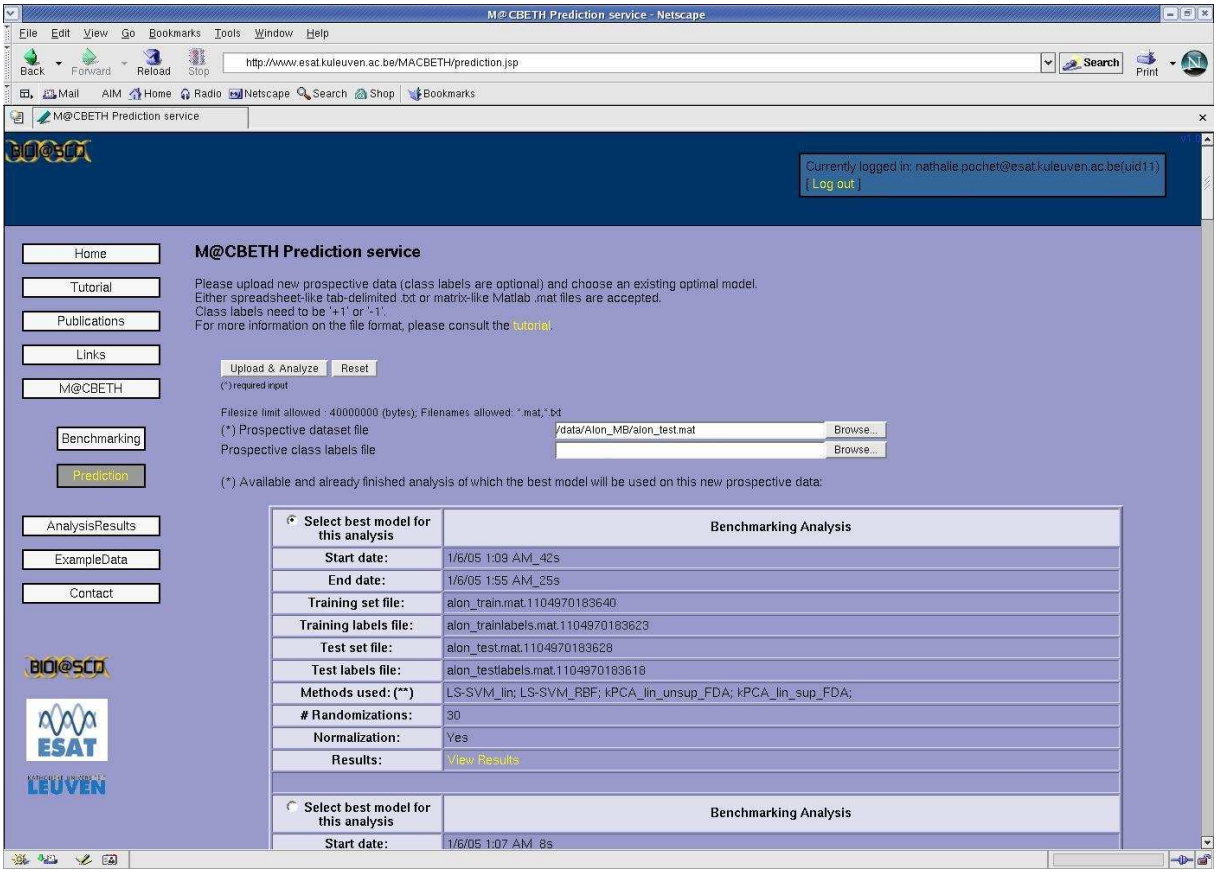

Figure 14: Completed benchmarking page for performance of a prediction analysis for later evaluation of prospective data (prediction of prospective samples).

The *results* page of this prediction analysis is presented in Figure 15. This consists of the evaluation of the selected optimal model onto the prospective data, given by the predicted class labels for all prospective samples.

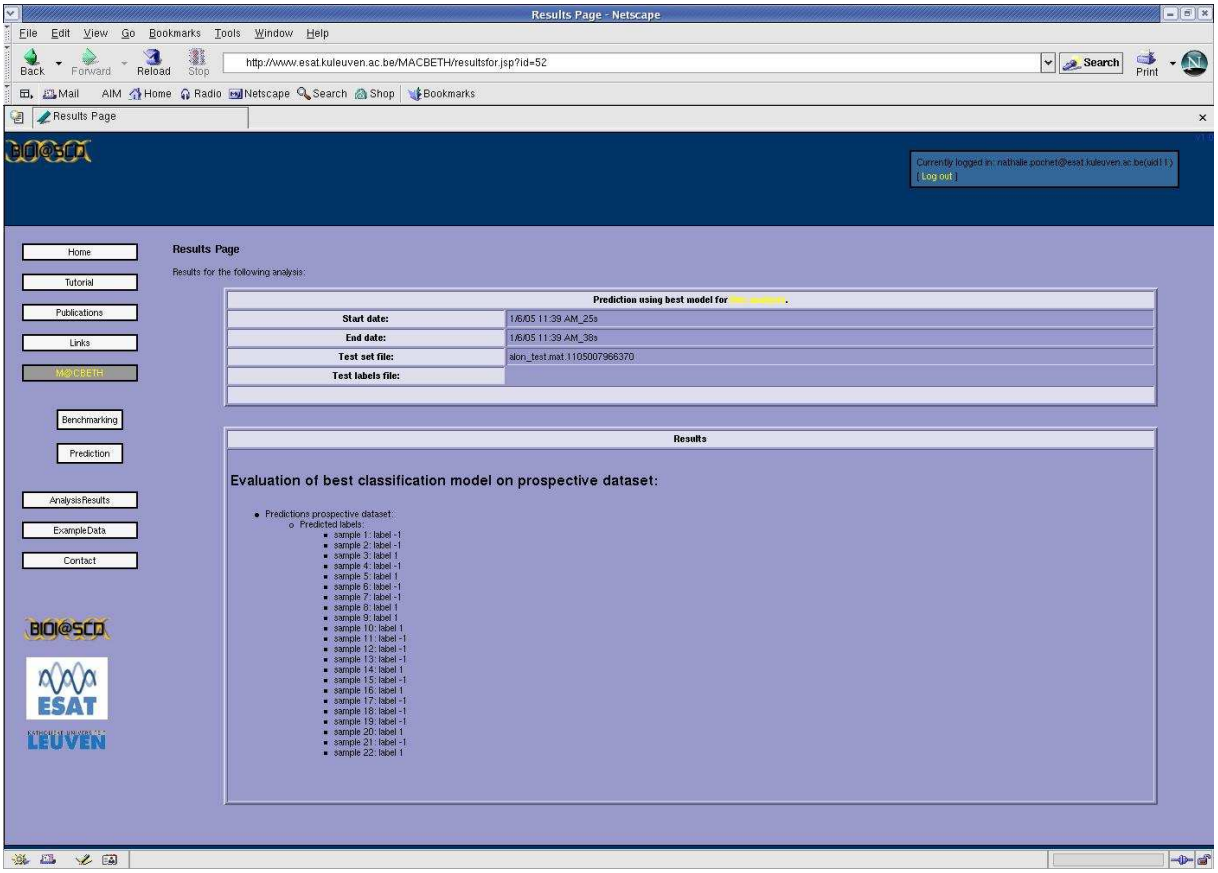

Figure 15: Results page for performance of a prediction analysis for later evaluation of prospective data (prediction of prospective samples).

Acknowledgements

Research supported by 1. Research Council KUL: GOA-AMBioRICS, IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; 2. Flemish Government: - FWO: PhD/postdoc grants, projects G.0115.01, G.0407.02, G.0413.03, G.0388.03, G.0229.03; - IWT: PhD Grants, STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA; 3. Belgian Federal Government: DWTC (IUAP V-22 (2002-2006)); 4. EU: CAGE; Biopattern.

References

Alon,A., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D., and Levine,A.J. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proc. Natl. Acad. Sci.* USA, 96,6745-6750.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286,531-537.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Raffeld,M., Yakhini,Z., Ben-Dor,A., Dougherty,E., Kononen,J., Bubendorf,L., Fehrle,W., Pittaluga,S., Gruvberger,S., Loman,N., Johannsson,O., Olsson,H., Wilfond,B., Sauter,G., Kallioniemi,O.-P., Borg,A. and Trent,J. (2001) Gene-Expression Profiles in Hereditary Breast Cancer, *The New England Journal of Medicine*, 344,539-548.

Iizuka,N., Oka,M., Yamada-Okabe,H., Nishida,M., Maeda,Y., Mori,N., Takao,T., Tamesa,T., Tangoku,A., Tabuchi,H., Hamada,K., Nakayama,H., Ishitsuka,H., Miyamoto,T., Hirabayashi,A., Uchimura,S. and Hamamoto,Y. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, *The Lancet*, 361,923-929.

Nutt,C.L., Mani,D.R., Betensky,R.A., Tamayo,P., Cairncross,J.G., Ladd,C., Pohl,U., Hartmann,C., McLaughlin,M.E., Batchelor,T.T., Black,P.M., von Deimling,A., Pomeroy,S.L., Golub,T.R. and Louis,D.N. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, *Cancer Research*, 63(7),1602-1607.

Pochet,N., De Smet,F., Suykens,J.A.K. and De Moor,B.L.R. (2004) Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction, *Bioinformatics*, 20,3185-3195.

Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P., Lander,E.S., Loda,M., Kantoff,P.W., Golub,T.R. and Sellers,W.R. (2002) Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, 1(2),203-209.

Suykens,J.A.K., Van Gestel,T., De Brabanter,J., De Moor,B. and Vandewalle,J. (2002) *Least Squares Support Vector Machines*. World Scientific, Singapore (ISBN 981-238-151-1).

van 't Veer,L.J., Dai,H., Van De Vijver,M.J., He,Y.D., Hart,A.A.M., Mao,M., Peterse,H.L., Van Der Kooy,K., Marton,M.J., Witteveen,A.T., Schreiber,G.J., Kerkhoven,R.M., Roberts,C., Linsley,P.S., Bernards,R. and Friend,S.H. (2002) Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, *Nature*, 415,530-536.